

Bogdan Gabrys
Robert J. Howlett
Lakhmi C. Jain (Eds.)

LNAI 4252

Knowledge-Based Intelligent Information and Engineering Systems

10th International Conference, KES 2006
Bournemouth, UK, October 2006
Proceedings, Part II

2 Part II

 Springer

Lecture Notes in Artificial Intelligence 4252

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Bogdan Gabrys Robert J. Howlett
Lakhmi C. Jain (Eds.)

Knowledge-Based Intelligent Information and Engineering Systems

10th International Conference, KES 2006
Bournemouth, UK, October 9-11, 2006
Proceedings, Part II

Volume Editors

Bogdan Gabrys
Bournemouth University
School of Design, Engineering and Computing
Computational Intelligence Research Group
Talbot Campus, Fern Barrow, Poole, GH12 5BB, UK
E-mail: bgabrys@bournemouth.ac.uk

Robert J. Howlett
University of Brighton
School of Engineering
Centre for SMART Systems, Brighton BN2 4GJ, UK
E-mail: r.j.howlett@brighton.ac.uk

Lakhmi C. Jain
University of South Australia
School of Electrical and Information Engineering
Knowledge-Based Intelligent Information and Engineering Systems Centre
Adelaide, Mawson Lakes Campus, South Australia SA 5095, Australia
E-mail: Lakhmi.Jain@unisa.edu.au

Library of Congress Control Number: 2006933827

CR Subject Classification (1998): I.2, H.4, H.3, J.1, H.5, K.6, K.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-46537-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-46537-9 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11893004 06/3142 5 4 3 2 1 0

Preface

Delegates and friends, we are very pleased to extend to you the sincerest of welcomes to this, the 10th International Conference on Knowledge Based and Intelligent Information and Engineering Systems at the Bournemouth International Centre in Bournemouth, UK, brought to you by KES International.

This is a special KES conference, as it is the 10th in the series, and as such, it represents an occasion for celebration and an opportunity for reflection. The first KES conference was held in 1997 and was organised by the KES conference founder, Lakhmi Jain. In 1997, 1998 and 1999 the KES conferences were held in Adelaide, Australia. In 2000 the conference moved out of Australia to be held in Brighton, UK; in 2001 it was in Osaka, Japan; in 2002, Crema near Milan, Italy; in 2003, Oxford, UK; in 2004, Wellington, New Zealand; and in 2005, Melbourne, Australia. The next two conferences are planned to be in Italy and Croatia. Delegate numbers have grown from about 100 in 1997, to a regular figure in excess of 500. The conference attracts delegates from many different countries, in Europe, Australasia, the Pacific Rim, Asia and the Americas, and may truly be said to be 'International'. Formed in 2001, KES International has developed into a worldwide organisation that provides a professional community for researchers in the discipline of knowledge-based and intelligent engineering and information systems, and through this, opportunities for publication, networking and interaction. Published by IOS Press in the Netherlands, the KES Journal is organised by joint Editors-in-Chief R.J. Howlett and B. Gabrys. There are Associate Editors in the UK, the US, Poland, Australia, Japan, Germany and the Czech Republic. The Journal accepts academic papers from authors in many countries of the world and has approximately 600 subscribers in about 50 countries. KES produces a book series, also published by IOS Press, and there are plans for the development of focus groups, each with associated publications and symposia, and other ventures such as thematic summer schools.

The KES conference continues to be a major feature of the KES organisation, and KES 2006 looks set to continue the tradition of excellence in KES conferences. This year a policy decision was made not to seek to grow the conference beyond its current size, and to aim for about 450-500 papers based on their high quality. The papers for KES 2006 were either submitted to Invited Sessions, chaired and organised by respected experts in their fields, or to General Sessions, managed by Track Chairs and an extensive International Programme Committee. Whichever route they came through, all papers for KES 2006 were thoroughly reviewed. There were 1395 submissions for KES 2006 of which 480 were published, an acceptance rate of 34%.

Thanks are due to the very many people who have given their time and goodwill freely to make the conference a success.

We would like to thank the KES 2006 International Programme Committee who were essential in providing their reviews of the papers. We are very grateful for this service, without which the conference would not have been possible. We thank the high-profile keynote speakers for providing interesting and informed talks to catalyse subsequent discussions.

An important distinction of KES conferences over others is the Invited Session Programme. Invited Sessions give new and established researchers an opportunity to present a “mini-conference” of their own. By this means they can bring to public view a topic at the leading edge of intelligent systems. This mechanism for feeding new blood into the research is very valuable. For this reason we must thank the Invited Session Chairs who have contributed in this way.

The conference administrators, Maria Booth and Jo Sawyer at the Universities of Brighton and Bournemouth respectively, and the local committees, have all worked extremely hard to bring the conference to a high level of organisation, and our special thanks go to them too.

In some ways, the most important contributors to KES 2006 were the authors, presenters and delegates without whom the conference could not have taken place. So we thank them for their contributions.

In less than a decade, KES has grown from a small conference in a single country, to an international organisation involving over 1000 researchers drawn from universities and companies across the world. This is a remarkable achievement. KES is now in an excellent position to continue its mission to facilitate international research and co-operation in the area of applied intelligent systems, and to do this in new and innovative ways. We hope you all find KES 2006 a worthwhile, informative and enjoyable experience. We also hope that you will participate in some of the new activities we have planned for KES, for the next 10 years, to benefit the intelligent systems community.

August 2006

Bob Howlett
Bogdan Gabrys
Lakhmi Jain

KES 2006 Conference Organization

Joint KES 2006 General Chairs

Bogdan Gabrys
Computational Intelligence Research Group
School of Design, Engineering and Computing
University of Bournemouth, UK

Robert J. Howlett
Centre for SMART Systems
School of Engineering
University of Brighton, UK

Conference Founder and Honorary Programme Committee Chair

Lakhmi C. Jain
Knowledge-Based Intelligent Information and Engineering Systems Centre
University of South Australia, Australia

Local Organising Committee (University of Brighton)

Maria Booth, KES Administrator
Shaun Lee, Simon Walters, Anna Howlett
Nigel Shippam, PROSE Software Support
Anthony Wood, Web Site Design

Local Organising Committee (University of Bournemouth)

Jo Sawyer, KES 2006 Local Committee Coordinator
Michael Haddrell, Mark Eastwood, Zoheir Sahel, Paul Rogers

International Programme Committee and KES 2006 Reviewers

Abbass, Hussein	University of New South Wales, Australia
Abe, Akinori	ATR Intelligent Robotics and Communication Labs, Japan
Adachi, Yoshinori	Chubu University, Japan
Alpaslan, Ferda	Middle East Technical University, Turkey
Ang, Marcello	National University of Singapore, Singapore
Angelov, Plamen	Lancaster University, UK
Anguita, Davide	DIBE - University of Genoa, Italy
Anogeianakis, Giorgos	Aristotle University of Thessaloniki, Greece
Arotaritei, Dragos	Polytechnic Institute of Iasi, Romania

Arroyo-Figueroa, Gustavo	Electrical Research Institute of Mexico, Mexico
Asano, Akira	Hiroshima University, Japan
Asari, Vijayan	Old Dominion University, USA
Augusto, Juan	University of Ulster at Jordanstown, UK
Baba, Norio	Osaka Kyoiku University, Japan
Bajaj, Preeti	G.H. Rasoni College of Engineering, Nagpur, India
Bargiela, Andrzej	Nottingham Trent University, UK
Berthold, Michael	University of Konstanz, Germany
Berthouze, Nadia	University of Aizu, Japan
Binachi-Berthouze, Nadia	University of Aizu, Japan
Bod, Rens	University of St Andrews, UK
Bosc, Patrick	IRISA/ENSSAT, France
Bouvry, Pascal	University of Applied Sciences, Luxembourg
Brown, David	University of Portsmouth, UK
Burrell, Phillip	London South Bank University, UK
Cangelosi, Angelo	University of Plymouth, UK
Ceravolo, Paolo	University of Milan, Italy
Chakraborty, Basabi	Iwate Prefectural University, Japan
Chen, Yen-Wei	Ritsumeikan University, Japan
Chen-Burger, Yun-Heh	University of Edinburgh, UK
Chung, Jaehak	Inha University, Korea
Cios, Krzysztof	University of Colorado at Denver and Health Sciences Center, USA
Coello, Carlos A.	LANIA, Mexico
Coghill, George	University of Auckland, New Zealand
Corbett, Dan	SAIC, USA
Corchado, Emilio	University of Burgos, Spain
Correa da Silva, Flavio	University of Sao Paulo, Brazil
Cuzzocrea, Alfredo	University of Calabria, Italy
Damiani, Ernesto	University of Milan, Italy
Deep, Kusum	Indian Institute of Technology, Roorkee, India
Deng, Da	University of Otago, Dunedin, New Zealand
Dubey, Venky	University of Bournemouth, UK
Dubois, Didier	University Paul Sabatier, France
Duch, Wlodzislaw	Nicolaus Copernicus University, Poland
Eldin, Amr Ali	Delft University of Technology, The Netherlands
Far, Behrouz	University of Calgary, Canada
Finn, Anthony	DSTO, Australia
Flórez-López, Raquel	University of Leon, Spain
Fortino, Giancarlo	Università della Calabria, Italy
Fuchino, Tetsuo	Tokyo Institute of Technology, Japan
Fyfe, Colin	University of Paisley, UK
Gabrys, Bogdan	University of Bournemouth, UK

Galitsky, Boris	Birkbeck College University of London, UK
Ghosh, Ashish	Indian Statistical Institute, Kolkata, India
Girolami, Mark	University of Glasgow, UK
Gorodetski, Vladimir	St. Petersburg Institute of Informatics, Russia
Grana, Manuel	Universidad Pais Vasco, Spain
Grana Romay, Manuel	Universidad Pais Vasco, Spain
Grzech, Adam	Wroclaw University of Technology, Poland
Grzymala-Busse, Jerzy	University of Kansas, USA
Gu, Dongbing	University of Essex, UK
Håkansson, Anne	Uppsala University, Sweden
Hanh H., Phan	State University of New York, USA
Hansen, Lars Kai	Technical University of Denmark, Denmark
Harrison, Robert	The University of Sheffield, UK
Hasebrook, Joachim. P	University of Luebeck, Germany
Hatem, Ahriz	The Robert Gordon University, Aberdeen, UK
Hatzilygeroudis, Ioannis	University of Patras, Greece
Helic, Denis	Technical University of Graz, Austria
Hildebrand, Lars	University of Dortmund, Germany
Hirai, Yuzo	Institute of Information Sciences and Electronics, Japan
Hong, Tzung-Pei	National University of Kaohsiung, Taiwan
Honghai, Liu	The University of Aberdeen, UK
Hori, Satoshi	Institute of Technologists, Japan
Horio, Keiichi	Kyushu Institute of Technology, Japan
Howlett, Robert J.	University of Brighton, UK
Huaglory, Tianfield	Glasgow Caledonian University, UK
Illuminada, Baturone	University of Seville, Spain
Imada, Akira	Brest State Technical University, Belasus
Ishibuchi, Hisao	Osaka Prefecture University, Japan
Ishida, Yoshiteru	Toyohashi University of Technology, Japan
Ishida, Yoshiteru	Toyohashi University, Japan
Ishii, Naohiro	Aichi Institute of Technology, Japan
Jacquetnet, François	University of Saint-Etienne, France
Jadranka, Sunde	DSTO, Australia
Jain, Lakhmi C.	University of South Australia, Australia
Jarvis, Dennis	Agent Oriented Software Pty. Ltd., Australia
Jesse, Norbert	University of Dortmund, Germany
Kacprzyk, Janusz	Polish Academy of Sciences, Poland
Karacapilidis, Nikos	University of Patras, Greece
Karny, Miroslav	Institute of Information Theory and Automation, Czech Republic
Kasabov, Nik	Auckland University of Technology, New Zealand
Katarzyniak, Radoslaw	Wroclaw University of Technology, Poland

Kazuhiko, Tsuda	University of Tsukuba, Japan
Keskar, Avinash	Visvesvaraya National Institute of Technology, India
Kim, Dong-Hwa	Hanbat National University, Korea
Kim, Jong Tae	SungKyunKwan University, Republic of Korea
Kim, Sangkyun	Yonsei University, South Korea
Kim, Tai-hoon	Korea
Kittler, Josef	University of Surrey, UK
Kóczy, Tamás, László	Budapest University of Technology and Economics, Hungary
Koenig, Andreas	Technical University of Kaiserslautern, Germany
Kojiri, Tomoko	Nagoya University, Japan
Konar, Amit	Jadavpur University, India
Koshizen, Takamasa	Honda Research Institute Japan Co., Ltd., Japan
Kunifuji, Susumu	School of Knowledge Science, Japan
Kurgan, Lukasz	University of Alberta, Canada
Kusiak, Andrew	The University of Iowa, USA
Lanzi, Pier Luca	Politecnico di Milano, Italy
Lee, Dong Chun	Howon University, Korea
Lee, Geuk	Hannam Howon University, Korea
Lee, Hong Joo	Dankook University, South Korea
Lee, Hsuan-Shih	National Taiwan Ocean University, Taiwan
Lee, Raymond	Hong Kong Polytechnic University, Hong Kong, China
Liu, Yubao	Sun Yat-Sen University, China
Lovrek, Ignac	University of Zagreb, Croatia
Lu, Hongen	La Trobe University, Australia
Luccini, Marco	University of Pavia, Italy
Mackin, Kenneth J.	Tokyo University of Information Sciences, Japan
Main, J.	La Trobe University, Australia
Mandic, Danilo	Imperial College London, UK
Maojo, Victor	Universidad Politécnica de Madrid
Martin, Trevor	Bristol University, UK
Masulli, Francesco	University of Pisa, Italy
Mattila, Jorma	Lappeenranta University of Technology, Finland
Mazumdar, Jagannath	University of South Australia, USA
McKay, Bob	University of New South Wales, Australia
Mera, Kazuya	University of Hiroshima, Japan
Mesiar, Radko	STU Bratislava, Slovakia
Mira, Jose	ETS de Ingeniería Informática (UNED), Spain
Monekosso, Dorothy	University of Kingston, UK
Montani, Stefania	Università del Piemonte Orientale, Italy
Morch, Anders	University of Oslo, Norway
Munemori, Jun	Wakayama University, Japan

Munemorim Jun	Wakayama University, Japan
Murthy, Venu K.	RMIT University, Australia
Nakamatsu, Kazumi	University of Hyogo, Japan
Nakano, Ryohei	Nagoya Institute of Technology, Japan
Nakao, Zensho	University of Ryukyus, Japan
Nakashima, Tomoharu	Osaka Prefecture University, Japan
Narasimhan, Lakshmi	University of Newcastle, Australia
Nauck, Detlef	BT, UK
Navia-Vázquez, Angel	Univ. Carlos III de Madrid, Spain
Nayak, Richi	Queensland University of Technology, Brisbane, Australia
Neagu, Ciprian	University of Bradford, UK
Negoita, Mircea	KES, New Zealand
Nguyen, Ngoc Thanh	Wroclaw University of Technology, Poland
Nishida, Toyooki	University of Kyoto, Japan
Niskanen, Vesa A.	University of Helsinki, Finland
O'Connell, Robert	University of Missouri-Columbia, USA
Ong, Kok-Leong	Deakin University, Australia
Palade, Vasile	University of Oxford, UK
Palaniswami, Marimuthu	The University of Melbourne, Australia
Pant, Millie	BITS-Pilani, India
Papis, Costas	University of Piraeus, Greece
Paprzycki, Macin	Warsaw School of Social Psychology, Poland
Park, Gwi-Tae	Korea University, Korea
Pedrycz, Witold	University of Alberta, Canada
Peña-Reyes, Carlos-Andrés	Novartis Institutes for Biomedical Research, USA
Piedad, Brox	University of Seville, Spain
Polani, Daniel	University of Hertfordshire, UK
Popescu, Theodor	National Institute for Research and Development Informatics, Romania
Rajesh, R.	Bharathiar University, India
Reusch, Bernd	University of Dortmund, Germany
Rhee, Phill	Inha University, Korea
Rose, John	Ritsumeikan Asia Pacific University, Japan
Ruan, Da	The Belgian Nuclear Research Centre, Belgium
Ruta, Dymitr	BT Exact, UK
Rutkowski, Leszek	Technical University of Czestochowa, Poland
Sato-Ilic, Mika	University of Tsukuba, Japan
Sawada, Hideyuki	Kagawa University, Japan
Seiffert, Udo	Leibniz-Institute of Plant Genetics, Gatersleben, Germany
Semeraro, Giovanni	Università degli Studi di Bari, Italy
Sharma, Dharmendra	University of Canberra, Australia

Sirlantzis, Konstantinos	University of Kent, UK
Skabar, A.	La Trobe University, Australia
Sobecki, Janusz	Wroclaw University of Technology, Poland
Soo, Von-Wun	National University of Kaohsiung, Taiwan
Sordo, Margarita	Harvard Medical School, USA
Stumptner, Markus	University of South Australia, Australia
Stytz, Martin	Institute for Defense Analyses, USA
Suetake, Noriaki	Yamaguchi University, Japan
Sujitjorn, Sarawut	Suranaree University of Technology
Sun, Zhahao	University of Wollongong, Australia
Szczerbicki, Edward	University of Newcastle, Australia
Takahash, Masakazu	Simane University, Japan
Taki, Hirokazu	Wakayama University, Japan
Tanaka, Takushi	Fukuoka Institute of Technology, Japan
Tanaka-Yamawaki, Mieko	Tottori University, Japan
Teodorescu, Horia-Nicolai	Romanian Academy, Romania
Thalmann, Daniel	EPFL, Switzerland
Thatcher, Steven	University of South Australia, Australia
Tolk, Andreas	Virginia Modeling Analysis & Simulation center, USA
Torresen, Jim	University of Oslo, Norway
Treur, Jan	Vrije Universiteit Amsterdam, Netherlands
Turchetti, Claudio	Università Politecnica delle Marche, Italy
Tweedale, J.	DSTO, Australia
Uchino, Eiji	Yamaguchi University, Japan
Unland, Rainer	University of Duisburg-Essen, Germany
Verdegay, JoseLuis	University of Granada, Spain
Virvou, Maria	University of Piraeus, Greece
Walters, Simon	University of Brighton, UK
Wang, Dianhui	La Trobe University, Australia
Wang, Lipo	Nanyang Tech University, Singapore
Wang, Pei	Temple University, USA
Watada, Junzo	Waseda University, Japan
Watanabe, Keigo	Saga University, Japan
Watanabe, Toyohide	Nagoya University, Japan
Wermter, Stefan	University of Sunderland, UK
Wren, Gloria	Loyola College in Maryland, USA
Yamashita, Yoshiyuko	Tohoku University, Sedai, Japan
Yoo, Seong-Joon	Sejong University, Korea
Zahlmann, Gudrun	Siemens Medical Solutions; Med Projekt CTB, Germany
Zambarbieri, Daniela	University of Pavia, Italy
Zha, Xuan	NIST, USA

Zharkova, Valentina	Bradford University, Bradford, UK
Zhiwen, Yu	Northwestern Polytechnical University, China
Zurada, Jacek	University of Louisville, USA

General Track Chairs

Generic Intelligent Systems Topics

Track Title	Track Chair
Artificial Neural Networks and Connectionists Systems	Ryohei Nakano, Nagoya Institute of Technology, Japan
Fuzzy and Neuro-Fuzzy Systems	Detlef Nauck, BT, UK
Evolutionary Computation	Zensho Nakao, University of Ryukyus, Japan
Machine Learning and Classical AI	Mark Girolami, University of Glasgow, UK
Agent Systems	Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland
Knowledge Based and Expert Systems	Anne Hakansson, Uppsala University, Sweden
Hybrid Intelligent Systems	Vasile Palade, Oxford University, UK
Miscellaneous Intelligent Algorithms	Honghai Liu, University of Portsmouth, UK

Applications of Intelligent Systems

Track Title	Track Chair
Intelligent Vision and Image Processing	Tuan Pham, James Cook University, Australia
Intelligent Data Mining	Michael Berthold, University of Konstanz, Germany
Knowledge Management and Ontologies	Edward Szczerbicki, University of Newcastle, Australia
Web Intelligence, Multimedia, e-Learning and Teaching	Andreas Nuernberger, University of Magdeburg, Germany
Intelligent Signal Processing, Control and Robotics	Miroslav Karny, Academy of Science, Czech Republic
Other Intelligent Systems Applications	Anthony Finn, Defence Science & Technology Organisation, Australia

Invited Session Chairs

Zhaohao Sun, University of Wollongong, Australia
Gavin Finnie, Bond University, Queensland, Australia
R.J. Howlett, University of Brighton, UK
Naohiro Ishii, Aichi Institute of Technology, Japan
Yuji Iwahori, Chubu University, Japan
Sangkyun Kim, Yonsei University, South Korea
Hong Joo Lee, Dankook University, South Korea
Yen-Wei Chen, Ritsumeikan University, Japan
Mika Sato-Ilic, University of Tsukuba, Japan
Yoshiteru Ishida, Toyohashi University of Technology, Japan
Dorothy Monekosso, Kingston University, UK
Toyoaki Nishida, The University of Kyoto, Japan
Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland
Rainer Unland, University of Duisburg-Essen, Germany
Tsuda Kazuhiko, The University of Tsukuba, Japan
Masakazu Takahash, Simane University, Japan
Daniela Zambarbieriini, Università di Pavia, Italy
Angelo Marco Luccini, Giunti Interactive Labs, Italy
Baturone Iluminada, Institute de Microelectronica de Sevilla, University of Seville, Spain
Brox Poedad, Institute de Microelectronica de Sevilla, University of Seville, Spain
David Brown, University of Portsmouth, UK
Bogdan Gabrys, University of Bournemouth, UK
Davide Anguita, DIBE - University of Genoa, Italy
P. Urlings, DSTO, Australia
J. Tweedale, DSTO, Australia
C. Sioutis, DSTO, Australia
Gloria Wren, Loyola College in Maryland, USA
Nikhil Ichalkaranje, UNISA, Australia
Yoshiyuki Yamashita, Tohoku University, Sendai, Japan
Tetsuo Fuchino, Tokyo Institute of Technology, Japan
Hirokazu Taki, Wakayama University, Japan
Satoshi Hori, Institute of Technologists, Japan
Raymond Lee, Hong Kong Polytechnic University, China
Tai-hoon Kim, Korea University of Technology and Education, Republic of Korea
Kazumi Nakamatsu, University of Hyogo, Japan
Hsuan-Shih Lee, National Taiwan Ocean University, Taiwan
Ryohei Nakano, Nagoya Institute of Technology, Japan
Kazumi Saito, NTT Communication Science Laboratories, Japan
Giorgos Anogeianakis, Aristotle University of Thessaloniki, Greece
Toyohide Watanabe, Nagoya University, Japan

Tomoko Kojiri, Nagoya University, Japan
Naoto Mukai, Nagoya University, Japan
Maria Virvou, University of Piraeus, Greece
Yoshinori Adachi, Chubu University, Japan
Nobuhiro Inuzuka, Nagoya Institute of Technology, Japan
Jun Feng, Hohai University, China
Ioannis Hatzilygeroudis, University of Patras, Greece
Constantinos Koutsojannis, University of Patras, Greece
Akinori Abe, ATR Intelligent Robotics & Communication Labs, Japan
Shoji, ATR Intelligent Robotics & Communication Labs, Japan
Ohsawa, ATR Intelligent Robotics & Communication Labs, Japan
Phill Kyu Rhee, Inha University, Korea
Rezaul Bashar, Inha University, Korea
Jun Munemori, Wakayama University, Japan
Takashi Yoshino, Wakayama University, Japan
Takaya Yuizono, Shimane University, Japan
Gwi-Tae Park, Korea University, South Korea
Manuel Grana, Universidad Pais Vasco, Spain
Richard Duro, Universidad de A Coruna, Spain
Daniel Polani, University of Hertfordshire, UK
Mikhail Prokopenko, CSIRO, Australia
Dong Hwa Kim, Hanbat National University, Korea
Vesa A. Niskanen, University of Helsinki, Finland
Emilio Corchado, University of Burgos, Spain
Hujun Yun, University of Manchester, UK
Jaehak Chung, Inha University Korea, South Korea
Da Deng, University of Otago, New Zealand
Mengjie Zhang, University of Wellington, New Zealand
Valentina Zharkova, University of Bradford, UK
Jie Zhang, George Mason University, USA
Richi Nayak, Queensland University of Technology, Australia
Lakhmi Jain, University of South Australia, Australia
Dong Chun Lee, Howon University, Korea
Giovanni Semeraro, Università degli Studi di Bari, Italy
Eugenio Di Sciascio, Politecnico di Bari, Italy
Tommaso Di Noia, Politecnico di Bari, Italy
Norio Baba, Osaka Kyoiku University, Japan
Takumi Ichimura, Hiroshima City University, Japan
Kazuya Mera, Hiroshima City University, Japan
Janusz Sobecki, Wroclaw University of Technology, Poland
Przemyslaw Kazienko, Wroclaw University of Technology, Poland
Dariusz Król, Wroclaw University of Technology, Poland

Kenneth J. Mackin, Tokyo University of Information Sciences, Japan
Susumu Kunifuji, Japan Advanced Institute of Science and Technology, Japan
Motoki Miura, Japan Advanced Institute of Science and Technology, Japan
Jong Tae Kim, SungKyunKwan University, Republic of Korea
Junzo Watada, Waseda University, Japan
Radoslaw Katarzyniak, Wroclaw University of Technology, Poland
Geuk Lee, Hannam Howon University, Korea
Il Seok Ko, Chungbuk Provincial Univ., Korea
Ernesto Damiani, University of Milan, Italy
Paolo Ceravolo, University of Milan, Italy
Dharmendra Sharma, University of Canberra, Australia
Bala Balachandran, University of Canberra, Australia
Wanla Ma, University of Canberra, Australia
Danilo P. Mandic, Imperial College London, UK
Tomasz Rutkowski, RIKEN, Japan
Toshihisa Tanaka, TUAT, Tokyo, Japan
Martin Golz, Schmalkalden, Germany

Keynote Lectures

Evolving Intelligent Systems: Methods and Applications

*Nikola Kasabov, Knowledge Engineering and Discovery Research Institute (KEDRI),
Auckland University of Technology, New Zealand*

Feature Selection in Pattern Recognition

Josef Kittler, University of Surrey, UK

Ant Colony Optimisation

*Luca Maria Gambardella, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
(IDSIA), Switzerland*

Evolvable Genetic Robots in a Ubiquitous World

*Jong-Hwan Kim, Korea Advanced Institute of Science and Technology (KAIST),
Republic of Korea*

Industrial Applications of Intelligent Systems

Ben Azvine, BT, UK

Towards Multimodal Computing: Extracting Information from Signal Nonlinearity
and Determinism

Danilo Mandic, Imperial College London, UK

Table of Contents – Part II

Computational Intelligence for Signal and Image Processing

Intensity Modulated Radiotherapy Target Volume Definition by Means of Wavelet Segmentation	1
<i>Tsair-Fwu Lee, Pei-Ju Chao, Fu-Min Fang, Eng-Yen Huang, Ying-Chen Chang</i>	
Enhancing Global and Local Contrast for Image Using Discrete Stationary Wavelet Transform and Simulated Annealing Algorithm	11
<i>Changjiang Zhang, C.J. Duanmu, Xiaodong Wang</i>	
An Efficient Unsupervised MRF Image Clustering Method	19
<i>Yimin Hou, Lei Guo, Xiangmin Lun</i>	
A SVM-Based Blur Identification Algorithm for Image Restoration and Resolution Enhancement	28
<i>Jianping Qiao, Ju Liu</i>	
Regularization for Super-Resolution Image Reconstruction	36
<i>Vivek Bannore</i>	
Dynamic Similarity Kernel for Visual Recognition	47
<i>Wang Yan, Qingshan Liu, Hanqing Lu, Songde Ma</i>	
Genetic Algorithms for Optimization of Boids Model	55
<i>Yen-Wei Chen, Kanami Kobayashi, Xinyin Huang, Zensho Nakao</i>	
Segmentation of MR Images Using Independent Component Analysis	63
<i>Yen-Wei Chen, Daigo Sugiki</i>	

Soft Data Analysis

Equi-sized, Homogeneous Partitioning	70
<i>Frank Klawonn, Frank Höppne</i>	
Nonparametric Fisher Kernel Using Fuzzy Clustering	78
<i>Ryo Inokuchi, Sadaaki Miyamoto</i>	

Finding Simple Fuzzy Classification Systems with High Interpretability Through Multiobjective Rule Selection 86
Hisao Ishibuchi, Yusuke Nojima, Isao Kuwajima

Clustering Mixed Data Using Spherical Representaion 94
Yoshiharu Sato

Fuzzy Structural Classification Methods 102
Mika Sato-Ilic, Tomoyuki Kuwata

Innovations in Soft Data Analysis 110
Mika Sato-Ilic, Lakhmi C. Jain

Immunity-Based Systems: Immunoinformatics

Tolerance Dependent on Timing/Amount of Antigen-Dose in an Asymmetric Idiotypic Network 115
Kouji Harada

Towards an Immunity-Based Anomaly Detection System for Network Traffic 123
Takeshi Okamoto, Yoshiteru Ishida

Migration Strategies of Immunity-Based Diagnostic Nodes for Wireless Sensor Network 131
Yuji Watanabe, Yoshiteru Ishida

Asymmetric Wars Between Immune Agents and Virus Agents: Approaches of Generalists Versus Specialists 139
Yoshiteru Ishida

Designing an Immunity-Based Sensor Network for Sensor-Based Diagnosis of Automobile Engines 146
Yoshiteru Ishida

Ambient Intelligence: Algorithms, Methods and Applications

Dynamic Cooperative Information Display in Mobile Environments 154
Christophe Jacquet, Yacine Bellik, Yolaine Bourda

Extracting Activities from Multimodal Observation 162
Oliver Brdiczka, Jérôme Maisonnasse, Patrick Reignier, James L. Crowley

Using Ambient Intelligence for Disaster Management	171
<i>Juan Carlos Augusto, Jun Liu, Liming Chen</i>	

Dynamic Scene Reconstruction for 3D Virtual Guidance	179
<i>Alessandro Calbi, Lucio Marcenaro, Carlo S. Regazzoni</i>	

Communicative Intelligence

An Introduction to Fuzzy Propositional Calculus Using Proofs from Assumptions	187
<i>Iwan Tabakow</i>	

Fuzzy-Neural Web Switch Supporting Differentiated Service	195
<i>Leszek Borzemski, Krzysztof Zatwarnicki</i>	

Image Watermarking Algorithm Based on the Code Division Multiple Access Technique	204
<i>Fan Zhang, Guosheng Yang, Xianxing Liu, Xinhong Zhang</i>	

Construction of Symbolic Representation from Human Motion Information	212
<i>Yutaka Araki, Daisaku Arita, Rin-ichiro Taniguchi, Seiichi Uchida, Ryo Kurazume, Tsutomu Hasegawa</i>	

Toward a Universal Platform for Integrating Embodied Conversational Agent Components	220
<i>Hung-Hsuan Huang, Tsuyoshi Masuda, Aleksandra Cerekovic, Kateryna Tarasenko, Igor S. Pandzic, Yukiko Nakano, Toyoaki Nishida</i>	

Flexible Method for a Distance Measure Between Communicative Agents' Stored Perceptions	227
<i>Agnieszka Pieczyńska</i>	

Harmonisation of Soft Logical Inference Rules in Distributed Decision Systems	235
<i>Juliusz L. Kulikowski</i>	

Assessing the Uncertainty of Communication Patterns in Distributed Intrusion Detection System	243
<i>Krzysztof Juszczyszyn, Grzegorz Kołaczek</i>	

An Algorithm for Inconsistency Resolving in Recommendation Web-Based Systems	251
<i>Michał Malski</i>	

Distributed Class Code and Data Propagation with Java	259
<i>Dariusz Król, Grzegorz Stanisław Kukla</i>	
Conflicts of Ontologies – Classification and Consensus-Based Methods for Resolving	267
<i>Ngoc Thanh Nguyen</i>	
Knowledge-Based Systems for e-Business	
Estimation of FAQ Knowledge Bases by Introducing Measurements	275
<i>Jun Harada, Masao Fuketa, El-Sayed Atlam, Toru Sumitomo, Wataru Hiraishi, Jun-ichi Aoe</i>	
Efficient Stream Delivery over Unstructured Overlay Network by Reverse-Query Propagation	281
<i>Yoshikatsu Fujita, Yasufumi Saruwatari, Jun Yoshida, Kazuhiko Tsuda</i>	
A Method for Development of Adequate Requirement Specification in the Plant Control Software Domain	289
<i>Masakazu Takahashi, Yoshinori Fukue, Satoru Takahashi, Takashi Kawasaki</i>	
Express Emoticons Choice Method for Smooth Communication of e-Business	296
<i>Nobuo Suzuki, Kazuhiko Tsuda</i>	
A New Approach for Improving Field Association Term Dictionary Using Passage Retrieval	303
<i>Kazuhiro Morita, El-Sayed Atlam, Elmarhomy Ghada, Masao Fuketa, Jun-ichi Aoe</i>	
Analysis of Stock Price Return Using Textual Data and Numerical Data Through Text Mining	310
<i>Satoru Takahashi, Masakazu Takahashi, Hiroshi Takahashi, Kazuhiko Tsuda</i>	
A New Approach for Automatic Building Field Association Words Using Selective Passage Retrieval	317
<i>El-Sayed Atlam, Elmarhomy Ghada, Kazuhiro Morita, Jun-ichi Aoe</i>	
Building New Field Association Term Candidates Automatically by Search Engine	325
<i>Masao Fuketa, El-Sayed Atlam, Elmarhomy Ghada, Kazuhiro Morita, Jun-ichi Aoe</i>	

Neuro-fuzzy Techniques for Image Processing Applications

Efficient Distortion Reduction of Mixed Noise Filters by Neuro-fuzzy Processing	331
<i>Mehmet Emin Yüksel, Alper Baştürk</i>	
Texture Segmentation with Local Fuzzy Patterns and Neuro-fuzzy Decision Support	340
<i>Laura Caponetti, Ciro Castiello, Anna Maria Fanelli, Przemyslaw Górecki</i>	
Lineal Image Compression Based on Lukasiewicz's Operators	348
<i>Nashaat M. Hussein Hassan, Angel Barriga</i>	
Modelling Coarseness in Texture Images by Means of Fuzzy Sets	355
<i>Jesús Chamorro-Martínez, Elena Galán-Perales, Daniel Sánchez, Jose M. Soto-Hidalgo</i>	
Fuzzy Motion Adaptive Algorithm for Video De-interlacing	363
<i>Piedad Brox, Iluminada Baturone, Santiago Sánchez-Solano, Julio Gutiérrez-Ríos, Felipe Fernández-Hernández</i>	

Knowledge-Based Interface Systems (1)

Web Site Off-Line Structure Reconfiguration: A Web User Browsing Analysis	371
<i>Sebastián A. Ríos, Juan D. Velásquez, Hiroshi Yasuda, Terumasa Aoki</i>	
New Network Management Scheme with Client's Communication Control	379
<i>Kazuya Odagiri, Rihito Yaegashi, Masaharu Tadauchi, Naohiro Ishii</i>	
Prediction of Electric Power Generation of Solar Cell Using the Neural Network	387
<i>Masashi Kawaguchi, Sachiyoshi Ichikawa, Masaaki Okuno, Takashi Jimbo, Naohiro Ishii</i>	
Text Classification: Combining Grouping, LSA and kNN vs Support Vector Machine	393
<i>Naohiro Ishii, Takeshi Murai, Takahiro Yamada, Yongguang Bao, Susumu Suzuki</i>	

Particle Filter Based Tracking of Moving Object from Image Sequence	401
<i>Yuji Iwahori, Toshihiro Takai, Haruki Kawanaka, Hidenori Itoh, Yoshinori Adachi</i>	

Nature Inspired Data Mining

Discrete and Continuous Aspects of Nature Inspired Methods	409
<i>Martin Macaš, Miroslav Burša, Lenka Lhotská</i>	
Social Capital in Online Social Networks	417
<i>Przemysław Kazienko, Katarzyna Musiał</i>	
Nature-Inspiration on Kernel Machines: Data Mining for Continuous and Discrete Variables	425
<i>Francisco J. Ruiz, Cecilio Angulo, Núria Agell</i>	
Testing CAB-IDS Through Mutations: On the Identification of Network Scans	433
<i>Emilio Corchado, Álvaro Herrero, José Manuel Sáiz</i>	
Nature Inspiration for Support Vector Machines	442
<i>Davide Anguita, Dario Sterpi</i>	

Intelligent Agents and Their Applications

The Equilibrium of Agent Mind: The Balance Between Agent Theories and Practice	450
<i>Nikhil Ichalkaranje, Christos Sioutis, Jeff Tweedale, Pierre Urlings, Lakhmi Jain</i>	
Trust in LORA: Towards a Formal Definition of Trust in BDI Agents	458
<i>Bevan Jarvis, Lakhmi Jain</i>	
Agent Cooperation and Collaboration	464
<i>Christos Sioutis, Jeffrey Tweedale</i>	
Teamwork and Simulation in Hybrid Cognitive Architecture	472
<i>Jinsong Leng, Colin Fyfe, Lakhmi Jain</i>	
Trust in Multi-Agent Systems	479
<i>Jeffrey Tweedale, Philip Cutler</i>	

AI for Decision Making

Intelligent Agents and Their Applications	486
<i>Jeffrey Tweedale, Nihkil Ichalkaranje</i>	
From Community Models to System Requirements: A Cooperative Multi-agents Approach	492
<i>Jung-Jin Yang, KyengWhan Jee</i>	
Scheduling Jobs on Computational Grids Using Fuzzy Particle Swarm Algorithm.	500
<i>Ajith Abraham, Hongbo Liu, Weishi Zhang, Tae-Gyu Chang</i>	
Twinned Topographic Maps for Decision Making in the Cockpit	508
<i>Steve Thatcher, Colin Fyfe</i>	
Agent-Enabled Decision Support for Information Retrieval in Technical Fields	515
<i>Gloria Phillips-Wren</i>	
Is There a Role for Artificial Intelligence in Future Electronic Support Measures?	523
<i>Phillip Fitch</i>	
Artificial Intelligence for Decision Making	531
<i>Gloria Phillips-Wren, Lakhmi Jain</i>	
Path Planning and Obstacle Avoidance for Autonomous Mobile Robots: A Review	537
<i>Voemir Kunchev, Lakhmi Jain, Vladimir Ivancevic, Anthony Finn</i>	

Intelligent Data Processing in Process Systems and Plants

Adaptive Nonlinearity Compensation of Heterodyne Laser Interferometer	545
<i>Minsuk Hong, Jaewook Jeon, Kiheon Park, Kwanho You</i>	
Study on Safety Operation Support System by Using the Risk Management Information	553
<i>Yukiyasu Shimada, Takashi Hamaguchi, Kazuhiro Takeda, Teiji Kitajima, Atsushi Aoyama, Tetsuo Fuchino</i>	
Analysis of ANFIS Model for Polymerization Process	561
<i>Hideyuki Matsumoto, Cheng Lin, Chiaki Kuroda</i>	

Semi-qualitative Encoding of Manifestations at Faults in Conductive Flow Systems	569
<i>Viorel Ariton</i>	
Design Problems of Decision Making Support System for Operation in Abnormal State of Chemical Plant	579
<i>Kazuhiro Takeda, Takashi Hamaguchi, Yukiyasu Shimada, Yoshifumi Tsuge, Hisayoshi Matsuyama</i>	
A Training System for Maintenance Personnel Based on Analogical Reasoning	587
<i>Takashi Hamaguchi, Meng Hu, Kazuhiro Takeda, Yukiyasu Shimada, Yoshihiro Hashimoto, Toshiaki Itoh</i>	
On-Line Extraction of Qualitative Movements for Monitoring Process Plants	595
<i>Yoshiyuki Yamashita</i>	

Skill Acquisition and Ubiquitous Human Computer Interaction

An Expansion of Space Affordance by Sound Beams and Tactile Indicators	603
<i>Taizo Miyachi, Jens J. Balvig, Jun Moriyama</i>	
Automatic Discovery of Basic Motion Classification Rules	611
<i>Satoshi Hori, Mizuho Sasaki, Hirokazu Taki</i>	
An Interpretation Method for Classification Trees in Bio-data Mining	620
<i>Shigeki Kozakura, Hisashi Ogawa, Hirokazu Miura, Noriyuki Matsuda, Hirokazu Taki, Satoshi Hori, Norihiro Abe</i>	
Adequate RSSI Determination Method by Making Use of SVM for Indoor Localization	628
<i>Hirokazu Miura, Junichi Sakamoto, Noriyuki Matsuda, Hirokazu Taki, Noriyuki Abe, Satoshi Hori</i>	

IATA – Intelligent Agent Technology and Applications

Ontia iJADE: An Intelligent Ontology-Based Agent Framework for Semantic Web Service	637
<i>Toby H.W. Lam, Raymond S.T. Lee</i>	

iJADE FreeWalker: An Ontology-Based Tourist Guiding System	644
<i>Toby H.W. Lam, Raymond S.T. Lee</i>	
iJADE Content Management System (CMS) – An Intelligent Multi-agent Based Content Management System with Chaotic Copyright Protection Scheme	652
<i>Raymond S.T. Lee, Eddie Chun Lun Chan, Raymond Yiu Wai Mak</i>	
Agent-Controlled Distributed Resource Sharing to Improve P2P File Exchanges in User Networks	659
<i>J.C. Burguillo-Rial, E. Costa-Montenegro, Francisco J. González-Castaño, J. Vales-Alonso</i>	
An Agent-Based System Supporting Collaborative Product Design	670
<i>Jian Xun Wang, Ming Xi Tang</i>	
Computational Intelligence Approaches and Methods for Security Engineering	
Application Presence Fingerprinting for NAT-Aware Router	678
<i>Jun Bi, Lei Zhao, Miao Zhang</i>	
Interaction for Intelligent Mobile Systems	686
<i>Gregory M.P. O’Hare, Stephen Keegan, Michael J. O’Grady</i>	
Design of a Intelligent SOAP Message Service Processor for Enhancing Mobile Web Service	694
<i>Gil-Cheol Park, Seoksoo Kim</i>	
Convergence Rate in Intelligent Self-organizing Feature Map Using Dynamic Gaussian Function	701
<i>Geuk Lee, Seoksoo Kim, Tai Hoon Kim, Min Wook Kil</i>	
Development of an Attack Packet Generator Applying an NP to the Intelligent APS	709
<i>Wankyung Kim, Wooyoung Soh</i>	
Class Based Intelligent Community System for Ubiquitous Education and Medical Information System	718
<i>Seoksoo Kim</i>	
Intelligent Anonymous Secure E-Voting Scheme	726
<i>Hee-Un Park, Dong-Myung Shin</i>	

Actively Modifying Control Flow of Program for Efficient Anomaly Detection 737
Kohei Tatara, Toshihiro Tabata, Kouichi Sakurai

Intelligent Method for Building Security Countermeasures 745
Tai-hoon Kim, Sun-myoung Hwang

Intelligent Frameworks for Encoding XML Elements Using Mining Algorithm 751
Haeng-Kon Kim

Graphical Knowledge Template of CBD Meta-model 760
Haeng-Kon Kim

Hybrid Information Technology Using Computational Intelligence

Two-Phase Identification Algorithm Based on Fuzzy Set and Voting for Intelligent Multi-sensor Data Fusion 769
Sukhoon Kang

*u*H-PMAC Model Suitable for Ubi-Home Gateway in Ubiquitous Intelligent Environment 777
Jong Hyuk Park, Sangjin Lee, Byoung-Soo Koh, Jae-Hyuk Jang

A One-Time Password Authentication Scheme for Secure Remote Access in Intelligent Home Networks 785
Ilsun You

An Intelligent and Efficient Traitor Tracing for Ubiquitous Environments 793
Deok-Gyu Lee, Seo Il Kang, Im-Yeong Lee

e-Business Agent Oriented Component Based Development for Business Intelligence 803
Ho-Jun Shin, Bo-Yeon Shim

New Design of PMU for Real-Time Security Monitoring and Control of Wide Area Intelligent System 812
Hak-Man Kim, Jin-Hong Jeon, Myoung-Chul Shin, Tae-Kyoo Oh

Security Intelligence: Web Contents Security System for Semantic Web 819
Nam-deok Cho, Eun-ser Lee, Hyun-gun Park

Performance Analysis of Location Estimation Algorithm Using an Intelligent Coordination Scheme in RTLS	829
<i>Seung-Hee Jeong, Hyun-Jae Lee, Joon-Sung Lee, Chang-Heon Oh</i>	

Security Requirements for Ubiquitous Software Development Site	836
<i>Tai-hoon Kim</i>	

Logic Based Intelligent Information Systems

Paraconsistent Artificial Neural Network: Applicability in Computer Analysis of Speech Productions	844
<i>Jair Minoro Abe, João Carlos Almeida Prado, Kazumi Nakamatsu</i>	

Intelligent Paraconsistent Logic Controller and Autonomous Mobile Robot Emmy II	851
<i>Jair Minora Abe, Cláudio Rodrigo Torres, Germano L. Torres, Kazumi Nakamatsu, Michiro Kondo</i>	

EVALPSN Based Intelligent Drivers' Model	858
<i>Kazumi Nakamatsu, Michiro Kondo, Jair M. Abe</i>	

The Study of the Robust Learning Algorithm for Neural Networks	866
<i>Shigenobu Yamawaki</i>	

Logic Determined by Boolean Algebras with Conjugate	871
<i>Michiro Kondo, Kazumi Nakamatsu, Jair Minoro Abe</i>	

An Intelligent Technique Based on Petri Nets for Diagnosability Enhancement of Discrete Event Systems	879
<i>YuanLin Wen, MuDer Jeng, LiDer Jeng, Fan Pei-Shu</i>	

Knowledge-Based Mult-criteria Decision Support

Fuzzy Logic Based Mobility Management for 4G Heterogeneous Networks	888
<i>Jin-Long Wang, Chen-Wen Chen</i>	

On-Line Association Rules Mining with Dynamic Support	896
<i>Hsuan-Shih Lee</i>	

A Fuzzy Multiple Criteria Decision Making Model for Airline Competitiveness Evaluation 902
Hsuan-Shih Lee, Ming-Tao Chou

Goal Programming Methods for Constructing Additive Consistency Fuzzy Preference Relations 910
Hsuan-Shih Lee, Wei-Kuo Tseng

A Multiple Criteria Decision Making Model Based on Fuzzy Multiple Objective DEA 917
Hsuan-Shih Lee, Chen-Huei Yeh

A Fuzzy Multiple Objective DEA for the Human Development Index 922
Hsuan-Shih Lee, Kuang Lin, Hsin-Hsiung Fang

Neural Information Processing for Data Mining

Visualization Architecture Based on SOM for Two-Class Sequential Data 929
Ken-ichi Fukui, Kazumi Saito, Masahiro Kimura, Masayuki Numao

Approximate Solutions for the Influence Maximization Problem in a Social Network 937
Masahiro Kimura, Kazumi Saito

Improving Convergence Performance of PageRank Computation Based on Step-Length Calculation Approach 945
Kazumi Saito, Ryohei Nakano

Prediction of the O-glycosylation Sites in Protein by Layered Neural Networks and Support Vector Machines 953
Ikuko Nishikawa, Hirotaka Sakamoto, Ikue Nouno, Takeshi Iritani, Kazutoshi Sakakibara, Masahiro Ito

A Bayesian Approach to Emotion Detection in Dialogist’s Voice for Human Robot Interaction 961
Shohei Kato, Yoshiki Sugino, Hidenori Itoh

Finding Nominally Conditioned Multivariate Polynomials Using a Four-Layer Perceptron Having Shared Weights 969
Yusuke Tanahashi, Kazumi Saito, Daisuke Kitakoshi, Ryohei Nakano

Sharing of Learning Knowledge in an Information Age

Development of Know-How Information Sharing System in Care Planning Processes – Mapping New Care Plan into Two-Dimensional Document Space	977
<i>Kaoru Eto, Tatsunori Matsui, Yasuo Kabasawa</i>	
Educational Evaluation of Intelligent and Creative Ability with Computer Games: A Case Study for e-Learning	985
<i>Yukuo Isomoto</i>	
The Supporting System for Distant Learning Students of Shinshu University	994
<i>Hisayoshi Kunimune, Mika Ushiro, Masaaki Niimura, Yasushi Fuwa</i>	
Reflection by Knowledge Publishing	1002
<i>Akihiro Kashihara, Yasuhiro Kamoshita</i>	
Self-learning System Using Lecture Information and Biological Data	1010
<i>Yurie Iribe, Shuji Shinohara, Kyoichi Matsuura, Kouichi Katsurada, Tsuneo Nitta</i>	
Hybrid Approach of Augmented Classroom Environment with Digital Pens and Personal Handhelds	1019
<i>Motoki Miura, Susumu Kunifuji</i>	
An Interactive Multimedia Instruction System: IMPRESSION for Multipoint Synchronous Online Classroom Environment	1027
<i>Yuki Higuchi, Takashi Mitsuishi, Kentaro Go</i>	
A System Framework for Bookmark Sharing Considering Differences in Retrieval Purposes	1035
<i>Shoichi Nakamura, Maiko Ito, Hirokazu Shirai, Emi Igarashi, Setsuo Yokoyama, and Youzou Miyadera</i>	
Development of a Planisphere Type Astronomy Education Web System Based on a Constellation Database Using Ajax	1045
<i>Katsuhiko Mouri, Mamoru Endo, Kumiko Iwazaki, Manabu Noda, Takami Yasuda, Shigeki Yokoi</i>	
Group Collaboration Support in Learning Mathematics	1053
<i>Tomoko Kojiri, Yosuke Murase, Toyohide Watanabe</i>	

Annotation Interpretation of Collaborative Learning History
for Self-learning 1062
Masahide Kakehi, Tomoko Kojiri, Toyohide Watanabe

A System Assisting Acquisition of Japanese Expressions Through
Read-Write-Hear-Speaking and Comparing Between Use Cases
of Relevant Expressions 1071
*Kohji Itoh, Hiroshi Nakamura, Shunsuke Unno,
Jun'ichi Kakegawa*

**Intelligent Information Management
in a Knowledge-Based Society**

A Web-Based System for Gathering and Sharing Experience
and Knowledge Information in Local Crime Prevention 1079
Masato Goto, Akira Hattori, Takami Yasuda, Shigeki Yokoi

The Scheme Design for Active Information System 1087
Ping Zong, Jun Qin

R-Tree Based Optimization Algorithm for Dynamic Transport
Problem 1095
Naoto Mukai, Toyohide Watanabe, Jun Feng

Knowledge-Based System for Die Configuration Design in Cold
Forging 1103
*Osamu Takata, Tsubasa Mitani, Yuji Mure, Masanobu Umeda,
Isao Nagasawa*

An Automatic Indexing Approach for Private Photo Searching Based
on E-mail Archive 1111
Taketoshi Ushiyama, Toyohide Watanabe

Analysis for Usage of Routing Panels in Evacuation 1119
Toyohide Watanabe, Naoki Harashina

**Knowledge/Software Engineering Aspects
of Intelligent Systems Applications**

Facial Expression Classification: Specifying Requirements
for an Automated System 1128
Ioanna-Ourania Stathopoulou, George A. Tsihrintzis

On the Software Engineering Aspects of Educational Intelligence 1136
Thanasis Hadzilacos, Dimitris Kalles

Feature Model Based on Description Logics	1144
<i>Shaofeng Fan, Naixiao Zhang</i>	

PNS: Personalized Multi-source News Delivery	1152
<i>Georgios Paliouras, Alexandros Mouzakidis, Christos Ntoutsis, Angelos Alexopoulos, Christos Skourlas</i>	

Knowledge-Based Interface Systems (2)

Relational Association Mining Based on Structural Analysis of Saturation Clauses	1162
<i>Nobuhiro Inuzuka, Jun-ichi Motoyama, Tomofumi Nakano</i>	

Examination of Effects of Character Size on Accuracy of Writer Recognition by New Local Arc Method	1170
<i>Masahiro Ozaki, Yoshinori Adachi, Naohiro Ishii</i>	

Study of Features of Problem Group and Prediction of Understanding Level	1176
<i>Yoshinori Adachi, Masahiro Ozaki, Yuji Iwahori</i>	

Intelligent Databases in the Virtual Information Community

Graph-Based Data Model for the Content Representation of Multimedia Data	1182
<i>Teruhisa Hochin</i>	

NCO-Tree: A Spatio-temporal Access Method for Segment-Based Tracking of Moving Objects	1191
<i>Yuelong Zhu, Xiang Ren, Jun Feng</i>	

The Reasoning and Analysis of Spatial Direction Relation Based on Voronoi Diagram	1199
<i>Yongqing Yang, Jun Feng, Zhijian Wang</i>	

Some Experiments of Face Annotation Based on Latent Semantic Indexing in FIARS	1208
<i>Hideaki Ito, Hiroyasu Koshimizu</i>	

Extended Virtual Type for a Multiple-Type Object with Repeating Types	1216
<i>Hideki Sato, Masayoshi Aritsugi</i>	

Spatial Relation for Geometrical / Topological Map Retrieval 1224
*Toru Shimizu, Masakazu Ikezaki, Toyohide Watanabe,
 Taketoshi Ushiyama*

Geographical Information Structure for Managing a Set of Objects
 as an Event 1232
*Masakazu Ikezaki, Toyohide Watanabe,
 Taketoshi Ushiyama*

**Hybrid Intelligent Systems in Medicine and Health
 Care**

Using Multi-agent Systems to Manage Community
 Care 1240
Martin D. Beer, Richard Hill

Predictive Adaptive Control of the Bispectral Index of the EEG (BIS)
 – Using the Intravenous Anaesthetic Drug Propofol 1248
*Catarina S. Nunes, Teresa F. Mendonça, Hugo Magalhães,
 João M. Lemos, Pedro Amorim*

Using Aggregation Operators to Personalize Agent-Based Medical
 Services 1256
David Isern, Aïda Valls, Antonio Moreno

Shifting Patterns Discovery in Microarrays with Evolutionary
 Algorithms 1264
Beatriz Pontes, Raúl Giráldez, Jesús S. Aguilar-Ruiz

Gene Ranking from Microarray Data for Cancer Classification–A
 Machine Learning Approach 1272
*Roberto Ruiz, Beatriz Pontes, Raúl Giráldez,
 Jesús S. Aguilar-Ruiz*

H³: A Hybrid Handheld Healthcare Framework 1281
Seung-won Hwang

Hybrid Intelligent Medical Tutor for Atheromatosis 1289
Katerina Kabassi, Maria Virvou, George Tsihrintzis

Evolutionary Tuning of Combined Multiple
 Models 1297
Gregor Stiglic, Peter Kokol

A Similarity Search Algorithm to Predict Protein Structures	1305
<i>Jiyuan An, Yi-Ping Phoebe Chen</i>	
Fuzzy-Evolutionary Synergism in an Intelligent Medical Diagnosis System	1313
<i>Constantinos Koutsojannis, Ioannis Hatzilygeroudis</i>	
Author Index	1323

Intensity Modulated Radiotherapy Target Volume Definition by Means of Wavelet Segmentation

Tsair-Fwu Lee¹, Pei-Ju Chao², Fu-Min Fang¹, Eng-Yen Huang¹,
and Ying-Chen Chang²

¹ Chang Gung Memorial Hospital-Kaohsiung, 83342, Taiwan, ROC

² Kaohsiung Yuan's General Hospital, Kaohsiung, 800, Taiwan, ROC
alf@adm.cgmh.org.tw, pjchao@bit.kuas.edu.tw

Abstract. This study aimed to develop an advance precision three-dimensional (3-D) image segmentation algorithm to enhance the blurred edges clearly and then introduce the result onto the intensity modulated radiotherapy (IMRT) for tumor target volume definition. This will achieve what physicians usually demand that tumor doses escalation characteristics of IMRT. A proposed algorithm flowchart designed for this precision 3-D treatment targeting was introduced in this paper. Different medical images were used to test the validity of the proposed method. The 3-D wavelet based targeting preprocessing segmentation allows physicians to improve the traditional treatments or IMRT much more accurately and effectively. This will play an important role in image-guided radiotherapy (IGRT) and many other medical applications in the future.

Keywords: intensity modulated radiotherapy, target volume, wavelet, segmentation.

1 Introduction

Three-dimensional (3-D) conformal therapy has been used in most cancer patients receiving radiotherapy. The role of intensity modulated radiotherapy (IMRT) is well established due to its tumor doses escalation characteristics. The goal is to deliver as much radiation as possible to a tumor while sparing nearby normal tissue—especially critical, but radiation-sensitive organs such as the spinal cord or rectum [1-4]. Physicians need 3-D renderings to help them make diagnoses, conduct surgery, and perform radiation therapy which two-dimensional (2-D) images usually cannot offer. However, without precision segmentation these 3-D renderings could lead to misleading results. The aim of this study is to provide a precision 3-D segmentation method to achieve what physicians demanded. They will also become the preprocessing reference data to intensity modulated radiotherapy systems.

Since conventional medical images of computed tomography (CT) or magnetic resonance of imaging (MRI) although appeared to be 3-D images, they are all composed by slice-based 3-D image datasets. One effective way to obtain precision 3-D segmentation reconstruction rendering is to process the sliced data with high precision first [5, 6]. In this study, the proposed algorithm is to apply the wavelet segmentation

approaches in a maximize entropy sense. This allows us to utilize all available information to achieve the most robust segmentation results for 3-D image reconstruction. We then apply the segmentation method to medical images including two CT scan image and one MRI image to test the validity of our method and to compare the precision with a conventional segmentation approach. We aim to show the 3-D segmentation method is superior in precision with only a reasonable amount of computing time. From the mathematical viewpoint, since images are 2-D arrays of intensity values with locally varying statistics, different combinations of abrupt features like edges and contrasting homogeneous regions are better to process with wavelet-based transformations which is known to have the advantage of multi-resolutions. We shall see this indeed was feasible and the results are quite satisfactory.

2 Wavelet Segmentations

In recent years, wavelet theory is widely used in various signal-processing problems. Its great flexibility makes it the most desired signal processing technique in many applications. In this section we introduced the idea of multiresolution first and then developed the Discrete Wavelet Transform (DWT). We then extended the DWT into two dimensions and derived a two-dimensional Discrete Wavelet Transform (2-D DWT) by separate algorithms. Finally, we should develop the wavelet edge detector from the 2-D DWT.

The Wavelet method is known to be one of the best gradient segmentation methods due to its multi-scale and multi-resolution capabilities. Assume $S_{2^j}[\]$ and $D_{2^j}[\]$ as the low pass signal and the high pass signal of $f(x)$ at resolution 2^j respectively, and $S_{2^j}[n] = \langle f(u), \phi_{2^j}(u - 2^{-j}n) \rangle$ is the projection coefficient of $f(x)$ on V_j , $D_{2^j}[n] = \langle f(u), \varphi_{2^j}(u - 2^{-j}n) \rangle$ is the projection coefficient of $f(x)$ on O_j . We can define an orthogonal complement subspace of V_j as O_j , in space V_{j+1} . In other words, $O_j \perp V_j$ and $O_j \oplus V_j = V_{j+1}$, where V_j is the expansion of $\phi(x)$ by basis $\sqrt{2^{-j}}\phi_{2^j}(x - 2^{-j}n)_{n \in \mathbb{Z}}$, O_j is the expansion of $\varphi(x)$ by basis $\sqrt{2^{-j}}\varphi_{2^j}(x - 2^{-j}n)_{n \in \mathbb{Z}}$, \oplus denotes the union of space (like the union of sets) and \perp denotes two sets are orthogonal. The scaling function $\phi(x)$ and wavelet function have the orthogonal properties as shown in [7-9]. By the properties of multi-resolutions, a signal can always be decomposed into higher resolutions until a desired result is reached. This can be interpreted by tree architectures known as the Pyramid architecture [9]. Hence we may create a 2-D filter for edge detection by replacing the traditional 2-D wavelet functions with a 2-D discrete periodic wavelet transform (2-D DPWT) [8]. The 2-D DPWT can be written in the matrix form as follows: $[SS_{j+1}]_{N_1 \times N_1} = W_{LL} \otimes [SS_j]_{N_2 \times N_2}$, $[SD_{j+1}]_{N_1 \times N_1} = W_{LH} \otimes [SS_j]_{N_2 \times N_2}$, $[DS_{j+1}]_{N_1 \times N_1} = W_{HL} \otimes [SS_j]_{N_2 \times N_2}$, and $[DD_{j+1}]_{N_1 \times N_1} = W_{HH} \otimes [SS_j]_{N_2 \times N_2}$. Where $N_1 = 2^{|j+1|}$, $N_2 = 2^{|j|}$, W_{LL} , W_{LH} , W_{HL} and W_{HH} are the four subband filters; \otimes denoted a convolution operation; $S_{2^j}[\]$ is the low pass signal, or the approximated signal; $D_{2^j}[\]$ is the high pass signal, or the detailed

signal of $f(x)$ at resolution 2^j respectively. For 2-D images, they are in the form of $[SS_{j+1}]_{N_1 \times N_1}$ which come from $[SS_j]_{N_2 \times N_2}$ convoluted with the corresponding subband filter W_{LL} , where $[SS_j]_{N_2 \times N_2}$ is a matrix form the expanded image. The definition of the four subband filters' operators of 2-D DPWT are [7-9]: $W_{LL} = [h(i) \cdot h(j)]_{i,j \in Z}$, $W_{LH} = [(-1)^{3-j} h(i) \cdot h(3-j)]_{i,j \in Z}$, $W_{HL} = [(-1)^{3-j} h(3-i) \cdot h(j)]_{i,j \in Z}$, and $W_{HH} = [(-1)^{i+j} h(3-i) \cdot h(3-j)]_{i,j \in Z}$. Where $h(i) = \langle \phi_{2^{-i}}(u) \cdot \phi(u-i) \rangle$. Clearly, since the coefficients of the filter have length d , the operator of 2-D DPWT formed a $d \times d$ matrix. We now use the coefficients of the four filters given by the above equations to generate a wavelet edge detector.

Let $f_h(i, j)$ be the horizontal high-pass filter function and $f_v(i, j)$ be the vertical high-pass filter function obtained from the four operators of 2-D DPWT

$$f_h(i, j) = W_{LL}(i, j) \otimes W_{LH}(i, j) \quad (1)$$

$$f_v(i, j) = W_{LL}(i, j) \otimes W_{HL}(i, j) \quad (2)$$

We now apply the different length coefficients of Daubechies wavelet transform to generate multi-scale masks for images segmentation in different scales [7-9].

Next, let the original image pass through these masks to produce a series of multi-scale images with different gradient strengths. In order to avoid distortions caused by noise and to define exact edge points, an edge thinning technique is then used to make effective determination of the images. This can be done by observe a line through each pixel along its gradient direction first. If the pixel is local maximum along that line, we retain the pixel, or else we suppress it. We shall complete the 2-D image of each slice and then join the edges to form a surface. The advantage of processing 2-D images this way is obvious. Both storage and computation time were much less and parallel processing can be easily done.

3 Flowchart of the Proposed Method

The goal of any precision image segmentation is to partition an image into disjoint regions of desired objects as accurate as possible. The multi-resolution nature of a wavelet operator allows us to detect edges at different scales. In the wavelet edge detection algorithms, the transformation uses DPWT and the filter searches for local maximal in the wavelet domain. Many earlier researches proved that the wavelet edge detector can detect very complicated edges [10-13]. Therefore if we improve the wavelet transform further in the maximal entropy manner, we expect better results beyond any segmentation method are capable alone. The following descriptions are our precision 3-D treatment targeting reconstruction renderings applied to medical images [14]. The treatment targeting rendering algorithm which utilized the wavelet segmentation approaches is shown in Fig. 1. Figure 2 is a slice of the CT images of human chest. We then compared the quality of the segmentation results with a traditional method named region growing to show how the proposed algorithm did better in precision targeting. The processing steps are:

- 1) 2-D CT slices of human chests are very common in medical examinations, but they usually differ greatly in illumination contrasts. Therefore normalization is needed. CT images of 512*512*16 bits in the DICOM (Digital Image and Communication in Medicine) format bits are now normalized between 0 and 1.
- 2) Next we make a Region-of-Interest (ROI) selection. The ROI shall be our *a priori* knowledge in the segmentation algorithm. It defined the object we desire to inspect, whereas the rest of the image is treated as backgrounds. We then select a mask roughly covers the ROI. In this case the right lung is our desired object in this experiment, ROI and a mask of an initial shape covers the right lung is formed. At the same time the image were send through a bank of homomorphic filters and lower-upper-middle filters to smooth and sharpen the object that we want to segment.
- 3) The ROI partitioned image is now cut out. It shall be processed separately to increase efficiency.
- 4) We process the partitioned image via wavelet edge detection and region-growing segmentation [15] separately for comparison.
- 5) The wavelet segmented images is now selected for comparing pixel-by-pixel with their *a posteriori* probability to find the true edge.

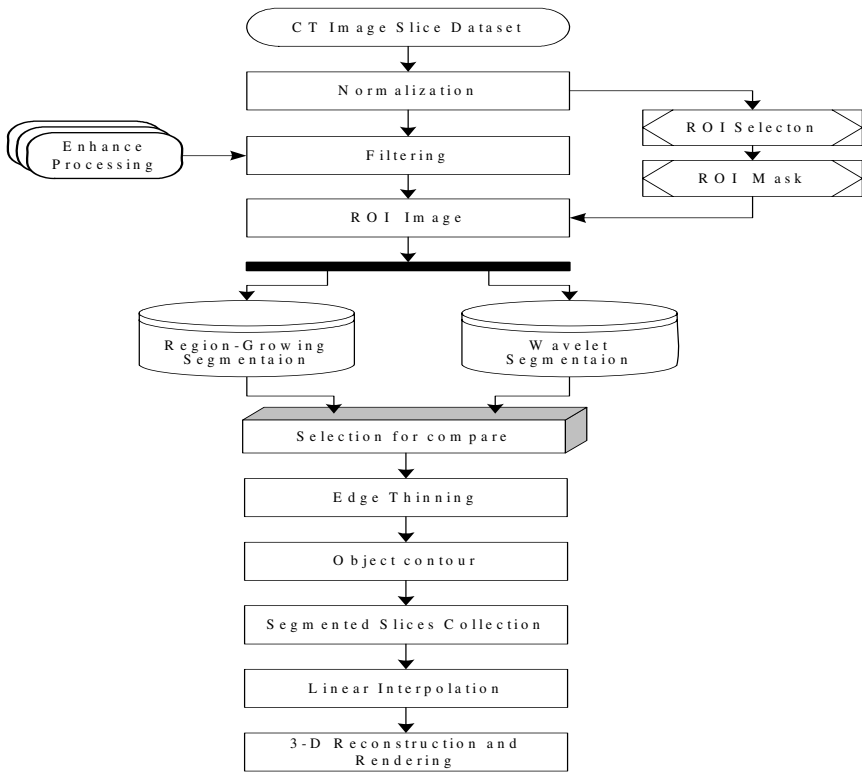


Fig. 1. The flowchart of the 3-D treatment target reconstruction rendering

- 6) Edge thinning comes next. Since our method pertains pixels with the same or similar strength, it may result in thick edges. But our objective is to segment the target precisely. Hence nearest neighborhood algorithm is next applied to create a thinning process to obtain the contour of the desired object. The contour is then inserted back into the original image.
- 7) As the contour is inserted back into the original image, there will be missing areas plus contrast differences. Dilation algorithm and Erosion algorithm were then applied to fill the gaps and to obtain the object respectively.
- 8) Slices with segmented objects are now collected for 3-D reconstruction.
- 9) We process the segmented slices with linear interpolations to form a 3-D rendering of the object.
- 10) Tiling algorithm which patches the stacking contours is now applied to form surfaces [13, 16]. The 3-D reconstruction is now completed and the target shows out for treatment procedure.

4 Experiment Results

In the first experiment, original slices of a human chest from CT scan are used, with one of the slices enlarged in Fig. 2 (a). We performed the 2-D segmentation first slice by slice. Since we aim to extract the right lung for feature extraction it becomes our natural ROI selected by using the cursor which is shown in the Fig. 2 (b). Fig. 2 (c) shows the region-growing contour; Fig. 2 (d) illuminated the wavelet contour; Figs. 3 (a) (b) showed the segmentation results of Figs. 2 (b), (c) respectively. On close inspection of these figures, shortcomings of lack of smoothness from both our wavelet segmentation method and that from the traditional region-growing method were obvious, but seemed to be harmless. However, as 3-D renderings were formed, they create errors, and shall not be tolerated if precision renderings were sought. Fig. 3 (c) shows the 3-D rendering from the region-growing segmentation. Fig. 3 (d) shows the 3-D rendering from the wavelet segmentation. If we look at them closely, we found that the 3-D renderings by the region-growing segmentation, seemed acceptable, but it has problems that at times a single slice will be very different from the others at some particular point due to noise and/or other disturbances, which makes the corresponding 3-D renderings appear with wrinkles. Clearly human lungs should be continuous and smooth in all directions always; hence we conclude that the region-growing method is unable to reconstruct the object precisely. On the other hand, our wavelet segmentation with maximum entropy approaches creates a 3-D reconstruction rendering much more correctly. Hence the proposed method is superior to many current methods.

The next experiment is processing a CT scan of a female patient with a pituitary tumor in her brain. The pituitary gland is about the size of a pea at the center of brain just at the back of the human nose. It makes hormones that affect growths and functions of other glands. Tumors that make infectious hormones are known as functioning tumors, which are most deadly. The choice of treatment on this kind of tumor uniquely depends on the position and orientation of the tumor. With a target so small and so vital, only position of the tumor pin down with the highest precision, treatments can then be effective, and ordinary brain cells can be spared the doses.

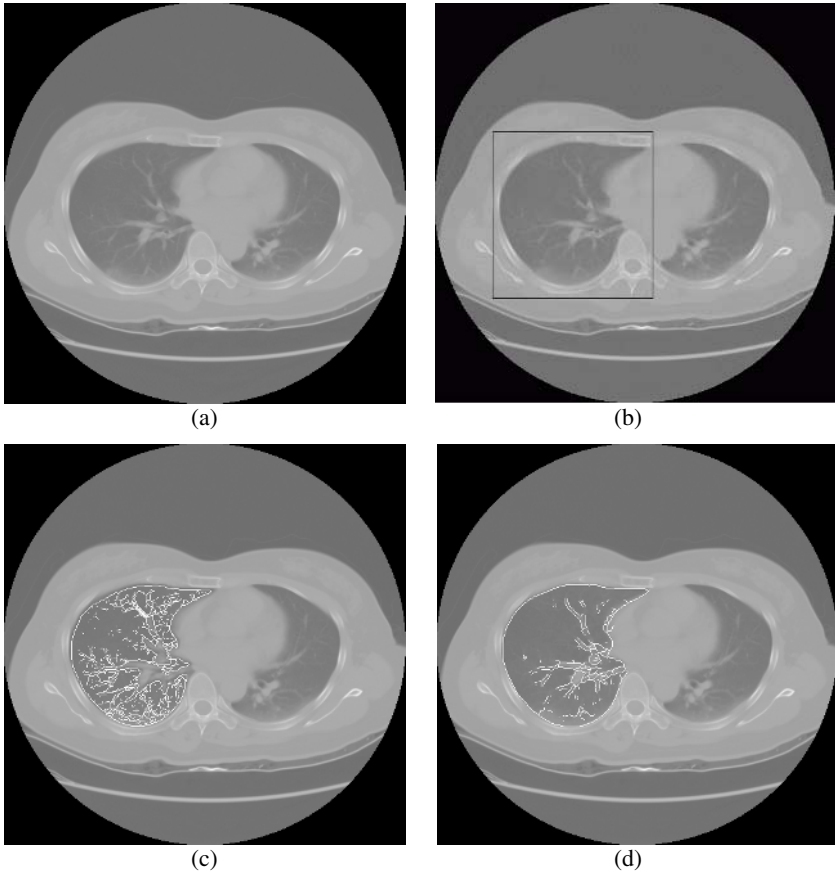


Fig. 2. (a) A selected original image. (b) A selected ROI in the original image. (c) Segmented contour by the region-growing method. (d) Segmented contour by the wavelet method (white line).

The original slices are shown in the Fig. 3 (a) where the tumor can hardly be seen visually. We first segmented it out with great precisions, and then insert it back into the original images with a strong contrast as shown in Fig. 3 (b). The tumor can now be clearly inspected. Comparison of Fig. 3 (a) and 3 (b) clearly demonstrated the advantage of our segmentation; it provided an outstanding positioning of the tumor which has not been achieved by other 3-D renderings so previously.

3-D treatment targeting segmentation renderings of the pituitary tumor are next reconstructed. Its various angles are shown in Fig. 5 (a-h). We shall test the accuracy of localization of our method by using coloring and a so-called transparency technique, we shall find not only we can identify the problem area precisely, but also their relative positions to other critical organs are all clearly shown. The size and shape of the tumor, its orientation with the brain, and the position relative to the head are all now

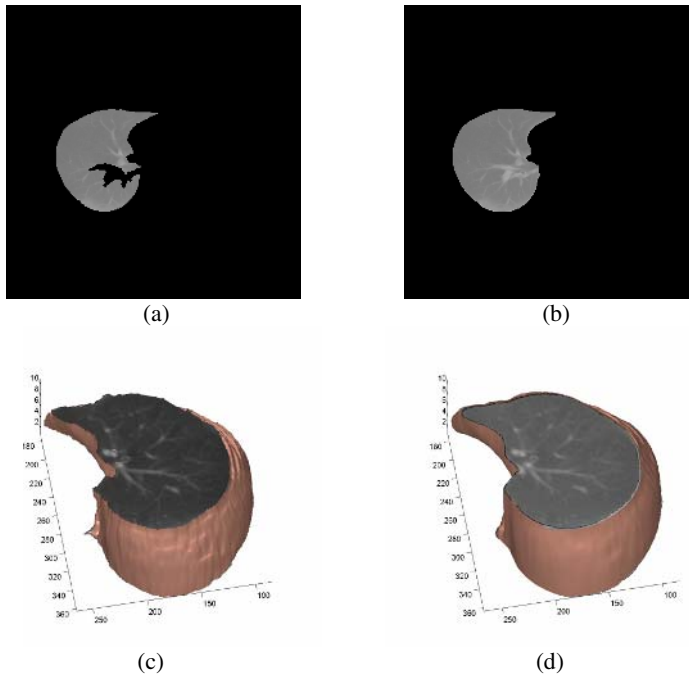


Fig. 3. (a) Segmented result of the region-growing method. (b) Segmented result of the wavelet method. (c) 3-D target reconstruction rendering by the region-growing method. (d) 3-D target reconstruction rendering by wavelet method.

clearly seen. These 3-D treatment targeting segmentation renderings can be rotated to any angle and with different colors for inspections. Transparency effect is now introduced in the last line. They shall be most useful for intensity modulated radiotherapies.

In the next example, we demonstrate further how colors were used in our 3-D wavelet based segmentation renderings to enhance the important parts. Slices of a human head from a MRI modality are used in this experiment. Our 3-D rendering method allows us to display the skull of the patient, or his brain, or both in all angles. We obtain a 3-D rendering of the skull quickly by noticing that the skull bone has a special feature of being dense and uniform, which clearly stands out in grey levels in each 2-D slice, segmentations hence becomes easy. The original slices of a human head from a MRI modality are reconstructed in Fig. 6(a). The result 3-D renderings of different angels of the skull are shown in Fig. 6(b).

To obtain a 3-D rendering of the brain, the skull rendering successfully constructed is now removed from the original image as shown in Fig. 6 (c). Finally, 3-D reconstruction of the brain in a partial skull is shown in Fig. 6 (d) with different colors to emphasize their relative positions. The superiority of our algorithm is now demonstrated by real medical images in terms of precision and efficiency.

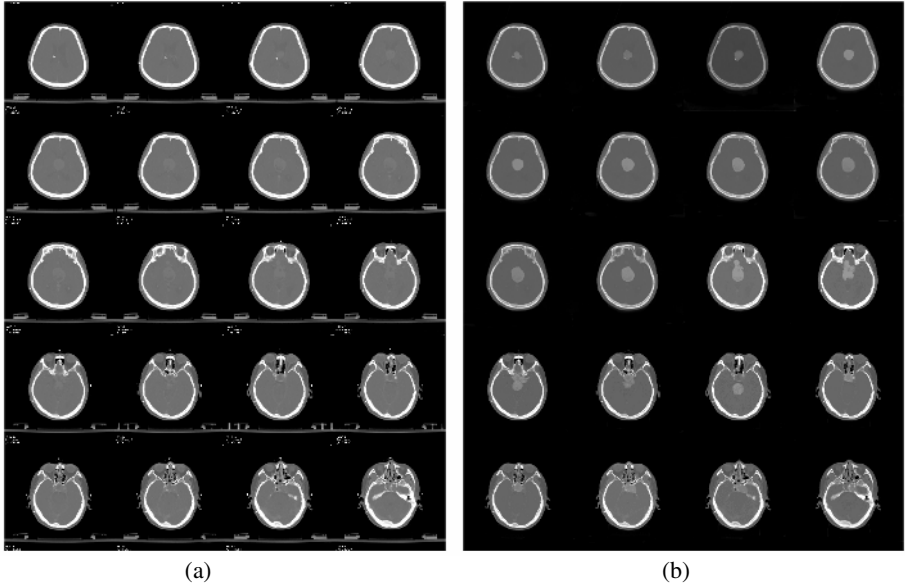


Fig. 4. (a) The original slices from a CT modality. (b) The enhanced tumor image inserted back into the original image slices.

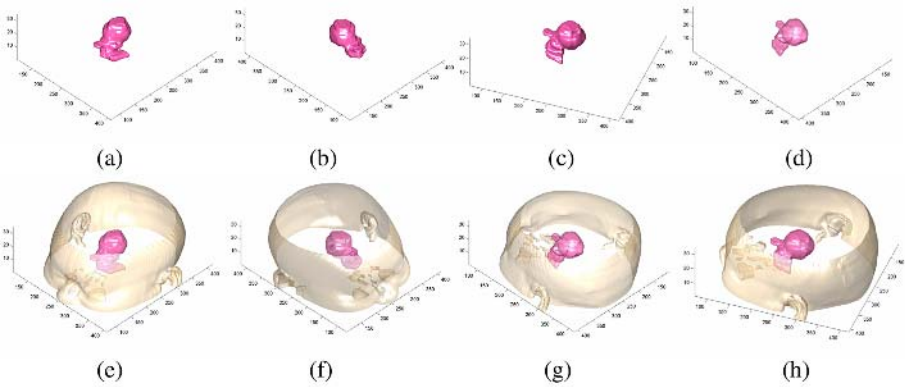


Fig. 5. 3-D treatment target reconstruction renderings of the tumor with transparent effect

We have successfully completed 3-D medical renderings with great precisions. These images although stacked by 2-D images, but resulting from processing by a maximum entropy method, hence mathematically optimal in statistical sense. We further prove our method is indeed practical by processing real medical images.

On all examples of medical images we processed, not only desired precision had been achieved, we were also able to create rotation of the objects to obtain its 3-D

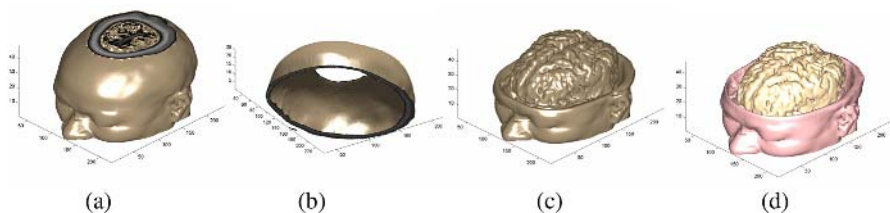


Fig. 6. (a) 3-D reconstruction of the original slices. (b) 3-D reconstruction of the skull. (c) 3-D reconstruction of the brain in an imaginary opened head. (d) An imaginary opened head in different colors.

images of different angles. We were also capable to show layers and features, organs and bones. We could emphasize the interested areas by adding various colors to them on purposes, or we could diminish the less important parts by applying the transparent effect. The 3-D renderings we created will allow physicians to conduct surgery or treatment much more accurately and effectively.

We have presented our results to seven doctors of the Department of Radiation Oncology. They examined them closely and all agreed they were superior than any rendering they have seen and encourage this line of research should proceed further as soon as possible.

Although our rendering method is more advance in many ways, but it still far from direct medical applications. The processing required only an ordinary PC but the processing time is considerable. This shortcoming makes it inapplicable medically for places like the emergency room where immediate results are required.

Applying our method for IMRT treatment for tumors has not been tested also. Our images no doubt shall be a great help for the physicians, but medical decisions and responsibilities lies on the shoulder of them. Whether they can totally rely on our images can be only evaluated after years of practical use.

5 Conclusions

For medical renderings, precision is our primary concern; we choose the wavelet method for its great multi-scale and multi-resolution characteristic to process 2-D slices in sequence. We further improve the wavelet method by introducing the maximum entropy sense then ensured improved accuracies. Linear interpolation was then used to form 3-D renderings, also proved to be effective and accurate. Many images of interest that physicians unable to visualize are now clearly identified, locations pinned down exactly, and relative orientations are now well understood. These are all vital for medical treatments. The preprocessing treatment targeting segmentation algorithm we developed can be extended to IMRT or image-guided radiotherapy (IGRT) easily. Features are now clearly identified, locations and size pinned down exactly, and relative orientations are now well understood. These are all definitely required for IMRT and IGRT treatments. We believe our precise 3-D treatment targeting segmentation method shall play an important role in future medical application.

References

1. Ezzell GA, Galvin JM, Low D, et al.: Guidance document on delivery, treatment planning, and clinical implementation of IMRT, Report of the IMRT Subcommittee of the AAPM Radiation Therapy Committee. *Med Phys* 2003; 30:2089-2115.
2. Intensity Modulated Radiation Therapy Collaborative Working Group (IMRTCWG). Intensity modulated radiotherapy: Current status and issues of interest. *International Journal Radiation Oncology Biology Phys* 2001; 51:880-914.
3. LoSasso T, Chui CS, and Ling CC: Comprehensive quality assurance for the delivery of intensity modulated radiotherapy with a multileaf collimator used in the dynamic mode. *Med Phys* 2001; 28:2209-2219.
4. Low DA: Quality assurance of intensity-modulated radiotherapy. *Semin Radiat Oncol* 2002; 12:219-228.
5. Chakraborty A: Feature and Module Integration for Image Segmentation. PhD thesis, Yale University, 1996; 89-185.
6. Gabriele Lohmann: Volumetric Image Analysis. Wiley & Teubner, 1998; 34-128.
7. Ku CT, Hung KC and Liag MC: Wavelet Operators for Multi-scale Edge and Corner Detection. Master thesis, Department of Electrical Engineering, I-Shou University, Taiwan, 1998; 4-65.
8. Leu YS, Chou CJ: Wavelet Edge Detection on Region-based Image Segmentation. Master thesis, Department of Computer & Communication Engineering, National Kaohsiung First University of Science and Technology, Taiwan, 2000; 8-27
9. Mallat SG: Multifrequency Channel Decompositions of Images and Wavelet Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, December 1989; 37. 12-17.
10. Canny JF: A Computational approach to edge-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986; 8:679-698.
11. Gonzalez RC, Woods RE: *Digital Image Processing*. Prentice Hall, 2nd ed. Edition, 2002; 349-405.
12. Kitchen L, and Rosenfeld A: Edge Evaluation Using Local Edge Coherence. *IEEE Transactions on Systems, Man, and Cybernetics*. 1981; SMC-11. 9, 597-605.
13. Russ JC: *The Image Processing Handbook*, Third ed. CRC Press & IEEE Press, 1999; 23-138.
14. Tsair-Fwu Lee, Ming-Yuan Cho: "Precise Segmentation Rendering for Medical Images Based on Maximum Entropy Processing", *Lecture Notes in Computer Science*, vol.3683, Springer-Verlag, 2005, pp.366-373.
15. Rafael C. Gonzalez, Richard E. Woods. *Digital Image Processing*. Prentice Hall, 2nd ed. Edition, 2002. pp.612-617.
16. John C. Russ. *The Image Processing Handbook*, Third ed. CRC Press & IEEE Press, 1999.

Enhancing Global and Local Contrast for Image Using Discrete Stationary Wavelet Transform and Simulated Annealing Algorithm

Changjiang Zhang, C.J. Duanmu, and Xiaodong Wang

Dept. of Information Science and Engineering, Zhejiang Normal University, Postcode
321004 Jinhua, China

{zcyj74922, duanmu, wxd}@zjnu.cn

Abstract. After the discrete stationary wavelet transform (DSWT) combined with the generalized cross validation (GCV) for an image, the noise in the image is directly reduced in the high frequency sub-bands, which are at the high-resolution levels. Then the local contrast of the image is enhanced by combining de-noising method with in-complete Beta transform (IBT) in the high frequency sub-bands, which are at the low-resolution levels. In order to enhance the global contrast for the image, the low frequency sub-band image is also processed by combining the IBT and the simulated annealing algorithm (SA). The IBT is used to obtain the non-linear gray transform curve. The transform parameters are determined by the SA so as to obtain the optimal non-linear gray transform parameters. In order to reduce the large computational requirements of traditional contrast enhancement algorithms, a new criterion is proposed with the gray level histogram. The contrast type for an original image is determined by employing the new criterion. The gray transform parameters space is respectively given according to different contrast types, which greatly shrinks gray transform parameters space. Finally, the quality of the enhanced image is evaluated by a new overall objective criterion. Experimental results demonstrate that the new algorithm can greatly improve the global and local contrast for an image while efficiently reducing gauss white noise (GWN) in the image. The new algorithm performs better than the histogram equalization (HE) algorithm, un-sharpened mask algorithm (USM), Tubbs's algorithm [2], Gong's algorithm [3] and Wu's algorithm [4].

1 Introduction

Traditional kinds of image enhancement algorithms include: point operators, space operators, transform operators and pseu-color contrast enhancement [1]. Tubbs gives a gray transform algorithm to enhance contrast for images [2]. However, the computation complexity of the algorithm is large. Based on Tubbs's algorithm, Zhou presented a new kind of genetic algorithm to optimize the non-linear transform parameters. Although it can well enhance the contrast for images, its computation requirement is still large. Many existing enhancement algorithms' intelligence and adaptability are bad and much artificial interference is required. To solve the above problems, a new algorithm

employing IBT, DSWT, and SA is proposed in this paper. To improve the optimization speed and intelligence of the algorithm, a new criterion is proposed based on the gray level histogram. The contrast type for an original image is determined by employing the new criterion, where the contrast for original images are classified into seven types: particular dark (PD), medium dark (MD), medium dark slightly (MDS), medium bright slightly (MBS), medium bright (MB), particular bright (PB) and good gray level distribution (GGLD). The IBT operator transforms an original image to a new space. A certain criterion function is used to optimize the non-linear transform parameters. SA is used to determine the optimal non-linear transform parameters. After the DSWT of the original image, the global contrast is enhanced directly by employing the IBT in the low frequency sub-band image. We expand the IBT in the DSWT domain so as to enhance the details in an original image. The noise is reduced directly with better resolution levels of the processed image by the de-noising algorithm. The de-noising asymptotic thresholds can be obtained by employing the GCV. Combining the de-noising algorithm with IBT enhances local contrast. In order to evaluate the quality of the enhanced image, a new overall objective criterion is proposed. Experimental results demonstrate that the proposed algorithm performs better than the histogram equalization (HE), un-sharpened mask algorithm (USM), Tubbs's algorithm.[2], Gong's algorithm [3], and Wu's algorithm.[4].

2 Transform Parameters Optimization with IBT and SA

In Tubbs's algorithm, it uses the unitary incomplete Beta function to approximate the non-linear gray transform parameters [2]. This algorithm searches the optimal parameter α and β in the whole parameter space, where $0 < \alpha, \beta < 10$. These parameters α and β control the shape of a non-linear transform curve. The incomplete Beta function can be written in the follows [2]:

$$F(u) = B^{-1}(\alpha, \beta) \times \int_0^u t^{\alpha-1} (1-t)^{\beta-1} dt, \quad 0 < \alpha, \beta < 10 \quad (1)$$

All the gray levels of an original image have to be unitary before implementing IBT. All the gray levels of the enhanced image have to be inverse-unitary after implementing IBT. In general, $\alpha < \beta$, when the image is particular dark; $\alpha > \beta$, when the image is particular bright, and $\alpha = \beta$ when all the gray levels of the image are centralized on the middle certain region. An objective function [1] is employed to evaluate the quality of the enhanced image. This function can be written in the following:

$$C_{contrast} = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N g'^2(i, j) - \left[\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N g'(i, j) \right]^2 \quad (2)$$

where M, N are, respectively, the width and height of the original image, $g'(i, j)$ is the gray level at (i, j) in the enhanced image. The larger the value of the function is, the better the distribution of the image gray levels is proportioned. Since an original

image has 255 gray levels, the whole gray level space is divided into six sub-spaces: $A_1, A_2, A_3, A_4, A_5, A_6$, where A_i ($i=1, 2, \dots, 6$) is the number of all pixels which distribute in the i th sub-space. Let,

$$M = \max_{i=1}^6 A_i, \quad B_1 = \sum_{k=2}^6 A_k, \quad B_2 = \sum_{k=2}^5 A_k, \quad B_3 = \sum_{k=1}^5 A_k, \\ B_4 = A_1 + A_6, \quad B_5 = A_2 + A_3, \quad B_6 = A_4 + A_5,$$

The following classification criterion can be obtained:

```

if (M = A1) & (A1 > B1)
    Image is PB;
elseif (B2 > B4) & (B5 > B6) & (B5 > A1) & (B5 > A6) & (A2 > A3)
    Image is MD;
elseif (B2 > B4) & (B5 > B6) & (B5 > A1) & (B5 > A6) & (A2 < A3)
    Image is MDS;
elseif (B2 > B4) & (B5 < B6) & (A1 < B6) & (A6 < B6) & (A4 > A5)
    Image is MBS;
elseif (B2 > B4) & (B5 < B6) & (A1 < B6) & (A6 < B6) & (A4 < A5)
    Image is MB;
elseif (M = A6) & (A6 > B3)
    Image is PB;
else
    Image is GGLD;
end
    
```

Where symbol $\&$ stands for logic “and” operator. We will employ the SA, which was given by William L. Goffe, to optimize transform parameters [5]. The range of α and β can be determined by Table 1.

Table 1. Range of α and β

Parameter	PD	MD	MDS	MBS	MB	PB
α	[0, 2]	[0, 2]	[0, 2]	[1, 3]	[1, 4]	[7, 9]
β	[7, 9]	[1, 4]	[1, 3]	[0, 2]	[0, 2]	[0, 2]

Let $\mathbf{x} = (\alpha, \beta)$, $F(\mathbf{x})$ is a function to be minimized, corresponding to (2). Where $a_i < \alpha, \beta < b_i$ ($i = 1, 2$), a_i and b_i ($i = 1, 2$) can be determined by Tab.1. The above parameters are determined according to $N_s = 20$, $N_T = 100$, $c_i = 2$, $i = 1, 2, T_0 = 5$, $r_T = 0.95$. Further details for SA can be found in [5].

3 Global Contrast Enhancement Based on DSWT and IBT

DSWT has been independently discovered several times, for different purposes and under different names [6]. Based on DSWT, IBT is employed to enhance the global and local contrast for an image. Employing IBT enhances the low frequency sub-band image. According to Section 2, proper non-linear gray transform parameter α and β are selected to enhance the global contrast for the image.

4 Local Contrast Enhancement Based on DSWT and IBT

We consider discrete image model as follows:

$$\mathbf{g} = \mathbf{f} + \boldsymbol{\varepsilon} \quad (3)$$

Where, $\mathbf{g} = \{g[i, j]\}_{i,j}$ shows the observation signal, $\mathbf{f} = \{f[i, j]\}_{i,j}$ indicates the un-corrupted original image, $\boldsymbol{\varepsilon} = \{\varepsilon[i, j]\}_{i,j}$, $i = 1, \Lambda, M; j = 1, K, N$ is a stationary signal. DSWT is implemented to Equation (3):

$$\mathbf{Y} = \mathbf{X} + \mathbf{V} \quad (4)$$

Where, $\mathbf{X} = \mathbf{S}\mathbf{f}$, $\mathbf{V} = \mathbf{S}\boldsymbol{\varepsilon}$, $\mathbf{Y} = \mathbf{S}\mathbf{g}$, and \mathbf{S} is the two-dimension stationary wavelet transform operator. ‘‘Soft-threshold’’ function is employed to reduce the noise in the image:

$$\mathbf{Y}_{\delta} = \mathbf{T}_{\delta} \circ \mathbf{Y} \quad (5)$$

$$\mathbf{X}_{\delta} = \mathbf{T}_{\delta} \circ \mathbf{X} \quad (6)$$

The total operator can be expressed as:

$$\mathbf{Z}_{\delta} = \mathbf{S}^{-1} \circ \mathbf{T}_{\delta} \circ \mathbf{S} \quad (7)$$

Where \mathbf{T}_{δ} is correlated to threshold δ and input signal \mathbf{g} . If the statistic properties of the noise are employed to approximate the optimal threshold δ , standard variance σ should be used. This is almost impossible in practice according to above discussion. Generalized cross validation principle is employed to solve the problem [7]. Consider $\tilde{g}[i, j]$ a linear combination of $g[k, l]$, thus special noise can be reduced. $g[i, j]$, which is the $[i, j]$ element of \mathbf{g} , is replaced by $\tilde{g}[i, j]$:

$$\tilde{\mathbf{g}} = \mathbf{Z} \cdot (g[1,1], K, \tilde{g}[i, j], K, g[M, N])^T \quad (8)$$

The same processing is repeated to all the components:

$$OCV(\delta) = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (g[i, j] - \tilde{g}_\delta[i, j])^2 \quad (9)$$

The forms of $\tilde{g}[i, j]$ are many kinds, here let $\tilde{g}_\delta[i, j] = \tilde{g}[i, j]$, we have:

$$g[i, j] - \tilde{g}_\delta[i, j] = \frac{g[i, j] - g_\delta[i, j]}{1 - \tilde{z}[i, j]} \quad (10)$$

Where, $\tilde{z}[i, j] = \frac{g_\delta[i, j] - \tilde{g}_\delta[i, j]}{g[i, j] - \tilde{g}_\delta[i, j]} \approx z'[m, n] = \frac{\partial g_\delta[i, j]}{\partial g[k, l]}$. However, in Equation (16), $z'[m, m]$ is either zero or 1. Thus “generalized cross validation” formula in the wavelet domain will be given as follows:

$$SGCV(\delta) = \frac{\frac{1}{MN} \cdot \|\mathbf{Y} - \mathbf{Y}_\delta\|^2}{\left[\frac{\text{trace}(\mathbf{I} - \mathbf{Z}'_\delta)}{MN} \right]^2} \quad (11)$$

Let $\delta^* = \arg \min MSE(\delta)$, $\tilde{\delta} = \arg \min GCV(\delta)$, M. Jansen has proved that $\tilde{\delta}$ is an asymptotic optimal solution [8].

5 Evaluation Criterion for Enhanced Image

The quality for enhanced image is evaluated by employing Equation (2). The larger the value of Equation (2) is, the better the contrast of the image is. Combining the ratio of signal-to-noise, an overall objective criterion is proposed as follows:

$$C_{total} = C_{contrast} * C_{snr} \quad (12)$$

where C_{snr} is the peak ratio of signal-to-noise for an enhanced image, it can be calculated by the following equation:

$$C_{snr} = 10 \cdot \log \left(\frac{MN \cdot \max(F_{ij}^2)}{\sum_{i=1}^M \sum_{j=1}^N (F_{ij} - G_{ij})^2} \right) \quad (13)$$

Where F_{ij} and G_{ij} are gray level value at (i, j) in original image and enhanced image respectively. M and N are width and height of the original image respectively. The larger the value of Equation (13) is, the better the overall visual quality is.

6 Experimental Results

Fig.1 shows gray transform curve, where $\alpha = 1.9819$, $\beta = 3.3340$. Fig.2 represents gray transform curve, where $\alpha = 6.9894$, $\beta = 9.9917$. The transform curve is employed to enhance the global and the local contrast of Fig.3 (b). Fig.3 (a) is a girl image and Fig.3 (b) is a corrupted image by GWN (the standard variance of noise is $\sigma = 8.8316$). In order to demonstrate the excellent performance of the new algorithm, two traditional contrast enhancement algorithms are compared with the proposed algorithm. They are the algorithms of HE and USM respectively. Fig.3 (c)-Fig.4 (f) show the enhanced images by HE algorithm, USM algorithm, Gong's algorithm [3], Wu's algorithm [4], Tubbs's algorithm [2] and the new algorithm (NEW), respectively.

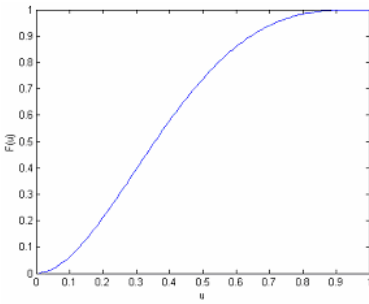


Fig. 1. Gray levels transform curve
($\alpha = 1.9819$, $\beta = 3.3340$)

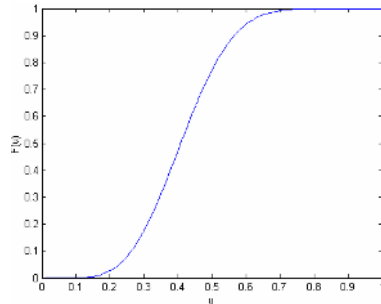


Fig. 2. Gray levels transform curve
($\alpha = 6.9894$, $\beta = 9.9917$)

According to section 2, the contrast type of Fig.3 (b) is “MD”. From the experimental results above, the noise in the image is enhanced greatly when using the algorithms of USM and HE to enhance the image, which is obvious in Fig.3 (c)-(d). Noise reduction is also considered in [3] and [4]. The total contrast is better by employing Gong's algorithm. However, it is not efficient to reduce the noise in the image. From Fig.3 (e), it can be seen that the noise in the image is enhanced greatly and the background clutter is also enlarged. Although Wu's algorithm can well reduce the noise in the image, the whole brightness of the image is so high that some details in the image are lost. Lots of burr is produced in Fig.3 (f), such as in the regions of hair and face of the girl. When Tubbs's algorithm [2] is used to enhance the contrast of Fig.3 (g), the global contrast is good. However, the local contrast is bad. For example, the regions of ear and neck of the girl in Fig.3 (g) is very blur. Compared with the above five algorithms, the new algorithm can efficiently reduce GWN in the image while well enhancing the contrast for the image. It is obvious that the new algorithm is more excellent in the total performance than the algorithms of USM, HE, Tubbs [2], Gong's algorithm [3] and Wu's algorithm [4].



(a) Girl image



(b) Girl image corrupted by AGWN



(c) Enhanced image by HE



(d) Enhanced image by USM



(e) Enhanced image by Gong



(f) Enhanced image by the Wu's



(g) Enhanced image by Tubbs



(h) Enhanced image by NEW

Fig. 3. Enhanced images by six algorithms

In order to further illustrate the efficiency of the new algorithm, Equation (13) is used to evaluate the quality of enhanced images. The total cost of enhanced images by the algorithms of HE, USH, Gong, Wu, Tubbs and NEW in Fig.3 are 19.1257, -89.1866, 13.4102, 0.2772, 15.0528 and 32.5503 respectively. The same conclusion can be drawn with the above analysis to enhance other images.

7 Conclusion

Employing GCV, the asymptotic optimal de-noising threshold can be obtained when the accurate statistic properties are not prior-known. Combining the IBT and SA directly enhances the global contrast of the image. Combining a de-nosing algorithm and IBT enhances the local contrast of the image. Experimental results show that the new algorithm can effectively enhance the global and local contrast for image while keeping the detail information of the original image. The total performance of the new algorithm is better than HE, USM, Tubbs, Gong and Wu's algorithm.

References

1. Azriel Rosenfield, Avinash C K.: Digital Picture Processing. New York: Academic Press, (1982)
2. Tubbs J D.: A note on parametric image enhancement. *Pattern Recognition*, **30** (1997) 616-621
3. GONG Wu-Peng, WANG Yong-Zhong: Contrast enhancement of infrared image via wavelet transforms. *Chinese Journal of National University of Defense Technology*, **22** (2000) 117-119
4. WU Ying-Qian, SHI Peng-Fei: Approach on image contrast enhancement based on wavelet transform. *Chinese J. Infrared and Laser Engineering*, **32** (2003) 4-7
5. William L. Goffe, Gary D. Ferrier, John Rogers: Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, **60** (1994) 65-99
6. M. Lang, H.Guo, J.E. odegend, C.S. Burrus, R.O.Wells, Jr.: Nonlinear processing of a shift-invariant DWT for noise reduction. In *SPIE Conference on wavelet applications*, **2491** (1995) 76-82
7. P. Hall and I. Koch: On the feasibility of cross-validation in image analysis. *SIAM J.Appl. Math*, **52** (1992) 292-313
8. Maarten Jansen, Geert Uytterhoeven, Adhemar Bultheel: Image de-nosing by integer wavelet transforms and generalized cross validation. Technical Report TW264, Department of Computer Science, Katholieke Universiteit, Leuven, Belgium, August 1997

An Efficient Unsupervised MRF Image Clustering Method

Yimin Hou¹, Lei Guo¹, and Xiangmin Lun²

¹ Department of Automation, Northwestern Polytechnical University, Xi'an 710072

² Xi'an Institute of Optics and Precision Mechanics of Cas, Xi'an 710068

Abstract. On the basis of Markov Random Field (MRF), which uses context information, in this paper, a robust image segmentation method is proposed. The relationship between observed pixel intensities and distance between pixels are introduced to the traditional neighbourhood potential function, which described the probability of pixels being classified into one class. To perform an unsupervised segmentation, the Bayes Information Criterion (BIC) is used to determine the class number. The K-means is employed to initialise the classification and calculate the mean values and variances of the classes. The segmentation is transformed to maximize a posteriori (MAP) procedure. Then, the Iterative Conditional Model (ICM) is employed to solve the MAP problem. In the experiments, the proposed method is adopted with K-means, traditional Expectation-Maximization (EM) and MRF image segmentation techniques, for noisy image segmentation applying on synthetic and real images. The experiment results and the histogram of signal to noise ratio (SNR)-miss classification ratio (MCR) showed that the proposed algorithm is the better choice.

1 Introduction

Image segmentation process, in which image pixels are classified into several classes, is the first step towards image automatic analysis and evaluation. Many segmentation methods have been presented so far[1][2], such as edge-based segmentation[3], region-based segmentation[4] and pixel labeling. In the early works of image segmentation, pixels were classified independently. These approaches could not produce satisfactory classifications if there was noise in the image. The correlation between adjacent pixels, called context or spatial information, shows the possibility that pixels are likely to come from the same class. But, the image spatial information was ignored in the traditional methods. To solve this problem, a statistical method, the MRF image models[5][6], is adopted in this paper. The relationships among the pixel intensities and distance values are introduced to the clique potential functions, which define the MRF probability distribution discussed here[7][8]. Compared with the classical potentials, the new function involves more spatial information of the images. For two pixels, the probability of being classified into the same class changes with the variation of intensity difference and distance between their positions. This is one of the differences between our method and the traditional MRF segmentation. The other one

is that the new method uses BIC to find out the optimization class number and uses the K-means method to obtain the mean values and variances of every class. Therefore, the algorithm proposed in this paper is absolutely an unsupervised segmentation method. The new potential is used in MAP procedure to get the segmentation result.

In Section 2 of the paper, the MRF image model is described and, based on Bayes theorem, the transformation from segmentation to MAP is discussed. The Iterative Conditional Model is used in this section to solve MAP problem. In Section 3, the part of algorithm improvement, new potential function, which involves the pixel intensity and distance values, is described. The model fitting is done according to K-means and BIC. In Section 4, some experiments are illustrated to show the segmentation results of proposed method. The method is compared with the K-means, classical EM algorithm and classical MRF algorithm. From the results, in section 5, a conclusion is drawn that segmentation method proposed in this paper does better than the others in either miss classification ratio or noise filtering.

2 MRF-MAP Segmentation

Let $Q = \{q(i, j) | 1 \leq i \leq W, 1 \leq j \leq H\}$ be a finite set of sites of a $W \times H$ rectangular lattice. Let the label field $X = \{X_1, \dots, X_m\}$ be a Markov Random Field defined on Q , m is the total number of the classes. In a MRF, the sites in Q are related to each other via a neighborhood system, $\psi = \{N_q, q \in Q\}$, where N_q is the set of neighbors of $q(i, j)$. A clique is a subset in N_q , $c \in N_q$ is a clique of distinct sites being neighbors in. Single site, pair-of-sites cliques, and so on, can be defined.

The random field X is considered to be a MRF on Q if and only if $P(X = x) > 0$ and $P(X_q = x_q | X_r = x_r, r \neq q) = P(X_q = x_q | X_r = x_r, r \in N_q)$. But it is difficult to determine the above characteristics in practice by computer. It is well known that a MRF is completely described by a Gibbs distribution using the Hammersley-Clifford theorem that established the relation between the MRF and Gibbs distribution[9][10]. So the MRF has the form

$$P(X = x) = \frac{1}{Z} e^{-U(X)/T} \quad (1)$$

where $X = x$ is a realization from $X = \{X_1, \dots, X_m\}$, and

$$U(X) = \sum_Q V_c(x) \quad (2)$$

is global energy function where $V_c(x)$ is the clique energy function. It is given by the sum of clique potentials over all possible cliques, c is a clique in N_q . And

$$Z = \sum_{x \in X} e^{-U(X)/T} \quad (3)$$

is the partition function. T is a constant called the temperature.

Image Y can be considered as a rectangular lattice Q . Let y_q denotes the intensity of the pixel at q and it correspond to the label x_q in X . Bayes theorem yields a complete model coupling intensities and labels. If we use $P(Y|X)$ and $P(X)$ to denote the, the conditional probability density of the image Y and the prior of the labeling X . The prior probability of the image $P(Y)$ is independent of the labeling X , and according MAP theorem, we have

$$X_{opt} = \arg \max_{X \in \Omega} \{P(Y|X)P(X)\} \quad (4)$$

It could be assumed that the image data are obtained by adding an identical independently distributed Gaussian noise[13][14]. And because the conditional probabilities of pixel y_q is independent in observed image Y , the conditional density $P(Y|X)$ takes the form of

$$P(Y|X) = \prod_Q P(y_q | x_q) = \prod_Q \left[\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(y_q - \mu_q)^2}{2\sigma_q^2}\right) \right] \propto \exp\left(-\sum_Q \frac{(y_q - \mu_q)^2}{2\sigma_q^2}\right) \quad (5)$$

where μ_q is the mean value and σ_q^2 is the variance of the class that site q belongs to.

Based on Eq.(1), $P(X)$, the prior density of MRF takes the form of

$$P(X) = \frac{1}{Z} \exp\left(-\sum_Q \sum_{c \in N_q} \frac{V_c(q(i, j), q(i_1, j_1))}{T}\right) \quad (6)$$

where $q(i, j)$ and $q(i_1, j_1)$ are sites belong to the same neighborhood, $V_c(q(i, j), q(i_1, j_1))$ is potential function. Then, the posteriori can be written as

$$\begin{aligned} P(X|Y) &\propto P(Y|X)P(X) \propto \left[\exp\left(-\sum_Q \frac{(y_q - \mu_q)^2}{2\sigma_q^2}\right) \right] \left[\frac{1}{Z} \exp\left(-\sum_Q \sum_{c \in N_q} \frac{V_c(q(i, j), q(i_1, j_1))}{T}\right) \right] \\ &\propto \frac{1}{Z} \exp\left\{-\sum_Q \left[\frac{(y_q - \mu_q)^2}{2\sigma_q^2} + \sum_{c \in N_q} \frac{V_c(q(i, j), q(i_1, j_1))}{T} \right]\right\} \end{aligned} \quad (7)$$

where Z is a constant parameter.

The segmentation of image Y is completed after finding the label field, X . According to the MAP criterion, the optimal X can be got by minimization the potential summation of conditional possibility and the prior.

$$X_{opt} = \arg \min_{x \in \Omega} U(X|Y) = \arg \min \left\{ \sum_Q \left[\frac{(y_q - \mu_q)^2}{2\sigma_q^2} + \frac{1}{T} \sum_{c \in N_q} V_c(q(i, j), q(i_1, j_1)) \right] \right\} \quad (8)$$

The optimization method, such as Simulated Annealing (SA), Iterated Conditional Models(ICM) and Highest Confidence First(HCF), are used to find the solution of Eq.(8) in the early works. The SA schedule is normally too slow for practical

applications. The ICM algorithm is likely to reach only local minima and there is no guarantee that a global minimum of energy function can be obtained, but it provides a much faster convergence than stochastic relaxation-based method. So it is adopted in this paper. The ICM iteratively decreases the energy by visiting and updating the pixels. For each pixel q , given the observed image and current labels of all the pixels in the neighborhood, the label of X_q is replaced with one that can maximize the probability

$$X_q^{(k+1)} = \arg \max P(X_q^{(k)} | Y, X_r^{(k)}, r \neq q) \quad (9)$$

Starting from the initial state, the algorithm will keep on running base on the procedure above until either the predefined number of iterations is reached or the label of X does not change.

3 New Potential Function and Model Fitting

The choice of the prior energy function is arbitrary and there are several definitions of $U(X)$ in the framework of image segmentation. A complete summary of them was done in some literatures where a general expression for the energy function is denoted by

$$U(X) = \sum_Q \left[V_c(q(i, j)) + \sum_{c \in N_q} V_c(q(i, j), q(i_1, j_1)) \right] \quad (10)$$

This is known as Potts model with an external field, $V_c(q(i, j))$, that weighs the relative importance of the different classes. Eq.(10) can be modelled by an Ising model at 2 states. However, the use of an external field includes additional parameter estimation, thus this model is less used in image segmentation. Instead, a simplified Potts model with no external energy, $V_c(q(i, j))=0$, is used. Then, only the local spatial transitions are taken into account and they can be defined for instance as

$$V_c(q(i, j), q(i_1, j_1)) = \begin{cases} -\beta & \text{if } q(i, j) = q(i_1, j_1) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

Intuitively, the equation above has not moreover involved spatial information of the image. A more effective function is used in this work as proposed:

$$V'_c(q(i, j), q(i_1, j_1)) = \begin{cases} \frac{\beta \sigma_{ij}^2}{(\sigma_q^2 + (y_{q(i,j)} - y_{q(i_1, j_1)})^2 \times d_{q(i,j)q(i_1, j_1)})} & \text{if } q(i, j) \neq q(i_1, j_1) \\ -\beta & \text{if } q(i, j) = q(i_1, j_1) \end{cases} \quad (12)$$

where $y_{q(i,j)}$ and $y_{q(i_1, j_1)}$ are the pixel intensities of $q(i, j)$ and $q(i_1, j_1)$; $d_{q(i,j)q(i_1, j_1)}$ represents the distance between the two pixels; β is a constant that controls the classification. We can draw a conclusion from the equation that with the decreasing of $d_{q(i,j)q(i_1, j_1)}$ and $(y_{q(i,j)} - y_{q(i_1, j_1)})^2$, $V'_c(q(i, j), q(i_1, j_1))$ decreased to $-\beta$,

that is $q(i, j)$ and $q(i_1, j_1)$ are right one pixel or their intensities are same, so, Eq.(8) can be written as

$$X_{opt} = \arg \min \left\{ \sum_Q \left[\frac{(y_q - \mu_q)^2}{2\sigma_q^2} + \frac{1}{T} \sum_{c \in N_q} V'_c(q(i, j), q(i_1, j_1)) \right] \right\} \quad (13)$$

The next step of the improvements is model fitting. A statistical model is completed only if both its functional form and its parameters are determined[11]. The procedure for estimating the unknown parameters is known as model fitting. For a MRF model, $\theta = \{\mu_1, \dots, \mu_m, \sigma_1, \dots, \sigma_m\}$ is one of the parameter set. K-means, an iterative method, was used to initialize the classification. After this step, we can get the initial parameter set $\theta = \{\mu_1, \dots, \mu_m, \sigma_1, \dots, \sigma_m\}$. We can sign a maximal value m_{\max} and a minimum value m_{\min} of the total class number m and K-means was used for every possible value between m_{\min} to m_{\max} . The K-means image segmentation method is composed of 4 steps.

Step1. If the model number is m_k ($m_{\min} \leq m_k \leq m_{\max}$), for the t th iterative operation, chose the initial mean values, $\mu_1^t, \mu_2^t, \dots, \mu_{m_k}^t$, $t = 0$, randomly.

Step2. For pixel x_l , if $\|x_l - \mu_g^i\| \leq \|x_l - \mu_h^i\|$, $g \in \{1, \dots, m_k\}$, $h \in H = \{1, \dots, m_k\} - g$, then x_l will be classified into the g th class.

Step3. Update the mean values, $\mu_1^t, \mu_2^t, \dots, \mu_{m_k}^t$.

Step4. Repeat from step2 until the $\mu_1^t, \mu_2^t, \dots, \mu_{m_k}^t$ do not change any more. In this paper, $m_{\min} = 2$ and $m_{\max} = 10$. Then we obtain 9 sets of $\theta = \{\mu_1, \dots, \mu_{m_k}, \sigma_1, \dots, \sigma_{m_k}\}$.

The Bayes Information Criterion was used to decide the class number. The following formulae described the Bayes Information Criterion.

$$BIC_{m_k} = 2 \log P(Y | \theta_{m_k}, m_k) - v_{m_k} \log(n) \quad (14)$$

where v_{m_k} is the number of independent parameters in the model. n is the pixel number, Y is the observation. BIC has given good results for choosing the number of components in a wide range of applications of mixture models. So, after this step, our algorithm is exactly an unsupervised segmentation method.

4 Experiments

Experiments were performed using synthetic image and real images. The algorithm is compared with the K-means, classical EM and classical MRF segmentation method.

A synthetic image shown in the Fig.1(a) was applied in the first experiment. Fig. 1(b) was obtained by adding noise to (a) and the signal to noise ratio(SNR) of it is 4.55. That is the image (b) contain a lot of noise and it is difficult to get very good segmentation result.

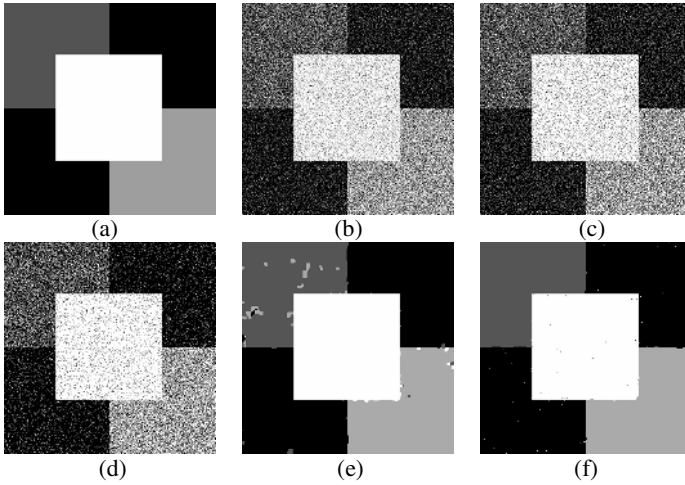


Fig. 1. Segmentation experiments on a synthesis image with 4 classes (a) The original image (b) Noisy synthetic image ($SNR=4.55$) (c) K-means segmentation result (d) Traditional EM segmentation result (e) Traditional MRF segmentation result (f) Proposed method segmentation result

Fig.2(c) to (f) showed different segmentation results, which was obtained by different methods, of noise image. The result displayed in Fig.1 (f) demonstrates the parameters of each class are properly estimated and the segmented regions are uniform respectively. This is great improvement over the K-means, classical EM and classical MRF whose results are shown as Fig.1 (c), (d) and (e).

Fig.2 shows the SNR-MCR curves of segmentation of Fig.2(a) with different method. We applied the classification using the four method to five noise images with different SNR. It is obvious that the method proposed in this paper is better than the other three in either high noise or low noise conditions.

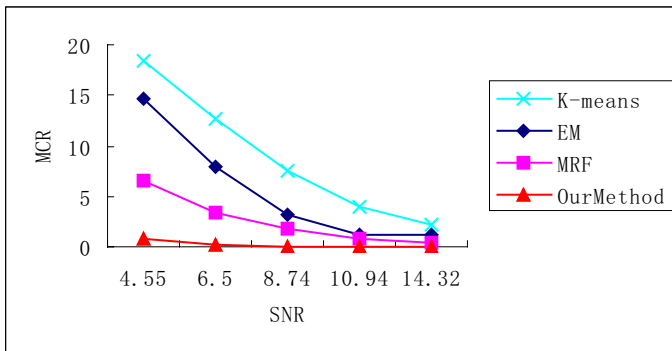


Fig. 2. The percentages of Miss Classified Ratio for different SNR images

In the second example, we use the house image. Fig.3 (a) is the original image. Fig.3 (b) is the image added Gaussian noise and its $SNR=8.55$. We can observe an improvement when applying our algorithm. No doubt, this model is better enough to capture some finer features of the house image than classical EM and MRF methods.

The third example, shown as Fig.4, employed a blood corpuscle image. It has three distinct clusters at three intensity levels. The segmentation result of our method shows more clear main body region of the corpuscle than the results of K-means, traditional EM and traditional MRF method.

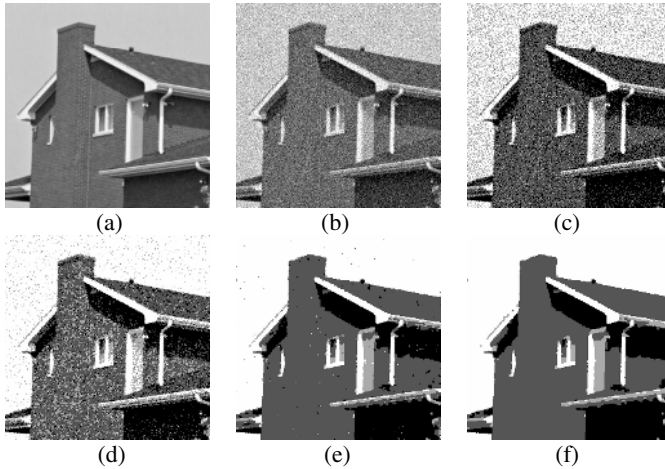


Fig. 3. The 4 classes segmentation experiments on house image (a) The original image (b) Noisy synthetic image ($SNR=8.07$) (c) K-means segmentation result (d) Traditional EM segmentation result (e) Traditional MRF segmentation result (f) Proposed method segmentation result

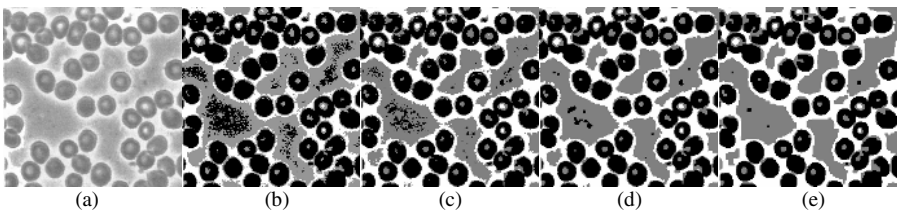


Fig. 4. Segmentation experiments on blood corpuscle image (a) The original image (b) The K-means segmentation result (c) Traditional EM segmentation result (d) Traditional MRF segmentation result (e) Proposed method segmentation result

We want to get an unsupervised segmentation. So the K-mean and BIC are used to fit the image model. Fig.5 shows the Class Number-BIC curves of the three experiments above. We can see that, in the first experiment image BIC values, the point at class 4 is the last highest one before the smooth phase of the curve. So, we regard 4 as

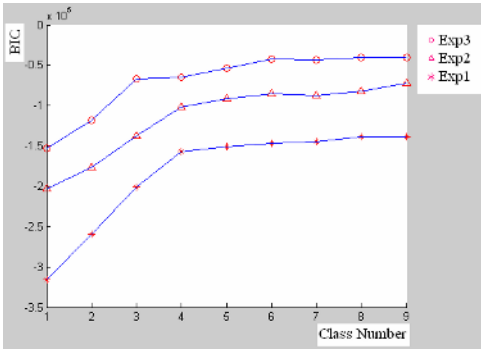


Fig. 5. Class number-BIC curves of experiments

Table. 1 Means and variance of every class

Experiments	Means	Variances	
Exp 1	Class 1	31.09	1974.15
	Class 2	87.15	330.89
	Class 3	154.02	4544.25
	Class 4	224.42	1992.05
Exp 2	Class 1	52.84	668.65
	Class 2	107.64	672
	Class 3	200.3	619.31
	Class 4	232.12	428.04
Exp 3	Class 1	136.63	195.05
	Class 2	164.57	73.19
	Class 3	251.68	14.98

the optimal class number of the experiment image. With this criterion, we can also find the optimal class numbers, 4 and 3, of the following two experiments segmentations. Table.(1) shows the means and variances of every class in every experiment.

5 Conclusions

Statistical algorithm is a very important way for image modeling in image segmentation and reversion, especially MRF methods. However, the classical MRF potential functions always do not involve the relationships of pixel intensities and the distances between pixels. For image segmentation, the results were thus negatively affected by these factors. To overcome this problem, a novel potential function, which involved the pixel intensity and distance values, is introduced to the traditional MRF image model. Then, the segmentation problem is transformed to MAP procedure and the ICM method is employed to obtain the MAP solution. On the other side, to complete an unsupervised segmentation, the model fitting is performed using K-means and BIC. The experiment results prove that the algorithm proposed in this paper is an efficient method for images unsupervised segmentation.

References

- [1] J.S. Wezka, A survey of threshold selection techniques, *Comput. Vision, Graphics Image Process.* 7 (1978) 259–265.
- [2] P.K. Sahoo, S. Soltani, A. Wong, Y. Chen, A survey of thresholding techniques, *Comput. Vision, Graphics Image Process.* 41 (1988) 233–260.
- [3] J. Basak, B. Chanda, D.D. Manjunder, On edge and line linking with connectionist models, *IEEE Trans. Systems, Man Cybernet.* 24 (3) 413–428, 1994.
- [4] S.A. Hojjatoleslami, J. Kittler, Region growing: a new approach, *IEEE Trans. Image Process.* 7 (7) 1079–1084, 1998.

- [5] S.Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer, New York, 2001.
- [6] Z. Tu and S.-C. Zhu, *Image Segmentation By Data-Driven Markov Chain Monte Carlo*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 657–673, 2002.
- [7] S.A. Barker, *Image segmentation using Markov random field models*, Ph.D. Thesis, University of Cambridge, 1998.
- [8] C.S. Won, H. Derin, *Unsupervised segmentation of noisy and textured images using Markov random fields*, *CVGIP: Graphical Models Image Process.* 54 (4) 308–328, 1992.
- [9] Geman, S., Geman, D, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.6, pp. 721–741, 1984.
- [10] T.Lei, *Gibbs ringing artifact spatial correlation and spatial correlation in MRI*, *SPIE Proceedings*, 5368, pp.837-847, 2004.
- [11] J. Besag, *Towards Bayesian image analysis*, *Journal of Applied Statistics*, vol.16, pp. 395-407, 1989.
- [12] Hurn, M.A., Mardia, K.V. et al., *Bayesian fused classification of medical images*. *IEEE Trans. Med. Imag.* 15 (6), 850–858, 1996.
- [13] Gath, I., Geva, A.B., *Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions*, *Pattern Recognition Letters*, vol.9 (3), Elsevier, pp. 77-78, 1989.
- [14] T.Lei and J.Udupa, *Performance evaluation of finite normal mixture model-based image segmentation Techniques*, *IEEE Trans. Image Processing* 12(10), pp. 1153-1169. 2003.

A SVM-Based Blur Identification Algorithm for Image Restoration and Resolution Enhancement

Jianping Qiao and Ju Liu*

School of Information Science and Engineering, Shandong University,
Jinan 250100, Shandong, China
{jpeqiao, juliu}@sdu.edu.cn

Abstract. Blur identification is usually necessary in image restoration. In this paper, a novel blur identification algorithm based on Support Vector Machines (SVM) is proposed. In this method, blur identification is considered as a multi-classification problem. First, Sobel operator and local variance are used to extract feature vectors that contain information about the Point Spread Functions (PSF). Then SVM is used to classify these feature vectors. The acquired mapping between the vectors and corresponding blur parameter provides the identification of the blur. Meanwhile, extension of this method to blind super-resolution image restoration is achieved. After blur identification, a super-resolution image is reconstructed from several low-resolution images obtained by different foci. Simulation results demonstrate the feasibility and validity of the method.

1 Introduction

Image restoration refers to the removal or reduction of degradations that were incurred while the digital image was being obtained. Restoration procedures that seek to recover information beyond the diffraction limit are referred to as super-resolution techniques which can be used in many applications such as remote sensing, military surveillance, medical diagnostics, HDTV and so on.

In classical image restoration, explicit knowledge about degradation process is usually assumed to be available. However, in many practical applications, the blurring process is not known or known only within a set of parameters. Therefore, blur identification and restoration must be implemented simultaneously or separately. Here we focus on the latter method because of its low computational complexity and flexibility. Reeves S et al. [1] simplified the identification problem by parameterizing the PSF and described a parameterized blur estimation technique using generalized cross-validation, but computation complexity is high. Nakagaki R et al. [2] proposed a VQ-based blur identification algorithm for image recovery in which the PSF estimate is chosen from a set of candidates according to minimum distortion criterion. But this method sometimes depends on the training images. In addition, these methods did not

* Corresponding author.

take into account the effect of some elements such as nonlinear camera response function, different exposures and white balancing.

In this paper we propose a parametric blur identification algorithm based on Support Vector Machines (SVM). In this method, blur identification was formulated as a machine learning problem. First, Sobel operator and local variance are used to extract feature vectors that contain information of various PSFs from training images. Then SVM is used to classify these vectors. Given a degraded image, the acquired mapping between the feature vectors and their corresponding blur parameter provides the identification of the blur. Optionally, a preprocessing of the degraded image is implemented to avoid the effect of the nonlinear sensor response, different exposures, and white balancing. Meanwhile, extension of this method to blind super-resolution image restoration is achieved. After blur identification, a super-resolution image is reconstructed from several low-resolution images obtained by different foci in different lightning conditions.

This paper is organized as follows: the SVM-based blur identification algorithm is presented in Section 2. Section 3 describes the blind super-resolution algorithm. Simulations in section 4 show the effectiveness of our method. Section 5 concludes this paper.

2 Blur Identification Model

Let y be the observed image, x be the original image, H be the linear degradation process which is formed by Point Spread Function (PSF) and n be an additive noise, then we have

$$y = Hx + n \quad (1)$$

However, the model in the above equation is not a complete model [3]. In real applications, the issues of nonlinear sensor response, different exposures and white balancing must be considered. Generally, exposure time and white balancing can be modeled as gain and offset terms. Then the imaging process is formulated as

$$y = f(\alpha Hx + \beta + n) + \eta \quad (2)$$

where α is the gain factor, β is the offset factor, η is the quantization errors, and $f(\cdot)$ is the nonlinear camera response function [4].

2.1 Basic Identification Scheme

The support vector machine (SVM) is a new universal learning machine proposed by Vapnik [5], which is applied to both regression and pattern recognition. Li et al. [6] describes a method of blind image deconvolution by support vector regression (SVR) in which SVR is used to learn the mapping between the blurred image and original image. In this paper, from a pattern recognition point of view, the blur identification is considered as a multi-class classification problem and SVM is used to identify the blur parameter.



Fig. 1. Edge detected images with different blur parameters

Assume that the blurred image is obtained from a camera with unknown degradation parameters, that is, the blurring process is partially known and blur identification can be reduced to finding the best estimation within a set of parameters. From a pattern recognition point of view, this problem can be considered as a multi-class classification problem, namely, selecting the blur parameter of a given image from certain candidates. The hypothesis behind the formulation is that we should extract feature vectors that could distinguish different Point Spread Functions (PSFs). If we could get such feature vectors, the mapping between feature vectors and the corresponding blur parameters can be learned by SVM. Given a blurred image, the acquired mapping can be used for blur identification.

In this method, Sobel operator and local variance are used to extract feature vectors. The basic idea is that the blurring function is considered as a low-pass filter and low-frequency regions contains little or no information about the PSF, so information about PSF is reserved after edge detection. Fig.1 illustrates the edge detected results by Sobel operator. We assume that the blur function is Gaussian which is parameterized by the variance σ^2 . Fig.1 (a) and (c) show the images blurred by different PSFs. The blur parameters are (a) $\sigma^2 = 2$; (c) $\sigma^2 = 5$. Fig.1 (b) and (d) show the edge detected results of (a) and (c) respectively. Evidently, the edge detected images are different which means that they contain information about PSFs. Meanwhile, Sobel operator enhances the robustness of the algorithm to different types of images. This is because what we need is the information that can distinguish different PSFs, the edge detected image by Sobel operator furthest preserves this information. After edge detection, the image is blocked into small patches. Feature vectors are composed by the lexicographically ordered pixels of the patches with larger local variance.

2.2 The Proposed SVM-Based Blur Identification Algorithm

According to the basic idea, the proposed detail procedure (as shown in Fig.2) contains two stages:

The first stage is training stage:

- 1) Choose a candidate set $\{I_i\}_{i=1}^R$ for the blurred image and some training images.
- 2) Degrade the training images according to the image acquisition model.

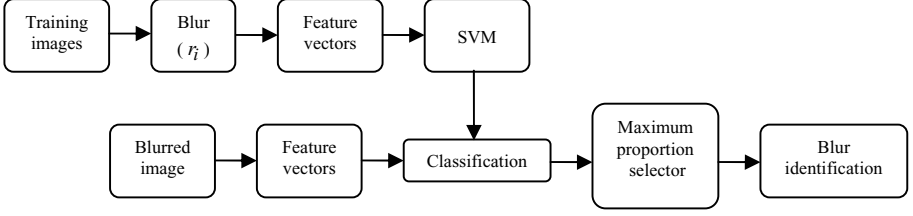


Fig. 2. Schematic diagram of the proposed identification method

- 3) Feature extraction. Detect the edge of each degraded image with Sobel operator and divide the edge detected image into blocks, then compute the local variance of each patch. Only the patches with larger local variance are selected to form the feature vectors. The threshold of local variance is varied adaptively according to the maximum local variance of the images. In our implementation, it equals to 0.2 times maximization. In this way, for each parameter candidate r_i , a number of feature vectors are extracted from the training images which will be used as the input of SVM and the output of SVM is the index of the corresponding candidate r_i .
- 4) SVM training. The parameter candidates are always more than two, so this is a multi-class classification problem. For the training set, we choose “one against one” approach in which $R(R-2)/2$ classifiers are constructed and each one trains data from two different classes, a voting strategy is used to designate the outcome of the classification. The kernel used in the implementation is the radial basis function (RBF) kernel, where γ is the parameter controlling the width of the Gaussian kernel.

$$k(X_i, X_j) = e^{-\gamma \| (x_i - x_j) \|^2} \quad (3)$$

The second stage is blur identification:

- 1) Given a degraded image y , an optional preprocessing is performed in order to avoid the effect of the nonlinear sensor response, different exposures and white balancing. Defining $g(\cdot) \equiv f^{-1}(\cdot)$, one may adjust y as follows

$$y' = \frac{g(y) - \beta}{\alpha} \quad (4)$$
- 2) Extracting feature vectors of y' in the same way as that in the first stage.
- 3) Perform classification using the trained SVM in the first stage. The output of with the largest proportion is identified as the blur parameter.

3 Blind Super-Resolution Image Reconstruction

Super-resolution reconstruction is the process of reconstructing a high-resolution (HR) image from several low-resolution (LR) images, and simultaneously reduce the

influence of system degrade and additive noise. According to the acquisition system, the relationship between the LR images and the HR image can be formulated as [3]:

$$y_k = f(\alpha_k D_k H_k F_k x + \beta_k + n_k) + \eta_k = f(\alpha_k B_k x + \beta_k + n_k) + \eta_k \quad (5)$$

where $B_k = D_k H_k F_k$, y_k is the k -th LR image, x is the unknown HR image, F_k, H_k, D_k are geometric warp, camera lens/CCD blur and down-sampling respectively. Principle of SR reconstruction is that the LR images contain similar but different information. SR reconstruction is also possible from differently blurred images without relative motion. In this case, F_k is the identity matrix. In this paper we consider the motionless model.

Super-resolution reconstruction is an ill-posed problem. Inversion of the problem can be stabilized by regularization. In this paper a ML/POCS [7] combined with Space-Adaptive Regularization (SAR) algorithm is used to reconstruct the HR image. The cost function can be expressed as

$$\begin{aligned} J(x) &= \frac{1}{2} \left(\sum_{k=1}^K \left\| \frac{g(y_k) - \beta_k}{\alpha_k} - B_k x \right\|_{w1}^2 + \lambda \|Cx\|_{w2}^2 \right) \\ &= \frac{1}{2} \left(\left[\sum_{k=1}^K \left(\frac{g(y_k) - \beta_k}{\alpha_k} - B_k x \right)^T w_{1k} \left(\frac{g(y_k) - \beta_k}{\alpha_k} - B_k x \right) \right] + \lambda (Cx)^T w_2 (Cx) \right) \quad (6) \\ &\quad \text{subject to } \{x \in \Omega_p, 1 \leq p \leq P\} \end{aligned}$$

where C is Laplace operator, Ω_p is additional constraint, λ is regularization parameter which can be determined by $\lambda = \frac{\sigma^2}{E^2}$, where σ^2 is the variance of the noise, E^2

limits high-frequency power of HR image among some power range. The space adaptive is achieved by the weight $w1$ and $w2$.

From equation (2) and (5), it can be seen restoration and SR reconstruction are closely related theoretically, and SR reconstruction can be considered as a second-generation problem of image restoration, therefore some methods of image restoration can be extended to solve SR problem.

In this method we first identify the blur parameters of LR images using SVM-based identification method, then reconstruct SR image by equation (6). The whole procedure can be described as follows:

In the training stage, we first blur and down-sample the training image according to the acquisition system. Then up-sample the LR image with bilinear or bicubic interpolation. Finally, the SVM is used to train the feature vectors that extracted from the up-sampled images. In the second stage, give a LR image, after preprocessing, interpolation and feature extraction, the trained SVM provides the identification of the blur. After identifying the blur parameter of every LR image, SR image is reconstructed using the hybrid ML/POCS method. The ordered subset conjugate gradient optimization algorithm is used to minimize the cost function in (6).

4 Simulations

In this section, we present several experiment to demonstrate the effectiveness of our method. First we describe some of the experimental results obtained with the proposed blur identification algorithm in section 2, then present some of the experimental results obtained with the blind SR restoration algorithm in section 3.

4.1 Blur Identification in Image Restoration

In this experiment we use *lena* as training image. Some synthetic and real images that used for test are shown in Fig. 3. The degradation includes defocus blurring, gain, offset and noise. We assume that the blurring process is out-of-focus blur and is parameterized by the radius r . Different types of noise were added into the simulated images, such as Gaussian white noise, Salt&Pepper noise and Poisson noise, etc. For the SVM, we use the LibSVM software package [8] with the RBF kernel. Because the process is similar for all test images, we only take one image as example to illustrate our experiment. Fig.4 shows the blurred noisy images under different exposure time and illumination conditions. Fig.5 shows the identification curve where X-axis shows the parameter candidates used for training and Y-axis shows the proportion of SVM output. The value corresponding to the maximum point of the plot in X-axis is identified as the correct parameter. For convenience, the proportion is scaled to [0 1]. The selection of kernel parameters is important, in our implementation the parameter γ in (3) is determined by cross-validation. In our experiment, the noisy images were pre-processed before identification in order to enhance the validity of the method. Different methods were used for different noise. For example, median filter was utilized when impulse noise exists and multi-frame mean preprocess was used when poison noise exists. The same results can be got when the blur function is Gaussian which is parameterized by σ^2 .



Fig. 3. Some of the test images

4.2 Blind Super-Resolution Reconstruction

In this experiment, four images of the same scene are captured with different exposure rates and illumination conditions. We generate four LR images by blurring and down-sampling the images by 2. We assume that the blur function is out of focus and



(a) Blurred LR image (Gaussian white noise) (b) Blurred LR image (Salt & Pepper noise) (c) Blurred LR image (Poisson noise) (d) Blurred LR image (Uniformly distributed noise)

Fig. 4. Degraded images

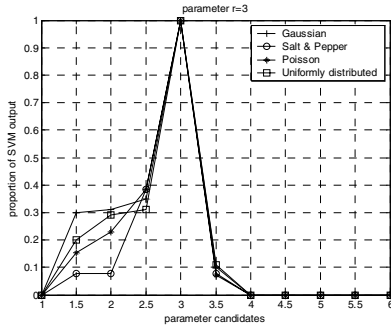
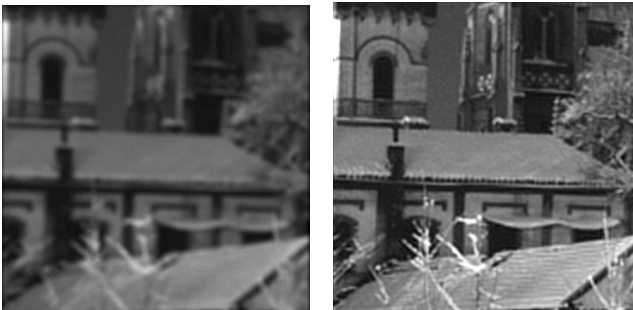


Fig. 5. Identification curve $r = 3$



(a) (b)

Fig. 6. (a) Bilinear interpolation; (b) Reconstructed HR image by hybrid ML/POCS approach

is parameterized by r . α_i and β_i are determined by least squares estimation after camera response function was estimated [9]. After blur identification, an HR image as shown in Fig. 6(b) was reconstructed. For comparison, one of the bilinear interpolated images is shown in Fig.6 (a). It can be seen the reconstructed image has more details and high dynamic range.

5 Conclusion

In this paper we propose a novel SVM-based blur identification method both for image restoration and resolution enhancement. SVM is used to classify feature vectors that extracted from the training images by Sobel operator and local variance, the acquired mapping between the vectors and corresponding blur parameter provides the identification of the blur. Meanwhile, the blind SR image reconstruction is achieved. Experimental results show that our method correctly identifies the blur parameters of images under different exposure rates and illumination conditions.

Acknowledgement. The work was supported by Program for New Century Excellent Talents in University (NCET), Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20050422017), Open Foundation of National Laboratory on Machine Perception, Peking University (0403) Open Foundation of National Laboratory of Networking and Switching, BUPT (07) and the Project-sponsored by SRF for ROCS, SEM ([2005]05).

References

1. Nguyen, N., Milanfar, P., Golub, G. Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. *IEEE Transactions on Image Processing*, 10(9) (2001) 1299 -1308.
2. Nakagaki R, Katsaggelos A K. A VQ-based blind image restoration algorithm. *IEEE Trans. Image Processing*, 12(9) (2003) 1044 -1053.
3. Gevrekci, M., Gunturk, B.K. Image Acquisition Modeling for Super-Resolution Reconstruction. *IEEE International Conference on Image Processing(ICIP)*, 2 (2005) 1058 -1061.
4. Robertson, M. A. High-Quality Reconstruction of Digital Images and Video from Imperfect Observations. Ph.D thesis (2001).
5. Vapnik, V. *The nature of statistical learning theory*. Springer-Verlag, New York, (1995)
6. Dalong Li, Mersereau, R.M., Simske, S. Blind Image Deconvolution Using Support Vector Regression. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2 (2005) 113 – 116.
7. Park, S.C., Park, M.K., Kang, M.G. Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Magazine*, 20 (2003) 21-36.
8. Chih-Chung Chang, Chih-Jen Lin, LIBSVM: a Library for Support Vector Machines, 2001. Available: <http://www.csie.ntu.edu.tw/~cjlin/ libsvm>.
9. Mann S., Mann R. Quantigraphic imaging: estimating the camera response and exposures from differently exposed images. *IEEE International Conference on Computer Vision and Pattern Recognition*, 1 (2001) 842–849.

Regularization for Super-Resolution Image Reconstruction

Vivek Bannore

School of Electrical and Information Engineering, University of South Australia,
Mawson Lakes, Adelaide, Australia

Vivek.Bannore@postgrads.unisa.edu.au

Abstract. Super-resolution image reconstruction estimates a high-resolution image from a sequence of low-resolution, aliased images. The estimation is an inverse problem and is known to be ill-conditioned, in the sense that small errors in the observed images can cause large changes in the reconstruction. The paper discusses application of existing regularization techniques to super-resolution as an intelligent means of stabilizing the reconstruction process. Some most common approaches are reviewed and experimental results for iterative reconstruction are presented.

1 Introduction

Under sampling of images occurs in many imaging sensors. It results in aliased imagery and, consequently, in partial loss of scene information. Super-resolution image reconstruction refers to image processing technique that attempts to reconstruct high quality, high-resolution images by utilising incomplete and degraded scene information contained in a sequence of aliased, low-resolution images. Super-resolution makes use of the fact that due to relative motion between the sensor and the scene each low-resolution image carries slightly different information about the scene. By fusing the partial information from many frames it is possible to reconstruct an image of higher spatial resolution [1, 2].

The problem of image reconstruction from noisy, aliased or otherwise degraded imagery occurs in a wide variety of scientific and engineering areas including civilian and military applications. Examples of these applications include: medical imaging, computer vision, target detection and recognition, radar imaging as well as surveillance applications. Many of these applications may also involve a related technique of image restoration. This technique, in contrast to super-resolution, does not attempt to increase pixel resolution but produces improved image from a degraded image at the same resolution scale.

In this paper we consider the problem of reconstructing a single high-resolution image X from N number of low-resolution, observed images Y_k ($k = 1 \dots N$) of the same scene. It is convenient to represent the images as vectors (as shown by an underscore) that are ordered column-wise lexicographically. Each observed image is the result of sampling, camera and atmosphere blur, motion effects, geometric

warping and decimation performed on the ideal high-resolution real scene. It is usually assumed that this imaging process can be represented by a linear operator H_k :

$$\underline{Y}_k = H_k \underline{X} + \underline{E} \quad \text{for } 1 \leq k \leq N . \quad (1)$$

where E is the additive noise present in any imaging system. The process of super-resolution is an inverse problem of estimating a high-resolution image from a sequence of observed, low-resolution images and it is now widely known to be intrinsically unstable or “ill-conditioned”. The common feature of such ill-conditioned problems is that small variations in the observed images Y_k can cause (arbitrary) large changes in the reconstruction. This sensitivity of the reconstruction process on the input data errors may lead to the restoration errors that are practically unbounded.

The important part of super-resolution process is thus to modify the original problem in such a way that the solution is a meaningful and close approximation of the true scene but, at the same time, it is less sensitive to errors in the observed images. The procedure of achieving this goal and to stabilize the reconstruction process is known as *Regularization*. The field of regularization has grown extensively [3-6] since the seminal paper by Tikhonov [7, 8] in 1963.

This paper is aimed at giving an overview of some most common regularization techniques and parameter estimations and their significance and application to the problem of super-resolution image reconstruction. We also show reconstruction results from a test sequence of images to illustrate our regularization procedure based on iterative approach. The paper is organized as follows: Section 2 describes the differences between well-posed and ill-posed problems and how regularization solves the problem of ill-conditioning. A few of the most common approaches to regularization parameter estimation and techniques are reviewed in section 3 and 4 respectively. Finally, we give our concluding remarks in section 5 along with our results.

2 Regularization

There are many ways of explaining well-posed and ill-posed problems. For example,

$$Hx = y . \quad (2)$$

where H is known. If y is determined by x , this is a well-posed problem whereas if x has to be determined from y , it's an inverse or ill-posed problem. The latter relates to super-resolution as explained in the introduction.

A problem whose solution exists, is unique and depends on the data continuously is known as a *well-posed* problem as defined by Hadamard [9] in 1902. On the contrary, the *ill-posed* problem is the one which disobeys the above given rules by Hadamard. In addition, as the solution of the ill-posed problem depends in a discontinuous fashion on the data, small errors such as round-off and measurement errors, may lead to a highly erroneous solution. The solution for an ill-posed problem is unstable and extremely sensitive to fluctuations in the data and other parameters. The classical example of an inverse and ill-posed problem is the Fredholm integral equation of the first kind, where, k is the kernel and g is the right-hand side.

$$\int_a^b k(t, s) f(s) dx = g(t) . \quad (3)$$

Both of these parameters are known, while f is the unknown function to be computed [10]. The theory on ill-posed problems is quite extensive and well developed. Engl [11] conducted a survey on a number of practical inverse problems in various applications such as computerised tomography, heat conduction, inverse scattering problems. Inverse problems are seen in various different fields, for example, medical imaging, astronomy, tomography, and many more. Ill-conditioning of inverse problems has always attracted a great deal of interest and research.

For many decades, it has been known that the best way to analyse a scientific problem is through its mathematical analysis. The most common analytical tool used in the case of ill-posed problems is *Singular Value Decomposition (SVD)*. This tool helps in diagnosing whether or not the singular values of a matrix are zero or decaying slowly towards zero (a number is so numerically small that due to the round off error it is rounded to zero). The SVD for a matrix A of dimension m by n where $m \geq n$, is given by:

$$A = U S V^T \Rightarrow A = \sum_{i=1}^n u_i s_i v_i^T . \quad (4)$$

For the above decomposition, $U (u_1, \dots, u_n)$ is an m by m and V^T is the transpose of matrix $V (v_1, \dots, v_n)$ which is n by n . The matrix S is a diagonal matrix containing the non-negative singular values of A arranged in descending order. The matrix U and V are orthogonal and their columns are orthonormal. The columns u_i and v_i of U and V are known as the left and right singular vectors. Also, for certain applications, as the dimension of matrix A increases, the numerical value of the singular values in S gradually decreases to zero which causes more oscillations in the left and right singular vectors. The greater the number of singular values in S tending to zero, the more singular is matrix A making it more ill-conditioned. Thus, SVD gives a good approximation on the ill-conditioning of the system.

Another easier way of testing a system for ill-conditioning is by computing the *condition number* of the matrix. The condition number can be defined as a ratio of the maximum and minimum singular values of the matrix in consideration, in our case, $H(2)$. A high condition number points to an ill-posed problem, whereas a low condition number points to a well-posed problem. If H is an m by n matrix:

$$condition(H) = \left| \frac{\sigma_{\max}(H)}{\sigma_{\min}(H)} \right|_{euclidean-norm} . \quad (5)$$

where, σ_{\max} and σ_{\min} represent the maximum and minimum singular values of matrix H . With ill-posed problems, the challenge is not of computing a solution, but computing a unique and stabilized solution. Thus, an ill-conditioned system requires

an intelligent method of mathematical computation to generate a meaningful solution, rather than the usual computational methods.

Referring to (1), the minimum norm solution for the estimation of high-resolution image would be:

$$\min \|Y_k - H_k X\|_2^2 \quad \text{for } 1 \leq k \leq N. \quad (6)$$

The matrix H is singular in nature and highly ill-conditioned. There is no uniqueness and stability in the solution for (6). Thus, to make the solution unique and stable, i.e. to make the above equation well-conditioned (as per Hadamard criteria), another term is added to (6) known as the *Regularization Term*. Most of the inverse problems (like super-resolution) are ill-posed and the solution is tremendously sensitive to the data. The solution can vary tremendously in an arbitrary manner with very small changes in the data. The solution to (6) would be highly sensitive and noise contaminated. The regularization term takes control of the ill-conditioned nature of the problem. The aim of this term is to make the solution more stable and less noise contaminated. The term also tries to converge the approximate solution as close as possible to the true solution. The modified version of (6) is:

$$\min \|Y_k - H_k X\|_2^2 + \lambda \|LX\|_2^2 \quad \text{for } 1 \leq k \leq N. \quad (7)$$

In (7), the parameter $\lambda > 0$, is known as the regularization parameter and L is a regularization / stabilization matrix. In [12], the stabilization matrix is referred to as the regularization operator. The regularization operator is given by an identity matrix ($L = I$), the regularization term is of *standard form* whereas when $L \neq I$, the term is in the *general form*. When treating problems numerically, it is easier to use the standard form rather than the general form as only one matrix, H , needs to be handled. In practical applications, however, it is recommended that the general form of the regularization term should be used.

The regularization term aims at filtering out the noise that contaminates the image and also makes it smoother. The regularization term can also include *a priori* information of the true solution which facilitates the minimization process to converge as close as possible. The regularization parameter controls the measure of smoothness in the final solution of (7). It is critical to choose the regularization parameter best suited to the particular application in which it is involved. If the regularization parameter is too small, the regularization term will have no effect on the solution and the noise will not be filtered out, thus leaving the approximate solution far from converging with the true solution. On the other hand, if the parameter is too large, the regularization term will have a dominating effect on the solution making it too smooth and there is a risk of losing important information from the solution. Hence, there needs to be a proper balance of smoothness and preservation of information when regularization is implemented.

There exist many techniques of regularization and parameter estimation in the literature. To discuss each of them is outside the scope of this paper. Thus, only those most commonly used will be discussed.

3 Estimating λ , the Regularization Parameter

Being the most critical part of regularization term, one has to carefully choose the appropriate technique based on their applications and expected results. The regularization parameter also depends on the properties of Y , H , X and noise (7). The parameter should balance the regularization and perturbation error in the computed solution. Over the years, many techniques have been proposed and discussed in relation to estimating the regularization parameter [12-14]. The techniques that will be discussed fall into two categories – one which require knowledge of error and the ones which do not require knowledge of error.

3.1 Method Which Require Error Knowledge – The Discrepancy Principle

In practical scenarios, considering (2) and (6), the right –hand side, Y , is never free from errors and contains various types of errors. Thus, Y can be written as $Y = Y_{true} + e$, where e is the errors and Y_{true} is the actual unperturbed right-hand side. Now, as per the discrepancy principle [15], the regularization parameter is chosen such that the residual norm of the regularized solution is equal to the norm of the errors.

$$\|Y - HX_{reg}\|^2 = \|e\|^2 . \quad (8)$$

If there is a rough estimate of the error norm, the discrepancy principle can be used to estimate a good regularization parameter. Unfortunately, in the practical world the knowledge about the error norm is not available and can be erroneous. Such data can lead to wrong estimations of the regularization parameter, thereby generating an unstable final solution.

3.2 Methods Which Do Not Require Error Knowledge – GCV and L-Curve

Generalized Cross-Validation (GCV)

GCV is one of the most popular methods used for estimating the regularization parameter [16]. It is based on the statistical cross-validation technique. In GCV, if a random element, Y_k , is left out of Y , then the estimated regularized solution should be able to predict the missing element, Y_k . The regularization parameter is chosen as the one which minimizes the prediction error and is independent of the orthogonal transformation of Y [17]. In this technique, no knowledge of the error norm is required. The GCV function is given as:

$$GCV = \frac{\|Y - HX_{reg}\|^2}{\tau^2} . \quad (9)$$

where, the numerator is the squared residual norm and the denominator is the squared effective number of degrees of freedom. For further details on this refer to chapter 7 from [18]. Although computation of regularization parameter using GCV technique

works for many applications, it should also be noted that GCV may have a very flat minimum, making it difficult to locate numerically [19].

L-Curve Criterion

The L-curve criterion proposed in [20, 21] was inspired from graphical analysis discussed in [22]. The L-curve is a plot of term 2 in (7) $\|LX\|^2$ or $\|X\|^2$ versus term 1 of (7) $\|Y - HX\|^2$ (which is the corresponding residual norm). This curve, when plotted on a log-log scale, takes the shape which resembles the alphabet ‘L’ and hence the name, L-Curve (see Figure 1. for illustration). This is the most powerful graphical tool for analysis as it shows the relationship between both the terms 1 and 2. The ‘corner’ of L-curve is the optimum point of balance between both the errors (one caused by regularization and the other by errors in Y). The value at this point (corner of L-curve) is chosen as the optimal regularization parameter. This is the L-curve criterion. The curve is continuous when the regularization parameter is continuous, but in the case when the parameter is discrete the curve is plotted as a set of points.

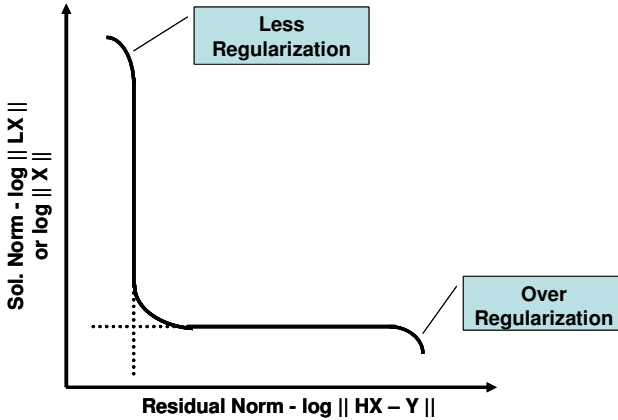


Fig. 1. A general graph of L-curve and its corner. The corner is the optimum point of balance between the regularization errors and errors in the right-hand side data, (Y). This corner can be taken as the regularization parameter.

4 Regularization Techniques

Regularization is an intelligent technique for computing a solution for an ill-posed problem. The main aim of this term is to make sure that the final solution is smooth and regularized with respect to the input data. It also makes sure that the final solution is less contaminated with errors and noise components. In the process of achieving this, the regularization term filters out the high-frequency components, thereby giving a smooth final solution. In the field of image restoration or super-resolution, a smooth approximate solution might not solve the purpose of being an appropriate solution. The high-frequency components filtered out by the regularization technique relate to the edges and discontinuities in the image. These components hold a significant value

in image restoration. As seen in section 2, the singular values of the matrix H , are of critical significance. These singular values relate to the high-frequency components. Thus, if there are too many small singular values (which can decay to zero), then the information relating to these is lost and only the information related to the large values is recoverable. There are various techniques for computing a regularized solution for ill-posed problems. The scope of this paper is limited and hence only the most common of these will be discussed.

4.1 Tikhonov Regularization

Tikhonov regularization was first introduced in 1963[7, 8]. It is defined as:

$$X_{reg.} = \min \|Y - HX\|_2^2 + \lambda \|LX\|_2^2 . \quad (10)$$

where, $\lambda > 0$, is known as the regularization parameter and L is a regularization/stabilization matrix. The regularization matrix can be $L = I$ or $L \neq I$ where I is an identity matrix. It is recommended to consider the regularization matrix as unequal to the identity matrix (see [12]). It should be noted that since the regularization matrix can also contain *a priori* knowledge, greater care must be taken in its selection. The regularization parameter is also of great importance as it is a trade-off between the smoothness and the accuracy of the solution. The above equation (10) can be also written as:

$$(H^T H + \lambda L^T L) X_{approx} = H^T Y . \quad (11)$$

From (11), it is evident how the regularization term manages to regularize the solution. It is also evident how the proper or improper selection of λ and L can lead to a good or bad approximation. A high value of λ diverts the solution to be very smooth, suppressing the high-frequency components even though the system has been regularized. Although Tikhonov regularization seems to be a straight forward technique, it has a high-computational cost and requires a lot of storage space when used in large-scale problems. Thus, this technique is more suitable to small-scale problems as compared to large-scale problems.

4.2 Maximum Entropy Method

The Maximum Entropy technique [23] of regularization is often used in astronomical image reconstruction. This technique is also known to preserve point edges in the estimated image, which makes it promising in the field of astronomical image restoration. The maximum entropy regularization term [18] is given as:

$$S(X) = \lambda^2 \sum_{i=1}^n x_i \log(w_i x_i) . \quad (12)$$

where, x_i are the positive elements of vector X and w_i are weights ($w_1 \dots w_n$). The above given function is negative of the entropy function Therefore, (10) is given as:

$$\|Y - HX\|_2^2 + S(X) . \quad (13)$$

The estimated solution from maximum entropy regularization is quite consistent as it is not related to the missing information of the right-hand side to a great extent. Although solving (12) and (13) is computationally intensive, there exist many iterative algorithms which are significantly less computationally intensive.

4.3 Conjugate Gradients (Iterative Regularization)

The conjugate gradient is one of the most commonly used numerical algorithms for symmetric positive definite systems. It is also known as the oldest and best known non-stationary method. The conjugate gradient can be computed as a direct method much like Tikhonov and maximum entropy but it proves to be much more efficient if it is used as an iterative method. Direct methods fail to perform when it comes to large-scale problems or huge sparse matrices, where only iterative technique comes to the rescue. The iterative conjugate gradient method can successfully compute solutions for large scale problems. Since the iterative method utilizes the property of matrix-vector multiplications between huge sparse matrices and vectors, computational time decreases and storage requirements for such matrices and vectors decreases tremendously. These advantages make iterative conjugate gradient regularization technique more favorable when compared with others. The iterative method generates successive approximations of the solution and their residuals. The conjugate gradient for a set of unregularized normal equations, $HX = Y$, is given as:

$$H^T HX = H^T Y . \quad (14)$$

It is seen that for (14), the low-frequency components of the estimated solution converge faster than the high-frequency components [12]. The iterative conjugate gradient technique generates X_K estimated solutions and calculates the residuals for each K . The number of iterations assigned is denoted by K . In this iterative technique of generating the regularized solution, K acts as the regularization parameter. It is very important to generate iterations up to an optimal number because the iterative solution can sometimes converge faster and if K is greater than K -optimal, the estimated solution might diverge from the true solution. The equation for the K^{th} iterative CG approach is given by:

$$X_{(K)} = X_{(K-1)} + \alpha_{(K)} P_{(K)} . \quad (15)$$

where, $X_{(K)}$ is the K^{th} iterative approximation of X . The conjugate gradient least squares is given by:

$$X_{reg.} = \min \|Y - HX\|_2^2 . \quad (16)$$

Equation (16) is similar to (10) – Tikhonov regularization technique only in (16), $\lambda = 0$, making the regularization term go to zero. Hence, in this technique, like in CG, the number of iterations, K , acts as the regularization parameter.

Computational cost and storage requirements are certainly the prime factors in choosing a particular regularization technique for a particular application. Iterative methods for estimating a regularized solution of an ill-posed problem are fast gaining popularity due to their low computational cost and low storage requirements as compared to direct methods of regularization.

5 Results and Concluding Remarks

It is a fact that super-resolution image reconstruction is an inverse problem which is highly ill-conditioned. If such a system (9) is solved, the image constructed would be highly sensitive and unstable. Thus, the term of regularization is introduced to make the final approximated solution less sensitive and more stable. The choice of technique used for regularization and estimation of the regularization parameter depends upon the application field and the expected output. In the field of super-resolution, the images are of band-limited nature, and hence, to restore the image, the Nyquist criterion needs to be fulfilled. The current regularization techniques are concentrated towards smoothing the final approximated or regularized image. The regularization matrix is taken in such a fashion that it blurs the regularized image by cutting off a major part of the high-frequency component.

We conclude with some experimental results of image reconstruction from simulated imagery. We have implemented an iterative technique for super-resolution that inherently stabilizes the reconstruction process without excessive blurring. In this approach the $K+1$ approximation to the high-resolution image is given by:

$$X_{K+1} = X_K + R_0(Y - H \cdot X_K), \rightarrow K = 0, 1, 2, \dots \quad (17)$$

where, H is the imaging operator and Y is the set of low-resolution images with X_0 being the first approximation input to the iterations algorithm. R_0 in (17) is an approximate reconstruction operator. In our approach the essential part of this operator is sub-pixel interpolation. In our initial experimentation we used truncated Sinc function for interpolation. From our experiments and figure 2, it can be seen that R_0 acts as a regularizing routine, cutting off the high-frequency components and noise and leaving a smooth approximated image. Truncated Sinc acts as an implicit regularization on the image. The extent of this regularization on the image can be controlled by the extent of the Sinc function. The Sinc function is also known as an ideal reconstruction filter, which in the frequency space, has a rectangular function. Although a true Sinc function cannot be used for reconstruction purposes, the regularization matrix can be chosen such that in frequency space it is like a rectangular function. Even if the rectangular function cuts off the high-frequency components, it doesn't blur the image and tries to preserve as much of the high-frequency components as possible from the band-limited image.

This paper signifies the role of regularization in the field of super-resolution and also reviews the most common regularization techniques. It is also recommended to use iterative regularization techniques rather than direct techniques for applications where computational cost and storage requirements are constraints.



Fig. 2. (a) – One of the 10 low-resolution images [42 x 42] simulated on the original image [512 x 512] using a sampling ratio of 12. (b) – The 20th final iterative super-resolution image generated [504 x 504] by our algorithm using truncated Sinc as the interpolation technique.

The plan is to take this research to the next step where we would implement the regularization technique with the new idea of considering the regularization matrix such that in frequency space, its response is more like the rectangular function. Such a technique will help to preserve the edges, rather than blurring it, thereby keeping the significant information intact. The problem of super-resolution is rewritten to combine the linear operator H which represents the imaging process along with the regularization term, such that (10) is given as:

$$\min \left\| \begin{bmatrix} H \\ \lambda L \end{bmatrix} X - \begin{bmatrix} Y \\ 0 \end{bmatrix} \right\|_2^2 \Rightarrow \min_x = \| \hat{H}X - \hat{Y} \|_2^2 . \quad (18)$$

The above equation is a least squares problem and can be solved using an iterative approach so as to tackle huge and sparse matrices.

Acknowledgements. This work is supported partially by Defence Science & Technology Organisation. The author would like to thank Lakhmi Jain, Noel Martin and Leszek Swierkowski for their guidance and support.

References

1. *Super-Resolution Imaging*. 1st ed, ed. S. Chaudhuri. 2001: Kluwer Academic Publishers. 279.
2. Hardie, R.C., et al., *High-Resolution Image Reconstruction from a Sequence of Rotated and Translated Frames and It's Application to an Infrared Imaging System*. Optical Engineering, 1998 Jan. **37**(1): p. 247-260.
3. Tikhonov, A.N., et al., *Numerical Methods for the Solution of Ill-Posed Problems*. 1995, Netherlands: Kluwer Academic.
4. Tikhonov, A.N. and V.Y. Arsenin, *Solutions of Ill-Posed Problems*. 1977, Washington, DC: John Wiley & Sons.
5. Morozov, V.A., *Regularization Methods for Ill-Posed Problems*. English ed. 1993, Boca Raton: CRC Press.

6. Tikhonov, A.N., ed. *Ill-Posed Problems in Natural Sciences*. 1991 Aug., TVP, Sci. Publishers: Moscow, Russia.
7. Tikhonov, A.N., *Regularization of Incorrectly Posed Problems*. Soviet Math. Dokl., 1963. **4**: p. 1624-1627.
8. Tikhonov, A.N., *Solution of Incorrectly Formulated Problems and the Regularization Method*. Soviet Math. Dokl., 1963. **4**: p. 1035-1038.
9. Hadamard, J., *Sur les problèmes aux dérivées partielles et leur signification physique. (On the problems with the derivative partial and their physical significance)*. Princeton University Bulletin, 1902: p. 49-52.
10. Groetsch, C.W., *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Research Notes in Mathematics:105. 1984, Research Notes in Mathematics:105, Boston: Pitman.
11. Engl, H.W., *Regularization methods for the stable solution of Inverse Problems*. Surveys on Mathematics for Industries, 1993. **3**: p. 77 - 143.
12. Hanke, M. and P.C. Hansen, *Regularization Methods For Large-Scale Problems*. Surveys on Mathematics for Industries, 1993. **3**: p. 253-315.
13. Hansen, P.C., *Numerical tools for analysis and solution of Fredholm integral equations of the first kind*. Inverse Problems, 1992. **8**: p. 849-872.
14. Galatsanos, N.P. and A.K. Katsaggelos, *Methods for choosing the Regularization Parameter and Estimating the Noise Variance in Image Restoration and Their Relation*. IEEE Transactions on Image Processing, 1992 Jul. **1(3)**: p. 322-336.
15. Morozov, V.A., *On the Solution of Functional Equations by the method of Regularization*. Soviet Math. Dokl., 1966. **7**: p. 414-417.
16. Golub, G.H., M.T. Heath, and G. Wahba, *Generalized Cross-Validation as a method for choosing a good Ridge Parameter*. Technometrics, 1979. **21**: p. 215-223.
17. Wahba, G., *Spline Model for Observational Data*. CBMS-NSF regional conference series in applied mathematics;59. Vol. 59. 1990, Philadelphia: Society for Industrial and Applied Mathematics.
18. Hansen, P.C., *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. 1998, Philadelphia: SIAM.
19. Varah, J.M., *Pitfalls in the Numerical Solution of Linear Ill-Posed Problems*. SIAM J. Sci. Stat. Comput., 1983 Jun. **4(2)**: p. 164-176.
20. Hansen, P.C., *Analysis of Discrete Ill-Posed Problems by means of the L-Curve*. Siam Review, 1992 Dec. **34(4)**: p. 561-580.
21. Hansen, P.C. and D.P. O'Leary, *The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems*. Siam J. Sci. Comput., 1993 Nov. **14(6)**: p. 1487-1503.
22. Lawson, C.L. and R.J. Hanson, *Solving least squares problems*. 1974, Englewood Cliffs, N.J.: Prentice Hall.
23. Zhuang, X., E. Ostevold, and M. Haralick, *A Differential Equation Approach To Maximum Entropy Image Reconstruction*. IEEE Trans. Acoust., Speech, Signal Processing, 1987 Feb. **ASSP-35(2)**: p. 208-218.

Dynamic Similarity Kernel for Visual Recognition

Wang Yan, Qingshan Liu, Hanqing Lu, and Songde Ma

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
P.O. Box 2728, Beijing, P.R. China, 100080
{wyan, qsliu, luhq, masd}@nlpr.ia.ac.cn

Abstract. Inspired by studies of cognitive psychology, we proposed a new dynamic similarity kernel for visual recognition. This kernel has great consistency with human visual similarity judgement by incorporating the perceptual distance function. Moreover, this kernel can be seen as an extension of Gaussian kernel, and therefore can deal with nonlinear variations well like the traditional kernels. Experimental results on natural image classification and face recognition show its superior performance compared to other kernels.

1 Introduction

Researchers on computer vision and pattern recognition, regularly confront the problem of nonlinearity in high-dimensional data, such as variation in image due to the change of illumination and viewpoint. Generally speaking, there are two schemes to solve the problem. The first one is to extend the existing linear method by so-called kernel trick [1], such as extensions of Principal Component Analysis (PCA) [2] and Linear Discriminant Analysis (LDA) [3]. Recently, these kernel methods attract much attention and show promising performance in many applications, such as image retrieval [4, 5, 6], face recognition [7, 8] and object recognition [9, 10]. Using nonlinear functions to model or to transform the nonlinear data is the second solution, such as neural network [11] and manifold learning [12, 13, 14]. We focus on the kernel methods in this paper.

The basic idea of the kernel methods is first mapping data into an implicit feature space with a nonlinear function, and then analyzing the image of data. Instead of explicitly computing the mapping, a kernel function is adopted to calculate the inner product of implicit feature vectors. This function determines the nonlinear mapping implicitly. If the solution of a linear method is able to be expressed by the inner products of its input samples, this method can be extended to the nonlinear version. The kernel trick is first introduced into Support Vector Machine (SVM) in [15]. Similar to SVM, PCA and LDA are extended to their nonlinear forms, i.e. Kernel PCA (KPCA) [16] and Generalized Discriminant Analysis (GDA) [17, 18]. PCA generates a set of orthonormal projections which maximizes the covariance over all samples, while LDA seeks a linear transformation which maximizes the inter-class scatter and minimizes the intra-class

scatter. Their nonlinear extensions, i.e. KPCA and GDA, do the same thing in feature space.

To measure (dis)similarity between images, Minkowski distance functions are commonly used, in which all coordinates of image feature vector are employed. But studies of cognitive psychology [19, 20] show that this approach does not model human perception well, which actually infers overall similarity based on the aspects that are similar among the compared objects, rather than based on the dissimilar ones. From this point of view, Dynamic Partial Function (DPF) [21, 22] is proposed, which can be seen as an extension of Minkowski distance. By dynamically selecting features with minimum differences in a pair-wise fashion, DPF works similarly to human process of similarity judgement. Extensive experiments in [21, 22] show its effectiveness.

A modified kernel function based on geodesic distance [12] is proposed in [23], and experimental results show its promising performance. But it still has some difficulties modeling human perception, because the estimation of geodesic distance is based on Minkowski distance. In this paper, we present a new kernel based on DPF, which has great consistency with human visual similarity judgement. Moreover, this kernel can be seen as an extension of Gaussian kernel, and therefore can deal with nonlinear variations well like the traditional kernels. By using KPCA and GDA on Corel and FERET database, we test this kernel and compare it with other traditional kernels, such as Gaussian and polynomial kernels. The experimental results show that the proposed kernel outperforms the others.

The rest of this paper is organized as follows. In Section 2, we describe Dynamic Partial Function for measuring image similarity. Dynamic Similarity Kernel is proposed in Section 3. Section 4 presents the experimental results on natural image classification and face recognition, followed by conclusions in Section 5.

2 Dynamic Partial Function

Minkowski distance is widely used for measuring similarity between images. Let $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ be two image feature vectors, the distance under Minkowski metric is defined as

$$D(X, Y) = \left(\sum_{i=1}^n \Delta d_i^r \right)^{\frac{1}{r}}, \quad (1)$$

where $\Delta d_i = |x_i - y_i|$, and r is the norm factor. When we set $r = 2$, the above is the well known Euclidean or L2 distance. When we set $r = 1$, it is the City-block or L1 distance.

Studies of cognitive psychology [20] show that human perception of similarity is the process that determines the respects for measuring similarity. Human infers overall similarity based on the aspects that are similar among the compared objects, rather than based on the dissimilar ones. [21] verifies this notion by extensive experiments on natural image database. From this point of view, the

similarity based on Minkowski distance which incorporates all aspects of the compared objects is questionable.

Dynamic Partial Function is proposed in [21, 22] to solve the above problem. It is a modified version of Minkowski distance. For the sake of clarity, we first assume Δd_i s to be ordered as $\Delta d_1 \leq \Delta d_2 \leq \dots \leq \Delta d_n$. The DPF is defined as

$$DPF(X, Y) = \left(\frac{1}{m} \sum_{i=1}^m \Delta d_i^r \right)^{\frac{1}{r}}, \quad (2)$$

where $m \leq n$ is the number of aspects activated in similarity measurement. Different from Minkowski distance, DPF dynamically selects the subset of features that are most similar for a given pair of images, and computes the similarity based on it. Hence, it works in a similar way as human visual perception system does. Empirical studies in [21] show its performance to be superior to other widely used distance functions, such as fractional function, histogram cosine distance function and Minkowski distance.

3 Dynamic Similarity Kernel

Kernel function is important for the kernel based methods. One can compute inner product in feature space efficiently by kernel function. Some kernels are based on Minkowski distance, such as Gaussian kernel and exponential kernel [24]. The former is defined as

$$K(X, Y) = \exp \left(-\frac{\|X - Y\|^2}{n\sigma^2} \right), \quad (3)$$

where σ^2 is the bandwidth, and n is the dimensionality of X and Y .

[23] substitutes geodesic distance for Euclidean distance in (3), and shows promising results. But the estimation of geodesic distance is based on Euclidean distance, and also employs all coordinates of the feature vector. So, it doesn't model human visual perception well, either.

To overcome this defect, we proposed a new kernel based on DPF as follows.

$$K(X, Y) = \exp \left(-\frac{DPF^2(X, Y)}{\sigma^2} \right). \quad (4)$$

A kernel dot product can be seen as a similarity measure. The similarity measured by the proposed kernel is more accurately than Gaussian kernel, because of DPF's consistency with human perception. It dynamically activates the most similar aspects when evaluating the similarity of image pair. For its dynamic nature, we call it Dynamic Similarity Kernel (DSK for short). Since DPF is an extension of Minkowski distance, Gaussian kernel can be seen as a special case of DSK. We set $r = 2$ for DPF in (4) for analogue with Gaussian kernel.

4 Experiments

To examine the effectiveness of the proposed kernel, we compare it with other traditional kernels in some classic kernel methods, i.e. KPCA and GDA. The results are also compared with state-of-the-art. Two types of experiments are conducted. One is natural image classification, and the other is face recognition. They are both current hot topics in computer vision and pattern recognition.

4.1 Natural Image Classification

The WANG dataset [25,26] is commonly used in the literature to evaluate image retrieval and classification methods. It consists of 10 image categories, each of which contains 100 images. The themes of these categories are African people and villages, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountain and glaciers, and food.

Three types of features are used to represent image: color histogram, color moment and wavelet based texture. Color histogram is taken in HSV space with quantization of $8 \times 8 = 64$ bins on H and S channels. The first three moments from each of the three color channels are used as color moment. Mean and variance of each channel in a 4-level PWT decomposition form the wavelet based texture feature. In total, a 105-dimensional feature vector is extracted for each image. Each feature component is normalized such that the variance of which equals $\frac{1}{3}$.

We use a leave-one-out cross-validation as the testing protocol. That is, choose one image one time and use the rest 999 images as training set, then test the image left out. The classification accuracy is averaged on the whole dataset. Nearest neighbor classifier is used for classification.

There are two parameters in DSK, i.e. m and σ^2 . For the latter one, we first compute pair-wise distances for all images in the dataset, and then adaptively set σ^2 to the mean of the distances' square. Different values of m are investigated, and the results are shown in Table 1. It can be seen that $m = 80$ gives the best classification accuracy, which means about 25 dimensions are dissimilar aspects for pair-wise matching. We set $m = 80$ in the following experiment.

Table 1. Accuracies (%) on image classification with different m values

m	65	70	75	80	85	90	95	100	105
Accuracy	82.8	84.0	84.7	85.4	85.2	84.4	84.0	83.7	78.3

Next, we compare DSK with Gaussian kernel and polynomial kernels in KPCA. By investigating values in $[0.01, 1]$, we find that Gaussian kernel with $\sigma^2 = 0.27$ achieves the best result of 78.3%, and use this setting in the following comparison. There are two forms of polynomial kernels,

$$K(X, Y) = (X \cdot Y)^p, \quad (5)$$

and

$$K(X, Y) = (X \cdot Y + 1)^p, \quad (6)$$

noted as PolyI and PolyII, respectively. We set $p = 2$ for both because it gives the best results among $\{2, 3, 4, 5\}$.

Nearest neighbor classifier using DPF directly has been investigated, too. It is found that $m = 89$ gives the highest accuracy, i.e. 83.9%.

In Fig. 1, the recognition accuracies of four kernels are plotted versus the number of dimensions used by the classifier, and the best result of direct DPF is also plotted as the baseline. Except for the accuracy of 85.4% achieved by DSK, the best results of Gaussian kernel, PolyI and PloyII are 78.3%, 74.0% and 75.5%, respectively. Gaussian kernel is better than polynomial kernels. Because of similarity to human visual perception, DPF gets better result than traditional kernels. DSK outperforms all other kernels with large margins. It is noticeable that DSK can get even better result than DPF by incorporating Gaussian function to deal with nonlinearity. The best result in the literature is 84.1% [26, 27], while DSK breaks this record.

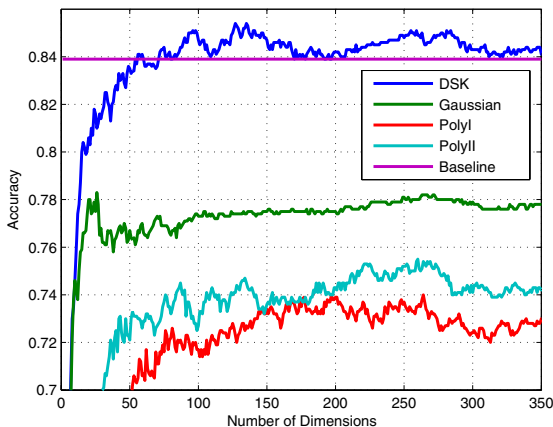


Fig. 1. Comparison of DSK, Gaussian kernel and polynomial kernels in image classification

4.2 Face Recognition

We conduct face recognition experiments on the FERET database. The experimental data include 1002 front view face images selected from the training CD, the FA set of 1196 subjects and the FB set of 1195 subjects. There's only one image per person in the FA and FB sets. Images from the training CD are used as training set. FA is used as gallery image set, and FB is taken as probe set. All images are normalized to 48×54 by eye locations, and histogram equalization is performed as preprocessing. The appearance of image, i.e. raw pixel value, is taken directly as the feature. Each feature component is normalized such that the variance of which equals 1.

GDA is adopted this time, for its popularity and excellent performance in face recognition [8, 28]. We compare DSK with Gaussian kernel and PolyI, while PloyII is excluded according to [28]. We set $\sigma^2 = 3.3$ in Gaussian kernel for the best performance, and set PolyI's parameter $p = 2$ as in [28]. σ^2 in DSK is set to 10 by investigating values from 0.1 to 20. Classification results under different values of m are listed in Table 2. It shows that $m/n = 0.8$ gives the best accuracy. These settings are used in the following experiment.

Table 2. Accuracies (%) on face recognition with different m values

m/n	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
Accuracy	93.56	93.97	95.15	96.23	96.99	97.41	97.15	96.99	95.98	92.22

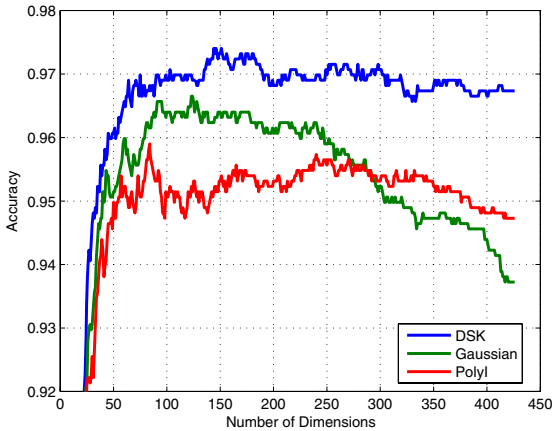


Fig. 2. Comparison of DSK, Gaussian kernel and polynomial kernels in face recognition

Fig. 2 shows the comparison of the three kernels. DSK gets the highest recognition rate of 97.41%, and best result of Gaussian kernel and PolyI are 96.65% and 95.9%, respectively. Once again, the superior performance implies that DSK is more consistent to human vision.

5 Conclusion

A kernel for visual recognition is proposed. By incorporating Dynamic Partial Function, the kernel has great consistency with human visual similarity judgment. Moreover, the kernel can be seen as an extension of Gaussian kernel. The adoption of Gaussian function enables its capability to deal with nonlinear variations effectively. Experimental results on natural image classification and face recognition show its superior performance compared to other classic kernels.

Acknowledgement

This work is partially supported by the National Key Basic Research and Development Program (973) under Grant No. 2004CB318107, and the Natural Sciences Foundation of China under Grant No. 60405005 and 60121302.

References

1. Aizerman, M., Braverman, E., Rozonoer, L.: Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* **25** (1964) 821–837
2. Jolliffe, I.: *Principal Component Analysis*. Springer-Verlag, New York (1986)
3. Fisher, R.: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** (1936) 179–188
4. Wang, L., Gao, Y., Chan, K., Xue, P., Yau, W.: Retrieval with knowledge-driven kernel design: An approach to improving SVM-based CBIR with relevance feedback. In: *Proc. of Int. Conf. Computer Vision*. (2005)
5. Wu, G., Chang, E., Panda, N.: Formulating context-dependent similarity functions. In: *Proc. of ACM Multimedia*. (2005)
6. Yan, W., Liu, Q., Lu, H., Ma, S.: Multiple similarities based kernel subspace learning for image classification. In: *Proc. of Asian Conf. Computer Vision*. (2006)
7. Yang, M., Ahuja, N., Kriegman, D.: Face recognition using kernel eigenfaces. In: *Proc. of Int. Conf. Image Processing*. (2000)
8. Liu, Q., Lu, H., Ma, S.: Improving kernel fisher discriminant analysis for face recognition. *IEEE Trans. on Circuits and Systems for Video Technology, Special Issue on Image and Video Based Biometrics* **14** (2004) 42–49
9. Lyu, S.: Mercer kernels for object recognition with local features. In: *Proc. of IEEE Computer Society Conf. Computer Vision and Pattern Recognition*. (2005)
10. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: *Proc. of Int. Conf. Computer Vision*. (2005)
11. Duda, R., Hart, P., Stork, D.: *Pattern Classification*. 2nd edn. John Wiley & Sons, Inc (2001)
12. Tenenbaum, J., Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290** (2000) 2319–2323
13. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** (2000) 2323–2326
14. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *Advances in Neural Information Processing System*. Volume 13., Cambridge, MA, MIT Press (2001)
15. Vapnik, V.: *Statistical Learning Theory*. John Wiley & Sons (1998)
16. Schölkopf, B., Smola, A., Müller, K.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10** (1998) 1299–1319
17. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.: Fisher discriminant analysis with kernels. In: *Proc. of IEEE Neural Networks for Signal Processing Workshop*. (1999)
18. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. *Neural Computation* **12** (2000) 2385–2404
19. Tversky, A.: Features of similarity. *Psychological Rev.* **84** (1977) 327–352

20. Goldstone, R.: Similarity, interactive activation, and mapping. *J. Experimental Psychology: Learning, Memory, and Cognition* **20** (1994) 3–28
21. Li, B., Chang, E., Wu, Y.: Enhancing DPF for near-replica image recognition. *ACM Multimedia J.*, special issue on content-based image retrieval **8** (2003) 512–522
22. Qamra, A., Meng, Y., Chang, E.: Enhanced perceptual distance functions and indexing for image replica recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **27** (2005) 379–391
23. Yong, Q., Jie, Y.: Modified kernel functions by geodesic distance. *EURASIP J. Applied Signal Processing* **16** (2004) 2515–2521
24. Gunn, S.: Support vector machines for classification and regression. Technical Report, School of Electronics and Computer Science, University of Southampton (1998)
25. Chen, Y., Wang, J.: Image categorization by learning and reasoning with regions. *J. Machine Learning Research* **5** (2004) 913–939
26. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: A quantitative comparison. In: *Proc. of 26th DAGM Symposium on Pattern Recognition*. (2004)
27. Marée, R., Geurts, P., Piater, J., Wehenkel, L.: Random subwindows for robust image classification. In: *Proc. of IEEE Computer Society Conf. Computer Vision and Pattern Recognition*. (2005)
28. Liu, Q., Huang, R., Lu, H., Ma, S.: Face recognition using kernel based Fisher discriminant analysis. In: *Proc. of Int. Conf. Automatic Face and Gestures Recognition*. (2002)

Genetic Algorithms for Optimization of Boids Model

Yen-Wei Chen^{1,2}, Kanami Kobayashi¹, Xinyin Huang³, and Zensho Nakao⁴

¹ School of Information Science and Eng., Ristumeikan Univ., Shiga 525-8577, Japan

² College of Elect. and Information Eng., Central South Forest Univ., Changsha 410004, China

³ School of Education, Soochow University, Suzhou, Jiangsu 215006, China

⁴ Faculty of Eng., Univ. of the Ryukyus, Okinawa 903-0213, Japan

Abstract. In this paper, we present an extended boids model for simulating the aggregate moving of fish schools in a complex environment. Three behavior rules are added to the extended boids model: following a feed; avoiding obstacle; avoiding enemy boids. The moving vector is a linear combination of every behavior rule vector, and the coefficients should be optimized. We also proposed a genetic algorithm to optimize the coefficients. Experimental results show that by using the GA-based optimization, the aggregate motions of fish schools become more realistic and similar to behaviors of real fish world.

1 Introduction

Simulating the aggregate moving of a fish school or a bird flock is an important issue in the areas of computer animation and artificial life. In 1986, Reelond proposed a computer model of coordinated animal motion such as bird flocks and fish schools, which is called as boids [1]. The Boids model has three basic behavior rules, which are avoiding collision against neighbors; matching and coordinating own moves with neighbors; gathering together. The boids model has been used for modeling of fish [2]. In this paper, we present an extended boids model for simulating the aggregate moving of fish schools in a complex environment. Three behavior rules are added to the extended boids model: following a feed; avoiding obstacle; avoiding enemy boid. Each rule is represented by a vector. The direction and amplitude of the vector are adaptive to the environment. The moving vector of the boid (fish) is a linear combination of every behavior rule vector. As increasing the behavior rules, the setting of the coefficients becomes complex and difficult. We also proposed a genetic algorithm[3,4] to optimize the coefficients. Experimental results show that by using the GA-based optimization, the aggregate motions of fish schools become more realistic and similar to behaviors of real fish world.

The paper is organized as following: the extended boids model is presented in Sec.2, the genetic algorithm for optimization of coefficients is presented in Sec.3 and the experimental results are shown in Sec.4. Finally, the conclusion is given in Sec.5.

2 Extended Boids Model

The boids model is an example of an individual-based model. Each simulated boid (fish) is implemented as an independent actor that navigates according to its local perception of the dynamic environment. The global behavior of the school is

simulated by a large number of interacting individual boid (fish). In the extended boids model, each boid is an agent that follows following five behavior rules: avoiding collision against schoolmates; gathering together; following a feed; avoiding obstacle; avoiding enemy boids. The first two rules are Reynold's and the last three rules are our newly proposed ones.

2.1 Avoiding Collision Against Schoolmates

The first rule is avoiding collision against schoolmates. The rule is illustrated in Fig.1. The vector determined by the first rule is shown in Eq.(1).

$$\mathbf{V}_1 = \begin{cases} \left(\frac{|\mathbf{BoidVec}|}{fKeepDist} - 1 \right) \cdot \frac{\mathbf{BoidVec}}{|\mathbf{BoidVec}|} & (|\mathbf{BoidVec}| \leq fVisibleDist) \\ 0 & (|\mathbf{BoidVec}| > fVisibleDist) \end{cases}, \quad (1)$$

where $fVisibleDist$ is the visible distance of the boid (fish), $fKeepDist$ is the safe distance for avoiding collision against schoolmates, and $\mathbf{BoidVec}$ is the vector from the boid to the nearest schoolmate. As shown in Eq.(1), when the distance to the nearest schoolmate is smaller than $fKeepDist$, a vector (force) is acted in opposite direction in order to keep away from the schoolmate. On the other hand, when the distance to the nearest schoolmate is larger than $fKeepDist$, a vector (force) is acted in the same direction in order to close to the schoolmate.

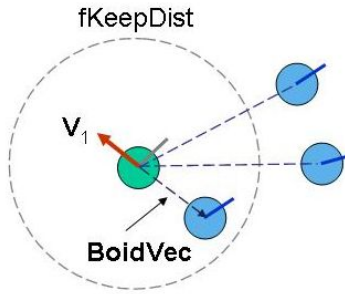


Fig. 1. Rule 1: avoiding collision against schoolmates

2.2 Gathering Together

The second rule is gathering together. A vector (force) is acted in the direction to the center (average position) of the neighborhood (fish school) in the view as shown in Fig.2. The vector is given by Eq.(2).

$$\mathbf{V}_2 = \begin{cases} \frac{\mathbf{CenterVec}}{|\mathbf{CenterVec}|} & (|\mathbf{CenterVec}| \leq fVisibleDist) \\ 0 & (|\mathbf{CenterVec}| > fVisibleDist) \end{cases}, \quad (2)$$

where $\mathbf{CenterVec}$ is the vector from the boid to the center of the neighborhood.

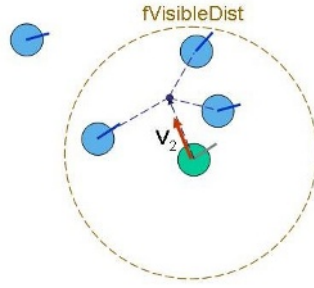


Fig. 2. Rule 2: gathering together

2.3 Following a Feed

The third rule is following a feed. A vector (force) is acted in the direction to the feed as shown in Fig.3. The vector is given by Eq.(3).

$$\mathbf{V}_3 = \frac{\mathbf{FoodVec}}{|\mathbf{FoodVec}|}, \quad (3)$$

where $\mathbf{FoodVec}$ is the vector from the boid to the feed.

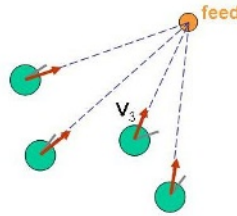


Fig. 3. Rule 3: following a feed

2.4 Avoiding Obstacles

The fourth rule is avoiding obstacles. Obstacle avoidance allowed the boids to fly through simulated environments while dodging static objects. The rule is illustrated in Fig.4. Assuming the avoiding angle be α and the size of obstacle be $ObsMag$.

$$\cos \alpha = \frac{\sqrt{|\mathbf{ObsVec}|^2 - \left(\frac{ObsMag}{2}\right)^2}}{|\mathbf{ObsVec}|}, \quad (4a)$$

where \mathbf{ObsVec} is the vector from the boid to the center of obstacle as shown in Fig.4. The vector acted for avoiding obstacle is given as

$$\mathbf{V}_4 = \begin{cases} -\cos\theta \cdot \left(1 - \frac{|\mathbf{ObsVec}|}{fVisibleDist}\right) \cdot \frac{\mathbf{ObsVec}}{|\mathbf{ObsVec}|} & (\cos\theta \geq \cos\alpha) \\ 0 & (\cos\theta < \cos\alpha) \end{cases}, \quad (4b)$$

where θ is the angle of current direction of the boid with the obstacle.

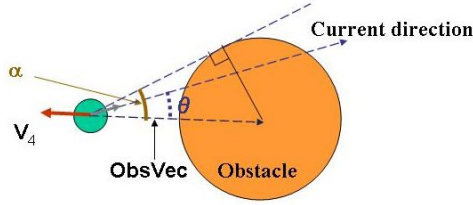


Fig. 4. Rule 4: Avoiding obstacles

2.5 Avoiding Enemy Boids

The fifth rule is avoiding enemy boid. When the boid finds an enemy boid in the visible distance, a vector (force) is acted in the opposite direction to the enemy boid as shown in Fig.5. The vector is given by

$$\mathbf{V}_5 = \begin{cases} \left(\frac{|\mathbf{OtherVec}|}{fVisibleDist} - 1 \right) \cdot \frac{\mathbf{OtherVec}}{|\mathbf{OtherVec}|} & (|\mathbf{OtherVec}| \leq fVisibleDist) \\ 0 & (|\mathbf{OtherVec}| > fVisibleDist) \end{cases}, \quad (5)$$

where $\mathbf{OtherVec}$ is the vector from the boid to the enemy boid.

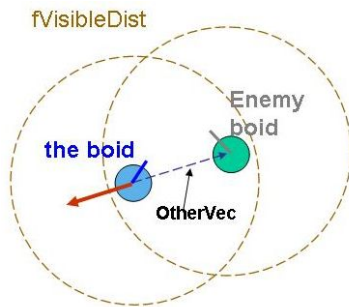


Fig. 5. Rule 5: Avoiding enemy boids

2.6 The Moving Vector

The moving vector of each boid is determined by above five rules. The moving vector can be considered as a linear combination of the five vectors as

$$\mathbf{V} = w_1 \mathbf{V}_1 + w_2 \mathbf{V}_2 + w_3 \mathbf{V}_3 + w_4 \mathbf{V}_4 + w_5 \mathbf{V}_5 \quad (6)$$

where w_i is the coefficients used to balance the five rules and the coefficients should be optimized.

3 Genetic Algorithms for Optimization of the Moving Vector

As shown in Eq.(6), the moving vector of each boid is a linear combination of five vectors which are determined by each rule and the coefficients should be optimized. In this paper, we propose to use a genetic algorithm (GA) [3,4] for optimization of coefficients. GA applies the principles of evolution found in nature to the problem of finding an optimal solution. Since GA starts with a population of candidate solutions, it is easy to find a global optimum. In our previous works, we have applied the GA to image processing [5-7].

The flowchat of GA is shown in Fig.6. We use a real coding to represent chromosomes. The chromosome has five bits and each bit corresponds to w_1 , w_2 , w_3 , w_4 and w_5 , respectively. A roulette wheel selection is used as a selection operator. A two points crossover is used to generate two children from two selected parents. In the two points crossover, two points are randomly selected and everything between the two points is swapped between the parent organisms. We use Eq.(7) for mutation.

$$x' = x_l + \beta(x_u - x_l) \quad (7)$$

where x_u and x_l are upper limit and lower limit of the coefficients, respectively. β is a random value between 0 and 1. The chromosome and the bit for mutation are randomly selected.

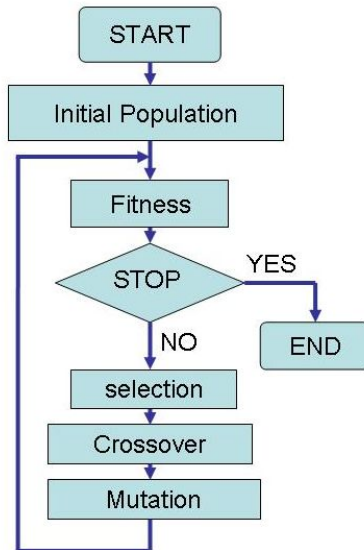


Fig. 6. Flowchat of GAs

The cost function and fitness function are shown in Eqs (8)-(14).

$$Cost1 = \begin{cases} \frac{fVisibleDist \cdot (|\mathbf{BoidVec}| - fKeepDist)^2}{fKeepDist^2} & (|\mathbf{BoidVec}| \leq fKeepDist) \\ \frac{fVisibleDist \cdot (|\mathbf{BoidVec}| - fKeepDist)^2}{(fVisibleDist - fKeepDist)^2} & (|\mathbf{BoidVec}| > fKeepDist) \end{cases}, \quad (8)$$

$$Cost2 = \begin{cases} \frac{fVisibleDist \cdot \left(|\mathbf{FoodVec}| - \frac{1}{3} \cdot fVisibleDist \right)^2}{\left(\frac{1}{3} \cdot fVisibleDist \right)^2} & \left(|\mathbf{FoodVec}| \leq \frac{1}{3} \cdot fVisibleDist \right) \\ 0 & \left(\frac{1}{3} \cdot fVisibleDist < |\mathbf{FoodVec}| < \frac{2}{5} \cdot fVisibleDist \right) \\ \frac{fVisibleDist \cdot \left(|\mathbf{FoodVec}| - \frac{2}{5} \cdot fVisibleDist \right)^2}{\left(\frac{2}{5} \cdot fVisibleDist \right)^2} & \left(|\mathbf{FoodVec}| \geq \frac{2}{5} \cdot fVisibleDist \right) \end{cases}, \quad (9)$$

$$Cost3 = \begin{cases} \left(\frac{|\mathbf{ObsVec}| - 3 \cdot fKeepDist}{\frac{ObsMag}{2} - 3 \cdot fKeepDist} \right)^2, & \left(\left(\frac{ObsMag}{2} - fFishSize \right) < |\mathbf{ObsVec}|, |\mathbf{ObsVec}| \leq 3 \cdot fKeepDist \right) \end{cases}. \quad (10)$$

$$Cost4 = \begin{cases} \frac{(|\mathbf{OtherVec}| - 2 \cdot fKeepDist)^2}{2 \cdot fKeepDist^2} & (|\mathbf{OtherVec}| \leq 2 \cdot fKeepDist), \end{cases} \quad (11)$$

$$Cost5 = \begin{cases} \frac{2 \cdot (fDirVecLen - 1.5 \cdot UniqueSpeed)^2}{fVisibleDist} & (1.5 \cdot UniqueSpeed < fDirVecLen, fDirVecLen \leq 10 \cdot UniqueSpeed) \\ 0 & (0.5 \cdot fSchoolSpeed \leq fDirVecLen \leq 1.5 \cdot fSchoolSpeed) \\ \frac{fVisibleDist \cdot |fDirVecLen - 0.5 \cdot UniqueSpeed|}{0.5 \cdot UniqueSpeed} & (fDirVecLen < 0.5 \cdot UniqueSpeed) \end{cases}, \quad (12)$$

$$Cost = Cost1 + Cost2 + Cost3 + Cost4 + Cost5. \quad (13)$$

$$Fitness = \frac{1}{1 + Cost}. \quad (14)$$

4 Experimental Results

We have made an interactive fish school system [8] based on the extended boids model and the system is made by Open GL [9]. The examples of the system are shown in Fig.7. Two fish schools are simulated in the system. Figure 7(a) is a result without GA-based optimization and Fig.7(b) is a result with GA-based optimization

(after 100 generations). It can be seen that by using the GA-based optimization, the aggregate motions of fish schools become more realistic and similar to behaviors of real fish world.

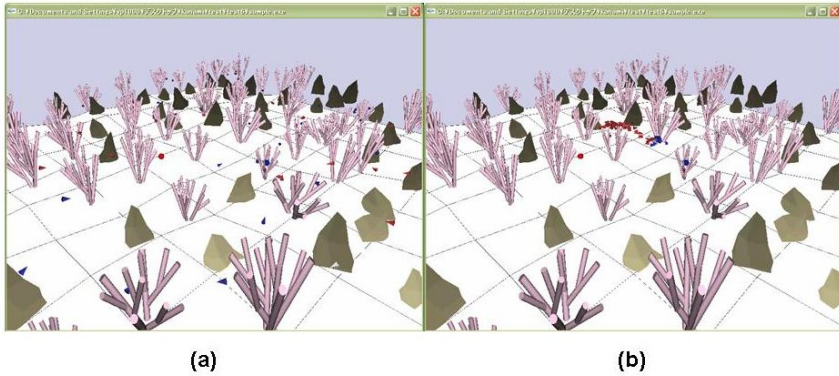


Fig. 7. Experimental results. (a) without GA, (b) with GA.

5 Conclusions

In this paper, we proposed an extended boids model for simulating the aggregate moving of fish schools in a complex environment. Three behavior rules were added to the extended boids model: following a feed; avoiding obstacle; avoiding enemy boids. We also proposed a genetic algorithm to optimize the coefficients. Experimental results showed that by using the GA-based optimization, the aggregate motions of fish schools become more realistic and similar to behaviors of real fish world.

This work was supported in part by the "Open Research Center" Project for Private Universities: matching fund subsidy from MEXT (Ministry of Education, Culture, Sports, Science, and Technology).

References

1. Reynolds, C. W.: Flocks, Herds, and Schools: A Distributed Behavioral Model. *Computer Graphics*, 21(1987) 25-34.
2. DeAngelis, D. L., Shuter, B. J., Ridgeway, M. S., Blanchfield, P., Friesen, T., and Morgan, G. E.: Modeling early life-history stages of smallmouth bass in Ontario lakes. *Transaction of the American Fisheries Society*, (1991) 9-11.
3. Goldberg, D.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, (1989).
4. Forrest, S.: Genetic algorithms: principles of natural selection applied to computation. *Science*, 261(1993) 872-878.
5. Chen, Y. W., Nakao, Z., Arakaki, K., Fang, X., and Tamura, S.: Restoration of Gray Images Based on a Genetic Algorithm with Laplacian Constraint. *Fuzzy Sets and Systems*, 103 (1999)285-293.

6. Chen, Y. W., Nakao, Z., Arakaki, K., and Tamura, S.: Blind Deconvolution Based on Genetic Algorithms. *IEICE Trans. Fundamentals*, E-80-A (1997)2603-2607.
7. Mendoza, N., Chen, Y. W., Nakao, Z., and Adachi, T.: A hybrid optimization method using real-coded multi-parent EA, simplex and simulated annealing with applications in the resolution of overlapped signals. *Applied Soft Computing*, 1(2001)225-235.
8. Kobayashi, K.: Interactive fish school generation system using GA. Graduation thesis of Ritsumeikan Univ., (2005).
9. Shreiner, D.: *Open GL Reference Manual: The Official Reference Document to Open GL, Version 1.4*. Addison-Wesley, (2004).

Segmentation of MR Images Using Independent Component Analysis

Yen-Wei Chen^{1,2} and Daigo Sugiki¹

¹ College of Electronic and Information Engineering, Central South Forest University, Changsha 410004, China

² College of Information Science and Eng., Ritsumeikan Univ., Shiga 525-8577, Japan
chen@is.ritsumei.ac.jp

Abstract. Automated segmentation of MR images is a difficult problem due to complexity of the images. In this paper, we proposed a new method based on independent component analysis (ICA) for segmentation of MR images. We first extract three independent components from the T1-weighted, T2-weighted and PD images by using ICA and then the extracted independent components are used for segmentation of MR images. Since ICA can enhance the local features, the MR images can be transformed to contrast-enhanced images by ICA. The effectiveness of the ICA-based method has been demonstrated.

1 Introduction

Automated segmentation of MR images is an important step for quantitative studies and 3-D visualization of anatomical structures. It has been studied from different viewpoints [1-4]. As a statistical classification task, the segmentation includes a strategy of feature extraction and classification [5]. For single spectral images, multiscale features obtained by local filtering have been used to represent the spatial features of MR images [6,7]. The effectiveness, however, is limited by the crucial choice of filter parameters. In MR analysis, we usually use several MR images (multi-spectral MR images) taken with different conditions, such as T1-weighted image, T2-weighted image, and proton density (PD) image. Most traditional segmentation methods based on spectral features use the spectral images independently [8,9]. They utilize the contrast between different brain tissues in the three images, for instance, the CSF has uniquely bright intensity in the T2 image and the cerebrum region is relative bright in the PD image. Nakai [10] applied independent component analysis (ICA) to MR imaging for enhancing the contrast of gray matter and white matter. In this paper, we propose an ICA-based segmentation method to improve the segmentation performance. In the proposed method, we first extract three independent components, which are contrast-enhanced images [10], from the T1-weighted, T2-weighted and PD images by using ICA and then the extracted independent components are used for segmentation of MR images. Experimental results show that the segmentation performance can be significantly improved by using ICA.

The paper is organized as following: the proposed ICA-based segmentation method is presented in Sec.2 and the experimental results with real MR images are shown in Sec.3. Finally, the conclusion is given in Sec.4.

2 Independent Component Analysis for Segmentation of MR Images

In MR image segmentations, T1, T2 and PD images are usually used as shown in Fig.1(a). The classifier will be neural networks (supervised segmentation) or the K-means algorithm (un-supervised segmentation) and other clustering algorithms. In this paper, we proposed a new method based on independent component analysis (ICA) for segmentation of MR images as shown in Fig.1(b). We first extract three independent components from the T1-weighted, T2-weighted and PD images by using ICA and then the extracted independent components are used for segmentation of MR images. Since ICA can enhance the local features, the MR images can be transformed to contrast-enhanced images (independent components) by ICA.

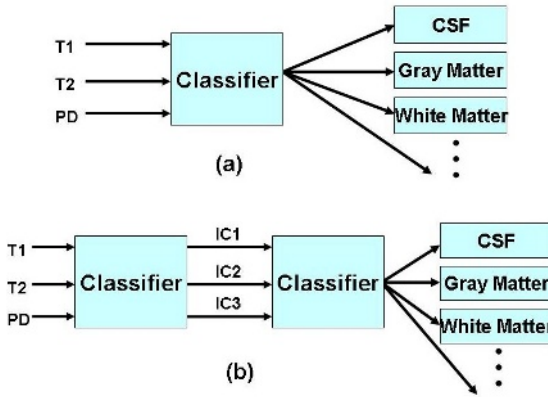


Fig. 1. (a) Conventional pixel-based method; (b) proposed region-based method

Independent component analysis (ICA) is an algorithm by which the transformed variable are not only decorrelated, but also as statistically independent from each other as possible [11, 12].

In the simplest model of ICA, we observe m scalar random variables x_1, x_2, \dots, x_m which are assumed to be linear combinations of n unknown independent components (sources) s_1, s_2, \dots, s_n that are mutually statistically independent, and zero-mean. In addition, we must assume $n \leq m$. Let us arrange the random variables into a vector $\mathbf{x}=(x_1, x_2, \dots, x_m)$ and the sources into $\mathbf{s}=(s_1, s_2, \dots, s_n)$; then the linear relationship is given by

$$\mathbf{x} = \mathbf{As} = \mathbf{a}_1 \cdot s_1 + \mathbf{a}_2 \cdot s_2 + \dots + \mathbf{a}_N \cdot s_N \tag{1}$$

where \mathbf{A} is a $M \times N$ un-mixing matrix and the columns $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N$ are called basis functions. The basis functions are consistent and the coefficients vary with data. The goal of ICA is to find a matrix \mathbf{W} defined in Eq.(2), so that the resulting vector \mathbf{y} as independent as possible. \mathbf{y} would actually correspond to the independent components \mathbf{s} and possibly permuted and rescaled.

$$\mathbf{y} = \mathbf{W}\mathbf{x} \quad (2)$$

Bell & Sejnowski have proposed a neural learning algorithm for ICA [11]. The approach is to maximize the joint entropy by stochastic gradient ascent. The updating formula for \mathbf{W} is:

$$\Delta \mathbf{W} = (\mathbf{I} + g(\mathbf{y})\mathbf{y}^T) \mathbf{W} \quad (3)$$

where $\mathbf{y} = \mathbf{W}\mathbf{x}$, and $g(y) = 1 - 2/(1 + e^{-y})$ is calculated for each component of \mathbf{y} . Before the learning procedure, \mathbf{x} is sphered by subtracting the mean \mathbf{m}_x and multiplying by a whitening filter:

$$\mathbf{x} = \mathbf{W}_0(\mathbf{x} - \mathbf{m}_x) \quad (4)$$

where $\mathbf{W}_0 = [(\mathbf{x} - \mathbf{m}_x)(\mathbf{x} - \mathbf{m}_x)^T]^{-1/2}$. Therefore the complete transform is calculated as the product of the whitening matrix and the learned matrix:

$$\mathbf{W}_I = \mathbf{W}\mathbf{W}_0 \quad (5)$$

As the spectral independent components, $\mathbf{y} = (y_1, y_2, \dots, y_n)$, are obtained, the unsupervised K-means algorithm [13] is used for clustering spectral features.

3 Segmentation of MR Brain images

The proposed method has been applied to segment MR brain images. The data is obtained from the Brain Web (<http://www.bic.mni.mcgill.ca/brainweb>). The typical 2D MR slice images (\mathbf{x}) are shown in Fig.2(a). We randomly selected 1000 pixels as training data to learn ICA transformation matrix \mathbf{W} . The learned ICA transformation \mathbf{W} is shown in Eq.(6).

$$\mathbf{W} = \begin{bmatrix} 0.84416 & -0.19045 & -0.16213 \\ -0.12476 & 0.70713 & -0.25485 \\ -0.12925 & -0.20911 & 0.6036 \end{bmatrix} \quad (6)$$

By using Eqs.(6) and (2), we extracted three independent components as shown in Fig.2(b). It can be seen that the contrast of white matter is enhanced in the first component (IC1), while the contrast of CSF is enhanced in the second component (IC2). In order to make a quantitative comparison, all images are normalized. We also show the profiles of Figs.2(a) and 2(b) in Fig.3.

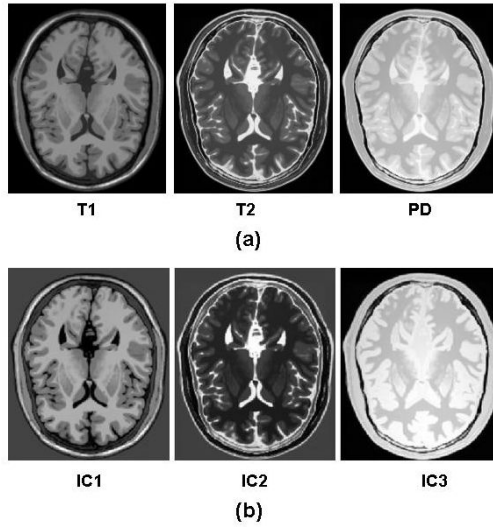


Fig. 2. Typical multi-spectral MR images

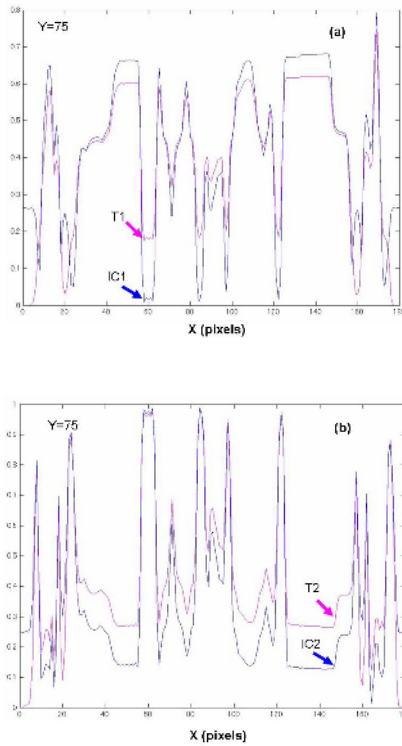


Fig. 3. Comparison of profiles. (a) T1 vs. IC1; (b) T2 vs. IC2.

The segmentation results by the proposed method is shown in Figs.4(a). In order to make a comparison, the ground truth, which is obtained from brain web database and is labeled by experts manually, is also shown in Fig.4(b). It can be seen that accuracy segmentations for CSF, white matter and gray matter have been done by the proposed method.

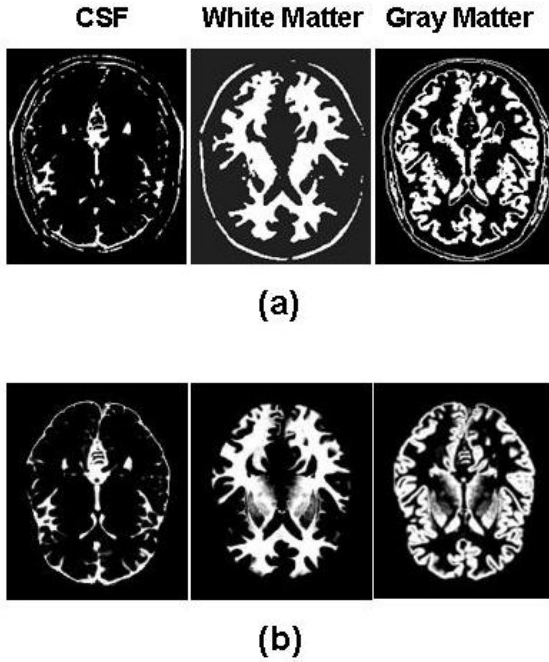


Fig. 4. Segmentation results of 2D MR brain images (a), and ground truths (b)

In order to make a quantitative evaluation, we compare the each segmented image with the corresponding ground truth. By taking the difference between the segmented image and the corresponding ground truth, we can easily count the number of pixels (N_p) that belong to the target class (white matter or gray matter or CSF) but are not segmented into the target class, and the number of pixels (N_n) that do not belong to the target class but are segmented into the target class. Segmentation performance is measured by two criteria of false positive rate δ_p and false negative rate δ_n , which are defined as

$$\delta_p = N_p / C$$

and

$$\delta_n = N_n / C$$

where C is the number of pixels belong to the target class in the ground truth.

The results of the proposed method (with ICA) and the conventional method (without ICA) are summarized in Table 1. It can be seen that both average false positive rate and average false negative rate by the proposed method are improved with the conventional method.

Table 1. Quantitative comparison of segmentation performance

Method	Averaged false positive rate (%)	Averaged false negative rate (%)
Proposed method (with ICA)	3.64	3.95
Conventional method (w/o ICA)	5.02	4.62

4 Conclusions

In this paper, we proposed an ICA-based method for segmentation of MR images. We first extract three independent components from the T1-weighted, T2-weighted and PD images by using ICA and then the extracted independent components are used for segmentation of MR images. Since ICA can enhance the local features, the MR images can be transformed to contrast-enhanced images (independent components) by ICA. The experimental results have shown that by using the proposed method, the segmentation performance can be significantly improved. Both averaged false positive rate and false negative rate are reduced to less than 4%.

This work was supported in part by the Strategic Information and Communications R&D Promotion Programme (**SCOPE**) under the Grand No. 051307017.

References

1. Bezdek, L.C., Hall, L.O., Clarke, L.P.: Review of MR image segmentation techniques using pattern recognition, *Medical Physics*, 20(1993)1033-1048.
2. Cohen, L.: On active contour models and ballons, *Computer vision, Graphics and Image Processing: Image Understanding*, 53(1991)211-218.
3. Hohne, K., Hanson, W.: Interactive 3d segmentation of MRI and CT volumes using morphological operations, *Journal of Computer Assisted Tomography*, 16(1992) 285-294.
4. Cline, H.E., Lorensen, W.E., Kikinis, R., Jolesz, F.A.: Three-dimensional segmentation of MR images of the head using probability and connectivity, *Journal of Computer Assisted Tomography*, 14(1990)1037-1045.
5. Warfield, S.K., Kaus, M., Jolesz, F.A., Kikinis, R.: Adaptive, template moderated, spatially varying statistical classification, 4(2000)43-55.
6. Gerig, G., Kubler, O., Kikinis, R., Jolesz, F.A.: Nonlinear anisotropic filtering of MRI data, *IEEE Transactions on Medical Imaging*, 11(1992)221-232.
7. Mohamed, N.A., Aly, A.F.: Volume segmentation of CT/MRI images using multiscale features, self-organizing principal component analysis, and self-organizing feature map, *Proc. of Int. Conf. on Artificial Neural Networks, Houston, (1997)*.

8. Clark, M., Hall, L., Goldgof, D., Clarke, L., Silbiger, L.: MRI segmentation using fuzzy clustering techniques: Integrating knowledge, *IEEE Eng. Med. & Biol. Mag.*, 13(1994)730-742.
9. Atkins, M.S., Mackiewicz, M.T.: Fully automatic segmentation of the brain in MRI, *IEEE Trans. Med. Imagin*, 17(1998)98-107.
10. Nakai, T., Muraki, S., Bagarinao, E., Miki, Y., Takehara, Y., Matsuo, K., Kato, C., Sakahara, H., Isoda, H.: Application of independent component analysis to magnetic resonance imaging for enhancing the contrast of gray and white matter. *NeuroImage*, 21(2004) 251-260.
11. Bell, A. J., Sejnowski, T. J.: An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(1995)1129-1159.
12. Chen, Y.-W.: Independent Component Analysis (1) –Cocktail Party Effect--. *Medical Imaging Technology*, 21(2003) 81-85. (in Japanese)
13. Duda, R., Hart, P.: *Pattern Classification and Scene Analysis*, John Wiley and Sons, Inc., (1973).

Equi-sized, Homogeneous Partitioning

F. Klawonn and F. Höppne

¹ Department of Computer Science
University of Applied Sciences Braunschweig /Wolfenbüttel
Salzdahlumer Str. 46/48
38302 Wolfenbüttel, Germany

² Department of Economics
University of Applied Sciences Braunschweig /Wolfenbüttel
Robert Koch Platz 10-14
38440 Wolfsburg, Germany
{f.klawonn, f.hoepfner}@fh-wolfenbuettel.de

Abstract. We consider the problem of partitioning a data set of n data objects into c homogeneous subsets (that is, data objects in the same subset should be similar to each other), such that each subset is of approximately the same size. This problem has applications wherever a population has to be distributed among a limited number of resources and the workload for each resource shall be balanced. We modify an existing clustering algorithm in this respect, present some empirical evaluation and discuss the results.

1 Introduction

Cluster analysis is a widely used technique that seeks for groups in data. The result of such an analysis is a set of groups or clusters where data in the same group are similar (homogeneous) and data from distinct groups are different (heterogeneous) [1]. In this paper, we consider a variation of the clustering problem, namely the problem of subdividing a set X of n objects into c homogeneous groups of *equal size*. In contrast to the clustering problem, we abandon the heterogeneity between groups and introduce the requirement of having equi-sized groups.

Applications for this kind of *uniform clustering* include for instance: (a) The distribution of n students into c groups of equal strength to obtain fair class sizes and with homogeneous abilities and skills to allow for teaching methods tailored to the specific needs of each group. (b) The distribution of n jobs to c machines or workers such that every machine has an identical workload and as similar jobs as possible to reduce the configuration time. (c) The placement of c sites such that goods from n locations can be transported to the c sites, while the total covered distance is minimized and queuing at the sites is avoided, that is, approximately the same number of goods should arrive at each site.

Due to the similarity of our problem with traditional clustering problems, we are going to modify an existing clustering algorithm, which will be reviewed in

section 2. This objective function-based clustering algorithm – a variant of k-means – transforms the discrete, combinatorial problem into a continuous one, such that numerical problem solving methods can be applied. We modify the objective function such that the equi-sized clusters are considered in section 3 and discuss the results in section 4.

2 The FCM Algorithm

The fuzzy c-means (FCM) clustering algorithm partitions a data set $X := \{x_1, \dots, x_n\} \subset \mathbf{R}^d$ into c clusters. A cluster is represented by a prototype $p_i \in \mathbf{R}^d$, $1 \leq i \leq c$. The data-prototype relation is not binary, but a membership degree $u_{ij} \in [0, 1]$ indicates the degree of belongingness of data object x_j to prototype p_i or cluster number i . All membership degrees form a membership matrix $U \in \mathbf{R}^{c \times n}$. We can interpret the membership degrees as “probabilistic memberships”, since we require

$$\forall 1 \leq j \leq n : \quad \sum_{i=1}^c u_{ij} = 1. \quad (1)$$

The clustering process is carried out by minimizing the objective function

$$J_m = \sum_{j=1}^n \sum_{i=1}^c u_{ij}^m d_{ij} \quad \text{with} \quad d_{ij} = \|x_j - p_i\|^2. \quad (2)$$

under constraint (1). If the Euclidean distance between datum x_j and prototype p_i is high, J_m is minimized by choosing a low membership degree near 0. If the distance is small, the membership degree approaches 1. J_m is effectively minimized by alternating optimisation, that is, we alternately minimize (2) with respect to the prototypes (assuming memberships to be constant) and then with respect to the membership degrees (assuming prototypes to be constant). In both minimization steps, we obtain closed form solutions, for the prototypes:

$$\forall 1 \leq i \leq c : \quad p_i = \frac{\sum_{j=1}^n u_{ij}^m x_j}{\sum_{j=1}^n u_{ij}^m} \quad (3)$$

and for the membership degrees:

$$u_{ij} = \begin{cases} \frac{1}{\sum_{i=1}^c \left(\frac{\|x_j - p_i\|^2}{\|x_j - p_i\|^2} \right)^{\frac{1}{m-1}}} & \text{in case } I_j = \emptyset \\ \frac{1}{|I_j|} & \text{in case } I_j \neq \emptyset, i \in I_j \\ 0 & \text{in case } I_j \neq \emptyset, i \notin I_j \end{cases} \quad (4)$$

where $I_j = \{k \in \mathbf{N}_{\leq c} \mid x_j = p_k\}$. The FCM algorithm is depicted in Fig. 1. For a more detailed discussion of FCM and examples we refer for instance to [2,3].

```

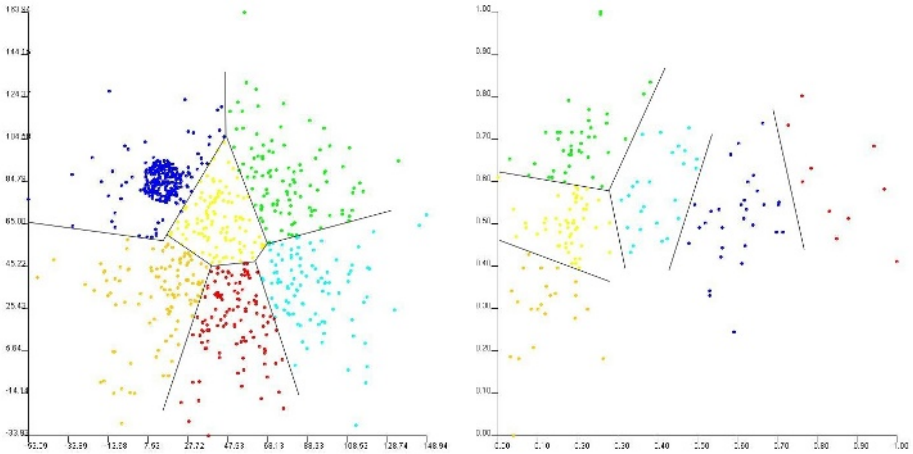
choose  $m > 1$  (typically  $m = 2$ )
choose termination threshold  $\varepsilon > 0$ 
initialize prototypes  $p_i$  (randomly)
repeat
  update memberships using (4)
  update prototypes using (3)
until change in memberships drops below  $\varepsilon$ 

```

Fig. 1. The FCM algorithm

3 Equi-sized Clusters

It is often said that the k-means (as well as the FCM) algorithm seeks for clusters of approximately the same size, but this is only true if the data density is uniform. As soon as the data density varies, a single prototype may very well cover a high-density cluster and thereby gains many more data objects than the other clusters. This leads to large differences in the size of the clusters. Examples for this phenomenon are shown in Fig. 2 for two data sets: On the left image, there is a very high density cluster in the top left corner, on the right image, the density decreases from left to right, so the rightmost cluster has only some data.

**Fig. 2.** Results of the FCM algorithm on two data sets

The idea of our modification is to include an additional constraint in the objective function (2) that forces the clusters to cover the same number of data objects. The size of cluster i (number of data objects) corresponds to the sum of the membership values $\sum_{j=1}^n u_{ij}$. In fact, since we have continuous membership degrees we may require

$$\sum_{j=1}^n u_{ij} = \frac{n}{c} \quad (5)$$

for all $i \in \{1, \dots, c\}$ even if n is not a multitude of c . This additional constraint (5) is – together with the constraint (1) – integrated into the objective function (2) via Lagrange multipliers. We then solve for the cluster prototypes and Lagrange multipliers by setting the partial derivatives to zero. This turns out to be a difficult problem for the general case of an arbitrary value of m , therefore we restrict ourselves to the case of $m = 2$, which is the most frequently used value of m in FCM. Given our Lagrange function

$$L = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^2 d_{ij} + \sum_{j=1}^n \alpha_j \left(1 - \sum_{i=1}^c u_{ij} \right) + \sum_{i=1}^c \beta_i \left(\frac{n}{c} - \sum_{j=1}^n u_{ij} \right) \quad (6)$$

we obtain as partial derivatives

$$\frac{\partial L}{\partial u_{ij}} = 2u_{ij}d_{ij} - \alpha_j - \beta_i = 0 \quad (7)$$

These equations, together with the constraints (1) and (5), lead to the following system of $(c \cdot n + c + n)$ linear equations for the variable u_{ij} , α_i and β_j ($i \in \{1, \dots, c\}$, $j \in \{1, \dots, n\}$). Empty entries indicate the value zero, RHS stands for the right hand side of the equation.

	$u_{1,1}$	\dots	$u_{1,n}$	\dots	$u_{c,1}$	\dots	$u_{c,n}$	α_1	\dots	α_n	β_1	\dots	β_c	RHS
$\frac{\partial L}{\partial u_{1,1}}$	$2d_{1,1}$							-1			-1			
\vdots		\ddots							\ddots		\vdots			
$\frac{\partial L}{\partial u_{1,n}}$			$2d_{1,n}$							-1	-1			
\vdots				\ddots					\ddots			\ddots		
$\frac{\partial L}{\partial u_{c,1}}$					$2d_{c,1}$			-1					-1	
\vdots						\ddots			\ddots				\vdots	
$\frac{\partial L}{\partial u_{c,n}}$							$2d_{c,n}$			-1			-1	
$\sum u_{i,1}$	1			\dots	1									1
\vdots		\ddots				\ddots								\vdots
$\sum u_{i,n}$			1				1							1
$\sum u_{1,j}$	1	\dots	1											n/c
\vdots				\ddots										\vdots
$\sum u_{c,j}$					1	\dots	1							n/c

In principle, this system of linear equations could be solved by a suitable numerical algorithm. Even for small data sets with 200 data objects and 5 clusters, this would mean that we have to solve a system of 1205 equations in each iteration step of the clustering algorithm, which is not acceptable in terms of computational costs. However, it is possible to solve this system of equations in a more efficient way. When multiplying the equations for u_{k1}, \dots, u_{kn} by $\frac{1}{2d_{k1}}, \dots, \frac{1}{2d_{kn}}$,

respectively, and then subtracting the resulting equations from the equation for $\sum_j u_{kj}$, we obtain

$$\sum_{j=1}^n \frac{\alpha_j}{2d_{kj}} + \beta_k \sum_{j=1}^n \frac{1}{2d_{kj}} = \frac{n}{c}. \quad (8)$$

From equation (7), we obtain

$$u_{ij} = \frac{\alpha_j + \beta_i}{2d_{ij}}. \quad (9)$$

Taking constraint (1) into account, yields

$$1 = \sum_{i=1}^c u_{ij} = \frac{\alpha_j}{2} \sum_{i=1}^c \frac{1}{d_{ij}} + \frac{1}{2} \sum_{i=1}^c \frac{\beta_i}{d_{ij}},$$

leading to

$$\alpha_j = \frac{2 - \sum_{i=1}^c \frac{\beta_i}{d_{ij}}}{\sum_{i=1}^c \frac{1}{d_{ij}}}. \quad (10)$$

Inserting (10) into (8), we obtain:
$$\sum_{j=1}^n \frac{2 - \sum_{i=1}^c \frac{\beta_i}{d_{ij}}}{2 \sum_{i=1}^c \frac{d_{kj}}{d_{ij}}} + \beta_k \sum_{j=1}^n \frac{1}{2d_{kj}} = \frac{n}{c}$$

and thus

$$- \sum_{j=1}^n \frac{\sum_{i=1}^c \frac{\beta_i}{d_{ij}}}{2 \sum_{i=1}^c \frac{d_{kj}}{d_{ij}}} + \beta_k \sum_{j=1}^n \frac{1}{2d_{kj}} = \frac{n}{c} - \sum_{j=1}^n \frac{1}{\sum_{i=1}^c \frac{d_{kj}}{d_{ij}}}. \quad (11)$$

This induces a system of c linear equations for the β_k with coefficients

$$a_{k\ell} = \begin{cases} - \sum_{j=1}^n \frac{\sum_{i=1}^c \frac{1}{d_{\ell j}}}{2 \sum_{i=1}^c \frac{d_{kj}}{d_{ij}}} & \text{if } k \neq \ell \\ - \sum_{j=1}^n \frac{\sum_{i=1}^c \frac{1}{d_{\ell j}}}{2 \sum_{i=1}^c \frac{d_{kj}}{d_{ij}}} + \sum_{j=1}^n \frac{1}{2d_{kj}} & \text{if } k = \ell. \end{cases} \quad (12)$$

This system of linear equations can be solved by a suitable numerical algorithm. The computation time is acceptable, since the number of equations is equal to the number of clusters and therefore independent of the number of data. Once the β_i have been determined, we can compute the α_j using equation (10) and finally obtain the membership degrees based on equation (9). After all, we arrive at the clustering algorithm depicted in Fig. 3. Note that the boundedness of the membership degrees $u_{ij} \in [0, 1]$ represents an additional constraint on the objective function of FCM as well as the objective function of our new algorithm. In the original FCM, however, it was not necessary to consider it explicitly, because one can easily see from the resulting membership degrees (4) that this condition is satisfied. It is not possible to conclude this boundedness for the new membership degrees (9). It is clear, however, that the influence of negative memberships will be rather small: Since the objective function (2) and (6) uses only positive weights u_{ij}^2 , large negative values cannot help in the minimization. We will comment on this in the following section.

```

choose termination threshold  $\varepsilon > 0$ 
initialise prototypes  $p_i$ 
repeat
  solve linear equation system (11) for  $\beta$ 
  using  $\beta$ , calculate  $\alpha$  using (10), update memberships using (9)
  update prototypes using (3)
until change in memberships drops below  $\varepsilon$ 

```

Fig. 3. The proposed algorithm

4 Examples and Discussion

To illustrate the impact of our modified objective function, we show the results of the new algorithm for the data sets shown in Fig. 2, where the standard FCM algorithm yielded a result with high variation in the cluster size. The results are shown in the left images of Figs. 4 and 6. By comparison to Fig. 2 we see, that the high-density cluster has been split into two clusters (Fig. 4) and that the data on the left of Fig. 6 is now distributed among four rather than three clusters, such that the rightmost cluster gains more data. As expected, the sum of membership degrees for each individual cluster equals $\frac{n}{c}$.

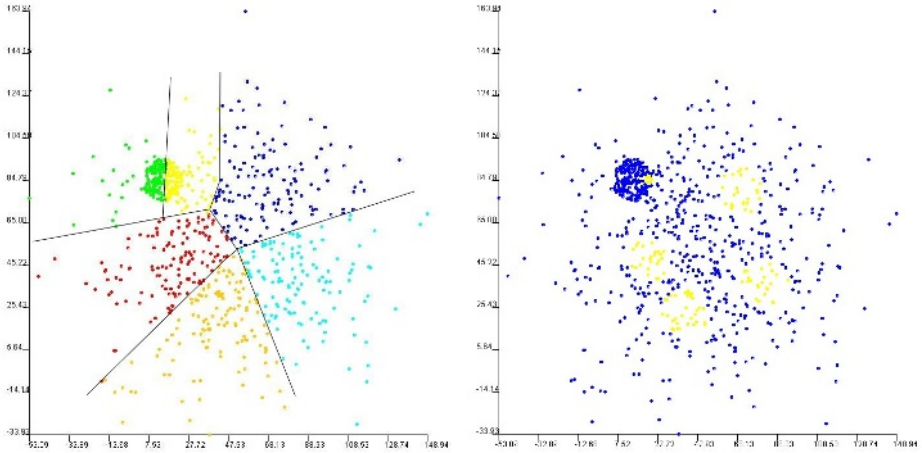


Fig. 4. Results of the new algorithm on the data set shown in Fig. 2. Left: Resulting partition. Right: Points with negative membership degrees (marked in lighter shading).

Regarding the boundedness of the membership degrees u_{ij} it turned out that they actually take negative values. This is, of course, an undesired effect, because then the interpretation of $\sum_{j=1}^n u_{ij}$ as the size or number of data objects is not quite correct. As conjectured in the previous section, it turned out on closer examination that the total sum of negative weights is rather small. In both data sets, the sum of all negative membership degrees was below 0.5% of the

total data set size n . We want to illustrate the kind of situation in which negative membership degrees occur with the help of the data set shown in Fig. 5. Consider the data set is partitioned into three clusters. Since the leftmost cluster has an additional data object x in the middle, it is not obvious how to distribute the data among all clusters in equal shares.

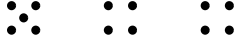


Fig. 5. A 'difficult' example data set

Regarding the minimization of the sum of weighted distances, it would be optimal to assign high membership degrees to all five data objects. This would, however, violate the constraint that all clusters must share the same size. To get membership degrees as high as possible, the membership of all other data objects (middle and right cluster) to this cluster are chosen slightly negative. Since the sum of all membership values is constrained to be one, negative values for the middle and right data allow us to have slightly higher degrees for the leftmost data. On the other hand, having a negative membership degrees for some data object x on the right forces us to increase other membership degrees of x (to guarantee a sum of 1). This is possible almost at no cost, if x is close to the centre of another cluster, because then we have a small distance value and increasing the membership degree to this cluster does no harm in the minimization of (2). (For a detailed discussion of the influence of the membership weight u_{ij}^m see [4].)

To summarise: In a situation where an equi-sized partition is difficult to obtain while minimizing at the same time the sum of weighted distances (2), the cluster with too many data objects 'borrows' some membership from data near the centres of the other clusters. Figs. 4 and 6 show this effect for the two example data sets. The data for which negative membership values occur are shown in a lighter shading. These data objects are all close to the respective cluster prototype. And there is always one cluster without any negative membership degrees, which corresponds to the rightmost cluster in our data set in Fig. 5.

In all our experiments, the side effects of this trade off between minimizing (2) and satisfying (5) were quite small, so we do not consider this as a major drawback of our approach. We can even make use of this information: By analysing which cluster has no negative membership degrees at all, we can find out which cluster tends to be 'too big'. When breaking ties in the final assignment of data to clusters, it should be this cluster that gets more data objects than the other. It should be noted that the assignment of a data object to the cluster with the highest membership degree does not guarantee that each cluster contains exactly the number of data. This is anyway impossible, except when n/c is an integer number. The small deviations from n/c resulting from our algorithm can be easily balanced by taking the more ambiguous membership degrees into account in order to re-assign a few data to other clusters.

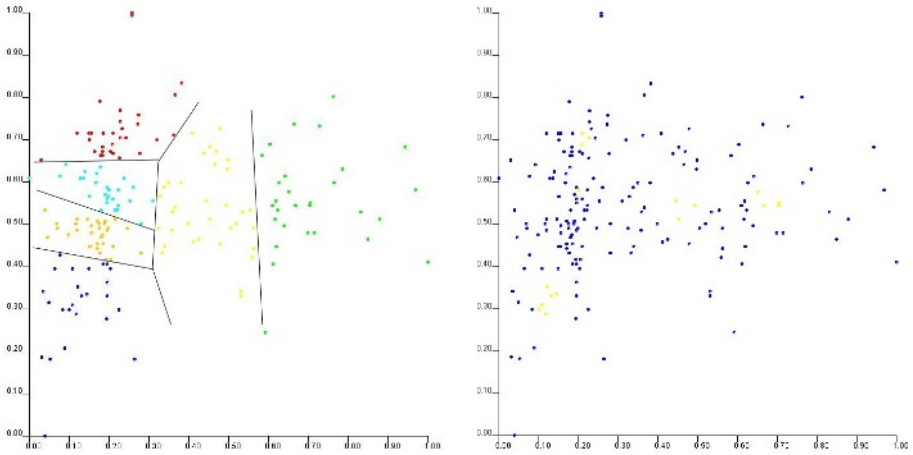


Fig. 6. Results on the wine data set shown as projections. Left: Derived partition. Right: Points with negative membership degrees (marked in lighter shading).

5 Conclusions

In this paper, we have considered the problem of subdividing a data set into homogeneous groups of equal size. Finding homogeneous groups is a typical task for clustering algorithms, however, if the data density is not uniform, such algorithms usually tend to deliver clusters of unequal size, which is inappropriate for some applications. We have proposed an algorithm that outperforms a popular variant of k-means in that respect. Although we have only discussed the case of equi-sized clusters, in principle it is also possible to subdivide the data set into groups of any predefined size, which makes our approach quite useful for a range of applications where capacity restrictions apply.

Another, slightly weaker approach to the problem of uniform clustering would be to replace the strict constraints (5) by adding a (weighted) penalty term of the form $\sum_{i=1}^c (\frac{n}{c} - \sum_{j=1}^n u_{ij})^2$ to the objective function (2). Due to the limited space here, we leave this discussion open for a subsequent paper.

References

1. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall (1988)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
3. Höppner, F., Klawonn, F., Kruse, R., Runkler, T.A.: Fuzzy Cluster Analysis. John Wiley & Sons, Chichester, England (1999)
4. Klawonn, F., Höppner, F.: What is fuzzy about fuzzy clustering? – Understanding and improving the concept of the fuzzifier. In: Advances in Intelligent Data Analysis, Springer (2003) 254–264

Nonparametric Fisher Kernel Using Fuzzy Clustering

Ryo Inokuchi¹ and Sadaaki Miyamoto²

¹ Doctoral Program in Risk Engineering, University of Tsukuba,
Ibaraki 305-8373, Japan

`inokuchi@soft.risk.tsukuba.ac.jp`

² Department of Risk Engineering, University of Tsukuba,
Ibaraki 305-8373, Japan

`miyamoto@risk.tsukuba.ac.jp`

Abstract. The Fisher kernel, which refers to the inner product in the feature space of the Fisher score, has been known to be a successful tool for feature extraction using a probabilistic model. If an appropriate probabilistic model for given data is known, the Fisher kernel provides a discriminative classifier such as support vector machines with good generalization. However, if the distribution is unknown, it is difficult to obtain an appropriate Fisher kernel. In this paper, we propose a new nonparametric Fisher-like kernel derived from fuzzy clustering instead of a probabilistic model, noting that fuzzy clustering methods such as a family of fuzzy c -means are highly related to probabilistic models, e.g., entropy-based fuzzy c -means and a Gaussian mixture distribution model. The proposed kernel is derived from observing the last relationship. Numerical examples show the effectiveness of the proposed method.

1 Introduction

The Fisher kernel [4] has been increasingly applied to discriminative classifiers in order to extract features from probabilistic models. It has been observed that the Fisher kernel classifiers have much successful results if appropriate distributions for data are already known, e.g., biological sequences [4].

The Fisher kernel refers to the inner product in the feature space of the Fisher score, and it allows to convert discrete data into continuous vectors. In the case when the true class distribution has a mixture distribution, the class information is fully preserved [9].

However, assume that the distribution is unknown, parameters cannot be estimated (or make no sense), and hence an appropriate Fisher score cannot be obtained. We then have to obtain a quantity simulating the Fisher score with no assumptions of the distribution.

Fuzzy clustering is one of answers to obtain the Fisher score from a nonparametric model. Recently it is elucidated that fuzzy clustering [6] is highly related to a probabilistic model. For example, an entropy-based fuzzy c -means (FCM) [3] is equivalent to a Gaussian mixture distribution model by setting parameters adequately. It has also been shown that clustering is able to construct mutual information (MI) kernel [7] which generalizes the Fisher kernel [2].

Observing these facts, we propose a new nonparametric Fisher kernel derived from fuzzy clustering instead of a probabilistic model. An illustrative example which has been discussed in [9] show the effectiveness of the proposed method.

2 Preliminaries

We first review kernel functions, the Fisher kernel, entropy-based fuzzy c -means, for discussing a nonparametric Fisher kernel.

2.1 Kernel Functions

Studies in support vector machines [10] often employ a high-dimensional feature space S for having nonlinear classification boundaries. For this purpose a mapping

$$\Phi : \mathbf{R}^p \rightarrow S$$

is used whereby an object x is mapped into S :

$$\Phi(x) = (\phi_1(x), \phi_2(x), \dots).$$

Although x is a p -dimensional vector, $\Phi(x)$ may have a higher or the infinite dimension.

In the nonlinear classification method an explicit form of $\Phi(x)$ is unavailable, but the inner product is denoted by

$$K_{ij} = K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle. \quad (1)$$

The function K , called a kernel function, is assumed to be known. If x is a continuous value,

$$K(x, y) = \exp(-c \text{nst} \|x - y\|^2), \quad (2)$$

$$K(x, y) = \langle x, y \rangle^d \quad (3)$$

are frequently used. The first is called the Gaussian kernel, while the second is called the polynomial kernel.

Otherwise, in the case when x is a discrete value, many kernels which handle discrete data have been proposed. Among them, we focus on the Fisher kernel.

2.2 The Fisher Kernel

Let \mathcal{X} denote the domain of data, which is discrete or continuous. Let us also assume that a probabilistic model $p(x|\theta)$, $x_1, \dots, x_n \in \mathcal{X}$, $\theta \in \mathbf{R}^p$ is available. Given a parameter estimate $\hat{\theta}$ from given data, the feature vector which is called the Fisher score is obtained as

$$\mathbf{f}_{\boldsymbol{\theta}}(x) = \left(\frac{\partial \log p(x|\hat{\boldsymbol{\theta}})}{\partial \theta_1}, \dots, \frac{\partial \log p(x|\hat{\boldsymbol{\theta}})}{\partial \theta_p} \right)^T. \quad (4)$$

The Fisher kernel refers to the inner product in this space.

$$K_{ij}^f = K^f(x_i, x_j) = \mathbf{f}_{\boldsymbol{\theta}}(x_i)^T I^{-1} \mathbf{f}_{\boldsymbol{\theta}}(x_j), \quad (5)$$

where I is the Fisher information matrix

$$I = \frac{1}{n} \sum_{k=1}^n \mathbf{f}_{\boldsymbol{\theta}}(x_k) \mathbf{f}_{\boldsymbol{\theta}}(x_k)^T. \quad (6)$$

In the following, we will discuss how to determine $p(x|\boldsymbol{\theta})$. Let us assume that true distribution $p(x|\boldsymbol{\theta})$ is determined as a mixture model of true class distributions:

$$p(x|\boldsymbol{\alpha}) = \sum_{i=1}^c \alpha_i p(x|y=i), \quad (7)$$

where $\sum_{i=1}^c \alpha_i = 1$. The Fisher score for $p(x|\boldsymbol{\alpha})$ is

$$\frac{\partial \log p(x|\boldsymbol{\alpha})}{\partial \alpha_i} = \frac{p(x|y=i)}{\sum_{i=1}^c \alpha_i p(x|y=i)} = \frac{p(y=i|x)}{\alpha_i}. \quad (8)$$

In the case when we know the true distribution, the class information is fully preserved, that is, the loss function L is

$$L(U) = \sum_{i=1}^c E_x (U_i(x) - p(y=i|x))^2 = 0, \quad (9)$$

where the function U is the best estimator of the posterior distribution. In general, however, the true distribution is unknown, in which case we should consider the method of fuzzy clustering in the next section.

2.3 Entropy-Based Fuzzy c -Means

Objects to be clustered are denoted by $x_k = (x_k^1, \dots, x_k^p) \in \mathbf{R}^p$, $k = 1, \dots, n$, a vector in the p -dimensional Euclidean space. Cluster centers are $v_i = (v_i^1, \dots, v_i^p)^T$, $i = 1, \dots, c$, where c is the number of clusters. An abbreviated symbol $V = (v_1, \dots, v_c)$ is used for the whole collection of cluster centers. The matrix $U = (u_{ik})$, ($i = 1, \dots, c$, $k = 1, \dots, n$) is used as usual, where u_{ik} means the degree of belongingness of object x_k to cluster i .

In the KL information-based fuzzy c -means, two more variables are employed here. One is a cluster volume size variable $\alpha = (\alpha_1, \dots, \alpha_c)$; the other is within-cluster covariance matrix $S_i = (s_i^{j\ell})$ ($1 \leq j, \ell \leq p$, $i = 1, \dots, c$). We write $S = (S_1, \dots, S_c)$ for simplicity. As is well-known, fuzzy c -means clustering [1] is based on the optimization of an objective function. The objective function [3] has four variables (U, V, α, S):

$$J(U, V, \alpha, S) = \sum_{k=1}^n \sum_{i=1}^c u_{ik} D_{ik} + \lambda \sum_{k=1}^n \sum_{i=1}^c u_{ik} \log \frac{u_{ik}}{\alpha_i} + \sum_{k=1}^n \sum_{i=1}^c \log |S_i| \quad (10)$$

in which

$$D_{ik} = \|x_k - v_i\|_{S_i}^2 = (x_k - v_i)^T S_i^{-1} (x_k - v_i).$$

The symbol λ is a nonnegative constant that should be determined beforehand.

Two constraints are assumed for U and α :

$$M = \{ U = (u_{ik}) : \sum_{i=1}^c u_{ik} = 1, u_{jk} \geq 0, \forall j, k \},$$

$$A = \{ \alpha = (\alpha_1, \dots, \alpha_c) : \sum_{i=1}^c \alpha_i = 1, \alpha_j \geq 0, \forall j \}.$$

An alternate minimization algorithm is the basic idea in fuzzy c -means [1]. As we have four variables, the algorithm has the corresponding four steps.

Algorithm FCM (KL-information based fuzzy c -means)

FCM0. Set an initial value $\bar{V}, \bar{\alpha}, \bar{S}$ for V, α, S .

FCM1. Find optimal solution of J with respect to U while other variables are fixed: put

$$\bar{U} = \arg \min_{U \in M} J(U, \bar{V}, \bar{\alpha}, \bar{S}).$$

FCM2. Find optimal solution of J with respect to V while other variables are fixed: put

$$\bar{V} = \arg \min_V J(\bar{U}, V, \bar{\alpha}, \bar{S}).$$

FCM3. If $\mu = 0$, skip this step; else find optimal solution of J with respect to α while other variables are fixed: put

$$\bar{\alpha} = \arg \min_{\alpha \in A} J(\bar{U}, \bar{V}, \alpha, \bar{S}).$$

FCM4. If $\nu = 0$, skip this step; else find optimal solution of J with respect to S while other variables are fixed: put

$$\bar{S} = \arg \min_S J(\bar{U}, \bar{V}, \bar{\alpha}, S).$$

FCM5. If the solution $(\bar{U}, \bar{V}, \bar{\alpha}, \bar{S})$ is convergent, stop; else go to **FCM1**.

End of FCM.

We show solutions of each step where we write, for simplicity, u_{ik} instead of \bar{u}_{ik} , v_i instead of \bar{v}_i , *etc.* without confusion.

The solution for U is

$$u_{ik} = \frac{\alpha_i \exp(-\frac{1}{\lambda} D_{ik}) |S_i|^{\frac{1}{\lambda}}}{\sum_{j=1}^c \alpha_j \exp(-\frac{1}{\lambda} D_{jk}) |S_j|^{\frac{1}{\lambda}}}. \quad (11)$$

The solution for V is

$$v_i = \frac{\sum_{k=1}^n u_{ik} x_k}{\sum_{k=1}^n u_{ik}}. \quad (12)$$

The solution for α is

$$\alpha_i = \frac{1}{n} \sum_{k=1}^n u_{ik}. \quad (13)$$

The solution for S is

$$S_i = \frac{\sum_{k=1}^n u_{ik} (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^n u_{ik}}. \quad (14)$$

After the convergence, we obtain fuzzy classification functions [6] which express cluster membership distributions like posterior distributions.

$$u_i(x) = \frac{\alpha_i \exp(-\frac{1}{\lambda} D_i(x)) |S_i|^{\frac{1}{\lambda}}}{\sum_{j=1}^c \alpha_j \exp(-\frac{1}{\lambda} D_j(x)) |S_j|^{\frac{1}{\lambda}}}. \quad (15)$$

In the case when S_i is fixed with the unit matrix I , KL information-based fuzzy c -means is equivalent to the entropy-regularized fuzzy c -means with a size control variable [5].

3 Nonparametric Fisher Kernel

We propose a nonparametric Fisher kernel derived from fuzzy clustering. It is derived from the relation between Gaussian mixture distribution models and KL information-based fuzzy c -means.

In the Gaussian distribution model, the density function is

$$p(x|\phi_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |A|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - v_i)^T A_i^{-1} (x - v_i)\right). \quad (16)$$

The EM algorithm estimates parameters minimizing log likelihood:

$$Q(\phi|\phi^*) = \sum_{i=1}^c \sum_{k=1}^n \log(\alpha_i p_i(x_k|\phi_i)) p(y = i|x_k). \quad (17)$$

On the E step,

$$p(y = i|x_k) = \frac{\alpha_i \exp(-\frac{1}{2}(x - v_i)^T A_i^{-1} (x - v_i)) |A_i|^{\frac{1}{2}}}{\sum_{j=1}^c \alpha_j \exp(-\frac{1}{2}(x - v_j)^T A_j^{-1} (x - v_j)) |A_j|^{\frac{1}{2}}}, \quad (18)$$

on the M step,

$$\alpha_i = \frac{1}{n} \sum_{k=1}^n p(y = i|x_k). \quad (19)$$

By contrast to KL information-based fuzzy c -means, eqn. (18) corresponds to (15), and (19) corresponds to (13). Hence, the nonparametric Fisher score is

$$\frac{\partial \log p(x_k|\alpha)}{\partial \alpha_i} = \frac{p(y = i|x_k)}{\alpha_i} \approx \frac{u_{ik}}{\alpha_i}. \quad (20)$$

where $\alpha_i = \frac{1}{n} \sum_{k=1}^n u_{ik}$. Once the Fisher score is obtained using a clustering algorithm to a large amount of unlabeled data, the Fisher kernel is derived from (5). This kernel can be applied to a kernel classifier such as that for support vector machines. In this paper, we will apply the Fisher kernel to kernel-based entropy-based fuzzy c -means in the next section.

4 Entropy-Based Fuzzy c -Means with the Fisher Kernel

The kernel-based entropy fuzzy c -means [8] has often been used to obtain non-linear cluster boundaries.

Now we consider entropy fuzzy c -means in the feature space. Cluster centers in the feature space is represented as

$$v_i = \frac{1}{U_i} \sum_{k=1}^n u_{ik} \Phi(x_k). \quad (21)$$

where $U_i = \sum_{k=1}^n u_{ik}$. Instead of the distance $\|x_k - v_i\|^2$ in the data space, the next distance in the feature space is considered.

$$D_{ik} = \|\Phi(x_k) - v_i\|_S^2. \quad (22)$$

It should be noted that v_i is the cluster center in the high-dimensional feature space. Substituting (21) into (22), the updating formula is obtained.

$$\begin{aligned} D_{ik} &= \|\Phi(x_k) - v_i\|^2 \\ &= \langle \Phi(x_k), \Phi(x_k) \rangle - 2\langle \Phi(x_k), v_i \rangle + \langle v_i, v_i \rangle. \end{aligned} \quad (23)$$

Now we apply the Fisher kernel (5) to (23), we obtain

$$D_{ik} = K_{kk}^f - \frac{2}{U_i} \sum_{k=1}^n u_{ij} K_{jk}^f - \frac{1}{U_i^2} \sum_{j=1}^n \sum_{l=1}^n K_{jl}^f. \quad (24)$$

Thus (15) and (24) are repeated until convergence, as in the next algorithm.

Algorithm K-FCM (Kernel-based entropy fuzzy c -means)

K-FCM0. Set an initial value D_{ik}, α_i .

K-FCM1. Update u_{ik} using (15).

K-FCM2. Update D_{ik} using (24).

K-FCM3. If $\mu = 0$, skip this step; else update α_i using (13).

K-FCM4. If the solution is convergent, stop; else go to **K-FCM1**.

End of K-FCM.

5 Illustrative Examples

Artificially generated 150 points on a plane were analyzed. This clustering problem is the almost same as that in [9]. First, seven clusters are obtained by the algorithm entropy-regularized fuzzy c -means as shown in Fig. 1, and then the Fisher score is derived from the membership values. We applied two algorithms of the hard c -means and the entropy-based fuzzy c -means to data mapped into the Fisher feature space. Fig. 2 shows the result from hard c -means algorithm, which is the same as that in [9]. On the other hand, Fig. 3 shows a successful result from the entropy-based fuzzy c -means. While Tsuda et. al [9] develops a new and complicated algorithm due to the failure of the hard c -means, we are successfully using the simple algorithm of fuzzy c -means. Fuzzy c -means algorithm has robustness for nuisance dimensions in the feature space of the Fisher score.

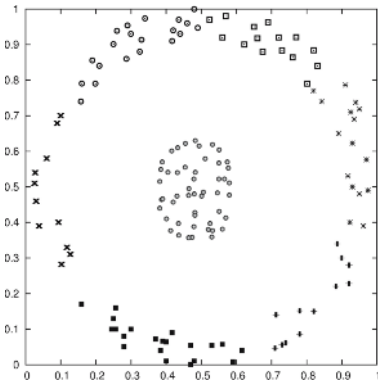


Fig. 1. Seven clusters obtained from entropy FCM

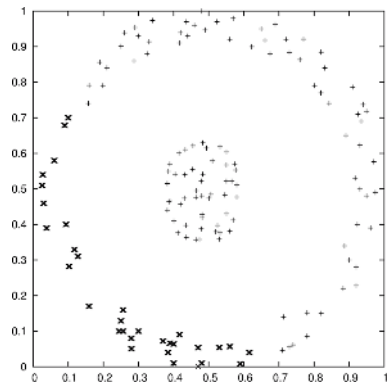


Fig. 2. Two clusters obtained from HCM in the feature space

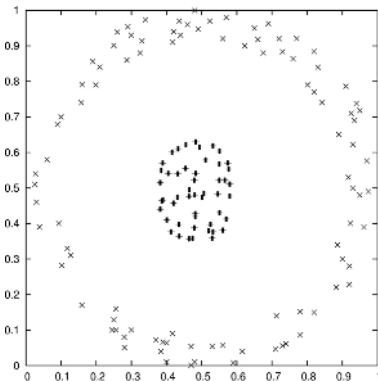


Fig. 3. Two clusters obtained from eFCM in the feature space

6 Conclusion

We have proposed a nonparametric Fisher kernel derived from fuzzy clustering instead of a probabilistic model. In the proposed kernel, the Fisher score using fuzzy clustering works as well as the Fisher score. The proposed kernel does not assume a probabilistic distribution. The numerical examples have shown nonlinearly separated clusters from kernelized fuzzy c -means using the nonparametric Fisher kernel.

Future studies include semi-supervised classification problems using a proposed method, and theoretical studies of relations between the proposed method and mutual information (MI) kernels including the Fisher kernel.

Acknowledgment. This study has partly been supported by the Grant-in-Aid for Scientific Research, Japan Society for the Promotion of Science, No.16300065.

References

1. J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.
2. O. Chapelle, J. Weston, B. Schölkopf, Cluster kernels for semi-supervised learning, *Advances in Neural Information Processing Systems*, Vol. 15, pp.585–592, 2003.
3. H. Ichihashi, K. Honda, N. Tani, Gaussian mixture PDF approximation and fuzzy c -means clustering with entropy regularization, *Proc. of the 4th Asian Fuzzy System Symposium*, May 31-June 3, 2000, Tsukuba, Japan, pp.217–221.
4. T. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, *Proc. of Neural Information Processing Systems. NIPS*, 1998.
5. S. Miyamoto, M. Mukaidono, Fuzzy c -means as a regularization and maximum entropy approach, *Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA'97)*, June 25-30, 1997, Prague, Czech, Vol.II, 1997, pp.86–92.
6. S. Miyamoto, *Introduction to Cluster Analysis: Theory and Applications of Fuzzy Clustering*, Morikita-Shuppan, Tokyo, 1999 (in Japanese).
7. M. Seeger, Covariance kernels from bayesian generative models, *Advances in Neural Information Processing Systems*, Vol.14, pp. 905–912, 2001.
8. S. Miyamoto and D. Suizu, Fuzzy c -means clustering using kernel functions in support vector machines, *J. of Advanced Computational Intelligence and Intelligent Informatics*, Vol.7, No.1, pp. 25–30, 2003.
9. K. Tsuda, M. Kawanabe, K.R. Muller, Clustering with the Fisher score, *Advances in Neural Information Processing Systems*, Vol.15, pp. 729–736, 2003.
10. V.N.Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

Finding Simple Fuzzy Classification Systems with High Interpretability Through Multiobjective Rule Selection

Hisao Ishibuchi, Yusuke Nojima, and Isao Kuwajima

Department of Computer Science and Intelligent Systems, Graduate School of Engineering
Osaka Prefecture University

1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan

hisaoi@cs.osakafu-u.ac.jp, nojima@cs.osakafu-u.ac.jp

kuwajima@ci.cs.osakafu-u.ac.jp

<http://www.ie.osakafu-u.ac.jp/~hisaoi/>

Abstract. In this paper, we demonstrate that simple fuzzy rule-based classification systems with high interpretability are obtained through multiobjective genetic rule selection. In our approach, first a prespecified number of candidate fuzzy rules are extracted from numerical data in a heuristic manner using rule evaluation criteria. Then multiobjective genetic rule selection is applied to the extracted candidate fuzzy rules to find a number of non-dominated rule sets with respect to the classification accuracy and the complexity. The obtained non-dominated rule sets form an accuracy-complexity tradeoff surface. The performance of each non-dominated rule set is evaluated in terms of its classification accuracy and its complexity. Computational experiments show that our approach finds simple fuzzy rules with high interpretability for some benchmark data sets in the UC Irvine machine learning repository.

1 Introduction

Evolutionary multiobjective optimization (EMO) is one of the most active research areas in the field of evolutionary computation [2], [3], [7]. The main advantage of EMO algorithms over classical approaches is that a number of non-dominated solutions can be obtained by a single run of EMO algorithms. EMO algorithms have been successfully applied to various application areas [2], [3], [7]. In the application to fuzzy logic, EMO algorithms have been used to find accurate, transparent and compact fuzzy rule-based systems [6], [16]-[18]. That is, EMO algorithms have been used to maximize the accuracy of fuzzy rule-based systems and minimize their complexity.

In this paper, we clearly demonstrate that simple fuzzy rule-based classification systems with high interpretability can be obtained by multiobjective fuzzy rule selection. We also demonstrate that a number of non-dominated fuzzy rule-based classification systems along the accuracy-complexity tradeoff surface can be obtained by a single run of an EMO-based fuzzy rule selection algorithm. Fuzzy rule selection for classification problems was first formulated as a single-objective combinatorial optimization problem [13], [14]. A standard genetic algorithm was used to optimize a weighted sum fitness function, which was defined by the number of correctly

classified training patterns and the number of fuzzy rules. A two-objective combinatorial optimization problem was formulated in [10] as an extension of the single-objective formulation. An EMO algorithm was used to find a number of non-dominated fuzzy rule-based classification systems with respect to the two objectives: maximization of the number of correctly classified training patterns and minimization of the number of selected fuzzy rules. The two-objective formulation was further extended in [11], [15] to a three-objective combinatorial optimization problem by introducing an additional objective: minimization of the total number of antecedent conditions (i.e., minimization of the total rule length). A number of non-dominated fuzzy rule-based systems with respect to the three objectives were found by an EMO algorithm.

This paper is organized as follows. First we explain an outline of multiobjective fuzzy rule selection in Section 2. Next we explain heuristic rule extraction for extracting candidate rules in Section 3. Then we show experimental results on some benchmark data sets in the UC Irvine machine learning repository in Section 4. Finally we conclude this paper in Section 5.

2 Multiobjective Fuzzy Rule Selection

Let us assume that we have N fuzzy rules as candidate rules for multiobjective fuzzy rule selection. We denote a subset of those candidate rules by S . The accuracy of the rule set S is measured by the error rate on the given training patterns. We use a single winner rule-based method [12] to classify each training pattern by S . The single winner rule for a training pattern has the maximum product of the rule weight and the compatibility grade with that pattern. We include the rejection rate into the error rate (i.e., training patterns with no compatible fuzzy rules in S are counted among errors).

On the other hand, we measure the complexity of the rule set S by the number of fuzzy rules in S and the total number of antecedent conditions in S . Thus our multiobjective fuzzy rule selection problem is formulated as follows:

$$\text{Minimize } f_1(S), f_2(S), f_3(S), \quad (1)$$

where $f_1(S)$ is the error rate on training patterns, $f_2(S)$ is the number of fuzzy rules, and $f_3(S)$ is the total number of antecedent conditions. It should be noted that each rule has a different number of antecedent conditions. This is because we use *don't care* as a special antecedent fuzzy set, which is not counted as antecedent conditions. That is, the third objective is the number of antecedent conditions excluding *don't care* conditions. The third objective can be also viewed as the total rule length since the number of antecedent conditions is often referred to as the rule length.

Any subset S of the N candidate fuzzy rules can be represented by a binary string of length N as $S = s_1s_2 \dots s_N$ where $s_j = 1$ and $s_j = 0$ mean that the j -th rule is included in S and excluded from S , respectively. Such a binary string is handled as an individual in multiobjective fuzzy rule selection. Since individuals are represented by binary strings, we can apply almost all EMO algorithms with standard genetic operations to our multiobjective fuzzy rule selection problem in (1). In this paper, we use the NSGA-II algorithm [8] due to its popularity, high performance and simplicity.

In the application of NSGA-II to multiobjective fuzzy rule selection, we use two heuristic tricks to efficiently find small rule sets with high accuracy. One trick is biased mutation where a larger probability is assigned to the mutation from 1 to 0 than that from 0 to 1. The other trick is the removal of unnecessary rules, which is a kind of local search. Since we use the single winner rule-based method for the classification of each pattern by the rule set S , some rules in S may be chosen as winner rules for no training patterns. By removing those rules from S , we can improve the second and third objectives without degrading the first objective. The removal of unnecessary rules is performed after the first objective is calculated and before the second and third objectives are calculated. NSGA-II with these two tricks is used to find non-dominated rule sets of the multiobjective fuzzy rule selection problem in (1).

Here we briefly explain some basic concepts in multiobjective optimization. Let us consider the following k -objective minimization problem:

$$\text{Minimize } \mathbf{z} = (f_1(\mathbf{y}), f_2(\mathbf{y}), \dots, f_k(\mathbf{y})) \text{ subject to } \mathbf{y} \in \mathbf{Y}, \quad (2)$$

where \mathbf{z} is the objective vector, $f_i(\mathbf{y})$ is the i -th objective to be minimized, \mathbf{y} is the decision vector, and \mathbf{Y} is the feasible region in the decision space.

Let \mathbf{a} and \mathbf{b} be two feasible solutions of the k -objective minimization problem in (2). If the following condition holds, \mathbf{a} can be viewed as being better than \mathbf{b} :

$$\forall i, f_i(\mathbf{a}) \leq f_i(\mathbf{b}) \text{ and } \exists j, f_j(\mathbf{a}) < f_j(\mathbf{b}). \quad (3)$$

In this case, we say that \mathbf{a} dominates \mathbf{b} (equivalently \mathbf{b} is dominated by \mathbf{a}).

When \mathbf{b} is not dominated by any other feasible solutions (i.e., when there exists no feasible solution \mathbf{a} that dominates \mathbf{b}), the solution \mathbf{b} is referred to as a Pareto-optimal solution of the k -objective minimization problem in (2). The set of all Pareto-optimal solutions forms the tradeoff surface in the objective space. Various EMO algorithms have been proposed to efficiently search for Pareto-optimal solutions [2], [3], [7]. Since it is very difficult to find the true Pareto-optimal solutions of a large-scale multiobjective optimization problem, non-dominated solutions among the examined ones during the execution of EMO algorithms are usually presented as a final solution set.

3 Heuristic Fuzzy Rule Extraction

Let us assume that we have m training patterns $\mathbf{x}_p = (x_{p1}, x_{p2}, \dots, x_{pn})$, $p = 1, 2, \dots, m$ from M classes in the n -dimensional continuous pattern space where x_{pi} is the attribute value of the p -th training pattern for the i -th attribute. For the simplicity of explanation, we assume that all the attribute values have already been normalized into real numbers in the unit interval $[0, 1]$.

For our pattern classification problem, we use fuzzy rules of the following type:

$$\text{Rule } R_q: \text{ If } x_1 \text{ is } A_{q1} \text{ and } \dots \text{ and } x_n \text{ is } A_{qn} \text{ then Class } C_q \text{ with } CF_q, \quad (4)$$

where R_q is the label of the q -th fuzzy rule, $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is an n -dimensional pattern vector, A_{qi} is an antecedent fuzzy set, C_q is a class label, and CF_q is a rule

weight (i.e., certainty grade). We denote the fuzzy rule R_q in (4) as “ $\mathbf{A}_q \Rightarrow \text{Class } C_q$ ” where $\mathbf{A}_q = (A_{q1}, A_{q2}, \dots, A_{qn})$.

Since we usually have no *a priori* information about an appropriate granularity of the fuzzy discretization for each attribute, we simultaneously use multiple fuzzy partitions with different granularities to extract candidate fuzzy rules. In computational experiments, we use four homogeneous fuzzy partitions with triangular fuzzy sets in Fig. 1. In addition to the 14 fuzzy sets in Fig. 1, we also use the domain interval $[0, 1]$ as an antecedent fuzzy set in order to represent a *don't care* condition. That is, we use the 15 antecedent fuzzy sets for each attribute in computational experiments.

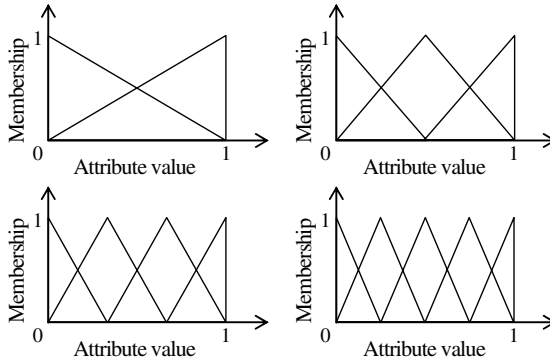


Fig. 1. Antecedent fuzzy sets used in computational experiments

Since we use the 15 antecedent fuzzy sets for each attribute of our n -dimensional pattern classification problem, the total number of combinations of the antecedent fuzzy sets is 15^n . Each combination is used in the antecedent part of the fuzzy rule in (4). Thus the total number of possible fuzzy rules is also 15^n . The consequent class C_q and the rule weight CF_q of each fuzzy rule R_q can be specified from compatible training patterns in a heuristic manner (for details, see Ishibuchi et al. [12]). That is, we can generate a large number of fuzzy rules by specifying the consequent class and the rule weight for each of the 15^n combinations of the antecedent fuzzy sets. It is, however, very difficult for human users to handle such a large number of generated fuzzy rules. It is also very difficult to intuitively understand long fuzzy rules with many antecedent conditions. Thus we examine only short fuzzy rules of length L_{\max} or less (e.g., $L_{\max} = 3$). This restriction on the rule length (i.e., the number of antecedent conditions) is to find rule sets of simple fuzzy rules with high interpretability.

Among short fuzzy rules, we generate a prespecified number of candidate fuzzy rules using heuristic rule evaluation criteria. In the field of data mining, two rule evaluation criteria (i.e., confidence and support) have been often used [1], [4], [5]. The fuzzy version of the confidence is defined as follows [12]:

$$c(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{\sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}{\sum_{p=1}^m \mu_{\mathbf{A}_q}(\mathbf{x}_p)}, \quad (5)$$

where $\mu_{\mathbf{A}_q}(\mathbf{x}_p)$ is the compatibility grade of each training pattern \mathbf{x}_p with the antecedent part \mathbf{A}_q of the fuzzy rules $\mathbf{A}_q \Rightarrow \text{Class } C_q$ in (4), which is defined as follows:

$$\mu_{\mathbf{A}_q}(\mathbf{x}_p) = \mu_{A_{q1}}(x_{p1}) \cdot \dots \cdot \mu_{A_{qn}}(x_{pn}). \quad (6)$$

In the same manner, the support is defined as follows [12]:

$$s(\mathbf{A}_q \Rightarrow \text{Class } h) = \frac{\sum_{\mathbf{x}_p \in \text{Class } h} \mu_{\mathbf{A}_q}(\mathbf{x}_p)}{m}. \quad (7)$$

A prespecified number of candidate fuzzy rules are extracted using the following ranking mechanisms of fuzzy rules (Of course, we can use other methods such as the SLAVE criterion [9]):

Support criterion with the minimum confidence level: Each fuzzy rule is evaluated based on its support when its confidence is larger than or equal to the prespecified minimum confidence level. Under this criterion, we never extract unqualified rules whose confidence is smaller than the minimum confidence level. Various values of the minimum confidence level are examined in computational experiments.

Confidence criterion with the minimum support level: Each fuzzy rule is evaluated based on its confidence when its support is larger than or equal to the prespecified minimum support level. Under this criterion, we never extract unqualified rules whose support is smaller than the minimum support level. Various values of the minimum support level are examined in computational experiments.

4 Computational Experiments

In the first stage of our multiobjective rule selection method, a prespecified number of candidate fuzzy rules are extracted. We extract 300 candidate fuzzy rules for each class using the above-mentioned two ranking mechanisms. Experimental results on some benchmark data sets in the UC Irvine machine learning repository are summarized in Table 1 where the classification rates on training patterns of 300M candidate fuzzy rules are shown (M : the number of classes in each data set). In the last column, we extract 30 candidate fuzzy rules using the ten specifications in the other columns and use all of them (i.e., 300 candidate fuzzy rules for each class in total). Bold face shows the best result for each data set. Bad results, which are more than 10% worse than the best result, are indicated by underlines. From Table 1, we can see that the choice of an appropriate specification in the rule ranking mechanisms is problem-specific. Since no specification is good for all the seven data sets, we use 30 candidate rules extracted by each of the ten specifications (i.e., 300 rules for each class in the last column of Table 1: 300M rules for each data set).

Then we apply NSGA-II to the 300M candidate fuzzy rules for each data set to search for non-dominated rule sets. Fig. 2 shows five rule sets obtained by a single run of NSGA-II for the iris data. It should be noted that the three plots show the same five rule sets using a different horizontal axis. We can observe in each plot of Fig. 2 the accuracy-complexity tradeoff with respect to a different complexity measure. One rule set with a 2.67% error rate (i.e., the simplest rule set in Fig. 2) and the

corresponding classification boundary are shown in Fig. 3. We can see that the rule set with the three fuzzy rules in Fig. 3 has high linguistic interpretability. For example, the first rule is linguistically interpreted as “If x_4 is *very small* then Class 1.” We can also see that the classification boundary in Fig. 3 is intuitively acceptable.

Table 1. Classification rates on training patterns of candidate fuzzy rules

Data set	Support with minimum conf.					Confidence with minimum sup.					Mixed
	0.6	0.7	0.8	0.9	1.0	0.01	0.02	0.05	0.10	0.15	
Iris	96.00	96.00	96.00	96.00	88.00	92.67	94.00	96.00	96.00	95.33	96.00
Breast W	95.90	95.90	95.90	95.90	<u>82.43</u>	90.48	94.58	96.78	96.78	96.24	96.34
Diabetes	69.40	69.92	73.05	78.26	<u>14.06</u>	<u>63.28</u>	<u>64.45</u>	77.34	75.52	71.61	70.44
Glass	69.63	65.89	<u>56.07</u>	<u>31.78</u>	<u>23.83</u>	<u>55.61</u>	69.16	62.62	64.49	<u>59.44</u>	68.69
Heart C	62.96	64.65	68.35	58.92	<u>48.82</u>	<u>54.21</u>	<u>55.62</u>	<u>50.17</u>	59.26	<u>52.46</u>	65.32
Sonar	<u>77.40</u>	<u>78.37</u>	<u>79.33</u>	90.38	83.65	81.25	87.98	88.94	<u>79.33</u>	<u>77.88</u>	87.02
Wine	97.19	97.19	96.63	97.19	98.88	95.51	96.07	98.88	96.63	96.07	96.63

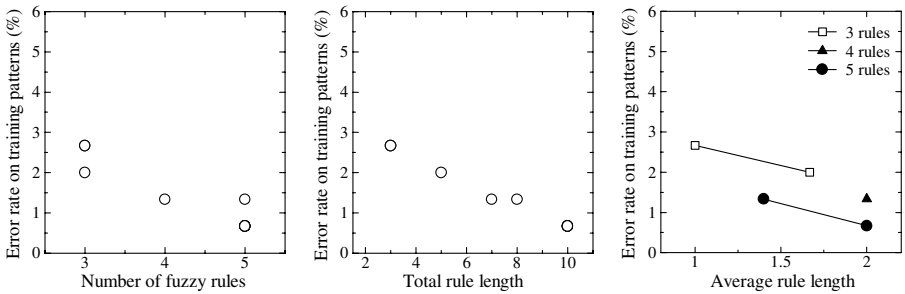


Fig. 2. Non-dominated rule sets by a single run of the NSGA-II algorithm for the iris data

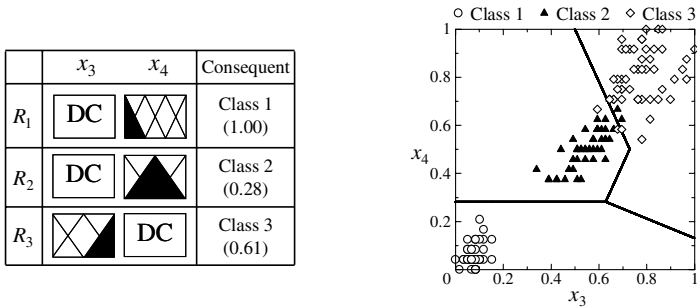


Fig. 3. One rule set obtained for the iris data and the corresponding classification boundary

We also obtain simple rule sets with high interpretability for the other data sets as shown in Fig. 4 where the error rate of each data set is shown as a figure caption.

	x_2	x_6	Consequent
R_1		DC	Class 1 (0.83)
R_2	DC		Class 2 (0.82)

	x_2	Consequent
R_1		Class 1 (0.57)
R_2		Class 2 (0.50)

	x_1	x_3	x_8	Consequent
R_1			DC	Class 1 (0.32)
R_2	DC	DC		Class 6 (0.80)

(a) Breast W: 6.73% error (b) Diabetes: 25.78% error (c) Glass: 58.89% error

	x_9	x_{10}	Consequent
R_1		DC	Class 1 (0.37)
R_2	DC		Class 1 (0.29)

	x_{11}	x_{12}	Consequent
R_1		DC	Class 1 (0.25)
R_2	DC		Class 2 (0.27)

	x_7	x_{10}	x_{13}	Consequent
R_1	DC	DC		Class 1 (0.85)
R_2	DC		DC	Class 2 (0.71)
R_3		DC	DC	Class 3 (0.62)

(d) Heart C: 46.13% error (e) Sonar: 22.59% error (f) Wine: 6.18% error

Fig. 4. Example of obtained rule sets for the other data sets

5 Concluding Remarks

In this paper, we demonstrated that simple rule sets with high interpretability can be obtained by multiobjective rule selection for some benchmark data sets in the UC Irvine machine learning repository. Since our approach selects a small number of short fuzzy rules using homogeneous fuzzy partitions, rule sets with high linguistic interpretability are obtained as shown in Fig. 3 and Fig. 4. Whereas all the rule sets in Fig. 3 and Fig. 4 have high interpretability, the classification accuracy of these rule sets is not necessarily high. Especially the rule sets in Fig. 4 (c) for the glass data with six classes and Fig. 4 (d) for the Cleveland heart disease data with five classes have poor classification accuracy. This is because the number of fuzzy rules is less than the number of classes. When emphasis should be placed on the classification accuracy, more complicated rule sets with higher accuracy can be chosen from non-dominated rule sets. For example, our approach found a rule set with 14 fuzzy rules for the glass data. The error rate of this rule set was 17.76% whereas the rule set with two fuzzy rules in Fig. 4 (c) has a 58.89% error rate. Another difficulty of our approach for multi-class problem is that fuzzy rules for all classes are not necessarily included in rule sets. When we need at least one fuzzy rule for each class, an additional constraint condition can be introduced to multiobjective rule selection.

Acknowledgement

This work was partially supported by Japan Society for the Promotion of Science (JSPS) through Grand-in-Aid for Scientific Research (B): KAKENHI (17300075).

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I.: Fast Discovery of Association Rules. In: Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park (1996) 307-328
2. Coello Coello, C. A., Lamont, G. B.: *Applications of Multi-Objective Evolutionary Algorithms*. World Scientific, Singapore (2004)
3. Coello Coello, C. A., Van Veldhuizen, D. A., Lamont, G. B.: *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, Boston, MA (2002)
4. Coenen, F., Leng, P.: Obtaining Best Parameter Values for Accurate Classification. *Proc. of the 5th IEEE International Conference on Data Mining* (2005) 549-552
5. Coenen, F., Leng, P., and Zhang, L.: Threshold Tuning for Improved Classification Association Rule Mining. *Lecture Notes in Computer Science 3518: Advances in Knowledge Discovery And Data Mining - PAKDD 2005*. Springer, Berlin (2005) 216-225
6. Cordon, O., Jesus, M. J. del, Herrera, F., Magdalena, L., and Villar, P.: A Multiobjective Genetic Learning Process for Joint Feature Selection and Granularity and Contexts Learning in Fuzzy Rule-based Classification Systems. In: Casillas, J., Cordon, O., Herrera, F., Magdalena, L. (eds.): *Interpretability Issues in Fuzzy Modeling*. Springer, Berlin (2003) 79-99
7. Deb, K.: *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Chichester (2001)
8. Deb, K., Pratap, A., Agrawal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Trans. on Evolutionary Computation* 6, 2 (2002) 182-197
9. Gonzalez, A., Perez, R.: SLAVE: A Genetic Learning System based on an Iterative Approach. *IEEE Trans. on Fuzzy Systems* 7, 2 (1999) 176-191
10. Ishibuchi, H., Murata, T., Turksen, I. B.: Single-Objective and Two-Objective Genetic Algorithms for Selecting Linguistic Rules for Pattern Classification Problems. *Fuzzy Sets and Systems* 89, 2 (1997) 135-150
11. Ishibuchi, H., Nakashima, T., Murata, T.: Three-Objective Genetics-based Machine Learning for Linguistic Rule Extraction. *Information Sciences* 136, 1-4 (2001) 109-133
12. Ishibuchi, H., Nakashima, T., Nii, M.: *Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining*. Springer, Berlin (2004)
13. Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H.: Construction of Fuzzy Classification Systems with Rectangular Fuzzy Rules Using Genetic Algorithms. *Fuzzy Sets and Systems* 65, 2/3 (1994) 237-253
14. Ishibuchi, H., Nozaki, K., Yamamoto, N., Tanaka, H.: Selecting Fuzzy If-Then Rules for Classification Problems Using Genetic Algorithms. *IEEE Trans. on Fuzzy Systems* 3, 3 (1995) 260-270
15. Ishibuchi, H., Yamamoto, T.: Fuzzy Rule Selection by Multi-Objective Genetic Local Search Algorithms and Rule Evaluation Measures in Data Mining. *Fuzzy Sets and Systems* 141, 1 (2004) 59-88
16. Jimenez, F., Gomez-Skarmeta, A. F., Sanchez, G., Roubos, H., Babuska, R.: Accurate, Transparent and Compact Fuzzy Models by Multi-Objective Evolutionary Algorithms. In Casillas, J., Cordon, O., Herrera, F., Magdalena, L. (eds.): *Interpretability Issues in Fuzzy Modeling*. Springer, Berlin (2003) 431-451
17. Wang, H., Kwong, S., Jin, Y., Wei, W., Man, K. F.: Multi-Objective Hierarchical Genetic Algorithm for Interpretable Fuzzy Rule-based Knowledge Extraction. *Fuzzy Sets and Systems* 149, 1 (2005) 149-186
18. Wang, H., Kwong, S., Jin, Y., Wei, W., Man, K. F.: Agent-based Evolutionary Approach for Interpretable Rule-based Knowledge Extraction. *IEEE Trans. on Systems, Man, and Cybernetics - Part C: Applications and Reviews* 35, 2 (2005) 143-155

Clustering Mixed Data Using Spherical Representaion

Yoshiharu Sato

Hokkaido University, Sapporo, Japan
ysato@main.ist.hokudai.ac.jp

Abstract. When the data is given as mixed data, that is, the attributes take the values in mixture of binary and continuous, a clustering method based on k -means algorithm has been discussed. The binary part is transformed into the directional data (spherical representation) by a weight transformation which is induced from the consideration of the similarity between binary objects and of the natural definition of descriptive measures. At the same time, the spherical representation of the continuous part is given by the use of multidimensional scaling on the sphere. Combining the binary part and continuous part, like the latitude and longitude, we obtained a spherical representation of mixed data. Using the descriptive measures on a sphere, we obtain the clustering algorithm for mixed data based on k -means method. Finally, the performance of this clustering is evaluated by actual data.

1 Introduction

The mixed data is defined such a data that each object is measured by the binary attributes and the continuous attributes simultaneously. Then the each object \mathbf{o}_i is denoted by

$$\mathbf{o}_i = (\mathbf{x}_i, \mathbf{y}_i) = (x_{i1}, x_{i2}, \dots, x_{ip}, y_{i1}, y_{i2}, \dots, y_{iq}), \quad (i = 1, 2, \dots, n) \quad (1)$$

where x_{ir} takes binary value 0 or 1, and y_{it} takes the continuous value.

Recently, the size of data goes on increasing by the development of information technology. Then the feasible clustering method seems to be k -means method or its modifications. In k -means method, the concept of mean and variance of the observed data play the essential role. Then the binary data is transromed into directional data in order to get the natural definition of descriptive measures.

When the mixed data is given, traditional cluster analysis has the essential problem in mixture of the distance between binary data and the distance between continuous data. Then a fundamental idea of this paper is that if we get the spherical representation of the binary data and the continuous data simultaneously, we may combine these two spherical data into one spherical data, that is, one is considered to be a latitude and the other to be a longitude. In order to get the spherical representation of q -dimensional continuous data, we use the concept of multidimensional scaling on q -dimensional sphere so as to keep a distance relation between q -dimensional continuous configuration and q -dimensional spherical configuration.

2 Transformation of Binary Data Into Directional Data

We assume that the following n binary objects with p attributes are given.

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}), \quad x_{ia} = 1 \text{ or } 2, \quad (i = 1, 2, \dots, n; a = 1, 2, \dots, p)$$

We suppose that each object \mathbf{x}_i is weighted by the sum of the value of attributes, i.e. sum of the component of the vector \mathbf{x}_i . When we denote the weighted vector as $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2}, \dots, \xi_{ip})$, the components are given by

$$\xi_{ia} = x_{ia} / \sum_{b=1}^p x_{ib}, \quad \sum_{a=1}^p \xi_{ia} = 1, \quad \xi_{ia} > 0, \quad (2)$$

Then the vectors $\boldsymbol{\xi}_i$ are located on $(p - 1)$ -dimensional hyperplane in the first quadrant of p -dimensional space. We must introduce a suitable metric function on this hyperplane. Since $\boldsymbol{\xi}_i$ has the property in expression 2, we can use an analogy of a discrete probability distribution, i.e. if we regard $\boldsymbol{\xi}_i$ as a probability, then we are able to introduce Kullback-Leibler divergence as a distance measure, which are defined as follows,

$$D(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j) = \frac{1}{2} \sum_{a=1}^p (\xi_{ia} - \xi_{ja}) \log \frac{\xi_{ia}}{\xi_{ja}}$$

When we evaluate Kullback-Leibler divergence between two points, $\boldsymbol{\xi}_i, \boldsymbol{\xi}_i + d\boldsymbol{\xi}_i$, up to the second order with respect to $d\boldsymbol{\xi}_i$, the line element in this space is given by

$$D(\boldsymbol{\xi}_i + d\boldsymbol{\xi}_i, \boldsymbol{\xi}_i) = \frac{1}{2} \sum_{a=1}^p d\xi_{ia} \log \frac{\xi_{ia} + d\xi_{ia}}{\xi_{ia}} = \frac{1}{4} \sum_{a=1}^p \frac{1}{\xi_{ia}} (d\xi_{ia})^2. \quad (3)$$

This is well known as a chi-square distance. However, since the dimension of this space (hyperplane) is $(p - 1)$, we get

$$D(\boldsymbol{\xi}_i + d\boldsymbol{\xi}_i, \boldsymbol{\xi}_i) = \frac{1}{4} \sum_{a=1}^{p-1} \sum_{b=1}^{p-1} \left(\delta_{ab} \frac{1}{\xi_{ia}} + \frac{1}{\xi_{ip}} \right) d\xi_{ia} d\xi_{ib}$$

Then we may consider the hyperplane should be a Riemannian space. The structure of the hyperplane will be discussed by the several geometrical quantities. But we know that the induced metric on a hypersphere in p -dimensional Euclidean is denoted as follows. Using a coordinate (u_1, u_2, \dots, u_p) and $u_1 + u_2 + \dots + u_p = 1, u_a > 0$, when we denote the hypersphere as follows,

$$\ell_1 = \sqrt{u_1}, \ell_2 = \sqrt{u_2}, \dots, \ell_{(p-1)} = \sqrt{u_{(p-1)}}, \ell_p = \left\{ 1 - \sum_{b=1}^{p-1} u_b \right\}^{1/2}, \quad (4)$$

the induced metric is given by

$$ds^2 = \sum_{a=1}^{p-1} \sum_{b=1}^{p-1} g_{ab} du_a du_b = \sum_{a=1}^{p-1} \sum_{b=1}^{p-1} \frac{1}{4} \left(\delta_{ab} \frac{1}{u_a} + \frac{1}{u_p} \right) du_a du_b. \quad (5)$$

Then we know that the structure of the hyperplane is a hypersphere.

From this result, we define a directional data, i.e. the data on the unit hypersphere using weighted ξ_i as

$$\ell_{ia} = \sqrt{\xi_{ia}}, \quad (a = 1, \dots, p)$$

The main advantage using the data on the hypersphere is easy to get a global geodesic distance, because we know the geodesic curve on the hypersphere is the great circle. If we discuss on the hyperplane, we must get the geodesic curve, which is a solution of the geodesic equation.

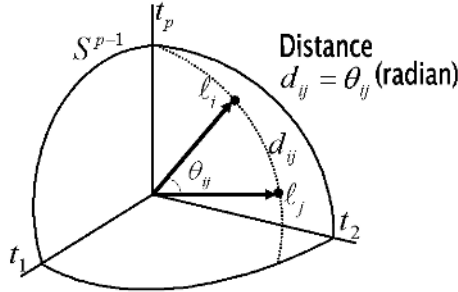


Fig. 1. Directional data and Distance

3 Spherical Representation of Continuous Data

Suppose a q -dimensional continuous data be given by

$$\mathbf{y}_i = (y_{i1}, y_{i2} \dots, y_{iq}), \quad (i = 1, 2, \dots, n)$$

Using the sample variance and covariance matrix \mathbf{S} , Mahalanobis distance between a pair of objects \mathbf{y}_i and \mathbf{y}_j is obtained as follows;

$$D = (d_{ij}^2) = (\mathbf{y}_i - \mathbf{y}_j)\mathbf{S}^{-1}(\mathbf{y}_i - \mathbf{y}_j), \quad (i, j = 1, 2, \dots, n) \tag{6}$$

This distance can be considered as a square of Euclidean distance when the original data \mathbf{y}_i is transformed that

$$\mathbf{z}_i = \mathbf{S}^{-\frac{1}{2}}(\mathbf{y}_i - \bar{\mathbf{y}}), \quad (i = 1, 2, \dots, n)$$

where $\bar{\mathbf{y}}$ denotes sample mean vector.

We are intended to get the spherical configuration such that the distance between the points on the sphere is consistent with the distance relation between continuous data \mathbf{z}_i as much as possible. Then we assign each \mathbf{z}_i to a positive quadrant in q -dimensional unit sphere.

q -dimensional unit hypersphere in $(q + 1)$ -dimensional Euclidean space is denoted as

$$\mathbf{x}(\theta_1, \theta_2, \dots, \theta_q) = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_q \\ x_{q+1} \end{bmatrix} = \begin{bmatrix} \sin \theta_1 \sin \theta_2 \cdots \sin \theta_q \\ \sin \theta_1 \sin \theta_2 \cdots \cos \theta_q \\ \dots \\ \sin \theta_1 \cos \theta_2 \\ \cos \theta_1 \end{bmatrix}. \quad (7)$$

Let $\boldsymbol{\alpha}$ be a center direction in positive quadrant in q -dimensional unit sphere, and $\boldsymbol{\beta}$ be a center direction perpendicular to the first axis.

$$\boldsymbol{\alpha} = \mathbf{x}\left(\frac{\pi}{4}, \frac{\pi}{4}, \dots, \frac{\pi}{4}\right), \quad \boldsymbol{\beta} = \mathbf{x}\left(\frac{\pi}{4}, \dots, \frac{\pi}{4}, 0\right). \quad (8)$$

Then, a point on the unit sphere is contained in the positive quadrant if the distance from $\boldsymbol{\alpha}$ is less than

$$\theta^* = \cos^{-1}(\boldsymbol{\alpha}'\boldsymbol{\beta})$$

Hence, the distance relation $D = (d_{ij})$ is transformed as

$$D^* = (d_{ij}^*) = \left\{ \frac{2\theta^*}{d_{\max}} \right\} d_{ij}, \quad (9)$$

where, $d_{\max} = \max_{i,j} d_{ij}$.

We suppose that the data point \mathbf{z}_i is assigned to a directional data ℓ_i . If the distance relation between assigned directional data reproduced the distance relation D^* completely, then

$$\cos d_{ij}^* = \ell_i' \ell_j.$$

When we denote the point on the unit sphere

$$\ell_i(\boldsymbol{\theta}_i) = \begin{bmatrix} \ell_{i1} \\ \ell_{i2} \\ \dots \\ \ell_{iq} \\ \ell_{i(q+1)} \end{bmatrix} = \begin{bmatrix} \sin \theta_{i1} \sin \theta_{i2} \cdots \sin \theta_{iq} \\ \sin \theta_{i1} \sin \theta_{i2} \cdots \cos \theta_{iq} \\ \dots \\ \sin \theta_{i1} \cos \theta_{i2} \\ \cos \theta_{i1} \end{bmatrix}, \quad \boldsymbol{\theta}_i = \begin{bmatrix} \theta_{i1} \\ \theta_{i2} \\ \dots \\ \theta_{iq} \end{bmatrix}, \quad (10)$$

and we put $Q = (q_{ij}) \equiv \cos d_{ij}^*$, the point ℓ_i is obtained so as to minimize

$$\eta = \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n (q_{ij} - \ell_i' \ell_j)^2 = \frac{1}{4} \sum_{i=1}^n \sum_{j=1}^n (q_{ij} - \sum_{k=1}^{q+1} \ell_{ik} \ell_{jk})^2 \quad (11)$$

provide that

$$0 \leq \theta_{ik} \leq \frac{\pi}{2},$$

because the point ℓ_i lies on the positive quadrant. In order to solve such a optimization problem, we must set an initial values of $\boldsymbol{\theta}_i$, denoted $\boldsymbol{\theta}_i^0$, which are

given as follows; When we denote T_α as the tangent space of the sphere on the point α , the dimension of T_α is q and the natural frame is given by

$$e_i = \frac{\partial \mathbf{x}}{\partial \theta_i}, \quad (i = 1, 2, \dots, q)$$

Normalizing each base e_i , we get

$$e_i^* = \frac{e_i}{\|e_i\|}, \quad (i = 1, \dots, q)$$

By the system $\{e_1^*, e_2^*, \dots, e_q^*\}$ and $\alpha = e_{q+1}^*$ is considered to be a orthonormal base of $q+1$ -dimensional Euclidean space. When we denote $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iq})$, the point on the tangent space is described as

$$\mathbf{z}_{(T_\alpha)_i} = z_{i1}e_1^* + z_{i2}e_2^* + \dots + z_{iq}e_q^*.$$

Then position vector in $(q + 1)$ -dimensional Euclidean space is denoted by

$$\mathbf{v}_i = \alpha + \mathbf{z}_{(T_\alpha)_i}.$$

Hence, we put the initial point ℓ_i^0 as

$$\ell_i^0 = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|}.$$

When the mixed data is given by

$$(\mathbf{x}_i, \mathbf{y}_i) = (x_{i1}, x_{i2}, \dots, x_{ip}, y_{i1}, y_{i2}, \dots, y_{iq}),$$

the binary data \mathbf{x}_i and the continuous data \mathbf{y}_i are represented as the directional data $\ell_i^B(\theta_i^B)$ and $\ell_i^C(\theta_i^C)$, respectively. Then the total $(p + q - 1)$ -dimensional polar coordinate is given by

$$(\theta_{i1}^B, \dots, \theta_{i(p-1)}^B, \theta_{i1}^C, \dots, \theta_{iq}^C) \equiv (\theta_{i1}, \dots, \theta_{i(p-1)}, \theta_{ip}, \dots, \theta_{i(p+q-1)}) \equiv \theta_i. \quad (12)$$

Using the polar coordinate θ_i , we get the spherical representation of mixed data, that is, the transformation the mixed data into directional data as follows;

$$\ell_i(\theta_i) = \begin{bmatrix} \sin \theta_{i1} \cdots \sin \theta_{ip} \sin \theta_{i(p+1)} \cdots \sin \theta_{i(p+q-1)} \\ \sin \theta_{i1} \cdots \sin \theta_{ip} \sin \theta_{i(p+1)} \cdots \cos \theta_{i(p+q-1)} \\ \cdots \\ \sin \theta_{i1} \cos \theta_{i2} \\ \cos \theta_{i1} \end{bmatrix}. \quad (13)$$

4 k -Means Method for Directional Data

The descriptive measures in directional data are given as follows. We suppose that the directional data on S^{p-1} with the size n is given by

$$\ell_i = (\ell_{i1}, \ell_{i2}, \dots, \ell_{ip}), \quad \ell_i' \ell_i = 1, \quad (i = 1, 2, \dots, n). \quad (14)$$

The mean direction is given by ([2])

$$\bar{\ell} = (\bar{\ell}_1, \dots, \bar{\ell}_p), \quad \bar{\ell}_a = \frac{\sum_{i=1}^n \ell_{ia}}{R}, \quad R^2 = \sum_{b=1}^p \left(\sum_{j=1}^n \ell_{jb} \right)^2.$$

The variance, called circular variance around mean is known ([2]) as

$$V = \frac{1}{n} \sum_{i=1}^n \{1 - \ell'_i \bar{\ell}\}.$$

By the natural extension the k -means algorithm to the directional data, we get the following algorithm (spherical k -means, in short). We suppose that a set of n directional objects on the hypersphere S^{p-1} is given by (14). When the number of clusters K is given, the criterion of spherical k -means algorithm is given by

$$\eta = \sum_{k=1}^K \sum_{\ell_i \in C_k} V^{(k)} = \sum_{k=1}^K \sum_{\ell_i \in C_k} \{1 - \ell'_i \bar{\ell}^{(k)}\},$$

$$\bar{\ell}^{(k)} = (\bar{\ell}_1^{(k)}, \dots, \bar{\ell}_p^{(k)}), \quad \bar{\ell}_a^{(k)} = \frac{\sum_{\ell_i \in C_k} \ell_{ia}}{\sqrt{\sum_{a=1}^p \left(\sum_{\ell_i \in C_k} \ell_{ia} \right)^2}}.$$

Minimization η is attained by the maximization of the term $\ell'_i \bar{\ell}^{(k)}$. This term denotes the cosine of the angle between ℓ_i and $\bar{\ell}^{(k)}$, that is, the distance between the points on the hypersphere ℓ_i and $\bar{\ell}^{(k)}$. Therefore, each point ℓ_i is assigned the cluster which has the nearest to its mean.

5 The Performance and Characteristic Feature of the Spherical Clustering

Here we discuss the characteristic feature of the spherical k -means, proposed here, using actual data set.

First example is a credit card approval data which is submitted by Quinlan, J. R. ([3]) to "The Machine Learning Database Repository". This dataset is interesting because there is a good mixture of attributes. We use 10 binary attributes and 6 continuous attributes. All attribute names and values have been changed to meaningless to protect confidentiality of the data. There two classes in this data, one is approved class the other is not approved class, these denoted "+" and "-". Number of observation of each class is 285 and 356, respectively. We transform this data into directional data, and applied the spherical k -means clustering. In k -means algorithm, we must set the initial seed points (initial class centers). Here we use two different observations which are select from the total observations randomly as the initial seed points. Since k -means algorithm could not guarantee the global optimum solution, this processes are repeated $10,000 \times 15$ times in order to get the local solutions. The result in Table 1.(a) has the minimum within variance in this experiment. Since this data is well-known, there

Table 1. Credit card data(KM: Spherical k -Means Method)

		Observed	
		+	-
KM	+	285	27
	-	0	329
Total		285	356

WV: 0.00887, Ac:95%

(a)

		Observed	
		+	-
KM	+	281	0
	-	4	356
Total		285	356

WV: 0.00891, Ac:99%

(b)

WV : Within Variance, Ac : Accuracy

are many reports on the result of discriminant analysis. But the accuracies are not so good. For the reference, in Table 2 (a), (b), (c), the results of discriminant analysis are shown. Table 1, (b) shows that these two classes are almost linear separable on the sphere. It will be understood that discriminant analysis and cluster analysis are the different criterion. Then the within variances of the discrimination are greater than the result of k -means. Moreover, the criterion of the discrimination is minimize a risk function, usually, the misdiscrimination rate, then the data is processed under the labeled data. But clustering does not take into account the label of the data. However, this result suggest that the classical prototype discrimination method seems to be useful when the data has some structure, that is, gather in clusters. And also we will obtain clusters for mixed data in a natural way by the spherical representaion.

Moreover, the spherical k -means method is essentially the same with ordinal k -means method. Then this property does not depend on the spherical representation. In order make sure that, we apply k -means to ordinal continuous data. The data is Wisconsin Diagnostic Breast Cancer. ([3]) This has 30 continuous attributes, namely the usual multivariate data. Total observations are 569. There are two classes, one is malignant cancer, 212 observations, and the other is benign cancer, 357 observations. The result of k -means method and discriminant analysis are shown in Table 3,(a), Table 4. The result of k -means method is almost the same with support vector machine. Most interesting point is Table 3, (b). This shows that this data is completely linear separable. However, this

Table 2. Credit card data(Discriminant Functions)

		Observed	
		+	-
SD	+	244	67
	-	41	289
Total		285	356

WV: 0.00986, Ac:83.2%

(a)

		Observed	
		+	-
LDF	+	253	90
	-	32	266
Total		285	356

WV: 0.120, Ac:81.0%

(b)

		Observed	
		+	-
SVM	+	269	55
	-	16	301
Total		285	356

WV: 0.0101, Ac:88.9%

(c)

SD : Bayse Discriminant function using Spherical Distribution.

LDF : Linear Discriminant Function, SVM : Support Vector Machine.

Table 3. Wisconsin Diagnostic Brest Cancer (KM: k -Means Method)

		Observed	
		M	B
KM	M	200	0
	B	12	357
Total		212	357

WV: 28.63, Ac:98%

(a)

		Observed	
		M	B
KM	M	212	0
	B	0	357
Total		212	357

WV: 29.49, Ac:100%

(b)

M : Malignant Cancer B : Benign Cancer

Table 4. Wisconsin Diagnostic Brest Cancer (Discriminant Functions)

		Observed	
		M	B
LDF	M	194	2
	B	18	355
Total		212	357

WV: 29.88, Ac:96.5%

(a)

		Observed	
		M	B
SVM	M	205	0
	B	7	355
Total		212	357

WV: 29.94, Ac:98.8%

(b)

LDF : Linear Discriminant Function, SVM : Support Vector Machine

solution is not the solution of k -means method but also the hyper plane which is the perpendicular bisector between means of two classes is not LDF function. It is natural that these classes are linear separable when we observed the attributes which are closely related to the discrimination.

References

1. MacQueen, J. : Some methods for classification and analysis of multivariate observations. In L.M.Le Can & J. Neyman (Eds), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, **1** (1967) 281–297
2. Mardia, K. : Statistics of Directional Data, Academic Press, (1972)
3. UCI Machine Learning Information / The Machine Learning Database Repository (<http://www.ics.uci.edu/~mllearn/>)

Fuzzy Structural Classification Methods

Mika Sato-Ilic¹ and Tomoyuki Kuwata²

¹ Faculty of Systems and Information Engineering, University of Tsukuba,
Tsukuba, Ibaraki 305-8573, Japan

`mika@risk.tsukuba.ac.jp`

² Faculty of Systems and Information Engineering, University of Tsukuba,
Tsukuba, Ibaraki 305-8573, Japan

`kuwata10@sk.tsukuba.ac.jp`

Abstract. This paper presents several fuzzy clustering methods based on self-organized similarity (or dissimilarity). Self-organized similarity (or dissimilarity) has been proposed in order to consider not only the similarity (or dissimilarity) between a pair of objects but also the similarity (or dissimilarity) between the classification structures of objects. Depending on how the similarity (or dissimilarity) of the classification structures cope with the fuzzy clustering methods, the results will be different from each other. This paper discusses this difference and shows several numerical examples.

1 Introduction

Recently, self-organized techniques were proposed and many algorithms have been developed and applied to many application areas. Clustering methods are no exception and much research based on this idea has been proposed [4], [8].

We also have proposed self-organized fuzzy clustering methods under an assumption that similar objects have similar classification structures [6], [7]. We have defined self-organized similarity and dissimilarity using weights which show the degree of similarity or dissimilarity between a pair of fuzzy classification structures. Given the self-organized similarity (or dissimilarity), it is known that the proposed methods tend to give clearer results. In this paper, we discuss several clustering methods based on self-organized similarity or dissimilarity and the features of these methods. These methods use the fuzzy clustering model [5] and the FANNY algorithm [3].

2 Additive Fuzzy Clustering Model

A fuzzy clustering model [5] has been proposed as follows:

$$s_{ij} = \sum_{k=1}^K \sum_{l=1}^K w_{kl} u_{ik} u_{jl} + \varepsilon_{ij}, \quad (1)$$

under the following conditions:

$$u_{ik} \in [0, 1], \forall i, k; \sum_{k=1}^K u_{ik} = 1, \forall i. \tag{2}$$

In this model, u_{ik} shows degree of belongingness of an object i to a cluster k . The weight w_{kl} is considered to be a quantity which shows the asymmetric similarity between a pair of clusters. That is, we assume that the asymmetry of the similarity between the objects is caused by the asymmetry of the similarity between the clusters. We assume the following condition:

$$0 \leq w_{kl} \leq 1. \tag{3}$$

s_{ij} shows the similarity between objects i and j . ε_{ij} is an error. K shows the number of clusters and n is the number of objects. The purpose of model (1) is to find $U = (u_{ik})$ and $W = (w_{kl})$ which minimize the following sum of squares error η^2 under the conditions (2) and (3),

$$\eta^2 = \sum_{i \neq j=1}^n (s_{ij} - \sum_{k=1}^K \sum_{l=1}^K w_{kl} u_{ik} u_{jl})^2. \tag{4}$$

3 FANNY Algorithm

Fuzzy c-means (FCM) [1] is one of the methods of fuzzy clustering. FCM is a method which minimizes the weighted within-class sum of squares:

$$J(U, \mathbf{v}_1, \dots, \mathbf{v}_K) = \sum_{i=1}^n \sum_{k=1}^K (u_{ik})^m d^2(\mathbf{x}_i, \mathbf{v}_k), \tag{5}$$

where $\mathbf{v}_k = (v_{ka})$, $k = 1, \dots, K$, $a = 1, \dots, p$ denotes the values of the centroid of a cluster k , $\mathbf{x}_i = (x_{ia})$, $i = 1, \dots, n$, $a = 1, \dots, p$ is i -th object with respect to p variables, and $d^2(\mathbf{x}_i, \mathbf{v}_k)$ is the square Euclidean distance between \mathbf{x}_i and \mathbf{v}_k . The exponent m which determines the degree of fuzziness of the clustering is chosen from $(1, \infty)$ in advance. The purpose of this is to obtain the solutions U and $\mathbf{v}_1, \dots, \mathbf{v}_K$ which minimize equation (5). The minimizer of equation (5) is shown as:

$$J(U) = \sum_{k=1}^K \left(\sum_{i=1}^n \sum_{j=1}^n ((u_{ik})^m (u_{jk})^m d_{ij}) / (2 \sum_{s=1}^n (u_{sk})^m) \right), \tag{6}$$

where $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$. The equation (6) is the objective function of the relational fuzzy c-means [2]. When $m = 2$, equation (6) is the objective function of the FANNY algorithm [3].

4 Self-organized Clustering Methods

4.1 Self-organized Clustering Methods Based on Additive Fuzzy Clustering Model

The method consists of the following three steps:

(Step 1) Apply the similarity data for model (1). If the similarity data is symmetric, we can obtain the symmetric matrix for $W = (w_{kl})$. Obtain the solutions $\hat{U} = (\hat{u}_{ik})$ and $\hat{W} = (\hat{w}_{kl})$.

(Step 2) Using the obtained \hat{U} , recalculate the following similarity:

$$\tilde{s}_{ij} = \frac{1}{\sum_{k=1}^K (\hat{u}_{ik} - \hat{u}_{jk})^2} s_{ij}, \quad i, j = 1, \dots, n. \tag{7}$$

Using \tilde{s}_{ij} , go back to Step 1 and obtain a new result for \tilde{U} and \tilde{W} .

(Step 3) Evaluate the fitness shown in equation (4) using \tilde{U} and \tilde{W} and compare with the fitness obtained by using \hat{U} and \hat{W} . If the fitness using \tilde{U} and \tilde{W} is smaller than the fitness using \hat{U} and \hat{W} , then replace \hat{U} and \hat{W} by \tilde{U} and \tilde{W} , and repeat Steps 1 to 3. If the fitness using \tilde{U} and \tilde{W} , and the difference between the fitness using \hat{U} and \hat{W} and the fitness using \tilde{U} and \tilde{W} are sufficiently small, then stop. Otherwise, repeat Steps 1 to 3, using new initial values for U and W .

Equation (7) shows that if \hat{u}_{ik} and \hat{u}_{jk} are similar to each other, then the similarity between objects i and j becomes larger. This similarity is self organizing according to the degree of belongingness for each of the clusters obtained in each iteration. In equation (7), we can rewrite \tilde{s}_{ij} when we assume W is a unit matrix I in equation (1) as follows:

$$\tilde{s}_{ij} = \frac{\sum_{k=1}^K \hat{u}_{ik} \hat{u}_{jk}}{\sum_{k=1}^K (\hat{u}_{ik} - \hat{u}_{jk})^2}, \quad i, j = 1, \dots, n. \tag{8}$$

If $W = I$ and $\sum_{k=1}^K (\hat{u}_{ik} - \hat{u}_{jk})^2$ is constant, then equation (8) is essentially the same as equation (1). $\sum_{k=1}^K (\hat{u}_{ik} - \hat{u}_{jk})^2$ has a bias because of condition (2). This method does not rectify the bias, it rather uses the bias to its advantage. Since this bias tends to make a cluster in which the objects do not have clear classification structures, we can obtain the defuzzified result, while ensuring the features of the fuzzy clustering result. Figure 1 shows the bias. We assume two degrees of belongingness, $\mathbf{u}_1 = (u_{11}, u_{12})$ and $\mathbf{u}_2 = (u_{21}, u_{22})$ corresponded to two objects 1 and 2. When we fix $\mathbf{u}_2 = (0.5, 0.5)$, the solid line shows the value of

$$\sum_{k=1}^2 u_{1k} u_{2k}. \tag{9}$$

The abscissa shows values of u_{11} . Due to the condition (2), it is enough to only determine the value of u_{11} . The dotted line shows the square Euclidean distance between the degree of the belongingness for objects 1 and 2 as follows:

$$\sum_{k=1}^2 (u_{1k} - u_{2k})^2. \tag{10}$$

From figure 1, we can see that a clearer result will make a large distance from the fixed point (0.5,0.5), even if the value of the inner product is the same. For example, if the values of u_{11} are 0.8 and 0.7, the values of equation (9) are the same for both 0.8 and 0.7. However, the values of equation (10) are different from each other. The classification structure is clearer, that is when $(u_{11}, u_{12}) = (0.8, 0.2)$ has a larger distance when compared with the case of the classification structure when $(u_{11}, u_{12}) = (0.7, 0.3)$. This is caused by a bias under the condition of $\sum_{k=1}^K u_{ik} = 1$.

Figure 2 shows the situation of the bias when the number of clusters is 3. Each axis shows each cluster. The solutions for objects 1, 2, 3, and 4 are on the triangle in figure 2. In this figure, the distance between u_1 and u_2 are larger than the distance between u_3 and u_4 , even if the angle θ are almost the same.

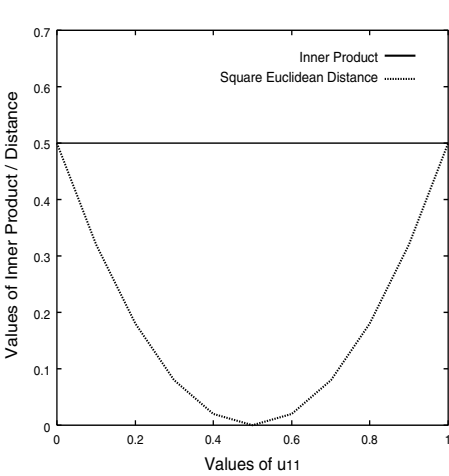


Fig. 1. Difference between Inner Product and Square Euclidean Distance

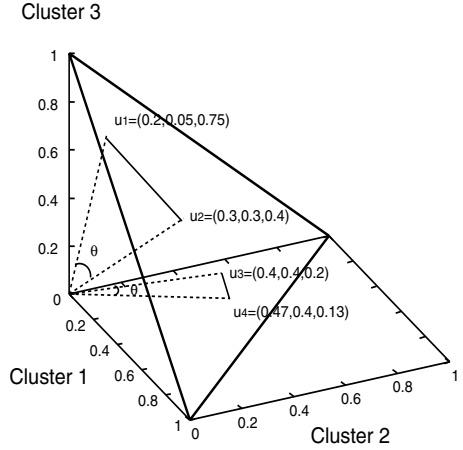


Fig. 2. Bias of the Distance

If we ignore the self-organization, then we can treat equation (8) as the following model.

$$s_{ij} = \frac{\sum_{k=1}^K u_{ik} u_{jk}}{\sum_{k=1}^K (u_{ik} - u_{jk})^2} + \varepsilon_{ij}, \quad i, j = 1, \dots, n. \tag{11}$$

4.2 Self-organized Clustering Methods Based on the FANNY Algorithm

Next, we use the FANNY algorithm shown in equation (6). We define the self-organized dissimilarity as

$$\tilde{d}_{ij} = \sqrt{\sum_{k=1}^K (\hat{u}_{ik} - \hat{u}_{jk})^2 d_{ij}}, \quad i, j = 1, \dots, n. \quad (12)$$

Here \hat{u}_{ik} is degree of belongingness of an object i to a cluster k obtained which minimize equation (6) when $m = 2$. The algorithm is as follows:

(Step 1) Apply the dissimilarity data for the equation (6) when $m = 2$. Obtain the solution $\hat{U} = (\hat{u}_{ik})$.

(Step 2) Using the obtained \hat{U} , recalculate the following dissimilarity:

$$\tilde{d}_{ij} = \sqrt{\sum_{k=1}^K (\hat{u}_{ik} - \hat{u}_{jk})^2 d_{ij}}, \quad i, j = 1, \dots, n.$$

Using \tilde{d}_{ij} , go back to Step 1 and obtain the new result for \tilde{U} .

(Step 3) Calculate the value of $\|\hat{U} - \tilde{U}\|$, where $\|\cdot\|$ shows the norm of matrix.

(Step 4) If $\|\hat{U} - \tilde{U}\| < \varepsilon$, then stop, or otherwise repeat Steps 1 to 4.

In equation (12), we assume $\hat{\mathbf{u}}_i = \hat{\mathbf{u}}_j$, ($\hat{\mathbf{u}}_i = (\hat{u}_{i1}, \dots, \hat{u}_{iK})$, $i = 1, \dots, n$). Then even if $d_{ij} \neq 0$, $\tilde{d}_{ij} = 0$. If $\hat{u}_{ik} \in \{0, 1\}$, then the possibility that the above situation occurs becomes larger. That is we stress the dissimilarity of the classification structures rather than the dissimilarity of the objects.

Considering the dissimilarity of degree of the belongingness for each object, we transform the data as follows:

$$\tilde{X} = (U|X).$$

Here \tilde{X} is a $n \times (K + p)$ matrix. Using this matrix \tilde{X} , we apply the conventional FCM or FANNY algorithm. Since the distance between objects i and j is

$$\tilde{d}_{ij} = \sqrt{\sum_{k=1}^K (\hat{u}_{ik} - \hat{u}_{jk})^2 + \sum_{a=1}^p (x_{ia} - x_{ja})^2},$$

the clustering is considering not only the dissimilarity of objects but also the dissimilarity of degree of belongingness. In this case, even if $\mathbf{u}_i = \mathbf{u}_j$, then $\tilde{d}_{ij} \neq 0$, unless $\mathbf{x}_i = \mathbf{x}_j$. This method is not self-organized using the weights of dissimilarity of the classification structures.

5 Numerical Examples

We now show an example which uses real observations. Landsat data observed over the Kushiro marsh-land is used. The value of the data shows the amount of reflected light from the ground with respect to 6 kinds of light for 75 pixels. We get data from mountain area, river area, and city area. The 1st to the 25th pixels show mountain area, the 26th to the 50th are river area, and 51st to the 75th show the city area. The results are shown in figures 3 and 4.

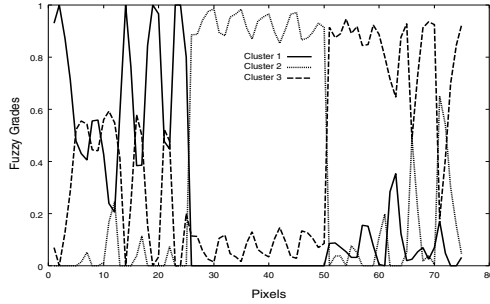


Fig. 3. Result of Landsat Data using the Fuzzy Clustering Model

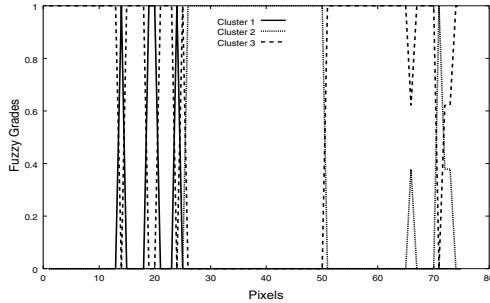


Fig. 4. Result of Landsat Data using the Self-Organized Fuzzy Clustering Method based on Equation (7)

Figure 3 shows the result using the conventional fuzzy clustering model shown in equation (1) and figure 4 is the result of the self-organized fuzzy clustering method using equation (7). In these figures, the abscissa shows each pixel and the ordinate shows the degree of belongingness for each cluster. The number of clusters is 3. From these results, we see that the result of the proposed method obtains a clearer result when compared with the result shown in figure 3.

Table 1. Comparison of the Fitness

Fitness	$\hat{\eta}^2$	$\tilde{\eta}^2$
	0.14	0.07

Table 1 shows the comparison of the fitness for both the methods. From this table, we see that a better solution is obtained by using the self-organized fuzzy clustering method using equation (7).

In table 1, $\hat{\eta}^2$ and $\tilde{\eta}^2$ are as follows:

$$\hat{\eta}^2 = \frac{\sum_{i \neq j=1}^n (s_{ij} - \sum_{k=1}^K \sum_{l=1}^K \hat{w}_{kl} \hat{u}_{ik} \hat{u}_{jl})^2}{\sum_{i \neq j=1}^n (s_{ij} - \bar{s}_{ij})^2}, \quad \bar{s}_{ij} = \frac{\sum_{i \neq j=1}^n s_{ij}}{n(n-1)}.$$

$$\tilde{\eta}^2 = \frac{\sum_{i \neq j=1}^n (\tilde{s}_{ij} - \sum_{k=1}^K \sum_{l=1}^K \tilde{w}_{kl} \tilde{u}_{ik} \tilde{u}_{jl})^2}{\sum_{i \neq j=1}^n (\tilde{s}_{ij} - \bar{\tilde{s}}_{ij})^2}, \quad \bar{\tilde{s}}_{ij} = \frac{\sum_{i \neq j=1}^n \tilde{s}_{ij}}{n(n-1)}.$$

We use the Kushiro marsh-land data for 4087 pixels over 6 kinds of light. We apply the self-organized fuzzy clustering discussed in section 4.2. The number of clusters is assumed to be 3. Figures 5 to 7 show the results of degree of belongingness for the FANNY method, our proposed method when one iteration is used, and our proposed method until it is convergent, respectively. These figures show the results of cluster 1. In these figures each color shows the range of degree of belongingness to each cluster. Red shows $0.8 < u_{ik} \leq 1.0$, yellow shows $0.6 < u_{ik} \leq 0.8$, light blue shows $0.4 < u_{ik} \leq 0.6$, blue is $0.2 < u_{ik} \leq 0.4$, and white is $0 < u_{ik} \leq 0.2$. From these figures, we can see that the degree of belongingness for the clusters is going to create a crisper situation.

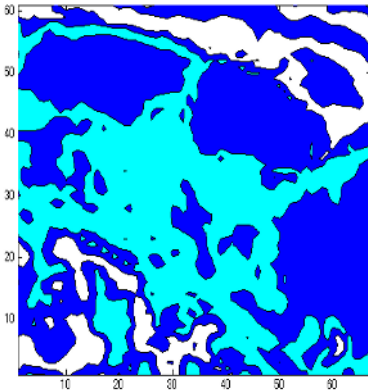


Fig. 5. Result of FANNY for Cluster 1

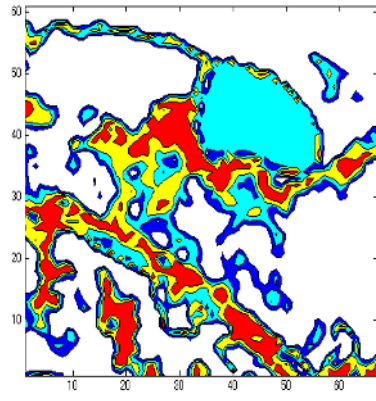


Fig. 6. Result of Self-Organized Method for Cluster 1 (after one iteration)

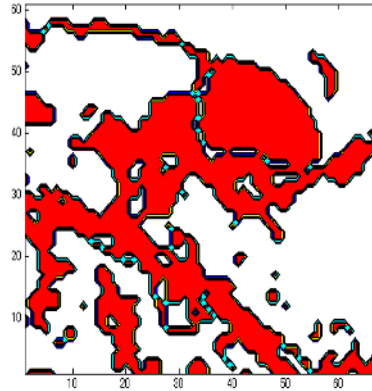


Fig. 7. Result of Self-Organized Method for Cluster 1 (after the convergence)

6 Conclusion

We discussed several methods using self-organized similarity (or dissimilarity) for fuzzy clustering methods. The concept of “self-organization” refers to the situation where the similarity of objects is affected by the similarity of classification structures of objects in each iteration. We consider two spaces. One is the observation space of objects and the other is a space of degree of belongingness of objects to clusters which show the classification structures corresponding to each object. As a result, we could obtain a clearer result and it ensures applicability as a defuzzification method.

References

1. Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press. (1987)
2. Hathaway, R. J., Davenport, J. W., and Bezdek, J. C.: Relational Duals of the C-Means Clustering Algorithms. *Pattern Recognition*. **22** (1989) 205–212
3. Kaufman L., Rousseeuw, P. J.: Finding Groups in Data. John Wiley & Sons. (1990)
4. Kohonen, T.: Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **43** (1982) 59–69
5. Sato, M., Sato, Y., and Jain, L. C.: Fuzzy Clustering Models and Applications. Springer. (1997)
6. Sato-Ilic, M.: Self Organized Fuzzy Clustering. *Intelligent Engineering Systems through Artificial Neural Networks*. **14** (2004) 579–584
7. Sato-Ilic, M., Kuwata, T.: On Fuzzy Clustering based Self-Organized Methods. *FUZZ-IEEE 2005*. (2005) 973–978
8. Van Hulle, M. M.: Faithful Representations and Topographic Maps -From Distortion to Information based Self-Organization. John Wiley & Sons. (2000)

Innovations in Soft Data Analysis

Mika Sato-Ilic¹ and Lakhmi C. Jain²

¹ Faculty of Systems and Information Engineering, University of Tsukuba,
Tsukuba, Ibaraki 305-8573, Japan
mika@risk.tsukuba.ac.jp

² Knowledge-Based Intelligent Engineering Systems Centre,
University of South Australia,
Adelaide, SA 5095, Australia
Lakhmi.Jain@unisa.edu.au

Abstract. The amount of data is growing at an exponential rate. We are faced with a challenge to analyze, process and extract useful information from the vast amount of data. Traditional data analysis techniques have contributed immensely in the area of data analysis but we believe that the soft data analysis techniques, based on soft computing techniques, can be complementary and can process complicated data sets. This paper provides an introduction to the soft data analysis paradigms. It summarizes the successful and possible applications of the soft computing analysis paradigms. The merits and demerits of these paradigms are included. A number of resources available are listed and the future vision is discussed. This paper also provides a brief summary of the papers included in the session on “Innovation in Soft Data Analysis”.

1 Introduction

Soft data analysis (SDA) is based on soft computing which is a consortium of methodologies including fuzzy logic, neural networks, probabilistic reasoning, genetic algorithms, and chaotic systems which complement each other.

Soft computing reflects the pervasiveness of imprecision and uncertainty which exists in the real world. On the other hand, hard computing does not reflect this imprecision and uncertainty. The guiding principal of soft computing is to exploit the tolerance of imprecision and uncertainty in order to achieve tractability, robustness, and low solution cost.

Recently, in the area of data analysis, many new methods have been proposed. One reason for this is that traditional data analysis does not adequately reflect the imprecision and uncertainty of real world data. For instance, analyses for uncertainty data including interval-valued data, modal data, functional data, categorical data, and fuzzy data have been proposed. In the area of neuroscience, data which have a larger number of variables than the number of objects are also important concerns in multivariable data analysis. Conventional multivariate analyses can not treat such data. Huge amounts of data are also a problem and data mining is placing an increased emphasis on revealing the latent structure existing in such data. For example, in the analyses of point of sale (POS) data in

the marketing area the key challenge is interfacing the emulator with the POS data traffic. In another area Deoxyribonucleic Acid (DNA) data in information biology also has a similar problem resulting from the amount of data.

The second reason is the increase of interest and recognition for the specific features of data which must be treated as non-ignorable features of the data. Several types of data have asymmetric features. Input and output data in an input-output table in the economic area or in human relationship data such as in psychometrics are typical examples. Traditional techniques treat the asymmetry feature as noise affecting the symmetric structure of the data. However, based on the premise that the systematic asymmetry is of value and the information contained should be captured, new methods have been proposed.

The third reason is that with the advance of computer technology and the resulting expansion of computer ability, improved visualization techniques of data, results of data analysis, and the features of data analysis have flourished. The features of complex data analysis involving imprecision and uncertainty have witnessed a crystallization of the exploratory visualization techniques for data.

The fourth reason is the limited precision in the data. In order to obtain precision from real data, we need to make many assumptions about the latent data structures. Under the many assumptions necessary for representing real data structures, even if we obtain a precise result for the data, since no one can know the real data structure, we can not prove that the assumptions actually represent the real data structure.

From this background, SDA is constituted as a consortium of data analyses aimed at reflecting imprecision and uncertainty existing in real data. Fuzzy multivariate data analysis is a popular current methodology of SDA. Examples include: fuzzy clustering, fuzzy regression analysis, fuzzy principal component analysis, fuzzy quantification analyses (types I, II, III, and IV). Hybrid methods over neural networks, genetic algorithms, support vector machines, and fuzzy data analysis are also types of SDA. The role of SDA techniques is growing and their potential is fully accepted in real world data analysis. In SDA, functional data analysis, non-linear generalized models, and symbolic data analysis, are also new and powerful methods that exhibit further progress with substantial reliance on the traditional statistical data analysis.

Several successful SDA applications have been presented. The use of grey-tone pictures in image classification is a typical example. X-ray pictures, images by satellite and particles registered by optical devices are examples of objective data. Although the data are represented by pixels as digital values, the real data is continuous, so the analyses considering the imprecision and uncertainty must be adaptable. Fuzzy clustering for this data achieves suitable representation of the result including the uncertainty of boundaries over clusters. The methods show the robustness and low solution cost.

Applications to temperature records, growth curves, time series records, and hand writing samples have also proved to be successful. Such as analog sources of data performed continuously in time or space. These data and the results are represented by functions or trajectory in hyper spaces, fuzzy data analysis or

functional data analysis, and fuzzy weighted regression analysis are suitable for these data.

The application of discrimination to neural networks and support vector machines are well known. Many applications have been presented in the cognitive neuroscience to develop the mechanisms underlying object vision.

The innovative SDA techniques are based on the idea that the universe may be chosen as a space of functions which can include extremely large dimensions. Techniques which use fuzzy data, functional data, symbolic data, kernel method, and spherics are typical examples.

2 Soft Data Analysis Paradigms

Statistical data analysis assumes “systematic” uncertainty for observational data. The amount of systemization is represented by statistical distribution.

The concept of exploratory data analysis [1] where the emphasize is on the idea that real data structure is on the multiple aspects of the data and that a model (or a structure) is not assumed have been prospered. For exploratory data analysis, the concept of statistical science was an ideal solution.

However, the essence of observed data is not always based on “systematic” uncertainty. The necessity for analysis allowing for the most comprehensive uncertainty have been increased. As an ideal solution for analysis capturing unique features of data with the intention of discovering uncertainty and a set of robust and modern methods, SDA has been proposed.

3 Resources on SDA

We introduce several of the latest references and software support for SDA.

3.1 Literatures for SDA

- H. Bandemer and W. Nather, *Fuzzy Data Analysis*, Kluwer Academic Publishers, 1992.
- J.C. Bezdek, J. Keller, R. Krisnapuram, and N.R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, 1999.
- H.H. Bock and E. Diday eds., *Analysis of Symbolic Data*, Springer, 2000.
- C. Brunson, S. Fotheringham, and M. Charlton, Geographically Weighted Regression-Modelling Spatial Non-Stationarity, *Journal of the Royal Statistical Society*, Vol. 47, Part 3, pp. 431-443, 1998.
- N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
- A.J. Dobson, *An Introduction to Generalized Linear Models*, Chapman and Hall, 1990.
- A.S. Fotheringham, C. Brunson, and M. Charlton, *Geographically Weighted Regression*, Wiley, 2002.

- A. Ghosh and L.C. Jain eds., *Evolutionary Computation in Data Mining*, Springer, 2005.
- T.J. Hastie and R.J. Tibshirani, *Generalized Additive Models*, Chapman & Hall, 1990.
- F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*, Wiley, 1999.
- H. Ishibuchi, T. Nakashima, and M. Nii, *Classification and Modeling with Linguistic Information Granules*, Springer, 2005.
- L.C. Jain ed., *Soft Computing Techniques in Knowledge-Based Intelligent Engineering Systems*, Springer, 1997.
- H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, and M. Schader eds., *Data Analysis, Classification, and Related Methods*, Springer, 2000.
- S. Miyamoto, *Fuzzy Sets in Information Retrieval and Cluster Analysis*, Cluser Academic Publishers, 1990.
- N. Pal and L. Jain, *Advanced Techniques in Knowledge Discovery and Data Mining*, Springer, 2005.
- W. Pedrycz, *Knowledge-Based Clustering*, Wiley, 2005.
- J.O. Ramsay and B.W. Silverman, *Applied Functional Data Analysis*, Springer, 2002.
- J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer, 2nd ed., 2005.
- Y. Sato, An Analysis of Sociometric Data by MDS in Minkowski Space, *Statistical Theory and Data Analysis II*, North Holland, pp. 385-396, 1988.
- M. Sato, Y. Sato, and L.C. Jain, *Fuzzy Clustering Models and Applications*, Springer, 1997.
- M. Sato-Ilic and L.C. Jain, *Innovations in Fuzzy Clustering*, Springer, 2006.
- B. Scholkopf and A.J. Smola, *Learning with Kernels*, The MIT Press, 2002.
- H. Tanaka and J. Watada, Possibilistic Linear Systems and their Application to the Linear Regression Model, *Fuzzy Sets and Systems*, Vol. 27, pp. 275-289, 1988.
- H. Tanaka and P. Guo, Possibilistic Data Analysis for Operations Research, Springer, 1999.
- Y. Yabuuchi and J. Watada, Fuzzy Principal Component Analysis and Its Application, *Journal of Biomedical Fuzzy Systems Association*, Vol. 3, No. 1, pp. 83-92, 1997.

3.2 Software Supports

- The R Project for Statistical Computing, <http://www.r-project.org/>
- MATLAB Fuzzy Logic Toolbox, <http://www.mathworks.com/products/fuzzylogic/>
- GWR for Geographically Weighted Regression, A.S. Fotheringham, C. Brunsdon, and M. Charlton: *Geographically Weighted Regression*, Wiley, 2002.
- SODAS (Symbolic Official Data Analysis System) for Symbolic Data Analysis, H.H. Bock and E. Diday eds.: *Analysis of Symbolic Data*, Springer, 2000.

4 Papers Included in This Session

Five papers are to be presented during this session. The first by Mr. Inokuchi and Prof. Miyamoto on a nonparametric fisher kernel using fuzzy clustering. Using a distribution of the degree of belongingness of objects to the clusters obtained by fuzzy clustering, a nonparametric fisher kernel is applied to the classification. As a new application of fuzzy clustering for fisher kernel, they present quite novel research.

The second by Prof. Ishibuchi, Dr. Nojima, and Mr. Kuwajima on a new methodology to find fuzzy classification systems. They demonstrate a high capability to find fuzzy classification systems with large interpretability using multiobjective rule selection. In reasoning for real world data, this technique is very progressive.

The third by Prof. Klawonn and Prof. Höppner on a new classification technique to create clusters of approximately the same size. Not only the similarity of objects in each cluster, but also the similarity of the sizes of clusters is considered. The idea is quite novel. For the use of real data application, this technique has enormous potential to solve a wide range of problems.

The fourth by Prof. Sato on a new method for clustering of mixed data using spherical representation. Several different transformations are introduced through the use of a probabilistic distance in probabilistic space, a distance in Riemannian space, and a map from a plane to the surface of the sphere. As a consortium of the several logical techniques involving over the probabilistic theory, differential geometry, spherics, and visual inspection for the representation for directional data, this research is quite novel.

The last paper by Prof. Sato-Ilic and Mr. Kuwata on a self-organized (dis)-similarity considering two spaces. One is the (dis)-similarity of classification structures of a pair of objects and the other is (dis)-similarity of a pair of objects. This is an attempt to discover “how to combine two spaces” which are a space of classification structures obtained as a result of fuzzy clustering and a space of observed data. Our final goal is to investigate the relation between the two spaces through a metric space defined (dis)-similarity.

5 Conclusion

This session provides an overview of SDA. Since SDA is based on traditional data analysis, the overview covers a wide range of topics. This overview of SDA described here is only the tip of an iceberg. However, most data analysts are aware of the present limitations for analyzing collected data. SDA provides a solution for these limitations.

Reference

1. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley Publishing. (1977)

Tolerance Dependent on Timing/Amount of Antigen-Dose in an Asymmetric Idiotypic Network

Kouji Harada

Tokuyama College of Technology,
3538, Kume, Shunan, Yamaguchi 745-8585, Japan
k-harada@tokuyama.ac.jp

Abstract. Physiological experiments demonstrate establishment of immunological tolerance is controlled by dose-timing and dose-amount of an antigen. My previous study reproduced the tolerance dependent on dose-timing of an antigen in a two Bcell clones network model with an “asymmetric” Bcell-Bcell interaction. This study first clarifies its mechanism using a dynamical system technique: nullcline analysis. Next, this study proposes a three Bcell clones network model with the same style of Bcell-Bcell interaction, and shows the model simulation can reproduce the tolerance dependent on dose-amount of an antigen: high and low zone tolerance. The success of this reproduction is worthy of attention. Because theoretical studies based on the traditional “symmetric” immune network modeling scheme have not been able to reproduce it well. This study would teach us the renewed recognition of “asymmetric” immune network modeling scheme.

1 Introduction

Tolerance is “negative” memory, which means an immune response against an antigen is “suppressed” by previously exposing an immune system to the antigen. It is known that the establishment of tolerance depends on 1) dose-timing of or 2) dose-amount of an antigen. The antigen dose-timing dependent tolerance was shown by Hilliard et al [1]. They observed that the tolerance was induced in EAE rats by re-dosing an antigen just 14 days after the primary antigen-dose, but not by re-dosing the antigen at the other timing. On the other hand, the antigen dose-amount dependent tolerance occurs when an immune system is previously exposed to high (more than $10^6 \mu\text{g/ml}$) or low (less than $10^2 \mu\text{g/ml}$) amount of the primary antigen. Especially, tolerance with the high/low dose amount is called high/low zone tolerance, respectively[2]. The antigen dose-amount or -timing dependent tolerance have fascinated us, but the mechanisms are poorly understood yet. Since Jerne proposed idiootype network approach[3], theoretical immunologists started examining the mechanism of the tolerances from a network-control viewpoint. However there was left a contravertial problem if style of lymphocyte-lymphocyte interaction was symmetric or asymmetric. Hoffmann involved validity of the interaction style with stability of fixed points representing

for virgin state and immune memory state, and concluded symmetric interaction style must be best[4], whereas De Boer and Hogeweg clarified a “symmetric” two Bcell clones model could not reproduce tolerance[5]. Also Neumann and Weisbuch succeeded in displaying tolerance using a “symmetric” network model with many Bcell clones more than three clones, but they had to assume an unrealistic network topology for the reproduction[6]. In recent years, a B-T model with a “symmetric” Bcell-Tcell interaction succeeded in reproducing the antigen dose-amount dependent tolerance, but had the problem that validity of the model’s assumption, Bcell-Tcell interaction was not identified experimentally [7].

In response to the results, my previous studies have applied an “asymmetric” style as a type of Bcell-Bcell interaction. In fact under this assumption, my previous study succeeded in reproducing antigen dose-timing dependent tolerance[8]. However its mechanism was under exam.

This study first clarifies the details of the mechanism using a dynamical system technique. Next this study reports an asymmetric three Bcell clones model can reproduce antigen dose-amount dependent tolerance so-called low and high zone tolerance. Lastly this study would indicate the asymmetrical immune network modeling scheme to be more promising on the realization of immune functions than the symmetrical one.

2 Description of a Model

Jerne proposed the original idea that a Bcell can recognize not only an antigen but also a Bcell through its own recognition molecular, and he called its specific three dimensional shape *idiotype*[3]. The idiotype is characterized by a “recognizing” receptor site and a “being-recognized” ligand site, called *paratope* and *idiotope* respectively. An antigen has its own proper ligand site called *antigenic idiotope*. The strength of interaction between a paratope and an (antigenic) idiotope is determined according to degree of *specificity* between their respective three-dimensional shapes.

This study presents two types of immune network models. The one is a two Bcell clones model, another is a three Bcell clones model. These models are described as a time-differential nonlinear equation, and are approximately solved using the Runge-Kutta method of order 4 with a time-step size, $h = 10^{-2}$.

2.1 Two Bcell Clones Model

A two Bcell clones model is composed of an antigen, an antigen-specific Bcell and an anti-idiotypic Bcell. Fig.1 illustrates mutual asymmetric idiotypic interactions of between the three species following the idiotype network hypothesis[3]. From Fig.1 the population dynamicses of the antigen-specific Bcell, $X_1(t)$ and its anti-idiotypic Bcell, $X_2(t)$ are respectively described by,

$$\dot{X}_1(t) = (b_1 - b_2)X_1(t)X_2(t) - dX_1(t) + s + bX_1(t)A(t), \quad (1)$$

$$\dot{X}_2(t) = (b_2 - b_1)X_1(t)X_2(t) - dX_2(t) + s. \quad (2)$$

Here $X_1(t)$, $X_2(t)$ and $A(t)$ represent population size of the antigen-specific Bcell, the anti-idiotypic Bcell and the antigen at the time, t , respectively. Also the antigen population, $A(t)$ follows the next dynamics,

$$\dot{A}(t) = -bX_1(t)A(t). \tag{3}$$

The parameters b , b_1 and b_2 mean the strength of the idiotypic interaction between a paratope and an (antigenic) idiotope. The parameters d and s are dumping and source rate of a Bcell, respectively.

The values of the system parameters go as follow: $b_1 = 0.1$, $b_2 = 15.0$, $b = 0.1$, $s = 10^{-6}$ and $d = 0.1$.

2.2 Three Bcell Clones Model

A three Bcell clones model is just a model newly added an anti-anti-idiotypic Bcell to the two Bcell clones model presented in Fig.1. Fig.2 as well as Fig.1 illustrates mutual asymmetric interactions between four species following the idiotype network hypothesis[3].

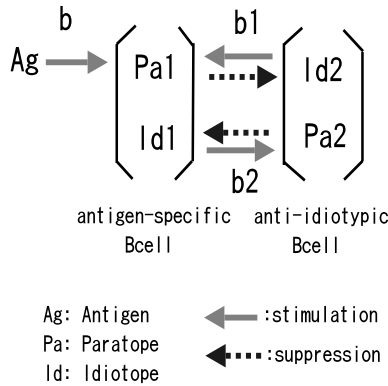


Fig. 1. Asymmetric two Bcell clones network model

From Fig.2 the population dynamicses of the antigen-specific Bcell, $X_1(t)$, the anti-idiotypic Bcell, $X_2(t)$ and the anti-anti-idiotypic Bcell, $X_3(t)$ are respectively described by,

$$\dot{X}_1(t) = (b_1 - b_2)X_1(t)X_2(t) - dX_1(t) + s + bX_1(t)A(t), \tag{4}$$

$$\dot{X}_2(t) = (b_2 - b_1)X_1(t)X_2(t) + (b_4 - b_3)X_3(t)X_2(t) - dX_2(t) + s. \tag{5}$$

$$\dot{X}_3(t) = (b_3 - b_4)X_2(t)X_3(t) - dX_3(t) + s. \tag{6}$$

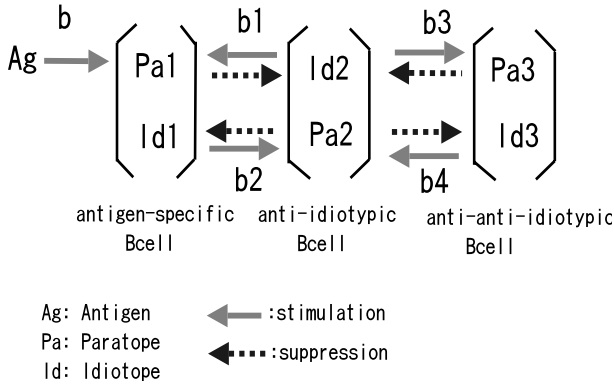


Fig. 2. Asymmetric three Bcell clones network model

Here $X_3(t)$ represent population size of the anti-anti-idiotypic Bcell at the time, t . Also the antigen population, $A(t)$ follows a dynamics,

$$\dot{A}(t) = -bX_1(t)A(t). \tag{7}$$

The values of the system parameters go as follow: $b_1 = 0.1, b_2 = 2.0, b_3 = 2.0, b_4 = 0.1, b = 0.1, s = 10^{-6}$ and $d = 0.1$.

3 Tolerance Dependent on Dose-Timing of an Antigen

A previous study presented tolerance dependent on dose-timing of an antigen developed in the two Bcell clones model[8]. However, its mechanism was under exam. One of purposes of this study is to clarify its mechanism using a dynamical system technique. First, this section starts from reviewing tolerance phenomenon dependent on dose-timing of an antigen. Fig.3 shows the time evolutions of the antigen-specific Bcell population, X_1 of when the antigen with 2.1 units dosed twice (dotted curve) and once as a control case (solid curves). Fig.3(a) displays a success example of tolerance because the secondary immune response is suppressed comparing to one of the control case. On the other hand, Fig.3(b) displays an un-success example of tolerance. The only difference between these two is the dose-timing of the secondary antigen (In fact in Fig.3(b), the secondary antigen is dosed at 821 steps delaying for 60 steps from the dose-timing of the secondary antigen in Fig.3(a)). Fig.3(a) and (b) show us that the dose-timing of the antigen is an important key for the establishment of this tolerance.

Next, I'll explain a mechanism of the antigen dose-timing dependent tolerance with the nullcline analysis of the model equations. A nullcline of a variable is given as a characteristic line which satisfies the condition of the time-derivative of the variable being zero. Thus X_1 -nullcline $X_1(A)$ as a function of the variable A , is given as,

$$X_1(A) = \frac{\gamma - \sqrt{\gamma^2 + 4(b_2 - b_1)(bA - d)sd}}{2(bA - d)(b_1 - b_2)}, \tag{8}$$

where

$$\gamma = 2(b_2 - b_1)s - (bA - d)d. \tag{9}$$

Also X_2 -nullcline $X_2(A)$ is solved similarly and given as,

$$X_2(A) = \frac{(bA - d)X_1(A) + 2s}{d}. \tag{10}$$

Also, a nullcline of the variable A is as follows,

$$A(X_1) = 0. \tag{11}$$

Fig.4(a) and (b) describes $X_1(A)$, $X_2(A)$ and $A(X_1)$ together with the orbits $X_2(t)$ at the success and un-success of the tolerance shown in Fig.3(a) and (b). Fig.4(a) indicates the tolerance develops if the secondary antigen is dosed at the timing of when the anti-idiotypic Bcell population, $X_2(t)$ is a maximum value. In fact the timing is crucial. Because when the X_2 is around a maximum value, the second dose of the antigen enable the orbit $X_2(t)$ shift near the $X_2(A)$ -nullcline,

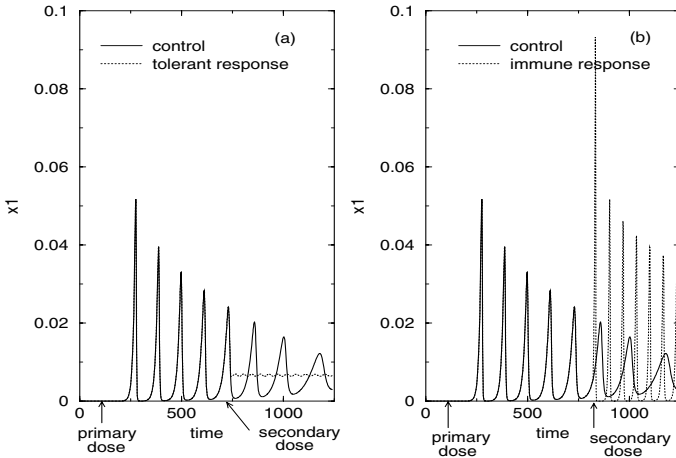


Fig. 3. Fig.(a) shows the immune response (dotted line) with the secondary antigen dose at the timing, 741 starts being suppressed comparing the control case (solid line) without doing a secondary antigen dose: a success of tolerance. Fig.(b) represents the immune response (dotted line) with the secondary antigen dose at the timing, 821 different from the one in Fig.(a) starts being boosted comparing with the control case(solid line): an un-success of tolerance.

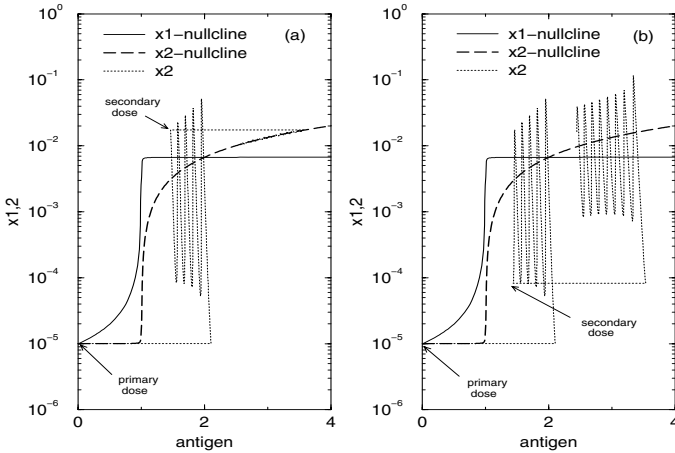


Fig. 4. Fig.(a) and (b) corresponds to Fig.3(a) and (b), respectively. Fig.(a) shows the secondary antigen dose at the timing of a maximum value of X_2 enables the X_2 -orbit shift to the X_2 -nullcline, and stops the X_2 -orbit oscillating. Fig.(b) shows the secondary antigen dose at the timing of a minimum value of X_2 locates the X_2 -orbit far from the X_2 -nullcline, so that activates the X_2 -orbit oscillating.

and it leads to dump down the primary immune response oscillating with a large amplitude. That is, the tolerance becomes established in the system. On the other hand, Fig.4(b) shows if the secondary antigen is dosed when the $X_2(t)$ is a minimum value, it locates the orbit far from the $X_2(A)$ -nullcline (in other words, X_2 velocity vector field becomes strong), so that the secondary immune response becomes boosted. From these considerations it is concluded that the optimal timing which leads the system to tolerance is when the anti-idiotypic Bcell population is around a maximum value. This result would support the experiment that anti-idiotypic B-cell clones involved in the suppression of the immune response against an antigen[9].

4 Tolerance Dependent on Dose-Amount of an Antigen

This section reports antigen dose-amount dependent tolerance in the three Bcell clones network model presented in Sect. 2.2. The reason why I newly introduce the three B-cell clones model is that the two Bcell clones model shows difficulty on reproducing dose-amount dependent tolerance because of the Bcell population oscillating. Fig.5(a) shows time serieses of the antigen-specific Bcell population, $X_1(t)$ when the dose amount of the primary antigen increases from 0.08 to 30.0; the dose-amount of the secondary antigen is always constant:1.1 units. The most important point we should focus on is the secondary immune response is strengthened most when the dose-amount of the primary antigen is 0.8 units, and is lessened below and above the dose-amount. This means the strength of the secondary immune response against the dose-amount of the primary

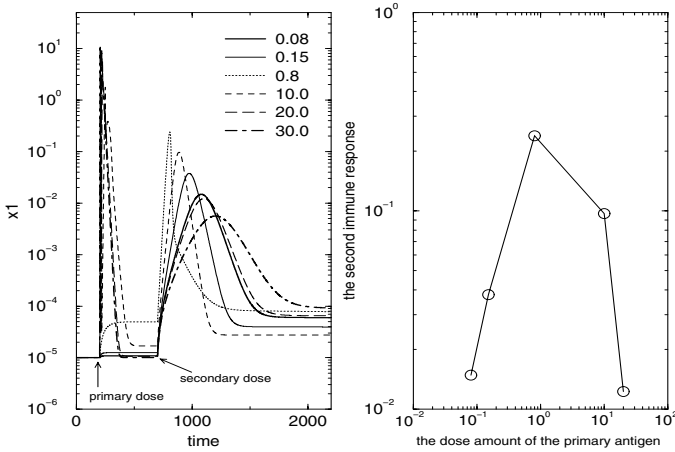


Fig. 5. Fig.(a) shows time serieses of the antigen-specific Bcell population, $X_1(t)$ when the dose amount of the primary antigen increases from 0.08 to 30.0; the dose-amount of the secondary antigen is always constant:1.1 units. The secondary immune response is strengthened most when the dose amount of the primary antigen is 0.8 units, and is lessened below and above the dose amount. Fig.(b) shows the strength of the secondary immune response against the dose-amount of the primary antigen.

antigen forms a bell-shaped function in Fig.5(b), therefore the function can reproduce high and low zone tolerance. However its establishment mechanism is under exam.

I think this finding is significant from a historical viewpoint on theoretical immunology but have to note that it has the problem that the primary immune response is strong comparing with the secondary immune response.

5 Conclusions

This study has discussed the antigen dose-timing and the antigen dose-amount dependent tolerance. My previous study succeeded in displaying the antigen dose-timing dependent tolerance in the two Bcell clones model but its mechanism were unknown. This study clarified its mechanism, and we understood an appropriate dose-timing for the establishment of the tolerance was when the anti-idiotypic Bcell population was at a maximum value. This fact is consistent with a physiological experimental result that the administration of an anti-idiotypic Bcell induces tolerance[9]. Also the optimal secondary antigen dose-timing (14 days after the primary antigen-dose) for the tolerance induction in Hilliard et al's experiment[1] might be determined by the population-level of anti-idiotypic Bcell.

It is noteworthy that this study succeeded in displaying the dose-amount dependent tolerance in the "asymmetrical" three Bcell clones model, and obtained the bell-shaped dose-response curve indicating high (low) zone tolerance. This is

an important finding for theoretical immunologists. Because it has been hard to demonstrate dose-amount dependent tolerance in “symmetrical” immune network modeling scheme. I think this result would make us reaffirm the “asymmetrical” immune network modeling scheme has the potentiality on realization of immune functions.

References

1. Hilliard, B., Ventura, E. S., Rostami, A.: Effect of timing of intravenous administration of myelin basic protein on the induction of tolerance in experimental allergic encephalomyelitis. *Multiple Sclerosis*. 5 (1999) 2–9.
2. Elgert, K.D.: *Immunology - Understanding the Immune System -*. Wiley-Liss, Inc, New York (1996) 261
3. Jerne, N.K.: Towards a network theory of the immune system. *Ann. Immunol.* 125C (1974) 373–389
4. Hoffmann, G.W.: Regulation of Immune Response Dynamics. DeLisi, C. and Hienaux, J.R.J(Eds), 1 (1982) 137–162
5. De Boer, R.J., Hogeweg, P.: Memory but no suppression in low-dimensional symmetric idiotypic network. *Bull. Math. Biol.* 51 (1989) 223–246
6. Neumann, A.U., Weisbuch, G.: Window automata analysis of population dynamics in the immune system. *Bull. Math. Biol.* 54 (1992) 21–44
7. Carneiro, J., Couthinho, A., Faro, J., Stewart, J.: A model of the immune network with B-T cell co-operation I. *J. Theor. Biol.* 182 (1996) 513–530
8. Harada, K.: Emergence of Immune Memory and Tolerance in an Asymmetric Idiotypic Network. *Proc. of Ninth International Conference on Knowledge-based Intelligent Information Engineering Systems*, Vol. 2. Springer-Verlag, Berlin (2005) 61–71
9. David, A.H., Wang, A-L., Pawlak, L., Nisonoff, A.: Suppression of Idiotypic Specificities in Adult Mice by Administration of Antiidiotypic Antibody. *J. Exp. Med.* 135 (1972) 1293–1300

Towards an Immunity-Based Anomaly Detection System for Network Traffic

Takeshi Okamoto¹ and Yoshiteru Ishida²

¹ Department of Network Engineering, Kanagawa Institute of Technology,
1030, Shimo-ogino, Atsugi, Kanagawa, 243-0292 Japan
`take4@nw.kanagawa-it.ac.jp`

² Department of Knowledge-Based Information Engineering,
Toyoashi University of Technology,
1-1, Tempaku, Toyohashi, Aichi, 441-8580 Japan
`ishida@tutkie.tut.ac.jp`

Abstract. We have applied our previous immunity-based system to anomaly detection for network traffic, and confirmed that our system outperformed the single-profile method. For internal masquerader detection, the missed alarm rate was 11.21% with no false alarms. For worm detection, four random-scanning worms and the simulated *metaserver worm* were detected with no missed alarms and no false alarms, while a simulated *passive worm* was detected with a missed alarm rate of 80.57%.

1 Introduction

Anti-virus systems protect computers and networks from malicious programs, such as computer viruses and worms, by discriminating between malicious programs and harmless programs, and by removing only the former. Therefore, anti-virus systems can be considered as the computer's immune system.

An innovative method, a “*virus throttle*,” was proposed by Williamson in 2002 [1]. The *virus throttle* slows and halts high-speed worms without affecting normal network traffic. In our previous study [2], we proposed a “*worm filter*” for preventing both slow- and high-speed worms from spreading. The *worm filter* limits the number of unacknowledged requests, rather than the rate of connections to new computers. However, the *worm filter* cannot stop *metaserver worms*, *topological worms*, or *passive worms* [3]. Not only do these worms attempt to connect to active servers that return a reply, but also their network traffic is similar to that of an actual user. In other studies [4,5], we proposed the immunity-based anomaly detection system for a UNIX command sequence that can discriminate between a legitimate user and a masquerader who abuses someone else's account. This system would be expected to detect the above worms, because the difference between a user and a worm is greater than those between different users.

We have applied this system to anomaly detection for network traffic to detect worms. The algorithm of the immunity-based anomaly detection system is described in detail in section 2. In section 3, we describe our experimental data and evaluation model. Section 4 presents a comparison of the performance evaluation between our immunity-based method and the conventional method. Sections

5 and 6 present performance evaluations against simulated masqueraders, real worms, and some simulated worms.

2 Immunity-Based Anomaly Detection for Network Traffic

At the heart of the immune system is the ability to distinguish between “self” (the body’s own molecules, cells, and tissues) and “nonself” (foreign substances, such as viruses or bacteria). We define an operation sequence of legitimate users on their own account as “self,” and all other operation sequences produced by masqueraders, worms, *etc.*, as “nonself.”

The immunity-based anomaly detection system (IADS) uses multiple user-specific agents. Each agent has a unique profile that is expressed by a parameter of the hidden Markov model (HMM) $\lambda = [\pi, A, B]$, as our previous study [6] indicated that it performs well. The parameters of the HMM, using the Baum-Welch algorithm, is estimated from sequences of each legitimate user. The agent computes a likelihood $P(O|\lambda)$ of the sequence O with the profile λ . The agent computes a high score (*i.e.*, a high likelihood) for only the sequences of the user corresponding to the agent. In this way, the agent is specialized so as to recognize the user.

In every operation, all agents compute their own score for a new operation sequence. The agent associated with that account is activated and compares it with the scores of all other agents. If the user of the account is the owner of the account, the score will be relatively high, but not necessarily the highest score compared with the scores of the other agents. Thus, we set a threshold value (Th), which is a percentage of the difference between the minimum (Min) and maximum scores (Max). If the activated agent computes a score higher than the threshold, obtained by the equation $Min + (Max - Min) \times Th$, the activated agent classifies the operation sequence as normal (*i.e.*, “self”). Otherwise, the agent classifies the operation sequence as abnormal (*i.e.*, “nonself”), and it raises an alarm. Furthermore, provided that all the scores are equal to the minimum value of all the computable $P(O|\lambda)$, the sequence is regarded as abnormal. Conversely, if all the scores are equal to the maximum value of all the computable $P(O|\lambda)$, the operation sequence is regarded as normal. Examples of discrimination between “self” and “nonself” are shown in Fig. 1.

In this study, the IADS monitored network traffic, more specifically outgoing TCP SYN packets. The IADS makes a sequence of destination IP addresses of these packets. However, the IP address space is too large to allow construction of a profile. Hence, pre-processing of the sequence is done to scale down the IP address space. The number of different IP addresses to which a user transmits a packet more than once is very small. Hence, the IADS assigns a unique number v to each IP address to which the user transmits a packet more than once, where v begins with 0. The assignment table is created at profile construction, and all the sequences are replaced according to this table. If the IP address does not exist in this table, the IP address is assigned a unique number $v_{max} + 1$, where $v_{max} + 1$

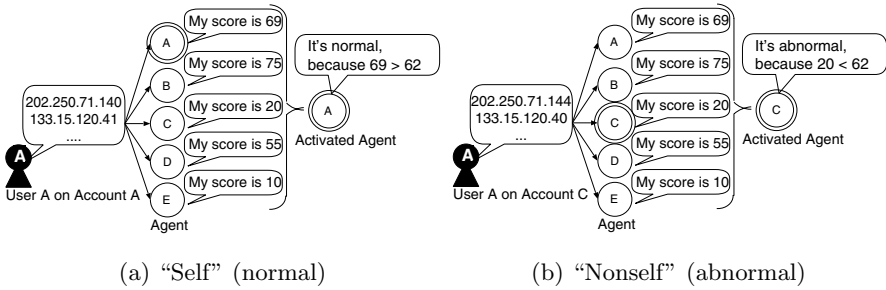


Fig. 1. Discrimination between “self” (a) and “nonself” (b). If we set the threshold value to 80%, and the different agents compute 10, 20, 55, 69, and 75, the effective threshold value is calculated to be 62 ($= 10 + (75 - 10) \times 0.80$). In the case of (a), if user A browses a website on his/her own account, agent A that is specialized to recognize user A is activated, and decides that the operation sequence is normal. In the case of (b), if user A browses a website on the account of user C, the agent C that is specialized to recognize user C is activated, and it decides that the operation sequence is abnormal.

is equal to the number of all IP addresses included in the table. Empirically, we assign a value of 100 ± 20 to v_{max} .

3 Experimental Data and Evaluation Model

To evaluate detection performance, we captured network traffic from 12 users for about one month. We focused on web traffic, as this accounts for the majority of network traffic. Hence, we extracted only outgoing TCP SYN packets with destination port 80. The web traffic of each user contains more than 3,000 requests. The first 500 requests for each user are used as training data to allow construction of a profile. The next 1,000 requests are test data to evaluate the detection performance. The test for the sequence is performed at every request. Each sequence length is set to 400, so that the total number of tests for each user is 601.

For evaluation of user traffic, we simulate anomalous behavior by testing one user’s request sequence against another user’s profile, as our data do not include anomalous behavior. This simulation corresponds to the evaluation of masquerader detection.

Anomaly detection is important to reduce false alarms, because too many false alarms can cause the “cry-wolf syndrome.” Hence, we evaluate the missed alarm rate with no false alarms.

4 Immunity-Based Method vs. Single-Profile Method

We compared our immunity-based method, which uses multiple profiles, with the conventional single-profile method using only one profile (*e.g.*, [6,7]). Each evaluation was performed for 12 users as described in section 3.

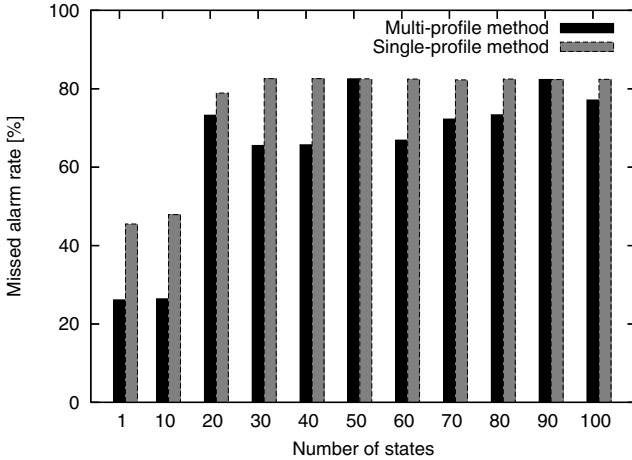


Fig. 2. Missed alarm rate with no false alarms for the single- and multi-profile methods as a function of the number of states in the HMM

Figure 2 shows the missed alarm rate with no false alarms for the single- and multi-profile methods as a function of the number of hidden states in the HMM, where the hidden states are assumed to be working states, such as searching, e-learning, blogging, *etc.* For all numbers of states, the immunity-based method outperformed the single-profile method. As shown in Fig. 2, the HMM for which the number of states = 1 showed the best performance. This HMM depends on only the request frequency of different websites. In addition, we confirmed that this HMM largely surpassed a method based on Markov chains. That is, the frequency property rather than the ordering property dominates the characteristics of request sequences. Therefore, we set the number of states to one in all performance evaluations after this section.

In our previous study [6], we confirmed that the missed alarm rate was inversely proportional to the number of states. In contrast, the missed alarm rate was proportional in Fig. 2. The reasons for this discrepancy are currently under investigation.

5 Performance Evaluation of User Traffic

Our approach may be less well-suited for detecting external masqueraders than for detecting internal masqueraders because our agents have no profiles for external masqueraders. Hence, we evaluated the performance of our detection scheme for internal and external masqueraders separately. We should evaluate all combinations of internal users among the 12 users because the performance of anomaly detection would depend on the combinations. We conducted 1,000 combinations chosen at random from among all possible combinations. For each combination, the first 6 users were assumed to be external masqueraders, while the remaining users were assumed to be internal users.

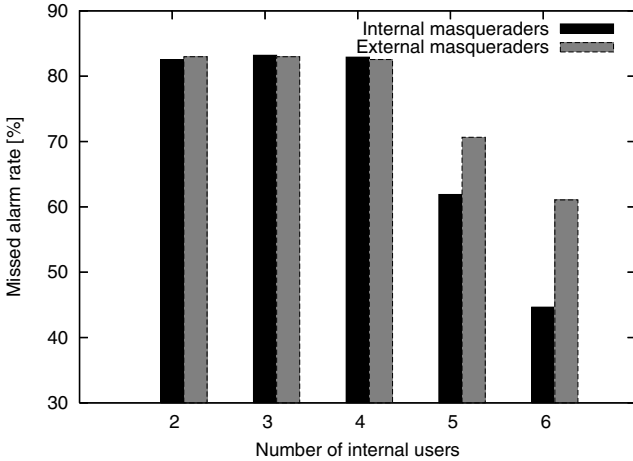


Fig. 3. Missed alarm rate with no false alarms for internal and external masqueraders. Each missed alarm rate is an average over 1,000 combinations.

Figure 3 shows the missed alarm rate with no false alarms for internal and external masqueraders, with varying numbers of internal users from two to six. As expected, the large difference in the missed alarm rates between internal and external masquerader detection was confirmed in the results with five and six internal users. In addition, both of the missed alarm rates were very high for small numbers of internal users. However, as the missed alarm rate seems to decrease with increasing number of internal users, the addition of internal users may reduce the missed alarm rate. In addition, as our previous study confirmed that the addition of diverse agents decreased the missed alarm rate [4,5], such diverse agents may reduce the number of missed alarms.

All the above evaluations were performed with only one threshold for all users to make evaluation easier. As fluctuations between agents' scores for each sequence were very large, we set a different threshold with no false alarms for each user. The results are shown in Fig. 4, in which there are no external users. As expected, the missed alarm rates decreased in the case with different thresholds. The average missed alarm rate was 11.21%. It is noteworthy that there were no missed alarms for the following users: E, F, I, K, and L.

6 Performance Evaluation of Worm Traffic

We evaluated worms, setting the threshold of the IADS to a different value with no false alarms for each user.

We have evaluated four random-scanning worms in the wild: `CodeRedv2`, `CodeRedII`, `Slammer`, and `Blaster`. Although `Slammer` and `Blaster` do not send a packet to TCP port 80, we assumed that they sent to this port. As a result, there were no missed alarms and no false alarms on any of the accounts for all the

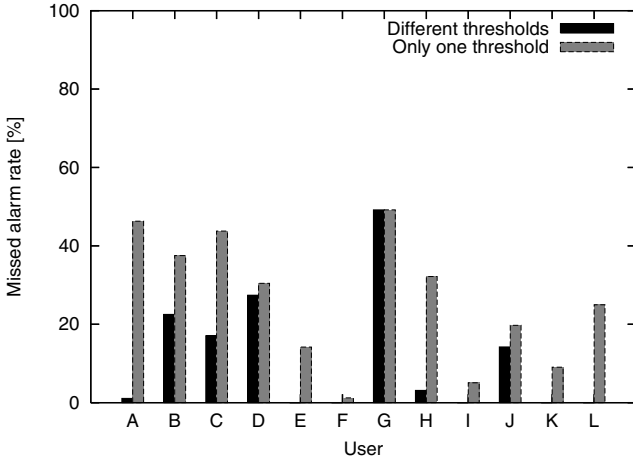


Fig. 4. Missed alarm rate with no false alarms for different thresholds and only one threshold. For the former, the thresholds were: 99.9%, 37.6%, 68.0%, 29.6%, 83.2%, 99.9%, 24.6%, 99.9%, 99.9%, 38.2%, 95.2%, and 89.6%. For the latter, the threshold was 24.6%, which coincided with the threshold of user G.

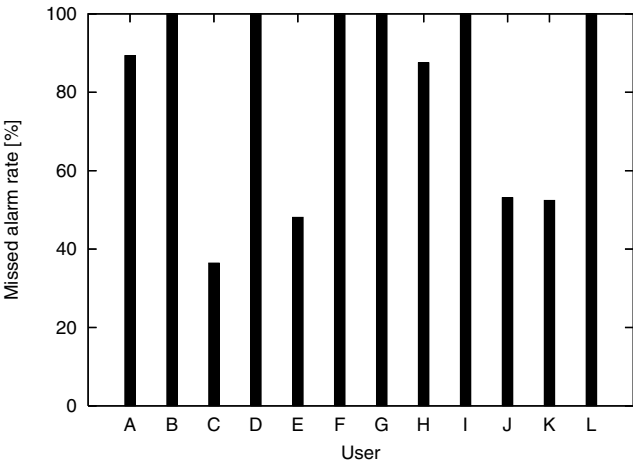


Fig. 5. Missed alarm rate with no false alarms for a simulated *passive worm*. The average missed alarm rate was 80.57%. User C showed the best missed alarm rate of 36.44%.

worms examined, because none of the users had ever connected to IP addresses generated randomly by these worms.

Three worms, a *metaserver worm*, a *topological worm*, and a *passive worm* [3], escaped our previous method (*i.e.*, worm filter) [2] by propagating to only the active servers that return a reply.

The IADS would be expected to detect the *metaserver worm* and the *topological worm*, because both of these worms are difficult to connect to IP addresses

to which the user has ever connected. We simulated a *metaserver worm*, such as **Santy** worm, which attempts to propagate to IP addresses in the search results provided by GoogleTM (www.google.com). The traffic of the simulated worm was evaluated and the results indicated that there were no false alarms and no missed alarms for all accounts.

The *passive worm*, which either waits for target computers to visit or follows user's requests into target computers, is more difficult for an anomaly detection system to detect, because its behavior is similar to that of the user. We simulated the *passive worm*, which propagates to servers to which the user has connected, immediately after the connection. For example, if a user browses websites: $A \rightarrow B \rightarrow B \rightarrow C$ in this order, the infected computer attempts to connect to $A \rightarrow A \rightarrow B \rightarrow B \rightarrow B \rightarrow B \rightarrow C \rightarrow C$. The traffic of the simulated worm was evaluated and the results are shown in Fig. 5. The average missed alarm rate was 80.37%. Although this rate was not good, it is notable that the worm was detected on six accounts. Investigation of request sequences on the six accounts indicated that the frequency of $v_{max} + 1$ (*i.e.*, the frequency of browsing new websites not included in the assignment table of websites) is relatively high. Hence, the detection of this worm would depend on the frequency of $v_{max} + 1$.

7 Conclusions

We applied our previous immunity-based system to anomaly detection for network traffic. The results of this study confirmed that our system outperformed the single-profile method. For internal masquerader detection, the missed alarm rate was 11.21% with no false alarms. For worm detection, four random-scanning worms and the simulated *metaserver worm* were detected with no missed alarms and no false alarms, while the simulated *passive worm* was detected with a missed alarm rate of 80.57%. Further studies will require the generation of diverse agents to reduce the missed alarm rate. Inspired by the mechanism of adaptation in the immune system, methods should be developed to update the user's profile so that each agent can adapt to recent behavior of the user.

Acknowledgements

This work was supported in part by Grants-in-Aid for Scientific Research (B) 16300067, 2004. This work was partly supported by the 21st Century COE Program "Intelligent Human Sensing" from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

1. Williamson, M.M.: Throttling viruses: restricting propagation to defeat malicious mobile code. In: ACSAC Security Conference 2002. (2002) 61–68
2. Okamoto, T.: A worm filter based on the number of unacknowledged requests. In: KES 2005, LNAI 3682 (2005) 93–99

3. Weaver, N., Paxson, V., Staniford, S., Cunningham, R.: A taxonomy of computer worms. In: The 2003 ACM Workshop on Rapid Malcode, ACM Press (2003) 11–18
4. Okamoto, T., Watanabe, T., Ishida, Y.: Towards an immunity-based system for detecting masqueraders. In: KES 2003, LNAI 2774 (2003) 488–495
5. Okamoto, T., Watanabe, T., Ishida, Y.: Mechanism for generating immunity-based agents that detect masqueraders. In: KES 2004, LNAI 3214 (2004) 534–540
6. Okamoto, T., Watanabe, Y., Ishida, Y.: Test statistics for a masquerader detection system – a comparison between hidden markov model and other probabilistic models. Transactions of the ISCIE **16**(2) (2003) 61–69
7. Schonlau, M., DuMouchel, W., Ju, W., Karr, A., Theus, M., Vardi, Y.: Computer intrusion: Detecting masquerades. Statistical Science **16**(1) (2001) 58–74

Migration Strategies of Immunity-Based Diagnostic Nodes for Wireless Sensor Network

Yuji Watanabe¹ and Yoshiteru Ishida²

¹ Graduate School of Natural Sciences, Nagoya City University,
Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya, Aichi 467-8501, Japan
yuji@nsc.nagoya-cu.ac.jp

² Dept. of Knowledge-based Information Eng., Toyohashi University of Technology,
Tempaku, Toyohashi, Aichi 441-8580, Japan
ishida@tutkie.tut.ac.jp

Abstract. In our previous studies, the immunity-based diagnostic model has been used by stationary agents in linked networks or by mobile agents on wired computer networks. We have not yet analyzed the performance of the diagnosis in wireless network where agents can move freely. In this paper, the diagnosis is applied to static and mobile sensor nodes in a 2-dimensional lattice space for wireless sensor network. Some simulation results show the strategy of going straight in the different direction can have the best detection rate. In addition, when the fraction of mobile nodes is changed, the transitions of the detection rate for the migration strategies are different.

1 Introduction

In recent year, sensor network, ad-hoc network, and ubiquitous computer have attracted much attention. Some keywords such as *wireless*, *mobile*, *distributed*, and *cooperative* in these fields are listed. These characteristics are endowed in the biological immune system. We have pursued the autonomous distributed diagnosis models inspired by the informational features of the biological immune system. The *immunity-based diagnostic model* based on the concept of the *idiotypic network hypothesis* [1] has been proposed in [2]. The diagnostic model is performed by mutual tests among agents and dynamic propagation of active states. In our previous studies, the diagnosis has been employed by stationary agents in linked networks [2] or by mobile agents on wired computer networks [3,4]. We have not yet analyzed the performance of the diagnosis in wireless network where agents can move freely.

In this paper, the immunity-based diagnostic model is applied to sensor nodes in a 2-dimensional lattice space for wireless sensor network. Each node is either static or mobile. Note that the term ‘node’, which is usually used in graph theory and sensor network community, is considered the same as the term ‘agent’ in our previous studies. Preliminary simulations are carried on both for wired networks and for wireless networks. When the immunity-based diagnosis is performed on random graph as a wired network, the capability of detecting abnormal nodes

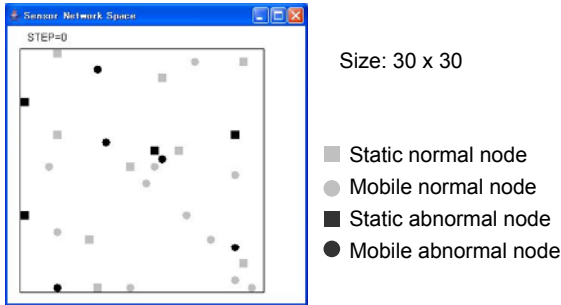


Fig. 1. 2-dimensional lattice space for wireless sensor network. There are four kinds of sensor nodes in the space.

relies on the number of edges. In wireless network where all the nodes are stationary, the detection rate depends on some environmental parameters: space size, visual distance, and the number of nodes. Next, we address some migration strategies and the fraction of mobile nodes. Some simulation results show the strategy of going straight in the different direction can have the best detection rate. Additionally, when the fraction of mobile nodes is changed, the different transitions of the detection rate for the migration strategies are observed.

2 Simulation Environment

To make it easy to analyze the performance of diagnosis, we use a simple environment for wireless sensor network. The environment is realized by a 2-dimensional lattice space with a periodic boundary condition. The size of space $S \times S$ is variable in preliminary simulations, and then is fixed in the next simulations of migration strategies.

The space consists of four kinds of sensor nodes as shown in Fig. 1. The total number of sensor nodes is defined by N . Each node is either *static* or *mobile*. Mobile nodes can move 1 distance per time step in a direction. The state of sensor node is simply represented as either *normal* or *abnormal*. The node can interact other nodes within a visual distance D . Each node can sense the state, not by itself, but only by comparisons with the others. The goal of diagnosis is to detect all the abnormal nodes by interactions among nodes.

3 Immunity-Based Diagnostic Model

The immunity-based distributed diagnostic model proposed by Ishida [2] is inspired by the concept of the *idiotypic network theory* [1]. The diagnostic model is performed by mutual tests among nodes and dynamic propagation of active states. In this study, each node has the capability of testing other nodes within the visual distance D , and being tested by the adjacent others as well. A state

variable R_i indicating the *credibility of node* is assigned to each node and calculated as follows:

$$\frac{dr_i(t)}{dt} = \sum_j T_{ji}R_j + \sum_j T_{ij}R_j - \frac{1}{2} \sum_{j \in \{k: T_{ik} \neq 0\}} (T_{ij} + 1), \tag{1}$$

$$R_i(t) = \frac{1}{1 + \exp(-r_i(t))}, \tag{2}$$

where the credibility $R_i \in [0, 1]$ is a normalization of $r_i \in (-\infty, \infty)$ using a sigmoid function. In equation (1), T_{ji} denotes binary test outcome from testing node j to tested node i as follows:

$$T_{ji} = \begin{cases} 1 & \text{if the states of nodes } i \text{ and } j \text{ are the same} \\ -1 & \text{if the states of nodes } i \text{ and } j \text{ are different} \\ 0 & \text{if node } j \text{ cannot test node } i \text{ out of view} \end{cases} . \tag{3}$$

The initial value of credibility $R_i(0)$ in the immunity-based diagnosis is 1.0. It means the diagnosis regards all the nodes as normal. The aim of the diagnosis is to decrease the credibility of all the abnormal nodes. The threshold of credibility between normal node and abnormal one is set to be 0.5.

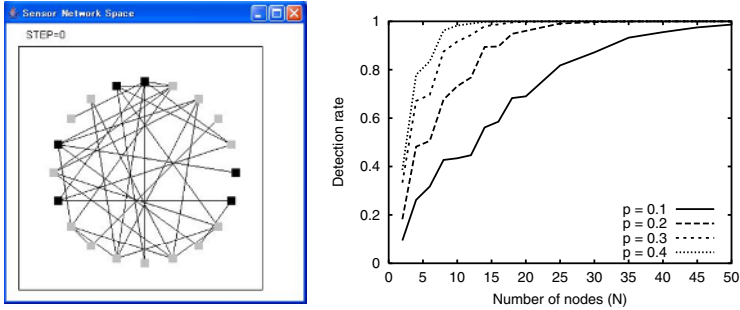
4 Preliminary Simulations

4.1 Simulation Conditions

We carry on some preliminary simulations both for wired networks and for wireless networks. We firstly describe conditions for the simulations. The previous studies [3,4] say that the immunity-based diagnosis can mostly detect abnormal nodes up to $0.5N$. In this study, the number of abnormal nodes is set to be $0.3N$. For a performance measurement, we record a detection rate, that is, the fraction of abnormal nodes detected by the diagnosis model. Since all the nodes are located randomly at the start of each simulation, the detection rate is averaged over 1000 trials. Furthermore, the credibility of the immunity-based diagnosis can converge almost by 20 time steps, so that we inspect the detection rate after 20 steps.

4.2 Wired Network

The immunity-based diagnosis is performed on random graph as a wired network. Although other network models such as *small-world* and *scale free* [5] have been already applied, the results will be described in another paper for the lack of space. The random graph model includes N nodes and each possible edge independently with probability p . When $N = 20$ and $p = 0.2$, an example of random graph is illustrated in Fig. 2 (a). For the summation operators in equation (1), the accurate calculation of the credibility relies on the number of edges for each node, which is averagely $p(N - 1)$ on random graph. Figure 2 (b) depicts



(a) random graph when $N = 20$ and $p = 0.2$. (b) average detection rate vs. the number of nodes N on random graphs with various p .

Fig. 2. An example of random graph and simulation result for random graphs

the average detection rate after 20 time steps over 1000 trials vs. the number of nodes N on random graphs with various p . From the result, as expected, the detection rate of random graph depends on both N and p . When the detection rate becomes over 0.99, $N = 51, 25, 17, 12$ for $p = 0.1, 0.2, 0.3, 0.4$, respectively, and then the average number of edges for each node $p(N - 1)$ is 5, 4.8, 4.8, 4.4, almost the same.

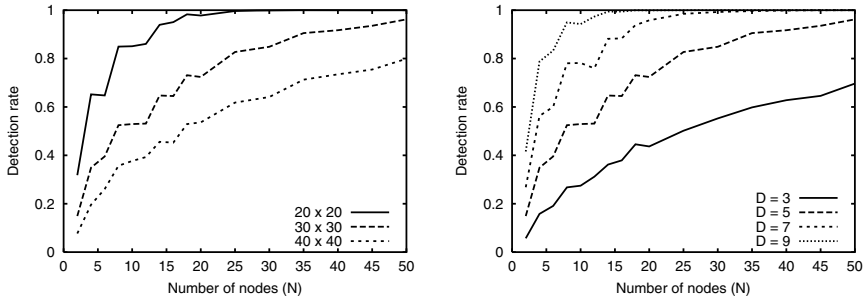
4.3 Wireless Network

In next simulations for wireless network where all the nodes are stationary, some environmental parameters are explored. It is easy to predict that the detection rate depends on the frequency of interactions between nodes, namely the density of nodes within the visual distance. In the preliminary simulations, the size of space $S \times S$, the visual distance D , and the number of nodes N are varied. Figure 3 illustrates the average detection rate vs. the number of nodes N changing S and D . From the results, as expected, the detection rate can increase when N and D increase, but S decreases. When the detection rate becomes over 0.99, the average number of adjacent nodes is 9.26, 7.25, 5.21 for $D = 5, 7, 9$, respectively. The numbers of necessary interactions in wireless network is not only scattered but also big compared with random graph. The reason is under study.

5 Simulations of Migration Strategies

5.1 Migration Strategies

When all the nodes cannot move, adjacent nodes are always identical. If mobile nodes are installed, each node would have a lot of opportunities of interactions. However, generally speaking, the mobile nodes require some additional mechanisms with respect to both hardware and software. The hardware items are not only moving mechanisms such as wheels and legs but also battery or energy for movement. Therefore, the number of mobile units would be as small as possible.



(a) space size S is varied while $D = 5$. (b) visual distance D is varied while $S = 30$.

Fig. 3. Average detection rate after 20 time steps over 1000 trials vs. the number of nodes N changing S and D when all the nodes are stationary

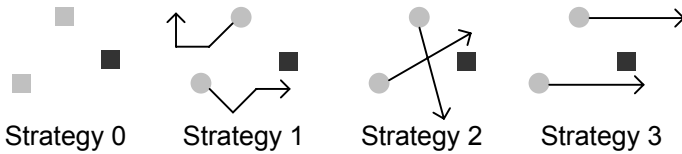


Fig. 4. Migration strategies indicated by the arrow. In strategy 0, all the nodes are stationary.

In addition, software mechanisms for migration and collision avoidance need to be implemented. Mobile node can also have more complicated capabilities, for example, learning and cooperation. The following simple migration strategies including static case as illustrated in Fig. 4 are firstly applied:

- Strategy 0:** All the nodes are stationary.
- Strategy 1:** Each mobile node can walk randomly.
- Strategy 2:** Each mobile node can go straight in a random direction.
- Strategy 3:** Each mobile node can go straight in the same direction.

5.2 Simulation Results

Based on the results of the preliminary simulations as shown in Fig. 3, two parameters S and D are fixed as $S = 30$ and $D = 5$ in the simulations of migration strategies. The other conditions are the same as the preliminary simulations.

Figure 5 depicts the average detection rate vs. the number of nodes N for each migration strategy when half or all the nodes are mobile. The results demonstrate that the detection rate of strategy 1 of randomly walking nodes is similar to strategy 0 of all static nodes, while the performance of strategy 2 can be improved most. In addition, the detection rate of strategy 3 is better than strategy 0 in Fig. 5 (a), but the same in Fig. 5 (b).

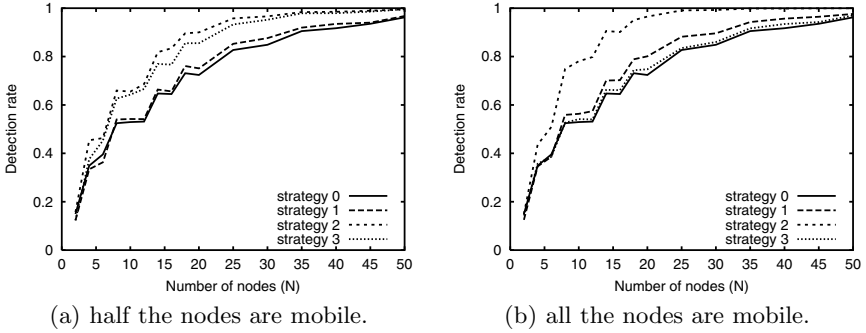


Fig. 5. Average detection rate after 20 time steps over 1000 trials vs. the number of nodes N for each migration strategy

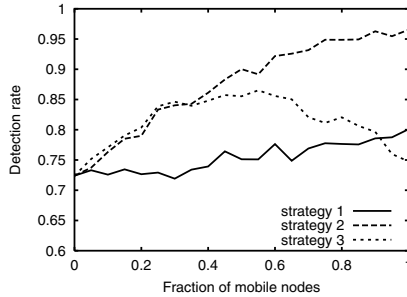


Fig. 6. Average detection rate vs. the fraction of mobile nodes for each migration strategy when $N = 20$

Relevant simulations are carried out changing the fraction of mobile nodes when N is set to be 20. Figure 6 illustrates the average detection rate after 20 time steps over 1000 trials vs. the fraction of mobile nodes for each migration strategy. From the result, the following points are observed:

- The detection rate of strategy 1 slightly increases as the fraction of mobile nodes become higher.
- The performance of strategy 2 grows, but keeps constant over the fraction of mobile nodes 0.8.
- In strategy 3, there is a peak of the detection rate near the fraction of mobile nodes 0.5.

The reasons for the first and third points can be easily explained. The detection rate relies on the number of testing and tested adjacent nodes during 20 steps. In strategy 1, randomly walking nodes stay almost near the initial location in 20 steps, so that adjacent nodes are not so varied. Since mobile nodes of strategy 3 move at the same speed in the same direction, the interaction between mobile nodes is always constant, and only the interaction between a mobile node and a static node is changed. Therefore, when all the nodes are stationary or

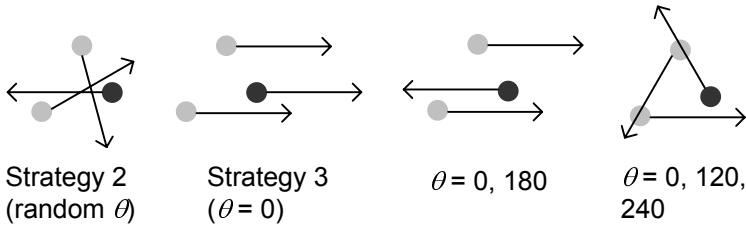


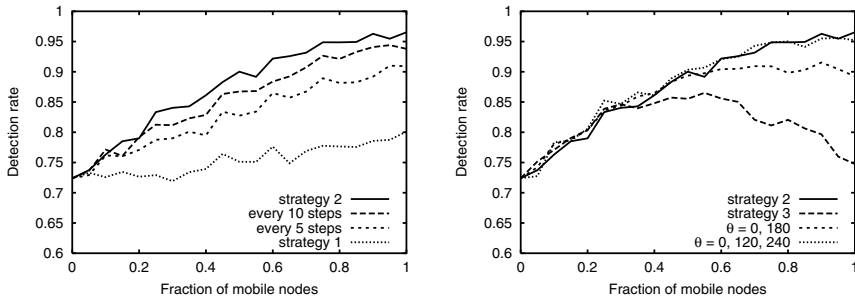
Fig. 7. Migration strategies with different assignment of direction θ

mobile, that is, the fraction of mobile nodes is 0 or 1, the detection rate marks the worst value because the interaction between nodes never changes. The reason for the second point is under investigation.

Some changes for strategy 2 with the best performance can be considered, for example, the following migration strategies:

- Each mobile node can change the direction every some steps. It is expected that the performance of the strategy would exist between strategy 1 and 2 because the interval of changing the direction is 1 in strategy 1 and ∞ (exactly 20) in strategy 2.
- Each mobile node can go straight in a differently assigned direction θ as shown in Fig. 7. Strategy 2 and 3 assign θ to mobile nodes randomly and identically, respectively. It is predicted that the detection rate of the other strategies would be between strategy 2 and 3.

To confirm the predictions given above, we perform additional simulations. From the simulation result in Fig. 8 (a), the first prediction comes true. However, the second prediction is a little different from the result as shown in Fig. 8 (b). The introduction of the opposite direction can highly improve the worst detection rate of strategy 3 when all the nodes are mobile. The strategy with $\theta = 0, 120, 240$ can realize the same performance as strategy 2, so that the random assignment of direction is not necessary. We will clarify the reason theoretically in future.



(a) the interval of changing the direction is changed. (b) the assignment of direction θ is changed.

Fig. 8. Average detection rate vs. the fraction of mobile nodes for the migration strategies when $N = 20$

6 Conclusions and Further Work

In this paper, the immunity-based diagnostic nodes with the simple migration strategies are implemented in the 2-dimensional lattice for wireless sensor network. Some simulation results show the strategy of going straight in the different direction can have the best detection rate. The random assignment of directions to mobile nodes is not necessary. Furthermore, when the rate of mobile nodes is changed, the different transitions of the performance by the migration strategies are observed.

In further work, we will go on analyzing the performance of migration strategies by both simulations and mathematical models. This paper has focused only on the mutual diagnosis between sensor nodes. Since real wireless sensor networks are given applications or tasks, the migration strategies should be examined in response to applications.

Acknowledgements

This work was partly supported by Grant-in-Aid for Research in Nagoya City University, by Grant-in-Aid for Young Scientists (B) No.15700050, and by Grant-in-Aid for Scientific Research (B) No.16300067 from the Ministry of Education, Culture, Sports, Science and Technology.

References

1. Jerne, N.: The immune system. *Scientific American* **229-1** (1973) 52–60
2. Ishida, Y.: Fully distributed diagnosis by PDP learning algorithm: towards immune network PDP model. *Proc. of IJCNN* (1990) 777–782
3. Watanabe, Y., Ishida, Y.: Mutual tests using immunity-based diagnostic mobile agents in distributed intrusion detection systems. *Journal of AROB* **8-2** (2004) 163–167
4. Watanabe, Y., Sato, S., Ishida, Y.: Mutual repairing system using immunity-based diagnostic mobile agent. *KES 2005 (LNAI 3682)* (2005) 72–78
5. Albert, R., Barabasi, A. L.: Statistical mechanics of complex networks. *Review of Modern Physics* **74** (2002) 47–97

Asymmetric Wars Between Immune Agents and Virus Agents: Approaches of Generalists Versus Specialists

Yoshiteru Ishida

Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology
Tempaku, Toyohashi 441-8580, Japan
<http://www.sys.tutkie.tut.ac.jp>

Abstract. This paper reports a multiagent approach to a basic model inspired by the asymmetric war between HIV and T-cells. The basic model focuses on the asymmetric interaction between two types of agents: Virus Agents (abstracted from HIV) and Immune Agents (abstracted from T-cells). Virus Agents and Immune Agents, characterized respectively as “generalists” and “specialists”, may be compared with asymmetric wars between computer viruses and antivirus programs, between guerrillas and armed forces, and so on. It has been proposed that antigenic diversity determines the war between HIV and T-cells. We also formalize the diversity of “generalists” that would determine whether generalists or specialists won. The multiagent simulations also suggest that there is a diversity threshold over which the specialist cannot control the generalist. In multiagent approaches, two spaces, Agent Space and Shape Space, are used to observe not only the spatial distribution of agent populations but also the distribution of antigenic profiles expressed by a bit string.

1 Introduction

The asymmetric wars dealt with in this paper are characterized by interactions between two distinct groups of agents: specialists who have an acute target acquisition capability, and generalists without such capability. This asymmetric situation can be found in many wars such as those between guerrillas and armed forces, between computer viruses and antivirus programs, and between HIV and T-cells in the immune system. Among them, we focus on the asymmetric interactions between HIV and T-cells, which have another aspect of *escaping-chasing* by mutations between parasites and hosts [1]. A similar aspect may be found in artificial information systems, where the cryptographic system must always be changed to fend off attempts to break the cryptographic system. Since a multiagent approach allows agents to replicate and mutate, this aspect can be included in the approach.

Regarding the escaping-chasing aspect of HIV and T-cells, the antigenic diversity threshold conjectured by Nowak-May [1] is worth investigating. In a specific modeling of differential equations [2, 3,4] and a stochastic model in a square lattice space [5], the antigenic diversity threshold has been investigated and its existence was

suggested. This paper again investigates the antigenic diversity threshold with the escaping-chasing aspect using multiagent approaches devised to observe not only the physical layout of agents in a two-dimensional square lattice (called Agent Space) but also the profile distribution of agent populations in a space (called Shape Space) [6].

2 The Basic Model

The model consists of two types of agents: Virus Agents (intruders characterized as generalists) and Immune Agents (defenders characterized as specialists) where a receptor is mounted on each agent for interaction with specific agents, hence agent populations can be divided into sub-classes similarly to the agent approach by Axelrod [7]. Each agent is assumed to be capable of moving (in an agent space), proliferating with possible mutation, and interacting with other agents.

We use two spaces:

- Agent Space, which expresses a physical layout of agents. That is, two agents placed in adjacent squares can interact.
- Shape Space, which expresses the “affinity” between two agents. That is, two agents placed close together in this space have a similar type and hence a high affinity in their profiles.

A physical sense of space is realized by the Agent Space where agents can move around. The Shape Space can be expressed by a lattice as in the models in the previous subsection or other shape space models [8, 9], however, an N-bit binary string is used to express the profile (a conventional approach), while closeness in the shape space is measured by the Hamming Distance (minimum distance in all the shifted comparisons). Diversity of virus agents is measured by the number of distinct profiles in the population. The Agent Space is realized by a two-dimensional lattice space with a periodic boundary condition.

Both Immune Agents and Virus Agents have Reproduction Rate, Error Rate in reproductions, and Kill Rate in interactions. The differences between these two agents are:

- **Virus Agents (Intruder)**
 - Motion: Walk randomly in the Agent Space.
 - Reproduction: Reproduce with a positive Reproduction Rate. In the simulations, one randomly selected string is flipped with an Error Rate in reproduction.
 - Interaction: Do not kill any agents (zero Kill Rate).
- **Immune Agents (Defender)**
 - Motion: Walk towards Virus Agents in the Agent Space.
 - Reproduction: Reproduce with a Reproduction Rate that depends on the distance between self and the closest (in the Agent Space) Virus Agent. No error in the reproduction (zero Error Rate) when the distance is far enough,

but a point mutation (one string flip) occurs to decrease the distance between its profile and that of the Virus Agent in a lattice adjacent to the immune agents (consistent with Clonal Selection Theory).

- Interaction: Kill the adjacent virus agents (positive Kill Rate) depending on the Hamming Distance between profiles of the two agents.

Virus Agents and Immune Agents are characterized as generalists and specialists, respectively. This study also examined the specificity of Immune Agents and its effect on virus extinction and diversity threshold. In this setting of escaping-chasing, Virus Agents do not kill the opponent Immune Agents for simplicity of investigating how long they can survive and how much they can spread, however, they could kill the opponents in a more realistic setting.

3 Multiagent Simulations

The agent rules specific to this simulation are as follows. Agents in the Agent Space can move one adjacent square per time step.

As for Virus Agents, they cannot move one time step after proliferation. Both reproduction rate and error rate are set to be constant in one shot of the simulation.

As for Immune Agents, adaptation occurs in the interaction phase, not in the proliferation phase: when failed in killing ($1 - \text{Killing Rate}$), one string in the profile is changed to decrease by one the Hamming Distance with the Virus Agent in a lattice adjacent to the immune agent. The Reproduction Rate of Immune Agents is set to be inversely proportional to the distance between self and the target (closest) Virus Agents: $1/(d + 1)$ where d is the distance.

Killing rate Kr of the Immune Agents depends on the Hamming Distance Hd between the profiles of self and the target Virus Agents. In this simulation, we used two types of Immune Agents: Specialist type and Generalist type. This difference is characterized by the Killing Rate: $Kr = 1$ when $Hd < 2$ and $Kr = 0$ otherwise for the Specialist type; $Kr = 9/256$ when $Hd < 4$ and $Kr = 0$ otherwise for the Generalist type.

Simulation parameters are set as follows: the size of the bit string is 8, and that of the Agent Space is 100×100 . Twenty agents each from both sides are placed randomly in the Agent Space.

The average number of time steps required for the extinction of Virus Agents is monitored when Immune Agents are the Specialist type (Fig. 1) and the Generalist type (Fig. 2). One trial of the simulation runs for 400 time steps. Each plot is the average of 25 trials.

Figures 1 and 2 indicate that the population of Virus Agents would not benefit from a high reproduction rate when the error rate is low. This fact suggests that merely the size of population does not matter in this asymmetric war between generalists and specialists. This aptitude is more significant when Immune Agents are specialists as in Fig. 1. This simulation also suggests the necessity of measuring the diversity of Virus Agents.

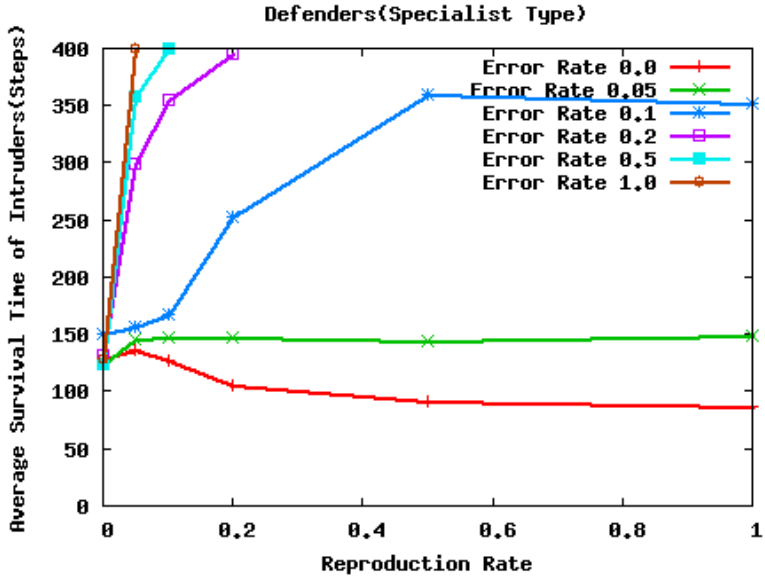


Fig. 1. Average number of time steps required for the extinction of Virus Agents (vertical axis) when the Reproduction Rate (horizontal axis) and Error Rate of Virus Agents (legend) changes when Immune Agents are the Specialist type

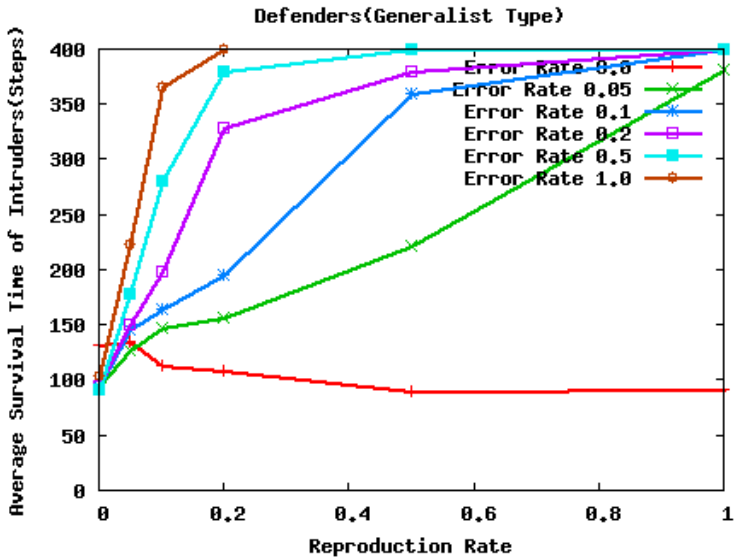


Fig. 2. Average number of time steps required for the extinction of Virus Agents (vertical axis) when the Reproduction Rate (horizontal axis) and Error Rate of Virus Agents (legend) changes when Immune Agents are the Generalist type

In the next simulation, we refer to *Survival* when the Virus Agents remain until 400 time steps and *Extinction* otherwise. We first identified that the trend of Intruder Survival Time changes when Error Rate $Erv = 0.05$ or 0.1 . We then monitored the survival fraction, which is defined to be the fraction of Survival cases where Error Rate $Erv = 0.1$ is fixed. The diversity of the Virus Agents Dv is defined to be the maximum value of distinct profiles (in the 256 profiles in 8-bit strings) throughout the 400 time steps. The survival fractions are plotted against the diversity of the Virus Agents when the Immune Agents are the Specialist type and the Generalist type (Fig. 3).

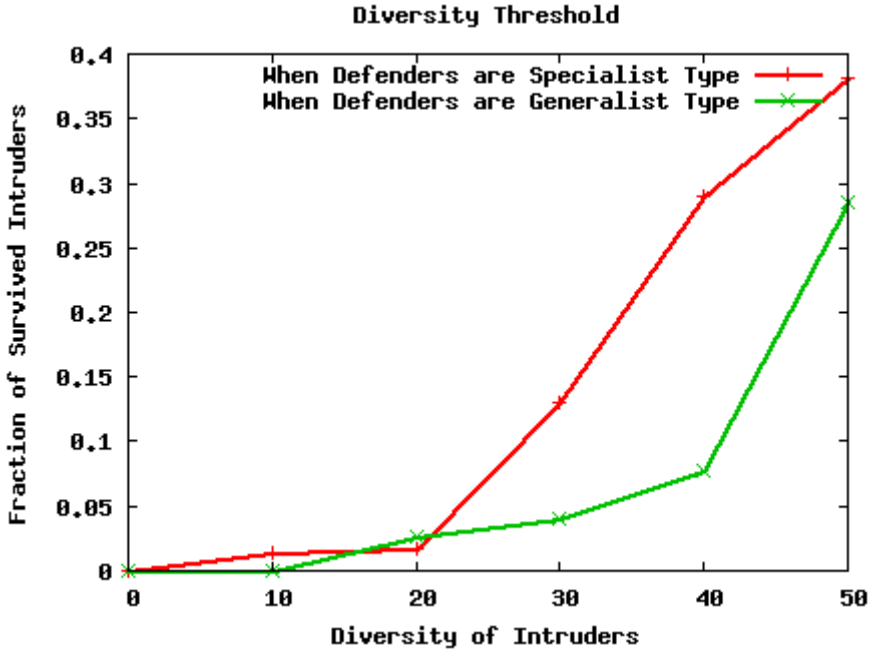


Fig. 3. Survival fraction against diversity of the Virus Agents when the immune agents are the specialist type

Figure 3 indicates the existence of a threshold when the diversity, as defined above, is measured. The threshold is higher when the Immune Agents tend to be generalists, hence the Virus Agents could be controlled more easily. However, total survival (survival fraction = 1) can occur even at lower values of diversity when the Immune Agents tend to be generalists. This issue requires closer investigation.

4 Discussion

Although the modest simulations above suggest the existence of a threshold to control the population of Virus Agents, more elaboration is needed even for multiagent approaches. Although the existence of a threshold seems robust against modeling

(differential equations, stochastic models, and multiagent modeling), it should be robust against different definitions of diversity. Further study on a minimum model that permits the existence of the diversity threshold is also needed.

In the framework of asymmetric wars, agents are clearly labeled into two sides (hence no internecine struggle). However, in many situations including interactions between parasites and immune systems, agents are not labeled beforehand and they must discriminate between *self* and *nonself*. Therefore, another framework is that agents are not clearly categorized into two sides, but tend to self-organize into one side or the other through interactions. In the framework of possible internecine struggle, lowering specificity (making more generalists) would not benefit the agent or the group, since lowering specificity increases the chance of internecine struggle. In our modeling, making Immune Agents more generalists (although the clearly labeled simulations above do not permit the internecine struggle) would cause the Immune Agents to attack not only the Virus Agents but also the Immune Agents themselves, which would lead to a situation called the “Double Edged Sword” in the immune systems [9] and information systems [10].

5 Conclusions

The diversity threshold in asymmetric interactions between HIV and T-cells had been investigated by models of differential equations or statistical models with a lattice space. Using a multiagent approach, this paper further suggested a diversity threshold in a more general framework of asymmetric wars between generalists and specialists. The multiagent approach allows both agents to include mutations: generalists involving “escaping” by mutation and specialists involving “chasing” by adaptation. This “escaping-chasing” situation can be found in many situations not only between parasites and immune systems but also between guerillas and modern armed forces, and between computer viruses and antivirus programs. This study focused on the escaping-chasing situation in the shape space as well as in the agent space.

Acknowledgements

This work was supported in part by Grants-in-Aid for Scientific Research (B) 16300067, 2004. We are indebted to Toshikatsu Mori and Yu Sudo who supported the multiagent simulations. This work was partly supported also by the 21st Century COE Program “Intelligent Human Sensing” of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Ridley M., *The Red Queen*. Felicity Bryan, Oxford, UK (1993)
2. Nowak, M.A. and May, R.M., Mathematical biology of *HIV* infections: Antigenic variation and diversity threshold, *Math. Biosci.*, 106 (1991), 1–21.
3. Nowak, M.A. and May, R.M., *Virus Dynamics Mathematical Principles of Immunology and Virology*, Oxford University Press (2000)

4. Ishida, Y., A Stability and Diversity Analysis on a Mathematical Model of the Interaction between HIV and T-Cell CD4+, in X.S. Gao and D. Wang ed., *Computer Mathematics, Lecture Notes Series on Computing Vol. 8* (2000) 276–279
5. Ishida, Y., A System Theoretic Approach to the Immune System – Adaptive System by Diversity, *Mathematical Science* (separate volume), October, (2001) (in Japanese)
6. Perelson A.S. and Oster G.F., Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol.* 81 (1979) 645–670
7. Axelrod, R. and Cohen, M.D., *Harnessing Complexity: Organizational Implications of a Scientific Frontier*, Basic Books Published (2001)
8. Bersini, H., Self-Assertion versus Self-Recognition: A Tribute to Francisco Varela, in Timmis, J. and Bentley, P.J. (eds). *Proceedings of the 1st International Conference on Artificial Immune Systems*, University of Kent at Canterbury Printing Unit (2002) 107–112
9. Ishida, Y., *Immunity-Based Systems: A Design Perspective*, Springer, Berlin & Heidelberg (2004)
10. Ishida, Y., A Critical Phenomenon in a Self-Repair Network by Mutual Copying, *Lecture Notes in Artificial Intelligence (LNAI 3682)*, (2005) 86–92

Designing an Immunity-Based Sensor Network for Sensor-Based Diagnosis of Automobile Engines

Yoshiteru Ishida

Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology
Tempaku, Toyohashi 441-8580, Japan
<http://www.sys.tutkie.tut.ac.jp>

Abstract. This paper reports on the construction and use of a dynamic relational network that has been studied as an immunity-based system based on the concept of immune networks. The network is constructed by an immunological algorithm that tunes the network not to react to normal sensor data, but to react to abnormal data thereafter. The tuning is not straightforward, since even normal sensor data involve many situations such as accelerating phase and cruise phase; on-road and off-road; and running at altitude. A case study on sensor systems for the combustion control system of an automobile engine is presented.

1 Introduction

Farmer, Packard, and Perelson proposed that the adaptation mechanism in the immune system can be yet another avenue to machine learning [1]. We have been studying the interaction of the network and proposed a dynamic relational network based on the immune network to increase robustness and adaptability (to the dynamic environment) and used it in diagnosis based on sensor data [2,3]. Here, we report on case studies on diagnosing the combustion system of automobile engines after briefly introducing the framework of immunity-based systems – the “semantics” of the network. In our framework, the network is double-sided and would raise the self-referential paradox in a flat logic without distribution, and hence subtle tuning is needed as in the immune system.

Sensor fusion has been studied extensively with the progress of device technology such as intelligent sensors and techniques for integrating them. In recent years, the technical innovation of sensor networks has been encouraged by the Internet, particularly ubiquitous computing and system integration with wireless devices [4]. Our studies have focused on sensor networks using autonomous distributed agents. The basic diagnostic model, called the *immunity-based diagnostic model*, is inspired by the informational features of biological immune systems: recognition of nonself by distributed and dynamically interacting cells, recognition by a simple comparison with the cells themselves, dynamic propagation of activation, and memory embedded as stable equilibrium states in the dynamic network [3]. The diagnostic model is mainly derived from the concept of the *immune (idiotypic) network hypothesis* proposed by Jerne [5]. The model implements network-level recognition by

connecting information from local recognition agents by dynamical evaluation chains. Although the diagnostic model can be considered a modification of the majority network, an important difference between the diagnostic model and the majority network is that the two (active or inactive) states in the diagnostic model are asymmetrical in determination of the next state, while states in the majority network are symmetrical in the sense that they can be exchanged without changing behavior.

As some examples of practical applications, the diagnostic model has been applied to sensor fault diagnosis in processing plants, and self-monitoring/self-repairing in distributed intrusion detection systems. In these applications, however, it is difficult to construct the sensor network from the time sequences of many real sensors. This paper reports on the construction and use of a network for diagnosis, focusing on a case study on sensor systems for the combustion control system of an automobile engine.

2 Recognition and Stimulation/Inhibition Among Sensors: An Immunity-Based System

2.1 An Introduction of Immunity-Based Systems for Computer Engineers

The *self-nonsel*f discrimination problem dealt with in the immune system would raise the self-referential problem, often-cited examples of which are statements such as: “I am a liar” and “This statement is false”. To resolve this paradox, hierarchical logic or distribution of subjects can be used. We use the latter approach (Fig. 1). Dividing the subject and placing them inside the system has been often discussed in *Autopoietic Systems* [6] and other complex systems. Placing the distributed subjects in the system implies that the subjects have only local and unlabelled information for solving problems. Also, and more importantly, the distributed subjects can be the objects on which the subjects operate and interact.

It is still controversial whether the immune system actually needs to discriminate self and *nonsel*f in order to eliminate *nonsel*f [7], however, elimination is done actually and hence the double-sided property that the subject must be the object is imperative. Thus, the immune system is a double-edged sword.

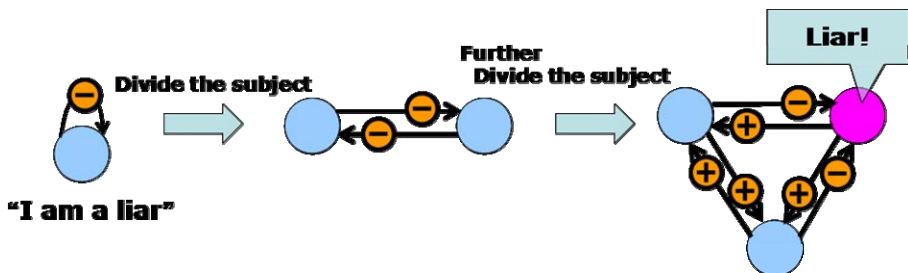


Fig. 1. The left figure corresponds to the statement: “I am a liar”. The middle figure distributes the subject into two. It is still not possible to identify the liar. By further distributing the subjects, it becomes possible to identify the liar. An arc with a minus sign indicates evaluating the target as unreliable, and with a plus sign as reliable.

Jerne proposed the immune network. In network theory, the immune system is not merely a “firewall” but a network of antigen–antibody reactions. That is, when antigen is administered, it stimulates the related cells and causes them to generate antibodies. However, the generated antibodies themselves are antigens to other cells and consequently result in another antibody generation. The antigen–antibody reaction percolates like a chain reaction and hence requires a regulation mechanism. An analogy of this problem in engineering design is the “alarm problem” of placing mutually activating and inactivating alarms whose sensitivity must be appropriately set to avoid mis-alarms (false negative) and false-alarms (false positive).

The challenge of immunity-based systems is to propose a design for placing these alarms which can autonomously tune themselves and adapt themselves to the environment. The immune system as an information system has been extensively studied, e.g. [1,8,9] to mention only a few. We have also studied the immune system as an information system, and proposed a design framework with the following three properties [2]:

- a *self-maintenance system* with monitoring not only of the *nonsel*f but also of the self
- a *distributed system* with autonomous components capable of mutual evaluation
- an *adaptive system* with diversity and selection

2.2 Algorithms for Evaluating Credibility of a Node in a Dynamic Relational Network [3]

Weighting the vote and propagating the information identifies the abnormal agents correctly. A continuous dynamic network is constructed by associating the time derivative of the state variable with the state variables of other agents connected by the evaluation chain. Further, considering not only the effect from evaluating agents, but also that from evaluated agents leads to the following dynamic network:

$$\frac{dr_i(t)}{dt} = \sum_j T_{ji} R_j + \sum_j T_{ij} R_j - \frac{1}{2} \sum_{j \in \{k: T_{ik} \neq 0\}} (T_{ij} + 1),$$

where

$$R_i(t) = \frac{1}{1 + \exp(-r_i(t))},$$

$$T_{ij} = \begin{cases} -1 & \text{if evaluating agent } i \text{ is normal and evaluated agent } j \text{ is faulty} \\ 1 & \text{if both agents } i \text{ and } j \text{ are normal} \\ 1, \text{ or } -1 & \text{if evaluating agent } i \text{ itself is faulty} \\ 0 & \text{if there is no evaluation from agent } i \text{ to agent } j. \end{cases}$$

In evaluating agents, agent j will stimulate (inhibit) agent i when $T_{ji}=1(-1)$. We call this model the *black and white model*, meaning that the network tries to separate an abnormal agent clearly from a normal agent; namely, the *credibility* (which differs from the probabilistic concept of *reliability*) of an agent tends to be 1 (fully credible)

or 0 (not credible), not an intermediate value. Moreover, we proposed several different variants of this dynamic network such as the *skeptical model* and the *gray model* for different engineering needs. The results of this paper are generated only from the *black and white model*.

2.3 An Algorithm for Building the Dynamic Relational Network

A dynamic relational network can be built in roughly two steps:

1. **Line up candidates of relational arcs:** Find causally related sensors by investigating correlation by checking indices such as coefficient of correlation.
2. **Narrow down the above candidates:** Remove those arcs from sensor A to B if the test from sensor A to B generates false positives or false negatives.

Building a dynamic relational network starts from statistical analysis on sensor data. Step 1 above can be done by calculating the coefficient of correlation in statistical analysis. In step 2, a regression line of sensor data B is first expressed with respect to sensor data A. The real data of sensor B when the target part is non-faulty are then compared with the regression line to check that the regression line does not cause false positives. Also in step 2, the real data of sensor B when the target part is faulty are compared with the regression line to check that the regression line does not cause false negatives. Removal of arcs causing false positive can be done when only normal data (data when no fault exists) are available, while removal of arcs causing false negative requires abnormal data (data when faults exist).

Depending on the required specifications of the diagnosis such as diagnostic resolution, diagnostic time and diagnostic accuracy, statistical analysis using the coefficient of correlation for step 1 and regression analysis for step 2 would suffice for building the network. However, if the time series pattern is critical and a more sophisticated diagnosis is required, time series analysis is needed for step 1 (using mutual correlation matrix) and/or for step 2 (prediction by the models of time series analysis). As reported below in the case of the combustion control system of an automobile engine and for a particular fault in an air-flow sensor, a statistical analysis of up to step 1 for building the network suffices. However, time series analysis (with VAR model) is used to determine the sign of an arc (evaluation from node i to node j) in online diagnosis.

3 Application to Diagnosis of Combustion Engine Sensor Systems: An Immunity-Based Sensor Network

3.1 Statistical Analysis of Sensor Data and Construction of a Sensor Network

In this section, a case study with statistical analysis for building the relational network is reported. S_a indicates the data from sensor A. In step 1, arcs between A and B are added if

$$|\text{coefficient of correlation between } S_a \text{ and } S_b| \geq \theta$$

Figure 2 shows a network built when $\theta = 0.4$ and only step 1 in the algorithm is used.

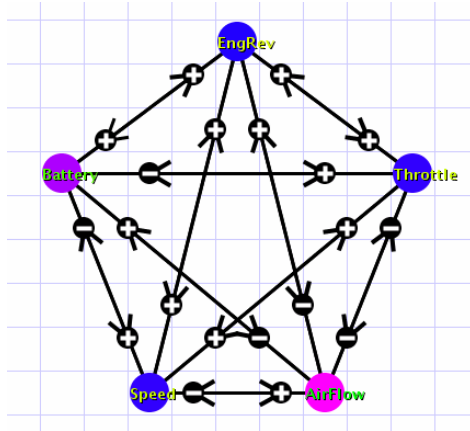


Fig. 2. A network with arcs added when $|\text{coefficient of correlation}| \geq 0.4$ in cruise phase. EngRev: Engine revolutions; Battery: Battery voltage. The network turned out to be complete. Signs are a snapshot of evaluation based on the sensor data. Gray level in the nodes indicates credibility. Dark nodes correspond to high credibility, while light nodes to low credibility (i.e. evaluated as faulty).

The signs of arcs in the network change dynamically in online diagnosis; Fig. 2 shows only a snapshot of signs. The network structure does not change during the diagnosis.

In online diagnosis using the network, the sign of a test from node i to node j is evaluated online. The statistical method is explained below, and a time series analysis is described in the next subsection. Let $x(t)$ and $y(t)$ be sensor data corresponding to nodes i and j respectively. When

$$|a + bx(t) - y(t)| < n\sigma$$

holds, node i evaluates node j as non-faulty (sign of the test from i to j is positive), and faulty (sign of the test from i to j is negative) otherwise, where

$a + bx(t)$: a regression line of $x(t)$ obtained with respect to $y(t)$ during the cruise phase of the automobile, and parameters a, b are determined in offline training using only sensor data without faults (normal sensor data),

- $y(t)$: real value,
- σ : standard deviation.

Further, $n = 3$ is set in this case study. Diagnosis is done by calculating the credibility of each sensor online. Figure 3 shows the time evolution of credibility calculated by the statistical method stated above. The diagnosis is not successful, since not only the credibility of the faulty sensor (Air Flow) but also those of other non-faulty sensors become 0.

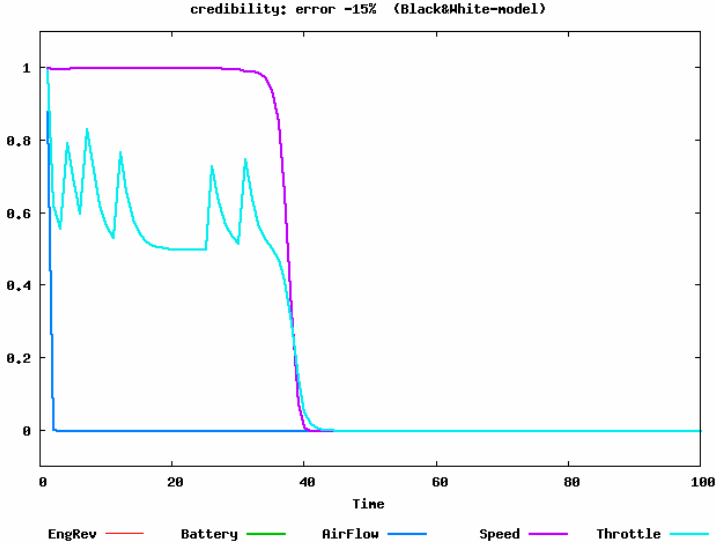


Fig. 3. Diagnosis by the network when the Air Flow sensor is faulty. The dotted line shows the time evolution of credibility for the sensor. Although the credibility for the faulty sensor is evaluated low (0), those of other sensors are also dragged to 0.

3.2 Time Series Analysis for Diagnosis on the Network

In this case study, time series analysis is used to eliminate arcs in step 2. As a model for the time series analysis, the VAR (vector autoregressive) model, which is a multivariate extension of the AR (autoregressive) model, is used. In the AR model, the target variable (explained variable) is estimated with respect to its past values (explaining variable). In the VAR model, however, not only its own past values but also those of related variables are involved. Let $x(t)$ and $y(t)$ be explained variables; $x(t-1), \dots, x(t-m)$; $y(t-1), \dots, y(t-m)$ be explaining variables; and a_1, \dots, a_m ; b_1, \dots, b_m ; c_1, \dots, c_m ; d_1, \dots, d_m be autoregressive coefficients. Then, the VAR model of order m is expressed as:

$$\underline{x(t)} = \underline{a_1 x(t-1) + \dots + a_m x(t-m) + b_1 y(t-1) + \dots + b_m y(t-m)} + \varepsilon_x$$

$$\underline{y(t)} = \underline{c_1 x(t-1) + \dots + c_m x(t-m) + d_1 y(t-1) + \dots + d_m y(t-m)} + \varepsilon_y$$

where the underlined parts ($x'(t), y'(t)$) represent predicted values while ε are the residual errors. In offline data handling before online diagnosis, autoregressive coefficients and the residual errors between the training data and predicted values $x'(t)$ are normalized with respect to $x(t)$. Let these normalized residual errors be $p'(t)$. In online diagnosis based on the network, tests corresponding to arcs generate plus or minus signs as follows:

1. Calculate the normalized residual errors between online data $x(t)$ and its predicted values $x'(t)$. Let these normalized residual errors be $p(t)$.
2. When $p(t)$ deviates from the already calculated $p'(t)$ by a predetermined extent (called the threshold), then the test to $x(t)$ is minus (evaluated as faulty), and plus otherwise.

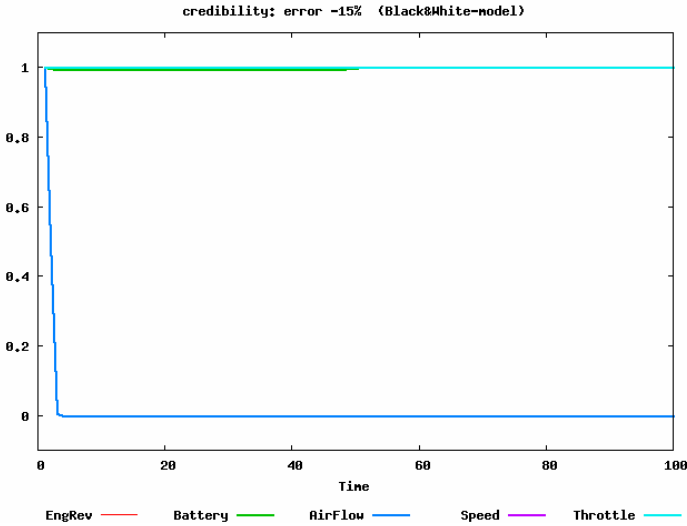


Fig. 4. Diagnosis by the evaluations calculated from the VAR model when the Air Flow sensor is faulty. The dotted line shows the time evolution of credibility for the sensor. Only the credibility of the faulty sensor becomes 0, hence the diagnosis is successful.

It should be noted that the above calculation is done only by normal sensor data. Figure 4 shows the time evolution of credibility calculated by the time series analysis stated above. Only the credibility of the faulty sensor (Air Flow) becomes 0, hence the diagnosis is successful.

4 Conclusion

We have demonstrated that a sensor network for online diagnosis can be built by a simple statistical analysis and time series analysis. A statistical analysis is used for building the network and a time series analysis is used for determining the thresholds for evaluating signs. Only normal sensor data are used for building the network and determining the thresholds. However, we only tested the case when a particular sensor (Air Flow sensor) is faulty and the sensor data used are restricted to the cruise phase. Other sensor faults in different phases should be tested.

Acknowledgements

This work was partly supported by grants from Toyota. We thank Toyota for providing the technical data for the engine sensors. We are also grateful to Toshiyuki Abe and Kazutaka Hattori at Design Department No. 21, Electronics Engineering Division II, Vehicle Engineering Group, Toyota Motor Corporation for incisive discussions on this project. Graduate students Masakazu Oohashi, Tokuya Hiraizumi, and Yuuki Sugawara also assisted the numerical simulations. We are indebted to Yuji Watanabe who supported the launching of the project. This work was also partly supported by Grants-in-Aid for Scientific Research (B) 16300067, 2004.

References

1. J.D. Farmer, N.H. Packard, and A.S. Perelson, The immune systems, adaptation and machine learning, *Physica D*, 22 (1986) 187–204.
2. Y. Ishida, *Immunity-Based Systems: A Design Perspective*, Springer, Berlin & Heidelberg, 2004.
3. Y. Ishida, An immune network approach to sensor-based diagnosis by self-organization, *Complex Systems*, vol. 10, pp. 73–90, 1996.
4. D. Estrin, R. Govindan, J. Heidemann, and S. Kumar, Next century challenges: scalable coordination in sensor networks. *Proc. of the ACM/IEEE International Conference on Mobile Computing and Networking*, 1999, pp. 263–270.
5. N. K. Jerne, The immune system, *Sci. Am.*, vol. 229, no. 1, 1973, pp. 52–60.
6. H. Maturana and F. Varela, *Autopoiesis and Cognition: the Realization of The Living*, D. Reidel, Dordrecht, 1980.
7. R.E. Langman and M. Cohn (eds.) *Seminars in Immunology*. (<http://www.idealibrary.com>), 2000.
8. G.M. Edelman, *Bright Air Brilliant Fire: on the Matter of the Mind*, Basic Books, New York, 1992.
9. A.I. Tauber, *The Immune Self*, Cambridge University Press, Cambridge, UK, 1997.

Dynamic Cooperative Information Display in Mobile Environments

Christophe Jacquet^{1,2}, Yacine Bellik², and Yolaine Bourda¹

¹ Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France
Christophe.Jacquet@supelec.fr, Yolaine.Bourda@supelec.fr

² LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
Yacine.Bellik@limsi.fr

Abstract. We introduce an interaction scenario in which users of public places can see relevant information items on public displays as they move. Public displays can dynamically collaborate and group with each other so as to minimize information clutter and redundancy. We analyse the usability constraints of this scenario in terms of information layout on the screens. This allows us to introduce a decentralized architecture in which information screens as well as users are modeled by software agents. We then present a simulator that implements this system.

1 Introduction

When people find themselves in unknown environments such as train stations, airports, shopping malls, etc., they often have difficulties obtaining information that they need. Indeed, public information screens show information for everybody: as a result, they are often cluttered by too many items. One given person is usually interested in only one item, so seeking it among a vast quantity of irrelevant items is sometimes long and tiresome.

To improve the situation, we aim at designing an ubiquitous information system that can use multiple output devices to give personalized information to mobile users. This way, information screens placed at random in an airport would provide passengers nearby with information about their flights. To reduce clutter, they would display information relevant to these passengers only.

However, if many people gather in front of a screen, they still will have to seek through a possibly long list of items to find relevant information. One possible solution would be to bring a second screen next to the first one to extend screen real estate. But in the absence of cooperation among the screens, the second one will merely copy the contents of the first one, both screens remaining very cluttered. The solution lies in the judicious *distribution of content* among the screens (see fig. 1).

In this article, we introduce an agent architecture in which neighboring output devices can cooperate to reduce clutter, *without having prior knowledge of each other*. Thus, no manual configuration is ever necessary, and in particular it is possible to move output devices at run time without changing the software setup. First, we present research work related to these topics. Then, we formally

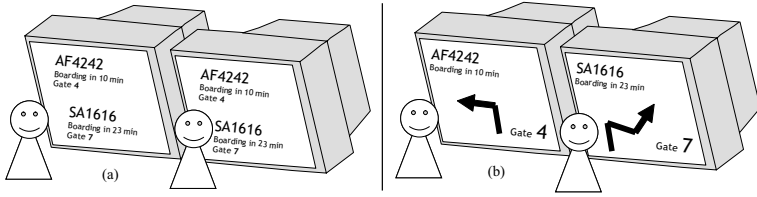


Fig. 1. Two screens, (a) only neighbors, merely duplicating content, and (b) cooperating with each other

introduce the problem, and draw a list of usability constraints for a cooperative display system. This allows us to propose a solution based on an agent algorithm distributed among output devices and users. The algorithm needs no centralized process: agents cooperate with each other *at a local level* to build a solution. We then present an implementation and a simulator that allow us to test this algorithm. Finally, we introduce perspectives for future work.

2 Related Work

Several systems have been designed to give contextual information to users as they move around. For instance, CoolTown [1] shows web pages to people, depending on their location. Usually, contextual information is given to users through small handheld devices: for example, Cyberguide [2], a tour guide, was based on Apple's Newton personal digital assistant (PDA). Indeed, most experiments in context-aware computing are based on portable devices that give people information about their environment [3].

However, some ubiquitous computing [4] applications no longer require that users carry PDAs: instead, public displays are used, for example the Gossip Wall [5]. In these systems, public displays can play several roles: providing public information when no-one is at proximity, and providing private information when someone engages in explicit interaction, which raises privacy concerns [6].

Though our system uses public displays, it does not provide *personal* information: actually, it provides *public* information *relevant to people located at proximity*. The originality of our system lies in the fact that public displays have not a priori knowledge of their conditions of operation: they can be placed anywhere without prior configuration, giving relevant information to people nearby, and collaborating with each other in order to improve user experience.

3 Problem Statement

For the sake of simplicity, we assume that output devices are screen displays, but this restriction could easily be lifted. We suppose that each user wishes to obtain a given information item called her *semantic unit* (s.u.), for instance her boarding gate. We call *load* of a screen the number of s.u.'s it displays.

We introduce a notion of *proximity* (also called *closeness*). A user is said to be *close to* a screen if he can see its contents. Therefore, this definition includes distance as well as orientation conditions. For instance, if a user turns his back to a monitor while talking in his cell phone, displaying information for her would be totally irrelevant, even if he is at a very short distance of the screen.

In introduction, we have seen that in our system, a user close to (at least) one display must be provided with his s.u. of interest. This is what we call the *completeness constraint*. We have also seen that information must be optimally distributed among screens. To do so, screen load must be minimal so as to reduce clutter (*display surface optimization constraint*). If we consider these two constraints only, the problem would boil down to resolving a distributed constraint system (first criterion) while minimizing the load parameter (second criterion). The problem could thus be seen as an instance of DCOP (Distributed Constraint Optimization Problem); several algorithms exist to solve a DCOP [7].

However, they are designed to find a solution all at once. In contrast, our problem is built step by step from an initial situation. Indeed, we can assume that at the beginning no user is close to a screen. Then, two kinds of events may occur: 1) a user comes close to a screen; 2) a user goes away from a screen. This way, any situation can be constructed by a sequence of events number 1 and number 2. If we assume that we have a suitable solution at one given moment, and if we know how to construct a new suitable solution after an event number 1 or 2 occurs, then we are able to solve the problem at any moment (recursion principle): an incremental algorithm would be highly efficient in this situation.

Optimizing the display surface is an important goal, but it may lead to obtaining unusable systems. Indeed, if a system tries to absolutely avoid clutter, and so always reorganizes screen layout to be optimal, users might end up seeing their s.u.'s leaping from one screen to another every time another user moves. They would then waste their time chasing their information items just to read them, which is worse than having to find an item in a cluttered screen.

Instead, information display should remain pretty much stable upon event occurrence so as not to confuse users. Indeed, if someone does not move, then they may be reading their s.u., so they expect it to remain where it is, not suddenly vanishing to reappear somewhere else. Conversely, when people move, they are generally not reading screens at the same time, so they do not mind if information items are migrated to a new place.

For all this, we take three constraints in consideration.

Constraint C_1 (completeness). When a user is close to a number of displays, his s.u. must be provided by (at least) one of these devices.

Constraint C_2 (stability). A user's s.u. must not move from one display to another, unless the user herself has moved.

Constraint C_3 (display surface optimization). To prevent devices from being overloaded, s.u. duplication must be avoided whenever possible. This means that the sum of device loads must be minimal.

For usability reasons, we consider that C_1 is stronger than C_2 , which in turn is stronger than C_3 . For instance, let us suppose that three displays show three

s.u.'s for three users (each display shows a different s.u.). Then, if the leftmost user leaves, and a new user arrives on the right, the s.u.'s will *not* be shifted leftwards. This breaks the surface optimization constraint, but we consider it less important than preserving display stability and not disturbing users. However, if at some point the rightmost screen becomes saturated, then s.u.'s will be shifted, so as to ensure completeness, considered to be more important than stability.

4 Solution

In this section, we introduce an algorithm to solve the problem of information display, while satisfying the above constraints. This algorithm is distributed among screens and users, each of them being represented by a software agent.

4.1 Mathematical Formalization

We introduce three costs, the *static cost* of a screen layout, the *dynamic cost* of the action of adding a s.u. to a screen, and the *migration cost* of moving a s.u. from one screen to another.

Let \tilde{c} be a function over \mathbb{R}^+ , strictly convex and strictly increasing (we will see the reason for this). $\tilde{c}(\ell)$ represents the *static cost* of a screen layout of load ℓ . Thus, for a screen s with load ℓ_s , we define $\mathbf{c}(s)$ to be $\tilde{c}(\ell_s)$. $\mathbf{c}(s)$ is called the *static cost* of the given screen s with its current layout.

Let us suppose that we want to add δ s.u.'s ($\delta \neq 0$) to a given screen s . The *dynamic cost* of the operation, written $\mathbf{d}(s, \delta)$, is defined to be: $\mathbf{d}(s, \delta) = \mathbf{c}(s)_{\text{after operation}} - \mathbf{c}(s)_{\text{before operation}} = \tilde{c}(\ell_s + \delta) - \tilde{c}(\ell_s)$.

Note that the dynamic cost increases as ℓ_s increases, because \tilde{c} is strictly increasing and strictly convex. Thus, δ being given, if $\ell_2 > \ell_1$ then $\tilde{c}(\ell_2 + \delta) - \tilde{c}(\ell_2) > \tilde{c}(\ell_1 + \delta) - \tilde{c}(\ell_1)$. With this definition of a *dynamic cost*, we convey the idea that *the more items are displayed on a screen, the more costly it is to add an incremental item*. Indeed, if there is one item (or even no item) on a screen, adding an item should not increase the time needed to find one's s.u. But if a screen is already overloaded, finding a new item among the multitude will be very long and tiresome.

For instance, let us assume that on a given system, \tilde{c} is defined by $\tilde{c}(x) = x^2$ for two screens, called a and b . Note that this choice for \tilde{c} is totally arbitrary: in practice, \tilde{c} must be chosen for every screen so as to match the screen's proneness to become overloaded. If screen a currently displays 2 s.u.'s, then $\mathbf{c}(a) = 2^2 = 4$; if screen b currently displays 4 s.u.'s, then $\mathbf{c}(b) = 4^2 = 16$. If we want to add one s.u. to these screens, what are the associated dynamic costs? $\mathbf{d}(a, 1) = (2 + 1)^2 - 2^2 = 9 - 4 = 5$; likewise, $\mathbf{d}(b, 1) = (4 + 1)^2 - 4^2 = 25 - 16 = 9$. So if we have the choice, we will then choose to display the s.u. on screen a , which seems to be reasonable. Note that here, both screens share the same static cost function, but in practice each display can define its own static cost function.

We also introduce *migration costs*. A migration cost is taken into account when a s.u. u is moved from one display to another one. Each user U_i interested in the given s.u. contributes a partial migration cost $m(U_i, u)$. The total migration cost is the sum of all partial costs contributed by each user: $m(u) = \sum_i m(U_i, u)$.

4.2 Agent-Based Architecture

Of course, it would have been possible to build a solution around a *centralized* architecture. However, we think that this has a number of shortcomings, namely fragility (if the central server fails, every display fails) and rigidity (one cannot move the displays at will). In contrast, we wish to be able to move displays, bring new ones in case of an event, etc., all this without having to reconfigure anything. Displays have to adapt to the changes themselves, without needing human intervention.

Our implementation is based on software agents that model physical entities: screens are modeled by display agents; users are modeled by user agents. We assume that each agent knows which agents are nearby, and can communicate with them. These assumptions are quite realistic. Proximity of users can for instance be detected by an RFID reader located on a screen, provided that users carry RFID tags stucked to their tickets¹. As for ubiquitous communications, they are now commonplace thanks to wireless networks.

The agents are *reactive*; they stay in an idle state most of the time, and react to two kinds of events: the appearance or disappearance of an agent at proximity, or the reception of a network message from an agent (which is not necessarily at proximity). In section 4.3, we will see some examples of such messages.

4.3 Algorithm

It is now possible to describe the general behavior of the algorithm. First, note that every user agent references a *main screen*, i.e. a screen where its s.u. is displayed. On startup, all main screens are undefined.

The general layout of the algorithm is as follows: when a user agent either comes close to (i) or goes away from (ii) a screen, it ponders on doing some operations (described below). So it sends *evaluation requests* to neighboring display agents, to know the costs of these operations. Display agents answer the requests (iii, iv) and remember the evaluated operations. The user agent has then the choice between either *committing* or *canceling* each of the previously evaluated operations. In practice, the agent commits the operation with the best cost, and cancels all the others.

[i] When a user agent with s.u. u comes close to a screen s :

- if its main screen is already defined, it sends a `migration-evaluation(s)` request to its main screen. If the result (dynamic cost) of the request is negative the user agent commits it, otherwise it cancels it. This way, a user walking along a row of screens will have her s.u. “follow” her on the screens,
- if not, it sends a `display-evaluation(u)` request to s , and systematically commits it (to satisfy constraint C_1 , completeness). The s.u. u is sent to the display agent through the network (serialized object).

¹ In this case, only monitors detect the closeness of users. However, the relationship can be made symmetric if a display agent which detects a user agent at proximity systematically sends a notification to it.

[ii] When a user agents with s.u. u goes away from its *main screen*, it first sends a *going-away* notification to its main screen, and then:

- *if some other screens are nearby*, it sends a `display-evaluation(u)` request to each of them, and chooses the one with the lowest dynamic cost as its main screen (constraint C_3 , display surface optimization). It then sends a commit message to this one, and cancel messages to the others,
- *if not*, its main screen is set as undefined.

[iii] When a display agent receives a `display-evaluation(u)` request:

- *if there is still room for s.u. u* , it adds it to its display list: when constraint C_1 (completeness) is satisfiable, the screen tries to satisfy C_2 (stability),
- *otherwise*, it tries to move one of its other s.u.'s to another screen. In practice, for each displayed s.u. v , it sends recursively a `display-evaluation(v)` to each screen seen by every user agent a_i interested in v . The cost of one possible migration (if the corresponding recursive call does not fail), is the cost returned by the call (d), plus the associated migration cost, i.e., $d + \sum_i m(a_i, v)$. If some of the recursive calls do not fail, the display agent chooses the least costly, commits it, and cancels the others. Otherwise, the call itself fails. If C_2 (stability) is not satisfiable, the screen still tries to enforce constraint C_1 (completeness), but while doing so, it still optimizes constraint C_3 (display surface optimization). Rule C_1 is broken only if all neighboring screens have no space left.

[iv] When a display agent receives a `migration-evaluation(s)` request to migrate a s.u. u :

- if more than one user agents are interested in u , the call fails,
- otherwise, the display agent sends to s a `display-evaluation(u)` request, and calls the associated cost d_1 . It evaluates the “cost” of suppressing u from its display layout. This cost, negative, is called d_2 . It calculates the associated migration cost, called m . Then, it returns $d_1 + d_2 + m$. The migration is considered useful if this quantity is negative.

This is the basic behavior of the algorithm. The other operations, such as commits and cancels, are defined in a quite straightforward manner.

5 Implementation and First Results

The algorithm, as well as a graphical simulator (fig. 2) have been implemented. On the figure, users are called H0, H1 and H2. Their s.u.'s are respectively A, B and C. There are two screens, called S0 and S1. They can each display at most two s.u.'s. The matrix of 2×3 “boxes” shows proximity relationships: if a user is not close to a screen, the box is empty. If a screen is a user's *main* screen, there is an “M” in the box. If a screen is close to a user, but it is not his main screen, there is a “c” in the box.

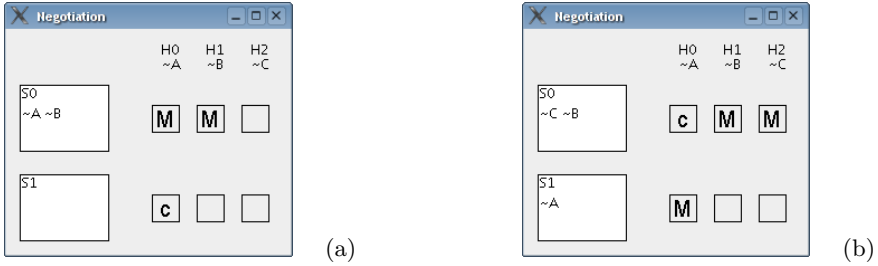


Fig. 2. The simulator introduces a GUI to manipulate proximity relationships

On figure 2a, H0 and H1 are close to screen S0, and S0 is their *main* screen. Thus, S0 displays s.u.'s A and B. H0 is also close to screen S1, but it is not his main screen. Then user H2 comes close to screen S0. To satisfy C_1 (completeness), S0 should display the s.u. C, but it can at most display two s.u.'s. So S0 chooses to break rule C_2 (stability) in favor of C_1 , and migrates A to screen S1 (fig. 2b).

The tests performed with this implementation were satisfying. Next, we plan to implement the system in real scale, so as to assess its practical usability.

The system can be used to provide information in an airport or train station, but also for instance to display examination results. In this case, people generally have to find their names in very long lists, which is very tedious. The task would be much easier if only the results of people located at proximity were displayed.

6 Future Work

In this paper, all information output devices were screens, thus favoring the visual modality. However, we are currently finalizing a generalization of the framework presented here to multimodal output devices, handling for instance speech output as well as text output. In this case, users have preferences not only about their s.u.'s, but also about their input modalities. For instance, blind users require information kiosks to provide them with audio information.

Moreover, within a given modality, people can express preferences about the *attributes* of output modalities. For instance, short-sighted people can indicate a minimum size for text to be rendered; people with hearing problems can indicate a minimum sound level for speech output. In short, the attributes are used when *instantiating* [8] semantic units. Note that this extension will have repercussions on cost calculations, since, for example, screen real estate depends on the size used to render s.u.'s textually.

On a given screen, it will be necessary to *sort* the various s.u.'s displayed. This could be done at random, but we think that a *level of priority* could be given to each s.u. This would for instance allow higher-priority s.u.'s (e.g. flights which are about to depart shortly, or information about lost children) to appear first. Similarly, there could be priorities among users (e.g. handicapped people, premium subscribers would be groups of higher priority). Therefore, s.u.'s priority levels would be altered by users' own priorities.

As seen above, priorities will determine the layout of items on a screen. Moreover, when there are too many s.u.'s so that they cannot fit all on the screens, priorities could help choose which ones are displayed.

In this paper, proximity was *binary*: agents are either close to each other, or away from each other. Actually, it is possible to define several degrees of proximity, or even a measure of distance. These degrees or distances could be used as parameters of the aforementioned instantiation process. For instance, text displayed on a screen could be bigger when people are farther away.

We plan to do real-scale experiments shortly, so the agents in the simulator already rely on Java RMI, so they will be easily deployable on a network. We also plan to test different proximity sensors that can be used to fulfill our needs.

7 Conclusion

In this paper, we have presented a novel mobile interaction scenario: as users are being given personalized information on public displays as they move, displays dynamically cooperate to reduce clutter and increase usability. We have analyzed the diverse constraints of this scenario, which has led us to propose a solution based on a decentralized multi-agent architecture.

This architecture appears to be efficient in simulation. The next step will be a real-scale implementation that will allow field trials with users in context.

References

1. Kindberg, T., Barton, J.: A Web-based nomadic computing system. *Computer Networks* **35** (2001) 443–456
2. Long, S., Kooper, R., Abowd, G.D., Atkeson, C.G.: Rapid Prototyping of Mobile Context-Aware Applications: The Cyberguide Case Study. In: *Mobile Computing and Networking*. (1996) 97–107
3. Hull, R., Neaves, P., Bedford-Roberts, J.: Towards Situated Computing. In: *ISWC '97*, Washington, DC, USA, IEEE Comp. Soc. (1997) 146
4. Weiser, M.: Some computer science issues in ubiquitous computing. *Communications of the ACM* **36** (1993) 75–84
5. Streitz, N.A., Rcker, C., Prante, T., Stenzel, R., van Alphen, D.: Situated Interaction with Ambient Information: Facilitating Awareness and Communication in Ubiquitous Work Environments. In: *HCI International 2003*. (2003)
6. Vogel, D., Balakrishnan, R.: Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In: *UIST '04*, New York, NY, USA, ACM Press (2004) 137–146
7. Mailler, R., Lesser, V.: Solving Distributed Constraint Optimization Problems Using Cooperative Mediation. In: *AAMAS '04*, IEEE Comp. Soc. (2004) 438–445
8. André, E.: The Generation of Multimedia Presentations. In Dale, R., Moisl, H., Somers, H., eds.: *A Handbook of Natural Language Processing*. M. Dekker (2000) 305–327

Extracting Activities from Multimodal Observation

Oliver Brdiczka, Jérôme Maisonnasse, Patrick Reignier, and James L. Crowley

INRIA Rhône-Alpes, Montbonnot, France

{brdiczka, maisonnasse, reignier, crowley}@inrialpes.fr

Abstract. This paper addresses the extraction of small group configurations and activities in an intelligent meeting environment. The proposed approach takes a continuous stream of observations coming from different sensors in the environment as input. The goal is to separate distinct distributions of these observations corresponding to distinct group configurations and activities. In this paper, we explore an unsupervised method based on the calculation of the Jeffrey divergence between histograms over observations. The obtained distinct distributions of observations can be interpreted as distinct segments of group configuration and activity. To evaluate this approach, we recorded a seminar and a cocktail party meeting. The observations of the seminar were generated by a speech activity detector, while the observations of the cocktail party meeting were generated by both the speech activity detector and a visual tracking system. We measured the correspondence between detected segments and labelled group configurations and activities. The obtained results are promising, in particular as our method is completely unsupervised.

1 Introduction

The focus of this work is analyzing human (inter)action in intelligent meeting environments. In these environments, users are collaborating in order to achieve a common goal. Several individuals can form one group working on the same task, or they can split into subgroups doing independent tasks in parallel. The dynamics of group configuration and activity need to be tracked in order to supply reactions or interactions at the most appropriate moment. Changes in group configuration need to be detected to identify main actors, while changes in activity within a group need to be detected to identify activities.

This paper proposes an unsupervised method for extracting small group meeting configurations and activities from a stream of multimodal observations. The method detects changes in small group configuration and activity based on measuring the Jeffrey divergence between adjacent histograms. These histograms are calculated for a window slid from the beginning to the end of a meeting recording and contain the frequency of observations coming from multi-sensory input. The peaks of the Jeffrey divergence curve between these histograms are used to segment distinct distributions of multimodal observations and to find the best model of observation distributions for the given meeting. The method has been tested on observations coming from a speech activity detector as well as a visual tracking system. The evaluation has been done with recordings of a seminar and a cocktail party meeting.

2 Previous and Related Work

Many approaches for the recognition of human activities in meetings have been proposed in recent years. Most work use supervised learning methods [2], [5], [6], [9], [10]. Some projects focus on supplying appropriate services to the user [9], while others focus on the correct classification of meeting activities [5] or individual availability [6]. Less work has been conducted on unsupervised learning of meeting activities [3] [11]. To our knowledge, little work has been done on the analysis of changing small group configuration *and* activity. In [2] a real-time detector for changing small group configurations has been proposed. This detector is based on speech activity detection and conversational hypotheses. We showed that different meeting activities, and especially different group configurations, have particular distributions of speech activity [2]. Detecting group configuration or activity (as in [2], [5], [6]) requires, however, a predefined set of activities or group configurations. New activities or group configurations with a different number of individuals cannot be detected and distinguished with these approaches. The approach proposed in this paper is a multimodal extension of [3] focusing on an unsupervised method segmenting small group meetings into consecutive group configurations and activities. These configurations and activities are distinguished by their distributions, but not labelled or compared. The method can thus be seen as a first step within a classification process identifying (unseen) group configurations and activities in meetings.

3 Approach

Our approach is based on the calculation of the Jeffrey divergence between histograms of observations. These observations are a discretization of events coming from multi-sensory input. The observations are generated with a constant sampling rate depending on the sampling rates of the sensors.

3.1 Observation Distributions

In [2], we stated that the distribution of the different speech activity observations is discriminating for group configurations in small group meetings. We assume further that in small group meetings distinct group configurations and activities have distinct distributions of multimodal observations. The objective of our approach is hence to separate these distinct distributions, in order to identify distinct small meeting configurations and activities.

As our observations are discrete and unordered (e.g. a 1-dimensional discrete code) and we do not want to admit any a priori distribution, we use histograms to represent observation distributions. A histogram is calculated for an observation window (i.e. the observations between two distinct time points in the meeting recording) and contains the frequency of each observation code within this window.

To separate different observation distributions, we calculate the Jeffrey divergence between the histograms of two adjacent observation windows. The Jeffrey divergence

[7] is a numerically stable and symmetric form of the Kullback-Leibler divergence between histograms. We slide two adjacent observation windows from the beginning to the end of the recorded meetings, while constantly calculating the Jeffrey divergence between these windows. The result is a divergence curve of adjacent histograms (Fig. 1).

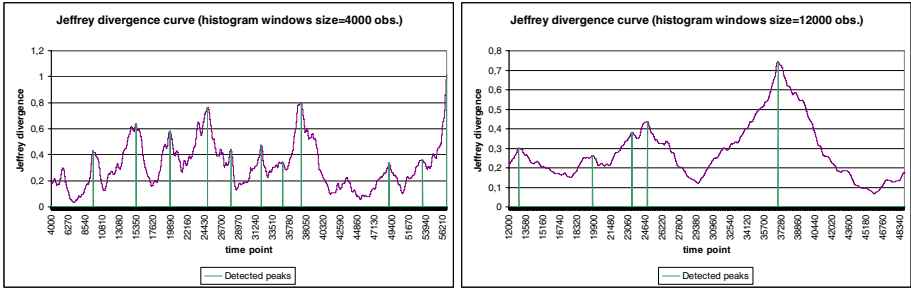


Fig. 1. Jeffrey divergence between histograms of sliding adjacent windows of 4000, and 12000 observations (64sec and 3min 12sec)

The peaks of the curves indicate high divergence values, i.e. a big difference between the adjacent histograms at that time point. The size of the adjacent windows determines the exactitude of the divergence measurement. The larger the window size, the less peaks has the curve. However, peaks of larger window sizes are less precise than those of smaller window sizes. Thus we parse the meeting recordings with different window sizes (e.g. in recording of the seminar: window sizes of between 4000 and 16000 observations, which corresponds to a duration between 64sec and 4min 16sec for each window). The peaks of the Jeffrey divergence curve can then be used to detect changes in the observation distribution of the meeting recording.

3.2 Peak Detection

To detect the peaks of the Jeffrey divergence curve, we use successive robust mean estimation. Robust mean estimation has been used in [8] to locate the center position of a dominant face in skin color filtered images. Mean and standard deviation are calculated repeatedly in order to isolate a dominant peak. To detect all peaks of the Jeffrey divergence curve, we apply the robust mean estimation process successively to the Jeffrey divergence values.

3.3 Merging and Filtering Peaks from Different Window Sizes

Peak detection using successive robust mean estimation is conducted for Jeffrey curves with different histogram window sizes. A global peak list is maintained containing the peaks of different window sizes. Peaks in this list are merged and filtered with respect to their window size and peak height.

To merge peaks of Jeffrey curves with different histogram window sizes, we calculate the distance between these peaks normalized by the minimum of the histogram

window sizes. The distance is hence a fraction of the minimum window size measuring the degree of overlap of the histogram windows. To merge two peaks, the histogram windows on both sides of the peaks need to overlap, i.e. the normalized distance needs to be less than 1.0.

We filter the resulting peaks by measuring peak quality. We introduce the relative peak height and the number of votes as quality measures. The relative peak height is the Jeffrey curve value of the peak point normalized by the maximum value of the Jeffrey curve (with the same window size). A peak needs to have a relative peak height between 0.5 and 0.6 to be retained. The number of votes of a peak is the number of peaks that have been merged to form this peak. A number of 2 votes are necessary for a peak to be retained.

The small number of peaks resulting from merging and filtering is used to search for the best allocation of observation distributions, i.e. to search for the best model for a given meeting.

3.4 Model Selection

To search for the best model for a given meeting recording, we examine all possible peak combinations, i.e. each peak of the final peak list is both included and excluded to the (final) model. For each such peak combination, we calculate the average Jeffrey divergence of the histograms between the peaks. As we want to separate best the distinct observation distributions of a meeting, we accept the peak combination that maximizes the average divergence between the peak histograms as the best model for the given meeting recording.

4 Evaluation and Results

The result of our approach is the peak combination separating best the activity distributions of a given meeting recording. We interpret the intervals between the peaks as segments of distinct group configuration and activity. To evaluate our approach, we recorded a seminar and a cocktail party meeting. The group configurations and activities of these meetings have been labeled. For the evaluation of the detected segments, we use the *asp*, *aap* and *Q* measure.

4.1 Evaluation Measures

For the evaluation, we dispose of the timestamps and durations of the (correct) group configurations and activities. However, classical evaluation measures like confusion matrices can not be used here because the unsupervised segmentation process does not assign any labels to the found segments. Instead, we use three measures proposed in [11] to evaluate the detection results: average segment purity (*asp*), average activity purity (*aap*) and the overall criterion *Q* (Fig. 2). The *asp* is a measure of how well a segment is limited to only one activity, while the *aap* is a measure of how well one activity is limited to only one segment. In the ideal case (one segment for each activity), $asp = aap = 1$. The *Q* criterion is an overall evaluation criterion combining *asp* and *aap*, where larger *Q* indicates better overall performance.

$$asp = \frac{1}{N} \sum_{i=1}^{N_s} p_{i\bullet} \times n_{i\bullet} \quad , \quad aap = \frac{1}{N} \sum_{j=1}^{N_a} p_{\bullet j} \times n_{\bullet j} \quad ,$$

$$Q = \sqrt{asp \times aap} \quad .$$

with

n_{ij} = total number of observations in segment i by activity j

$n_{i\bullet}$ = total number of observations in segment i

$n_{\bullet j}$ = total number of observations of activity j

N_a = total number of activities

N_s = total number of segments

N = total number of observations

$$p_{\bullet j} = \sum_{i=1}^{N_s} \frac{n_{ij}^2}{n_{\bullet j}^2}$$

$$p_{i\bullet} = \sum_{j=1}^{N_a} \frac{n_{ij}^2}{n_{i\bullet}^2}$$

Fig. 2. Average segment purity (*asp*), average activity purity (*aap*) and the overall criterion *Q*

4.2 Seminar

We recorded a seminar (duration: 25 min. 2 sec.) with 5 participants. The speech of the participants was recorded using lapel microphones. A speech activity detector was executed on the audio channels of the different lapel microphones. One observation was a vector containing a binary value (speaking, not speaking) for each individual that is recorded. This vector was transformed to a 1-dimensional discrete code used as input. Our automatic speech detector has a sampling rate of 62.5 Hz, which corresponds to the generation of one observation every 16 milliseconds.

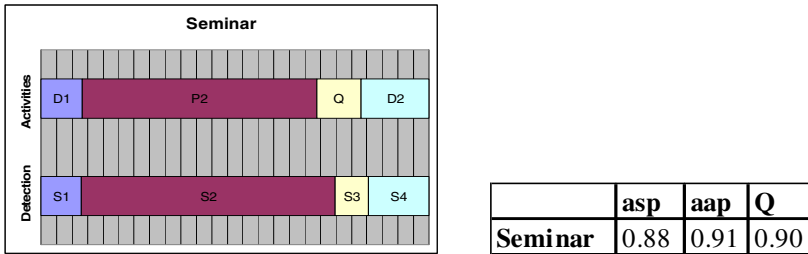


Fig. 3. Activities and their detection for the seminar (meeting duration = 25min 2sec)

The activities during the seminar were discussion in small groups (D1), presentation (P), questions (Q) and discussion in small groups (D2). Fig. 3 shows the labeled activities for the seminar and the segments detected by our approach as well as the corresponding *asp*, *aap* and *Q* values. The results of the automatic segmentation are very good; we obtain a *Q* value of 0.90.

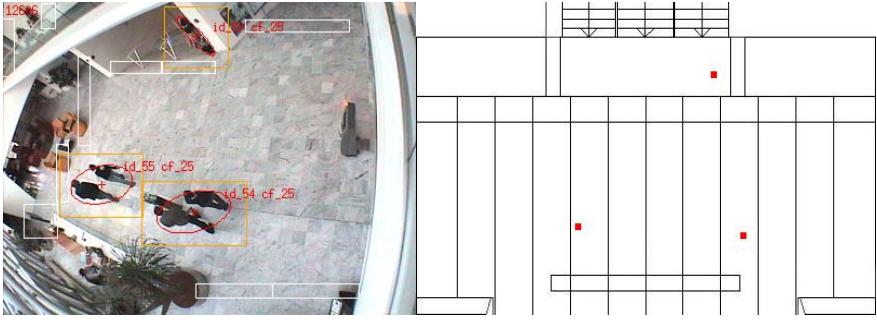


Fig. 4. Wide-angle camera image of INRIA Rhône-Alpes entrance hall with one individual and two small groups being tracked (left) and corresponding positions on the hall map after applying a homography (right)

4.3 Cocktail Party Meeting

We recorded a cocktail party meeting (duration: 30 min. 26 sec.) with 5 participants in the entrance hall of INRIA Rhône-Alpes. The speech of the participants was recorded using headset microphones. As for the seminar, a speech activity detector provided the speech activity observations for each individual. A wide-angle camera filmed the scene and a visual tracking system [4] based on background subtraction provided targets corresponding to individuals or small groups (Fig. 4 left). We used a homography to calculate the positions of these targets on the hall map (Fig. 4 right). The split and merge of the targets made it difficult to track small interaction groups directly, in particular when interaction groups are near to each other.

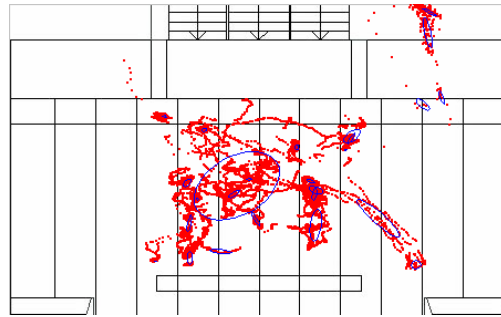


Fig. 5. Detected positions of the small groups for the cocktail party recording and the clusters learned by EM algorithm

To build up a visual model for the changing interaction groups in the scene, we applied a multidimensional EM clustering algorithm [1] to the positions on the hall map as well as the angle and the ratio of first and second moment of the bounding ellipses of all targets. The EM algorithm was initially run with a high number of clusters, while constantly eliminating those with too weak contribution to the whole model.

Finally, 27 clusters were identified for the cocktail party recording. Fig. 5 indicates the positions of all targets as well as the clusters learned by EM on the hall map.

The observations are provided by the automatic speech detector and by the visual model built up by EM. The observations provided by the visual model are the dominant clusters given the targets in the current video frame, i.e. the clusters of the model with the highest probability of having generated the targets. The tracking system has a frame rate of 16 frames per second, which corresponds to the generation of an observation every 62.5 ms. The histograms of our approach are calculated for the observations coming from the speech activity detector as well as from the visual model. The fusion is done by simply summing the Jeffrey divergence values of the speech detector and visual model histograms. Summing the Jeffrey divergence values of the histograms from different modalities is an easy and efficient way to fuse multimodal information because no data conversions or additional fusion calculations are necessary.

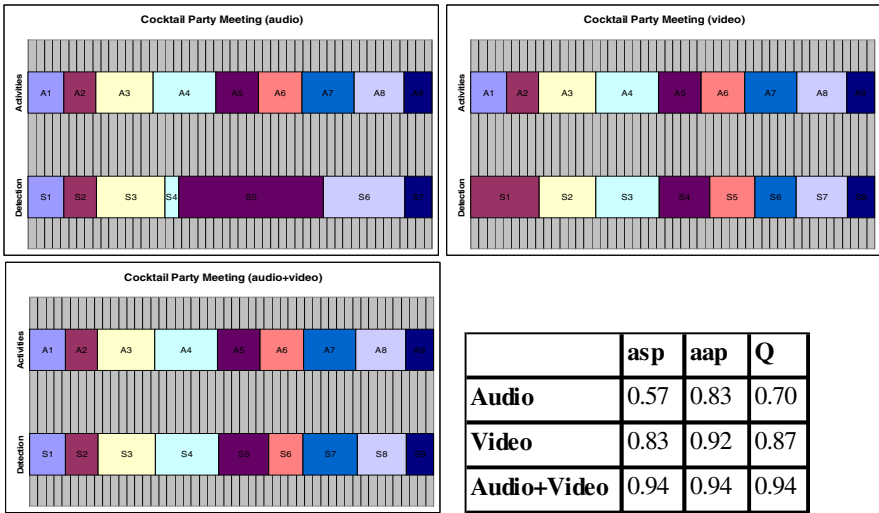


Fig. 6. Group configurations and their detection for the cocktail party (meeting duration=30min 26sec)

The participants formed different interaction groups during the cocktail party meeting. The interaction group configurations were labeled. Our approach has been applied to the speech detector observations, the visual model observations, and both the speech detector and the visual model observations. Fig. 6 shows the labeled group configurations and the detected segments as well as the corresponding *asp*, *aap* and *Q* values. The results of the segmentation of both audio and video are very good, outperforming the separate segmentations. The *Q* value of the video and audio segmentation is 0.94.

5 Conclusion

We proposed an approach for extracting small group configurations and activities from multimodal observations. The approach is based on an unsupervised method for segmenting meeting observations coming from multiple sensors. We calculate the Jeffrey divergence between histograms of meeting activity observations. The peaks of the Jeffrey divergence curve are used to separate distinct distributions of meeting activity observations. These distinct distributions can be interpreted as distinct segments of group configuration and activity. We measured the correspondence between the detected segments and labeled group configurations and activities for a seminar and a cocktail party recording. The obtained results are promising, in particular as our method is completely unsupervised.

The fact that our method is unsupervised is especially advantageous when analyzing meetings with an increasing number of participants (and thus possible group configurations) and a priori unknown activities. Our method then provides a first segmentation of a meeting, separating distinct group configurations and activities. These detected segments can be used as input for further classification tasks like meeting comparison or meeting activity recognition.

References

1. Bilmes, J. A., *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, Technical Report, University of Berkeley, 1998.
2. Brdiczka, O., Maisonnasse, J., and Reignier, P., *Automatic Detection of Interaction Groups*, Proc. Int'l Conf. Multimodal Interfaces, 2005.
3. Brdiczka, O., Reignier, P., and Maisonnasse, J., *Unsupervised segmentation of small group meetings using speech activity detection*, Proc. Int'l Workshop on Multimodal Multiparty Meeting Processing, 2005.
4. Caporossi, A., Hall, D., Reignier, P., Crowley, J.L., *Robust visual tracking from dynamic control of processing*, Proc. Int'l PETS Workshop, 2004.
5. McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D., *Automatic Analysis of Multimodal Group Actions in Meetings*, IEEE Trans. on Pattern Analysis and Machine Intelligence, March 2005.
6. Muehlenbrock, M., Brdiczka, O., Snowdon, D., and Meunier, J.-L., *Learning to Detect User Activity and Availability from a Variety of Sensor Data*, Proc. IEEE Int'l Conference on Pervasive Computing and Communications, March 2004.
7. Puzicha, J., Hofmann, Th., and Buhmann, J., *Non-parametric Similarity Measures for Unsupervised Texture Segmentation and Image Retrieval*. Proc. Int'l Conf. Computer Vision and Pattern Recognition, 1997.
8. Qian, R. J., Sezan, M. I., and Mathews, K. E., *Face Tracking Using Robust Statistical Estimation*, Proc. Workshop on Perceptual User Interfaces, San Francisco, 1998.
9. Stiefelhagen, R., Steusloff, H., and Waibel, A., *CHIL - Computers in the Human Interaction Loop*, Proc. Int'l Workshop on Image Analysis for Multimedia Interactive Services, 2004.

10. Zaidenberg, S., Brdiczka, O., Reignier, P., Crowley, J.L., *Learning context models for the recognition of scenarios*, Proc. IFIP Conf. on AI Applications and Innovations, 2006.
11. Zhang, D., Gatica-Perez, D., Bengio, S., McCowan, I., and Lathoud, G., *Multimodal Group Action Clustering in Meetings*, Proc. Int'l Workshop on Video Surveillance & Sensor Networks, 2004.

Using Ambient Intelligence for Disaster Management

Juan Carlos Augusto, Jun Liu, and Liming Chen

School of Computing and Mathematics,
University of Ulster at Jordanstown, UK
{jc.augusto, j.liu, l.chen}@ulster.ac.uk

Abstract. This paper presents an architecture to help the decision-making process of disaster managers. Here we focus on a core aspect of this process which is taking decisions in the presence of conflicting options. We exemplify this problem with three simple scenarios related to diverse contexts and provide an explanation on how our system will advice in all these cases.

1 Introduction

Disasters can occur at any time and in any context, from a house to a nuclear plant. Decisions taken in relations to a disaster can alter their development and its consequences dramatically. Usually decisions have been taken by humans but more and more computer-based decision-making support has been accepted and developed. Although humans are usually better than machines to judge complex situations and decide, computers can provide a stress-free view of the situation and compile important amounts of knowledge very quickly.

We provide in Section 2 a glimpse of a system under development which can provide assistance to human decision-makers. This assistance can be given by issuing alerts or offering an evaluation of the situation that can be used to double-check a forthcoming decision. We focus on what we consider one of the core aspects of such a system, their capacity to cope with conflicting decisions and we illustrate a variety of such situations in Section 3. If decisions involving conflicting alternatives have to be taken then a criteria is needed to decide which one is the best alternative. Section 4 describes a framework where this kind of criteria can be formalized and then in Section 5 we show how applying this criteria we can make decisions in relation to the scenarios introduced earlier on.

2 A General Architecture

In order to support disaster management in a broad range of real world cases, we propose a general architecture as shown in Figure 1. Central to the architecture is the notion of a decision support system, which takes in temporal and spatial information of an environment, assesses the situation and provides decision makers with suggestions and plans for tackling the emerging disaster in terms of

available resources. As a key feature of large-scale disasters is that information from different sources at different time is usually inconsistent and conflicting, our decision support system has placed special emphases on the assimilation of heterogeneous information and the mechanisms for resolving conflicts.

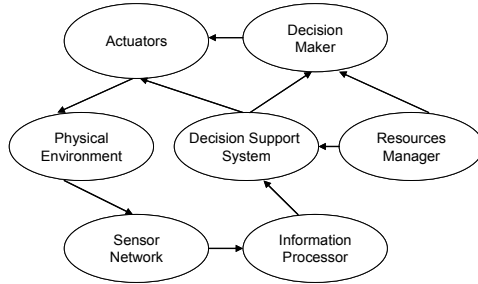


Fig. 1. The General Architecture

The architecture consists of a Sensor Network and Information Processor components. The Sensor Network is intended to capture as much information as possible by providing signals of different modalities and complexities. For example, it could be as simple as a temperature or a sound, and as complicated as video footage and scanning pictures. An Information Processor aims to collect a dedicated form of information, filter noises from raw sensor data and normalize data for the use of the decision support system.

An effective disaster management system should be able to prioritize reaction activities in terms of the value and importance of the affected entities and also the availability and constraints of resources. Both the Decision Support System and the DM refer to such factors when a suggestion or a plan is made. The decision Maker will have the final say by balancing and arbitrating among requirements and consequences. The Actuator will implement actions to minimize the loss or impact of the disaster.

We consider several different environments: a Smart Home, an airport and a paramedics unit doing assessing a victim of a nuclear disaster. We illustrate a loop in the system depicted in Figure 1 and how information is gathered from different sources from the *physical environment* by the *sensor network* and through the *information processor* arrives to the *decision support system* where is made available to the *decision maker* (DM). Here the word “sensor” has to be understood flexibly because the way information will be gathered or “sensed” will vary from a Smart Home to an airport area or the paramedics ambulance. The devices that collect information from the environment and the type of information they process will be different. Still the intelligent environment will be proactively evaluating the situation and advising to benefit the inhabitants of that environment. We focus on how AI can help the DM in the decision process to close the loop leading to *actuators* being applied.

3 Alternative and Conflicting Explanations

Decision-making is about evaluating alternatives, following a particular sequence of events the decision-maker is presented with a query and it is her/his task to assess the possible options to follow and make a decision. In this section we focus on the process of gathering the options available. The next section will consider how to assess these options.

Whenever a query about a particular state S of the system is passed to the DM our methodology will assist the analysis by considering the causal structure leading to that particular state at a particular point in time. The assistance provided comes in the form of an explanation, i.e., it provides details on what events e_1, e_2, \dots are meaningful for that state to be reached and what causal laws r_1, r_2, \dots governing the system are being exercised when events in the world can cause a state S to hold. We will call that a *causal explanation*. There may be more than one possible explanation for how the modelled system can reach S . These options can be consistent with each other in which case there is no conflict. But it may also be, due to the ambiguity or lack of information characteristic in real-time applications, that some of the explanations are contradictory or somehow antagonistic. Analyzing the quality of the explanations and why some of these are contradicting to each other is a difficult and time consuming task, to be avoided when a DM has to react to an imminent hazard. This module of our system is strongly related to previous developments in theory of argumentation ([1,2,3,4]).

We consider possible competing causal explanations $\langle c_1, s_1 \rangle$ and $\langle c_2, s_2 \rangle$ at time t , where c_1 is a causal structure (containing causal rules r_i) explaining why the system may reach state s_1 at time t and c_2 an alternative causal structure (most possibly based in a disjoint set of causal rules r_j) explaining why the system may reach state s_2 at time t . These causal explanations can be such that s_1 contradicts s_2 or contradicts the possibility that c_2 may exist. In this section we focus on the way that these possible causal explanations can be found and also on the potential scenarios that can cause two possible explanations to be mutually contradictory or undermining (similarly to *rebuttal* or *undercutting* arguments [5]).

In our Prolog-based implementation a predicate `holds(State, GoalTime)` gather all the meaningful trees to prove or disprove that a `State` is holding at a `GoalTime`. A specific procedure will also gather the alternative explanations for and against the position that `State` is holding at `GoalTime`. This process will be explained in more detail below with regards to different scenarios. Those scenarios will be represented in the implementation using the notation given at [6] where $\#$ represents negation, $\&$ represents logical 'and', `occurs(ingr(s), t1:t2)` represents the event of the system ingression to a state s at the instant in between $t1$ and $t2$. A *State-State rule* `ssr(a => b)` represents a being true causes state b to be true and a *State-Event rule* `ser(a -> ingr(b))` represents a being true causes an ingression to state b being true.

Hazards can arise in many contexts and for many reasons. The following scenarios describe situations involving hazards and possible explanations for them, each explanation will be suggesting different courses of action.

Scenario 1: automation of decision-making in Smart Homes. Let's assume a smart home is being used to care for a patient with senile dementia [7]. These type of patients tend to forget which activities they were immerse in. It is fairly common that they may go to the kitchen, turn the cooker on and then, as they continue with other activities, forget something was in the cooker. They may go to have a bath or even confuse the time of the day and go to bed to sleep. A system which aids these patients caring for their safety, but also try to allow them to develop their normal daily activities as much as possible without intervention, will have the dilemma of advising a carer to turn off the cooker automatically or to leave it on for more time. Lets assume we have states *cookerOn* (whether the cooker is or not on), *atKitchen* (movement sensor in the kitchen activated), *umt3u* (whether more than three units of time had elapsed), *cu* (cooker unattended) , *alarmOn* (status of the alarm), initially they are all false. We also consider a sequence of events and causal rules:

```
occurs(ingr(atKitchen), 1:2)      occurs(ingr(cookerOn), 2:3)
occurs(ingr(# atKitchen), 5:6)   occurs(ingr(umt3u), 9:10)}

ssr(atKitchen => #cu)
ssr(# atKitchen & cookerOn & umt3u => cu) ssr(cu => alarmOn)
ser(alarmOn -> ingr(#cookerOn))
ssr(#cookerOn => #hazzard) ssr(#cookerOn => #cu) ssr(#cookerOn => #umt3u)
```

For example, if we want to know whether the state *alarmOn* is true at 11 we can use `holds(alarmOn,11)` and the two contender explanations (*S1* and *S2*) will be produced:

```
[#atKitchen & cookerOn & umt3u => cu, cu => alarmOn]
[#atKitchen & cookerOn & umt3u => cu, cu => alarmOn,
 alarmOn -> ingr(#cookerOn), #cookerOn => #alarmOn]
```

Scenario 2: coping with airport disasters. Suppose an airport gathers information that a bomb has been placed there. To react to this specific event and avoid potential disaster, the system contacts national intelligence departments (NIDs) to collect the latest information on airport related terrorist threats. It also checks all surveillance information within the airport, such as video footage, and any reported suspected incidents. Lets assume no feedback has indicated attacks either being planned or underway. Facing with the bomb threat and the “everything is right” information from various security agents, the airport authority has to make a hard decision on what to do next. While keeping the airport operating normally critical, protecting human life from a bomb attack is certainly of paramount importance. Lets assume we have the following simplified description of the decision making process:

```
occurs(ingr(bombAlert), 1:2)      occurs(ingr(nidReportOK), 2:3)
occurs(ingr(localInfoOK), 5:6)   occurs(ingr( strongSource), 9:10)
occurs(ingr(intuitionSaysOK), 9:10)
```

```

ssr(bombAlert => investigateSource)
ssr(bombAlert => contactNIDs)
ssr(bombAlert => checkLocalInfo)
ser(# nidReportOK -> ingr(declareEmergency))
ser(# localInfoOK -> ingr(declareEmergency))
ser(nidReportOK & localInfoOK & strongSource -> ingr(declareEmergency))
ser(nidReportOK & localInfoOK & intuitionSaysOK -> ingr(#declareEmergency))

```

Based on these rules there are two possible explanations ($A1$ and $A2$) in relation to the declaration of emergencies:

```

[nidReportOK & localInfoOK & strongSource -> ingr(declareEmergency)]
[nidReportOK & localInfoOK & intuitionSaysOK -> ingr(#declareEmergency)]

```

Scenario 3: diagnosing level of exposure to radioactivity. This scenario addresses the problems paramedics can face in the context of a nuclear catastrophe. Lets consider a mobile hospital unit that is brought to an area where people has been exposed to dangerous levels of radioactive material. After the incident, time of exposure, distance from radioactive source, and duration of exposure is used as parameters to take decisions (which can be obtained from different sources, sometimes inconsistent and/or incomplete). Symptoms can be immediate or delayed, mild or severe, based on radiation levels. Nausea, vomiting may occur minutes to days after exposure. Time of onset of first symptoms is an important factor in diagnosis and treatment elaboration. Lets assume we have two rules capturing different levels of exposure:

*IF: Vomiting a few minutes after exposure, diarrhea in less than an hour, fever in less than an hour, severe headache, possibly unconsciousness
THEN: Supportive/palliative care is needed.*

*IF: Onset of vomiting less than 30 minutes after exposure incident, early fever, severe headache, possible parotid pain.
THEN: Intensive care needed beginning day 1.* Information about symptoms

can arrive to the DM at different times and the prescence/absence of some of the conditions are not guarantee to rule out or confirm a particular diagnosis:

```

occurs(ingr(immVomit), 0:1).      occurs(ingr(immFever), 0:1).
occurs(ingr(diarrheaSoon), 2:3).  occurs(ingr(headache), 2:3).
occurs(ingr(possUncons), 4:5).    occurs(ingr(possPP), 4:5).

```

```

ssr(immVomit & feverSoon & headache & diarrheaSoon & possUncons => #intCare)
ssr(vomitSoon & immFever & headache & possPP => intCare)
ser(immVomit -> ingr(vomitSoon))  ser(immFever -> ingr(feverSoon))

```

which leads to the two main alternative possible diagnosis ($N1$ and $N2$):

```

[immVomit->ingr(vomitSoon),
      vomitSoon & immFever & headache & possPP => intCare]
[immFever->ingr(feverSoon),
      immVomit & feverSoon & headache & diarrheaSoon & possUncons => #intCare]

```

4 Resolving Conflict

The knowledge and information available may often be uncertain, partial and possibly even contradictory. But still based on these information, we need to analyze the different options and figure out the way of integrating them for a final decision. As illustrated by the following example, agreement/disagreement among different explanations can be naturally modelled as a partial order, reflecting a system being unable or unwilling to order certain choices, or wishing to delay making such an ordering decision.

Example: We are trying to check if a nuclear disaster may occur at time t from three open sources A , B , and C . We may want to ensure that the decision made matches at least two of the three open sources. Define S to be the set $(0, \neg a, \neg b, \neg c, 1)$, with 1 meaning all three sources agree at time t , 0 meaning that at least two sources do not agree, and so the choice is inadequate, and e.g., $\neg a$ meaning that A disagree at time t but the other two agree. The order $<$ is defined by $0 < \neg a, \neg b, \neg c < 1$, so that $\neg a, \neg b, \neg c$ are said to be incomparable. Notice that here we use \neg for classical negation whilst in the previous section $\#$ represented negation in the implementation.

Hence the partially ordered algebraic structure could be a possible choice for modelling the above relationship. A *lattice* is one kind of partially ordered set with rich properties. The ordering of explanations usually forms a hierarchical structure in the form of a lattice. This lattice can be generated by the preference relations defined over the explanations.

Finally the ability to reason about inconsistent and incomplete models is needed when modelling disaster management. However, this cannot be done effectively using classical (two-valued) logic. Reasoning based on classical logic cannot solve the problem because the presence of a single contradiction results in trivialization—anything follows from $A \wedge \neg A$, and so all inconsistencies are treated as equally bad. We may be forced to make premature decisions about which information to discard. Multi-valued logics allow some contradictions to be true, without the resulting trivialization of classical logic. They are useful for merging information from inconsistent or contradictory viewpoints because they allow us to explicitly represent different levels of agreement, even if they cannot be totally ordered. Lattice-valued logic with truth-values in a lattice is a kind of multi-valued logic and provides a possible way to deal with a set of partially ordered values. When it is necessary to incorporate the notion of time, a temporal lattice-valued logic can be used.

A dynamic lattice-valued logic model framework is defined in terms of the following items: X be the set of propositional variables; L is a truth-value lattice; $L(X)$ is the propositional algebra of the lattice-valued calculus on the set of propositional variables X ; $F_L(L(X))$ is a finite set of L -assertions and premise axioms (a knowledge base); T is the time-related non-empty set of possible worlds; $\gamma_x : L(X) \rightarrow L$ is a valuation from $L(X)$ to L in the world $x \in T$.

In $F_L(L(X))$, L is a lattice structure to evaluate arguments which can be syntactically constructed, e.g., as follows: $< (1, 0), A, B, C >$ where “1” represents the top element and “0” represents the bottom element, A is a set

of arguments; B is a benchmark (what is the minimum value for an argument to be valid/useful?); and C a criteria to compare arguments and build a lattice.

The reasoning and the consistency resolving procedure can take into account both syntactical and semantic aspects. If we consider the $\{0, 1\}$ -valued semantic structure, then we only need to check and resolve the syntactical inconsistency. If we consider the multi-valued (even lattice-valued) semantic, e.g., multi-valued argumentation/explanation (here the semantic could be considered as the support of information source or the credibility of a certain state), then we have to check and resolve both syntactical and semantic inconsistency.

This module of our system is strongly related to previous developments in the theory of lattice-valued logic [8]. A dynamic lattice-valued logic model allows the use of appropriate reasoning methods to process the description of disaster situation and propose reaction plans.

5 Making Decisions

For each of the possible decisions and their associated explanations obtained in Section 3 we can apply the decision criteria embedded in the theory presented in section 4.

Scenario 1: the first alternative explanation lists the states that can be caused resulting in the alarm being on whilst the second one will list those supporting the possibility that the system can go through a sequence of states ending in the alarm being off. Lets call these explanations $S1$ and $S2$ respectively. Given they share the independent states which characterize these two possible developments they also share the list of interesting times to be investigated: $[10, 6, 3, 0]$. As $S1$ is entirely contained in $S2$ and furthermore $S2$ complements and extends $S1$ (is a “richer” explanation) then $S2$ should be preferred to $S1$. So the conclusion of the system will be that the alarm is off at 11 and the explanation is $S2$. This conflict was decided purely on syntactical basis, i.e., the structure of the explanation allow us to take a decision based on the structure of the competing explanations. $S2$ takes into account all that $S1$ has to offer and also brings extra information on what happens after the alarm is on the first time. It states that the alarm does not persist on but is turned off as the cause of concern (the cooker on) cease to be true.

Scenario 2: in this case, although confirmation from NIDs officials and ‘in situ’ evidence of danger suggest to ignore the call and even when the intuition of the DM tells everything is safe, information arrives that the source of the bomb alert can be trusted. If the DM is trying to collate all the supporting evidence from the system in terms of labelling the situation as an emergency or not the system will consider the two possible contending explanations $A1$ and $A2$ for declaring an emergency or ignoring the threat. But at time 10 there is evidence for both, however the system will rely on a partial order of strength for the elements of a possible explanation. Explanations based on reliable sources of information will be considered stronger than those based on intuitions so $A1$ will be preferred.

Scenario 3: at time 5, when all the required information is available, is also known that possible parotid pain is more likely than unconsciousness, so although both are likely to have occurred, given the previous symptoms there are more reasons to believe that the context related to explanation *N1* is the more likely to have occurred and therefore the more advisable decision will be to follow the procedure for intensive care.

The framework described in Section 4 can provide a partial order in between competing alternatives both in a syntactical and a semantic way. The first one is needed in Scenario 1 whilst the later is needed in Scenarios 2 and 3.

6 Conclusions

Here we described a system architecture which is under development. We illustrated a variety of contexts and problems where the decision-making system can operate to assist judgement. Space constraints forced us to consider simplified scenarios and only a few aspects of the decision-making, so we focused here on the illustration of the versatility of the system to consider the different alternatives and to evaluate them. The representation language is based on notions of temporal causality, the conflicting alternatives are possible outcomes associated with their supporting causal structure. The decision criteria is based on a lattice structure which naturally depicts a partial order of preference in between the contending options. This explicit consideration and formalization of the decision criteria is a fundamental difference and advantage with respect to [1,3].

Although we are aware that much development is still needed for our system to be used in real life decision-making, we conjecture in this paper those features listed above are fundamental to the kind of problems to be solved.

References

1. Ferguson, G., Allen, J.: Trips: An integrated intelligent problem-solving assistant. In: Proceedings of 15th National Conference on AI. (1998) 567–573
2. Chesñevar, C., Maguitman, A., Loui, R.: Logical models of argument. *ACM Computing Surveys* **32** (2000) 337–383
3. Fox, J., Das, S.: Safe and Sound, Artificial Intelligence in Hazardous Applications. AAAI Press/MIT Press (2000)
4. Augusto, J.C., Simari, G.R.: Temporal defeasible reasoning. *Knowledge and Information Systems* **3** (2001) 287–318
5. Pollock, J.: Cognitive Carpentry. A Blueprint for How to Build a Person. MIT Press (1995)
6. Galton, A., Augusto, J.C.: Stratified causal theories for reasoning about deterministic devices and protocols. In: Proceedings of TIME2002. (2002) 52–54
7. Augusto, J., Nugent, C.: The use of temporal reasoning and management of complex events in smart homes. In de Mántaras, R.L., Saitta, L., eds.: Proceedings of ECAI, IOS Press (2004) 778–782 August, 22–27.
8. Xu, Y., Ruan, D., Qin, K., Liu, J.: Lattice-Valued Logic: An Alternative Approach to Treat Fuzziness and Incomparability. Springer-Verlag (2003)

Dynamic Scene Reconstruction for 3D Virtual Guidance

Alessandro Calbi¹, Lucio Marcenaro², and Carlo S. Regazzoni¹

¹ DIBE, University of Genova, 16145 Genova, Italy
carlo@dibe.unige.it
<https://www.isip40.com>

² TechnoAware S.r.l, Via Greto di Cornigliano 6, 16100 Genova, Italy
lucio.marcenaro@technoaware.com
<http://www.technoaware.com>

Abstract. In this paper a system is presented able to reproduce the actions of multiple moving objects into a 3D model. A multi-camera system is used for automatically detect, track and classify the objects. Data fusion from multiple sensors allows to get a more precise estimation of the position of detected moving objects and to solve occlusions problem. These data are then used to automatically place and animate objects avatars in a 3D virtual model of the scene, thus allowing users connected to this system to receive a 3D guide into the monitored environment.

1 Introduction

Many algorithms have been studied during the last years for automatic 3D reconstruction [1,2] from image analysis for application in different fields. In [3] a semi-automatic system is described that is based on a 3D reconstruction of a museum environment, obtained by a stereo vision sensor: proposed system is able to detect interesting events and to guide the users into the museum. A visualization system for ambient intelligence based on an augmented virtual environment that fuses dynamic imagery with 3D models in a real-time display to help observers comprehend multiple streams of temporal data and imagery from arbitrary views of the scene is presented in [4].

One of the fundamental task for an ambient intelligence application is automatic objects tracking and classification. Researchers developed many specific solutions [5] but no optimal algorithm exists to solve the tracking problem in all real situations. As the complexity of the scene increases and occlusions between static and non-static objects occur [6], performances of standard tracking and classification algorithms typically decrease. Multi camera systems have been often used for overcoming the occlusion problem. Collins et al. in [7] propose understanding algorithms to automatically detect people and vehicles, seamlessly track them using a network of cooperating active sensors, determine their three-dimensional locations with respect to a geospatial site model, and present this information to a human operator who observes the system through a graphical user interface.

In this paper a multi sensor system is described that is able to detect, track and classify multiple moving objects. A three-dimensional model of the observed area is automatically updated by the tracking system and dynamic avatars are maintained within

the model. Such a system can be effectively used also for virtual guidance of users because it allows to reproduce the entire ambient evolution of the scene and to compute the path that users have to follow to reach their destination.

In section 2 processing modules for detecting and tracking moving objects are described; section 3 shows specific modules for multi-camera supervision, while section 4 deals with proposed 3D viewer and guidance module. Finally results are showed in section 5 and conclusions are drawn in section 6.

2 Detection and Tracking

In order to realistically reproduce a real environment and interactions between moving objects and to re-create them into a 3D model, the ambient intelligence system needs sophisticated modules for image analysis and object tracking and classification. Such processing modules allow the automatic comprehension of semantic contents of the image sequence. The primary objective of the system is the phase of *detection*, that is the automatic identification of the moving objects in the scene (entities perceived as different respect a reference background). Subsequently, the system has to evaluate and follow their position in time (*tracking*), being able to extract suitable information to describe the actions performed by the objects (*classification*) themselves. The last phase, therefore, will consist into recognize behaviors (see Figure 1).

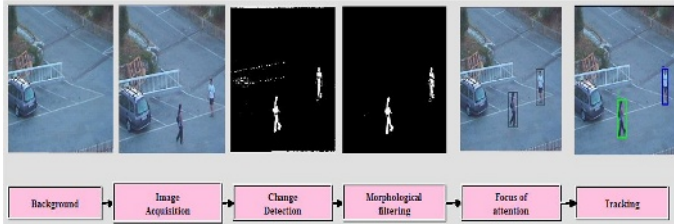


Fig. 1. Scheme of the principal modules of a tracking system

The algorithms adopted by the system to pursue previous objectives can be subdivided into several logical modules, on the basis of the task they have to complete: in particular, it is possible to subdivide basic processing modules into three different main categories:

- *Low level modules* are responsible of extracting interesting data from acquired raw images (image acquisition, change detection, morphological filter, background updating, focus of attention);
- *Middle level modules* are able to get contextual information previously extracted from video sequence and to derive a semantic description of the observed world (blobs matching, feature extraction);
- *High level modules* are responsible to track objects features to keep the history of the temporal evolution of each blob; through classification algorithms [8] these modules are able to classify detected objects.

3 Multi Camera Modules

In order to increase the area covered by a single sensor and to manage the situation of occlusion between the blobs a multi camera approach is adopted. The structure of a multi camera system is based on three steps [9]: Data alignment, Data association, State Estimation.

Data Alignment is needed in order to make the data comparable: dealing with video cameras, this step issues are related to:

- Temporal alignment: the sensors are synchronized to compare features referring to the same instant using a NTP (Network Time Protocol) server.
- Spatial alignment: through a joint cameras calibration procedure it's possible to obtain the correspondences between each image plane and the absolute *world coordinates*, exploiting geometric and optic features of each sensor. The calibration procedure is based on the Tsai algorithm described in [10].

Data Association consists on the *m-ary* decision process among the objects in the fields of view of the used cameras. Many different features are extracted to let the system autonomously adapt the association to different situation occurring in the scene. The use of different features has the advantage to extract in every instant and among the others the better discriminating feature, which will be responsible of the greater separation among the classes we are trying to distinguish in the decision process.

In order to be able to manage such different data and obtain a coherent representation, we define independent similarity functions connected to the information obtaining from each feature; each function provides an autonomous similarity coefficient yielding continuous values distributed between 0 and 1. The feature functions are based on measures in the map reference system (in term of position and speed of the blobs) and in the image plane (valuating the shape factor and chromatic characteristics). The results provided by each function leads to define an *Object Similarity Coefficient* (OSC), calculated as the mean value of the previous coefficients. To apply the criterion and choose the correct associations we seek for the highest values for each object in a camera field of view compared to all the objects in all the cameras image planes with a field of view overlapped with the first.

Once data are aligned and objects associated, the *state estimation* phase performs the actual redundant information exploitation: when the single cameras positioning data are available, they can be fused simply through the use of mean values. But when objects are not well separated in the image plane, a little more care must be put in the estimation phase.

In our system we consider 3 cases:

- if the objects to associate are well separated in both the fields of view, we use the position mean value;
- if the objects result occluded in the field of view of one of the sensors, we use the position computed by the other;
- if both the fields of view present occlusions, we apply the location data related to the objects' couple with the *strongest* OSC value in the association phase.



Fig. 2. 3D vision of the scene

4 3D Viewer and Guide

A visual 3D modality has been developed for the appliance of localization, detection and navigation. The idea is to provide to the users entering into the monitored areas a virtual ambient that reproduces in details the rooms and the ambient interested by the system of ambient intelligence. This ambient has the aim to represent in real time isolated zones of the areas where, for instance, the system detects the presence of other users. This result is performed following the next steps:

1. Creation of a three-dimensional map of the ambient of interest;
2. Importation of the model into a 3D ambient engine;
3. Real time acquisition of the spatial coordinates of the objects and them classification;
4. Implementation of virtual cameras;
5. Equipment the system of the required intelligence to evaluate the minimum path from the current position to the destination.

Data provided by the modules of tracking and classification have to be filtered as a consequence of the error's propagation inherent to the image processing. Therefore, these data are stabilized by using Kalman [11] and median filters [12]. Once the position, speed and class of each blob are filtered, it's possible to send them to a server machine which task is to acquire and elaborate data from the sensors and to forward them to the 3D maker. At this time, the system is able to represent the virtual model: figure 2 shows the virtual 3D representation of the scene.

In figure 3(a) is represented the 3DGuide interface where, under the buttons of connection, selection of the destination and of the virtual camera, is presented the textual guidance message. One of the most relevant benefit of the 3D virtual viewer is surely the possibility to change the point of view (figure 3(b)): in this manner, the users can use the 3D model itself placing virtual cameras into the model selecting the best point of view to reach the destination or to see other users moving into the environment.

The three-dimensional vectorial model has been generated by using Autodesk AutoCAD and 3D Studio Max [13] software using precise measurements of the environment; afterward, this model is imported into a 3D graphical engine. We adopted open-source libraries: the library OpenSceneGraph [14] provides the rules to build the model; with

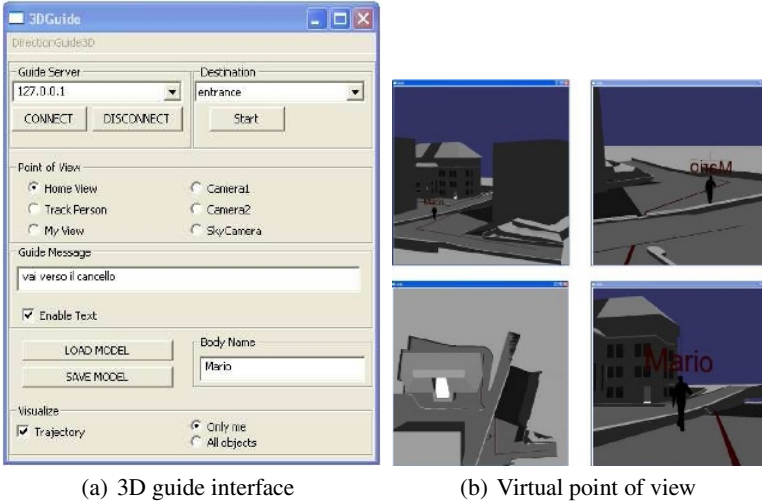


Fig. 3. 3DGuide interface and Virtual generated points of view of the scene

the library Cal3D [15] each object can perform a lot of movements as walk, run, turn, stand, etc.; eventually the last library used by the proposed system, ReplicantBody [16], allows to animate the human model by integrating [14] and [15].

5 Results

In this section we present the most significant results related to the multi-camera data fusion process and the effects of tracking errors to 3D virtual reality rendering. In first instance, we evaluate the goodness of the strategy of data association examining sequences with 4 mutually occluding moving objects, in the form of association rate confusion matrices: in the principal diagonal cells the rate of correct association is reported while the crossing values in the other cells define the wrong associations. Some associations were discarded defining the data belonging to the NC class.

Table 1(a) and table 1(b) contain the result matrix for the 4 objects sequences respectively reporting results with the use of position feature and with the complete set of the chosen 4 features.

Table 1. Association rate confusion matrix: 2 cameras, 4 objects sequences

(a) $OSC \equiv f(P)$						(b) $OSC \equiv f(P, V, S, C)$					
	1i	2i	3i	4i	NC		1i	2i	3i	4i	NC
1j	90.0	3.3	0.0	0.0	6.7	1j	97.7	0.0	0.0	0.0	2.3
2j	3.3	87.7	0.0	0.0	9.0	2j	0.0	97.7	0.0	0.0	9.0
3j	0.0	0.0	85.5	4.5	10.0	3j	0.0	0.0	93.3	1.1	5.6
4j	0.0	0.0	3.3	91.1	5.6	4j	0.0	1.1	0.0	98.9	0.0
NC	6.7	9.0	11.2	4.5		NC	2.3	1.2	6.7	0.0	

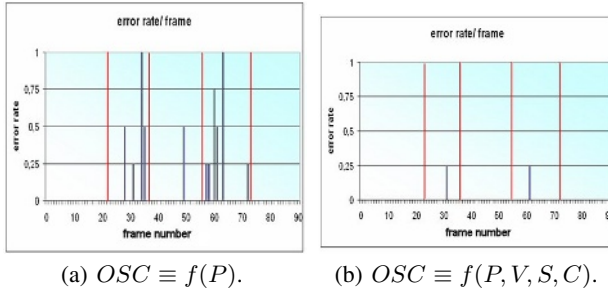


Fig. 4. Histogram of wrong objects associations

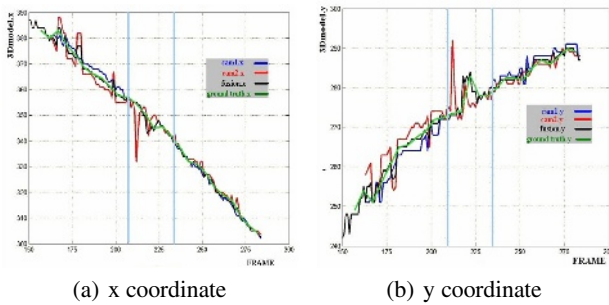


Fig. 5. 3D coordinates of the same object observed by camera1 and camera2, fused and filtered coordinate and ground truth

Presented association results are referred to a sample of 4 objects sequence observed along time: 90 frames contain two occlusion phases where association errors are much more frequent due to the failing of the position and often of the color features. The Y axis has discrete values $\in [0, 1/4, 2/4, 4/4]$ with the meaning of the number of erroneous associations on the total of four. It is easy to be convinced of the higher number of errors presented in figure 4(a) where the sole position is used, comparing to figure 4(b), where the 4-feature set is exploited.

Another interesting result regards the evaluation of the position of the objects estimated in the 3D model and the true position of the objects in the real coordinates (ground truth). As we can see in figure 5(a) and 5(b), the error between the coordinates of the same object observed by camera 1 (in blue) and by camera2 (in red) vs. the ground truth is major than the correspondingly error between fused filtered and filtered position vs. the ground truth. Also in this case, the graphs show that the worst situation for a single camera model happens in the situation of occlusion of the blobs (represented by the azure lines in the pictures): instead, this error is smaller for the fused coordinates.

Lastly, it's possible to compare the computational cost of the 3D vision with the video analysis of each cameras in term of bandwidth and CPU computational load. The 3D model updating requires the reception of a packet composed by integer values: three integers for the identifier label, three for position coordinates, three for speed

components and one for the class of the object. So, each object requires 13×32 bit = 416 bit. For ambient intelligence applications we can consider a transmission of 3 packets/second, that implies a bandwidth of around 1248bps.

Instead, if we consider the transmission of the video sequences acquired by a single camera, using colored images with size 720×480 pixels, 24 frame/second and using an MPEG2 coding, we need a rate from 4 to 6 Mbps. Obviously, n cameras require $n \times (4 - 6)$ Mbps.

Another result is evident in the comparison of the computational load of the CPU. Using a pc configured with a Pentium 4 processor, 2.66 Ghz and 512 MB of RAM, the 3D reconstruction of the scene requires the 35.3% of the CPU load; instead, single camera and dual camera tracking demand 68.6 and 93.4 CPU load percentage.

Table 2. Comparison of the bandwidth occupation and of the CPU load between 3D, single and camera cameras vision

	bandwidth (Kbps)	% CPU load
3D vision	1.5	35.3
Single camera	4000 – 6000	68.6
Dual camera	8000 – 12000	93.4

The previous results imply that, while multi camera tracking is possible only using pc with high performances, the 3D reconstruction of the scene is allowed also with less capable devices as tablet pc, with the great advantage of portability.

6 Conclusions

In this paper algorithms able to process images from a multi-camera ambient intelligence system and extract features of detected moving objects have been presented. Semantic information extracted from the scene is used by the system for update a dynamic virtual 3D model of the guarded environment. Synthetic automatically-generated 3D scene can be used by an user to be guided into the environment by selecting an arbitrary point of view of the considered area.

Acknowledgments

The work was partially supported by the project Virtual Immersive COMMunication (VICOM) founded by the Italian Ministry of University and Scientific Research (FIRB Project).

References

1. T. Rodriguez, P. Sturm, P. Gargallo, N. Guilbert, A. Heyden, J. M. Menendez, and J. I. Ronda, "Photorealistic 3d reconstruction from handheld cameras," *Machine Vision and Applications*, vol. 16, no. 4, pp. 246–257, sep 2005. [Online]. Available: <http://perception.inrialpes.fr/Publications/2005/RSGGHMR05>

2. M. Fiocco, G. Boström, J. G. M. Gonçalves, and V. Sequeira, "Multisensor fusion for volumetric reconstruction of large outdoor areas." in *3DIM*, 2005, pp. 47–54.
3. S. Bahadori and L. Iocchi, "A stereo vision system for 3d reconstruction and semi-automatic surveillance of museum areas," in *AI*IA 2003: Advances in Artificial Intelligence, 8th Congress of the Italian Association for Artificial Intelligence, Pisa, Italy, September 23-26, 2003, Proceedings of Workshop Intelligenza Artificiale per i Beni Culturali*, Pisa, Italy, Sept. 2003.
4. I. O. Sebe, J. Hu, S. You, and U. Neumann, "3d video surveillance with augmented virtual environments," in *IWVS '03: First ACM SIGMM international workshop on Video surveillance*. New York, NY, USA: ACM Press, 2003, pp. 107–112.
5. R. T. Collins, C. Biernacki, G. Celeux, A. J. Lipton, G. Govaert, and T. Kanade, "Introduction to the special section on video surveillance." *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 7, pp. 745–746, 2000.
6. D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift." in *CVPR*, 2000, pp. 2142–.
7. R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance," in *Proceedings of the IEEE*, vol. 89, no. 10, October 2001, pp. 1456–1477. [Online]. Available: citeseer.ist.psu.edu/collins01algorithms.html
8. T. H. Reiss, "The revised fundamental theorem of moment invariants," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 8, pp. 830–834, 1991.
9. S. Piva, A. Calbi, D. Angiati, and C. S. Regazzoni, "A multi-feature object association framework for overlapped field of view multi-camera video surveillance systems," in *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, Como, Italy, 15-16 September 2005, pp. 505–510.
10. R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," pp. 221–244, 1992.
11. G. Welch and G. Bishop, "An introduction to the kalman filter," Chapel Hill, NC, USA, Tech. Rep., 1995.
12. Y. Zhao and G. Taubin, "Real-time median filtering for embedded smart cameras." in *ICVS*. IEEE Computer Society, 2006, p. 55.
13. (2006) The AutoDesk website. [Online]. Available: <http://www.autodesk.com/>
14. (2006) The Open Scene Graph website. [Online]. Available: <http://www.openscenegraph.org/>
15. (2006) The Cal3D website. [Online]. Available: <http://cal3d.sourceforge.net/>
16. (2006) The ReplicantBody website. [Online]. Available: <http://sourceforge.net/projects/replicantbody/>

An Introduction to Fuzzy Propositional Calculus Using Proofs from Assumptions

Iwan Tabakow

Institute of Applied Informatics, Wrocław University of Technology, Poland
iwan.tabakow@pwr.wroc.pl

Abstract. The subject of this paper is fuzzy propositional calculus. The proposed approach is related to the basic fuzzy propositional logics, i.e. to each of the following three most important such systems (in short: BL): Łukasiewicz's, Gödel's, and product logic. The logical calculi considered here are based on a system of rules that define the methods used in proofs from assumptions.. To simplify the considered proofs some set of laws called also 'primitive rules' is next introduced. It was shown that any fuzzy propositional formula provable under Hájek's axioms of the logic BL is also provable under the above-proposed approach.

1 Introduction

In general, the following two main directions in fuzzy logic can be distinguished: fuzzy logic in the broad sense and fuzzy logic in the narrow sense, i.e. in *senso stricto*. The *broad sense fuzzy logic* has become a common buzzword in machine control. The *basic strict fuzzy logic* (in short: BL) is a strict fuzzy logic system using the logic of continuous t-norms. This system was developed in [5]. The most of the studies in this area focus attention on methodological problems, e.g: compactness, consistency, decidability or satisfiability of t-tautologies [1,5,7,11], on various proving techniques [2,8] or also introducing some new t-norms [11]. A survey of different such systems was given in [9].

The subject of this paper is fuzzy propositional calculus. Without loss of generality, the proposed approach is related to the basic fuzzy propositional logics, i.e. to each of the following three most important such systems: Łukasiewicz's, Gödel's, and product logic (in short: Ł-, G-, and π -BL systems). There exist two classical approaches in constructing of the propositional calculus: the *axiomatic approach* and the *approach from assumptions*. In general, the actual research methodology and extensions have been related to the Hájek's axiomatic approach. A new system based on assumptions is presented below. An inductive definition of the notion of a fuzzy propositional formula is first introduced. The proposed method is next presented. To simplify the considered proofs, as in the classical approach, some set of laws called also 'primitive rules' is next introduced. And finally, it was shown that any fuzzy propositional formula provable under Hájek's axioms of the logic BL is also provable under the above-proposed approach.

2 Basic Notions

Let \otimes be a binary operation over $[0,1]$ which is commutative, associative, monotonic, and has 1 as unit element. Any such operation is called to be a *t-norm*. The t-norm operator provides the characterisation of the AND operator. The dual t-conorm \oplus (called also: *s-norm*), characterising the OR operator, is defined in a similar way having 0 as unit element (a more formal treatment is omitted, see : [12]).

The following t-norms are assumed below (for any $a, b \in [0,1]$):

$a \& b =_{df} \max\{0, a + b - 1\}$	the Łukasiewicz's strong conjunction,
$a \wedge b =_{df} \min\{a, b\}$	the logical operation minimum, and
$a \cdot b =_{df} ab$	the usual arithmetic product
with the corresponding dual t-conorms:	
$a \vee b =_{df} \min\{1, a + b\}$	the Łukasiewicz's strong disjunction,
$a \vee b =_{df} \max\{a, b\}$	the logical operation maximum, and
$a \nabla b =_{df} a + b - ab$	the algebraic sum.

Next we shall say that \otimes and \oplus are *t-conjunction* and *t-disjunction*, respectively. The *fuzzy implication* connective is sometimes disregarded but is of fundamental importance for fuzzy logic in the narrow sense. The *implication* and *negation connectives* are assumed under [3,4,5,6].

The used names for the primitive and/or derived rules given below are in accordance with the Łukasiewicz's symbols of negation, conjunction, disjunction, implication, and equivalence denoted by N, K, A, C, and E, respectively. The following (generalised *primitive*) rules are considered below: $-C$ (*rule of detachment for implication or omitting an implication*), $\pm K$ (*rules of joining/ omitting a t-conjunction*), $\pm A$ (*rules of joining/omitting a t-disjunction*), and $\pm E$ (*rules of joining/omitting an equivalence*). The *rule of substitution for equivalence* is denoted by SR. Some additional rules are also used, such as: $\pm N$ (*rules of joining / omitting double negation*), CR (*implication rule*), CC (*the law of transposition or contraposition of implication*), NA (*rule of negating a t-disjunction*), NK (*rule of negating a t-conjunction*), NC (*rule of negating an implication*), Toll (*rule modus tollendo tollens*), TC (*the law of transitivity for implication*), MC (*the law of multiplication of consequents of two or more implications having the same antecedents*). Some additional inference rules of the first-order predicate logic calculus are also used below, such as: the rule of *negating an universal quantifier* (denoted by $N\forall$), the rules of *omitting an universal and an existential quantifiers* (denoted by: $-\forall$ and $-\exists$, respectively). The following abbreviations are also introduced below: a, aip, ada, and contr., denoting: *assumption(s)*, *assumption(s) of indirect proof*, *additional assumption of a proof*, and *contradiction*, respectively.

2 The Approach from Assumptions

In the fuzzy propositional calculus any formula is constructed by using the following three kinds of symbols: (i) propositional variables (denoted by $p, q, r, s, \dots, p_1, p_2, \dots$),

(ii) some fuzzy connectives (depending on the used system), e.g. such as: the Łukasiewicz's (strong) fuzzy conjunction, disjunction, implication, logical equivalence, and negation, denoted by: $\&$, $\underline{\vee}$, \Rightarrow , \Leftrightarrow , and \sim , respectively or the Gödel's (weak) fuzzy conjunction, disjunction, implication, logical equivalence, and negation, denoted by: \wedge , \vee , \Rightarrow , \Leftrightarrow , and \sim , respectively or also the product logic's fuzzy conjunction, disjunction, implication, logical equivalence, and negation, denoted by: \cdot , $\bar{\vee}$, \Rightarrow , \Leftrightarrow , and \sim , respectively and also (iii) parentheses (left: '(' and right: ')'). The following two truth and falsity constants are used below: $\underline{1}$ and $\underline{0}$ (denoted also by \top and \perp respectively, i.e. the truth degree 1 and 0, corresponding to 'T' and 'F' in the classical case). To minimise the number of used parentheses in an expression, some priorities for logical connectives can be introduced. The following convention is assumed below [13]: \sim , \otimes , \oplus , \Rightarrow , \Leftrightarrow (i.e. the symbol of negation binds more strongly than the symbol of t-conjunction, the last binds more strongly than the symbol of t-disjunction, etc.), where $\otimes \in \{\&, \wedge, \cdot\}$ and $\oplus \in \{\underline{\vee}, \vee, \bar{\vee}\}$ are depending on the used system (Ł- , G- or $\pi\text{-BL}$). The set of *fuzzy propositional formulae* (called equivalently *fuzzy propositional expressions*, in short: *expressions* or also *sentential formulae*) of this propositional calculus can be considered as the smallest set which includes propositional variables, and which is closed under the operations of forming the negation, conjunction, disjunction, implication and equivalence. Hence, any propositional variable can be considered as an expression and also the compound formulae formed from them by means of the corresponding logical functors. More formally, the following inductive definition can be used (a generalisation of the classical one [13]).

Definition 1

A *fuzzy propositional formula* is:

1. Any propositional variable,
2. If φ and ψ are some fuzzy propositional formulae, then such formulae are also: $\sim(\varphi)$, $(\varphi) \otimes (\psi)$, $(\varphi) \oplus (\psi)$, $(\varphi) \Rightarrow (\psi)$, and $(\varphi) \Leftrightarrow (\psi)$,
3. Every fuzzy propositional formula in this propositional calculus either is a propositional variable or is formed from propositional variables by a single or multiple application of rule (2). And this should be in accordance with the used definitions of fuzzy connectives, depending on the considered system, where $\otimes \in \{\&, \wedge, \cdot\}$ and $\oplus \in \{\underline{\vee}, \vee, \bar{\vee}\}$.

Any evaluation of fuzzy propositional variables can be considered as a mapping v assigning to each fuzzy propositional variable p its truth-value in $[0,1]$. This extends to each fuzzy propositional formula φ as an evaluation of propositional variables in φ by truth degrees in $[0,1]$ [5,6]. Below by $v_{\otimes}(\varphi) \in [0,1]$ (in short: $\varphi \in [0,1]$, e.g. $\varphi = a \in [0,1]$) we shall denote the *logical value of the fuzzy propositional formula* φ with respect to \otimes (in short: wrt \otimes). In a similar way, e.g. by $\varphi \leq \psi$ we shall denote: $\varphi = a$, $\psi = b$, and $a \leq b$ ($a, b \in [0,1]$).

Definition 2 [4]

Let \otimes be a given continuous t-norm and $v_{\otimes}(\varphi) \in [0,1]$ be the *logical value of the fuzzy propositional formula* φ wrt \otimes . So, we shall say φ is *t-tautology*, *t-thesis* or

also *standard BL-tautology* of that calculus if $v_{\otimes}(\varphi) = 1$ for each evaluation of propositional variables in φ by truth degrees in $[0,1]$ and each continuous t-norm.

Let φ be a fuzzy propositional formula obtained under Definition 1. Hence, as in the classical case, the main task is to verify if φ is a t-thesis. Unfortunately, the usefulness of Definition 2 seems to be limited considering arbitrary t-norms. Next we shall assume only t-norms related to the basic fuzzy propositional logics. Any such t-thesis is said to be a *strong t-thesis* (or equivalently: *strong t-tautology*, *strong standard BL-tautology*). The last definition can be modified assuming “t-norm dependence”, i.e. the following definition can be introduced.

Definition 3

Let \otimes be a given continuous t-norm and $v_{\otimes}(\varphi) \in [0,1]$, or $v(\varphi)$ if \otimes is understood, be the *logical value of the fuzzy propositional formula φ wrt \otimes* . We shall say φ a *weak t-thesis* if $v_{\otimes}(\varphi) = 1$ for each evaluation of propositional variables in φ by truth degrees in $[0,1]$.

The *proof* in the fuzzy propositional calculus can be interpreted as a process of joining new lines by using some primitive or derived rules and/or other theses in accordance with the used assumptions. The proposed approach is an extension of the classical one [13]. An illustration is given in the next example.

Example 1

Consider the following well-known classical law (of addition an arbitrary proposition to the antecedent and consequent of a given implication):

$$(p \Rightarrow q) \Rightarrow (p \vee r \Rightarrow q \vee r)$$

This law can be proved both using a direct or also an indirect proof. In general, the indirect proof is a more universal approach, but corresponding to more proof lines than the direct one (if it exists). The following indirect proof can be obtained.

Proof:

- (1) $p \Rightarrow q$ {1,2 / a}
- (2) $p \vee r$
- (3) $\sim (q \vee r)$ {aip}
- (4) $\sim q$
- (5) $\sim r$ {4,5 / NA: 3}
- (6) p { $\neg A$: 2,5}
- (7) q { $\neg C$: 1,6}
- contr. \square {4,7}

Since any $\varphi \in \{a\} \cup \{aip\}$ is assumed to be a true formula (i.e. true in any interpretation), the following proof technique can be equivalently introduced.

- (1) $\forall p, q \in \{0,1\} (p \Rightarrow q = 1)$ {1,2 / a}
- (2) $\forall p, r \in \{0,1\} (p \vee r = 1)$
- (3) $\sim \forall q, r \in \{0,1\} (q \vee r = 1)$ {aip}

- (4) $\exists q, r \in \{0,1\} (q \vee r \neq 1)$ $\{N\forall: 3\}$
(5) $q_0 \vee r_0 = 0$ $\{-\exists: 4\}$
(6) $(q_0 = 0) \wedge (r_0 = 0)$ $\{df.'\vee': 5\}$
(7) $q_0 = 0$
(8) $r_0 = 0$ $\{7,8 / -K: 6\}$
(9) $p_0 \Rightarrow q_0 = 1$ $\{-\forall: 1\}$
(10) $p_0 \vee r_0 = 1$ $\{-\forall: 2\}$
(11) $p_0 \leq q_0$ $\{df.'\Rightarrow': 9\}$
(12) $(p_0 = 1) \vee (r_0 = 1)$ $\{df.'\vee': 10\}$
(13) $p_0 = 0$ $\{7,11\}$
(14) $r_0 = 1$ $\{-A: 12,13\}$
contr. \square $\{8,14\}$

The above proof technique can be easily extended to the whole interval $[0,1]$. Hence, the following implication is satisfied.

Thesis 1 (law of addition an arbitrary fuzzy proposition to the antecedent and consequent of a given implication)

$$(p \Rightarrow q) \Rightarrow (p \oplus r \Rightarrow q \oplus r)$$

Proof (e.g. Ł-BL: $\oplus =_{df} \underline{\vee}$):

- (1) $\forall p, q \in [0,1] (p \Rightarrow q = 1)$ $\{1,2 / a\}$
(2) $\forall p, r \in [0,1] (p \underline{\vee} r = 1)$
(3) $\sim \forall q, r \in [0,1] (q \underline{\vee} r = 1)$ $\{aip\}$
(4) $\exists q, r \in [0,1] (q \underline{\vee} r \neq 1)$ $\{N\forall: 3\}$
(5) $q_0 \underline{\vee} r_0 \neq 1$ $\{-\exists: 4\}$
(6) $q_0 + r_0 < 1$ $\{df.'\underline{\vee}': 5\}$
(7) $p_0 \Rightarrow q_0 = 1$ $\{-\forall: 1\}$
(8) $p_0 \underline{\vee} r_0 = 1$ $\{-\forall: 2\}$
(9) $p_0 \leq q_0$ $\{df.'\underline{\Rightarrow}': 7\}$
(10) $p_0 + r_0 \geq 1$ $\{df.'\underline{\vee}': 8\}$
(11) $p_0 + r_0 \leq q_0 + r_0$ $\{+r_0: 9\}$
(12) $q_0 + r_0 \geq 1$ $\{10,11\}$
contr. \square $\{6,12\}$

In accordance with our considerations, T1 is a strong t-thesis. Also, the following example strong t-theses are satisfied (the corresponding proofs are omitted here).

Thesis 2 (law of compound constructive dilemma)

$$(p \Rightarrow q) \otimes (r \Rightarrow s) \otimes (p \oplus r) \Rightarrow q \oplus s. \square$$

Thesis 3 (law of compound destructive dilemma)

$$(p \Rightarrow q) \otimes (r \Rightarrow s) \otimes \sim (q \oplus s) \Rightarrow \sim (p \oplus r). \square$$

Thesis 4 (De Morgan's law of negating a t-disjunction)

$$\sim (p \oplus q) \Leftrightarrow \sim p \otimes \sim q. \square$$

Thesis 5 (De Morgan's law of negating a t-conjunction)

$$\sim (p \otimes q) \Leftrightarrow \sim p \oplus \sim q. \square$$

Thesis 6 (rule modus tollendo tollens)

$$(p \Rightarrow q) \otimes \sim q \Rightarrow \sim p. \square$$

Thesis 7 (law of transitivity for implication)

$$(p \Rightarrow q) \otimes (q \Rightarrow r) \Rightarrow (p \Rightarrow r). \square$$

Thesis 8 (laws of exportation and importation)

$$p \otimes q \Rightarrow r \Leftrightarrow p \Rightarrow (q \Rightarrow r). \square$$

Thesis 9 (law of reduction ad absurdum)

$$(p \Rightarrow q \otimes \sim q) \Rightarrow \sim p. \square$$

Thesis 10 (law of transposition or contraposition of implication)

$$p \Rightarrow q \Leftrightarrow \sim q \Rightarrow \sim p. \square$$

Thesis 11 (law of the hypothetical, called also conditional, syllogism)

$$(p \Rightarrow q) \Rightarrow ((q \Rightarrow r) \Rightarrow (p \Rightarrow r)). \square$$

Thesis 12 (A4 [4,5,6])

$$p \otimes (p \Rightarrow q) \Rightarrow q \otimes (q \Rightarrow p). \square$$

Thesis 13 (Hauber's law of converting implications)

$$(p \Rightarrow q) \otimes (r \Rightarrow s) \otimes (p \oplus r) \otimes \sim (q \otimes s) \Rightarrow (q \Rightarrow p) \otimes (s \Rightarrow r). \square$$

Thesis 14

$$(p \oplus q \Rightarrow r) \Rightarrow (p \Rightarrow r) \otimes (q \Rightarrow r). \square$$

Thesis 15 (law of multiplication of consequents)

$$(p \Rightarrow q) \otimes (p \Rightarrow r) \Leftrightarrow (p \Rightarrow q \otimes r). \square$$

The following example weak t-theses are satisfied (\mathcal{L} -BL only): $\sim \sim p \Leftrightarrow p$ (the law of double negation, and hence the rules $\pm N$), $p \Rightarrow q \Leftrightarrow \sim p \vee q$ (the law of implication, i.e. the rule CR), $\sim (p \Rightarrow q) \Leftrightarrow p \& \sim q$ (the law of negating an implication, i.e. NC), $p \Rightarrow p \wedge p$ (idempotence of t-conjunction: G-BL only, the opposite implication is strong t-thesis related to $-K$), the axiom [4]: $\sim \sim p \Rightarrow ((p \Rightarrow p \otimes q) \Rightarrow q \otimes \sim \sim q)$ is not satisfied for G-BL, the well-known absorptive and distributive axioms are satisfied only in G-BL, the law of addition of antecedents ($p \Rightarrow r) \otimes (q \Rightarrow r) \Leftrightarrow p \oplus q \Rightarrow r$ is satisfied only in G- and π -BL (strong t-thesis is the opposite implication: see T14), etc. The corresponding proofs are omitted. The following strong t-theses are considered as a generalisation or extension of the classical primitive and derived rules. In general, any strong t-thesis can be considered as a new derived rule.

$$\begin{array}{l}
 \text{-C: } \frac{\varphi \Rightarrow \psi}{\varphi}, \quad \text{+K: } \frac{\varphi}{\psi}, \quad \text{-K: } \frac{\varphi \otimes \psi}{\varphi \wedge \psi \wedge \frac{\varphi}{\psi}}, \quad \text{+A: } \frac{\varphi}{\varphi \oplus \psi}, \\
 \\
 \text{-A: } \frac{\varphi \oplus \psi}{\sim \varphi}, \quad \text{+E: } \frac{\varphi \Rightarrow \psi}{\psi \Rightarrow \varphi}, \quad \text{-E: } \frac{\varphi \Leftrightarrow \psi}{\varphi \Rightarrow \psi \wedge \psi \Rightarrow \varphi \wedge \frac{\varphi \Rightarrow \psi}{\psi \Rightarrow \varphi}}, \\
 \\
 \text{Toll: } \frac{\varphi \Rightarrow \psi}{\sim \psi}, \quad \text{CC: } \frac{\varphi \Rightarrow \psi}{\sim \psi \Rightarrow \sim \varphi}, \quad \text{NA: } \frac{\sim(\varphi \oplus \psi)}{\sim \varphi \wedge \sim \psi \wedge \frac{\sim \varphi}{\sim \psi}}, \quad \text{NK: } \frac{\sim(\varphi \otimes \psi)}{\sim \varphi \oplus \sim \psi}, \\
 \\
 \text{SR: } \frac{\varphi \Leftrightarrow \psi}{\chi \Leftrightarrow \chi(\varphi // \psi)}, \quad \text{TC: } \frac{\varphi \Rightarrow \psi}{\varphi \Rightarrow \chi}, \quad \text{MC: } \frac{\varphi \Rightarrow \psi}{\varphi \Rightarrow \psi \otimes \chi}.
 \end{array}$$

Example 2

Consider the proof of T1 under the above-introduced primitive rules. This proof can be realised as follows.

- | | | |
|-----|---------------------------------|-------------------|
| (1) | p \Rightarrow q | {1,2 / a} |
| (2) | p $\underline{\vee}$ r | |
| (3) | \sim (q $\underline{\vee}$ r) | {aip} |
| (4) | \sim q | |
| (5) | \sim r | {4,5 / NA, NK: 3} |
| (6) | p | {-A : 2,5} |
| (7) | q | {-C : 1,6} |
| | contr. \square | {4,7} |

The following proposition was shown (using the notion of t-consequence).

Proposition 1 (BL-provability)

Any t-thesis provable under the Hájek's axiomatic approach of the logic BL is also provable under the above-proposed approach from assumptions. \square

4 Conclusions

The above-considered approach in constructing of the fuzzy propositional calculus can be considered as an extension of the classical one. The presented technique of proofs from assumptions can be simplified by using some rules, i.e. a set of strong t-theses, as a generalisation or extension of the classical primitive and derived rules. Since some generalised classical rules are satisfied only as weak t-theses, the general study of consistency and completeness, i.e. the provability by the assumed rules may

be a difficult problem (whether every t-thesis provable by the rules of the assumptional system of this calculus is a true formula). On the other hand, the provability properties of the Hájek's axiomatic approach of the logic BL can be preserved by using the classical first-order predicate logic calculus. Finally, the presented approach from assumptions seems to be more attractive, more simpler and natural than the axiomatic one wrt the practical use, e.g. in the case of approximative reasoning or automated theorem proving (using Gentzen's sequents or also Robinson's resolution), etc. And so, essentially most remains to be done there.

References

- [1] Cintula P., Navara M., Compactness of fuzzy logics. *Fuzzy Sets and Systems*, vol. 143 (2004) 59 – 73.
- [2] Di Lascio L., Analytic fuzzy tableaux. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*. Springer-Verlag (2001) 434 – 439.
- [3] Gottwald S., *Many-valued logic*. The Stanford Encyclopedia of Philosophy (Zalta E.N. ed), The Metaphysics Research Lab at the Center for the Study of Language and Information, Stanford University, Stanford, CA (2000) 14pp.
- [4] Hájek P., *Fuzzy Logic*. The Stanford Encyclopedia of Philosophy (Zalta E.N. ed), The Metaphysics Research Lab at the Center for the Study of Language and Information, Stanford University, Stanford, CA (2002) 7pp.
- [5] Hájek P., *Metamathematics of fuzzy logic*. Kluwer Acad.Publ., Dordrecht (1998) 297pp.
- [6] Hájek P. and Godo L., *Deductive systems of fuzzy logic (a tutorial)*. Tatra-mountains mathematical publications, vol. 13 (1997) 35 – 66.
- [7] Horčík R., Navara M., Consistency degrees in fuzzy logics. *Proc 9th Int. Conf. Information Processing and Management of Uncertainty, ESIA – Université de Savoie, Annecy, France* (2002) 399 – 403.
- [8] Horčík R., Navara M., Validation sets in fuzzy logics. *Kybernetika* vol.38 (2002) 319 – 326.
- [9] Klement, E.P., Navara, M., *A survey of different triangular norm-based fuzzy logics*. *Fuzzy Sets and Systems*, vol. 101 (1999) 241 – 251.
- [10] Klement E.P., Navara M., Propositional fuzzy logics based on Frank t-norms. In: Dubois D., Klement E.P., Prade H (eds.). *Fuzzy sets, Logics and Reasoning about Knowledge*. Applied Logic Series, vol.15, Kluwer, Dordrecht (1999) 17 – 38.
- [11] Navara M., Satisfiability in fuzzy logics. *Neural Network World* vol.10 (2000) 845 – 858.
- [12] Schweizer B. and Sklar A., *Associative functions and abstract semi-groups*. *Publ. Math. Debrecen* 10 (1963) 69 – 81.
- [13] Słupecki J. and Borkowski L., *Elements of mathematical logic and set theory*. Oxford, New York, Pergamon Press (1967) 349pp.

Fuzzy-Neural Web Switch Supporting Differentiated Service

Leszek Borzemski¹ and Krzysztof Zatwarnicki²

¹ Institute of Information Science and Engineering
Wroclaw University of Technology, Wroclaw, Poland
leszek.borzemski@pwr.wroc.pl

² Department of Electrical Engineering and Automatic Control,
Technical University of Opole, Opole, Poland
KZatwarnicki@po.opole.pl

Abstract. New designs of the Web switches must incorporate a client-and-server-aware adaptive dispatching algorithm to be able to optimize multiple static and dynamic services providing quality of service and service differentiation. This paper presents such an algorithm called FNRD (Fuzzy-Neural Request Distribution) which operates at layer-7 of the OSI protocol stack. This algorithm assigns each incoming request to the server with the least expected response time estimated using the fuzzy approach. FNRD has ability for learning and adaptation by means of a neural network feedback loop. We demonstrate through the simulations that our dispatching policy is more effective than state-of-the-art layer-7 reference dispatching policies CAP (Client-Aware Policy) and LARD (Locality Aware Request Distribution).

1 Introduction

A cluster-based Web system is commonly used in locally distributed architecture for Web sites. Web servers in a cluster work collectively as a single Web resource in the network. Typical Web cluster architecture includes a Web switch that distributes user requests among Web servers. Web switch plays an important role in WWW performance boosting. It dispatches user requests among Web servers forming a Web-server cluster (briefly, Web cluster) distributed within a local area, providing at the same time a single user interface that is one URL and one virtual IP address. From the point of view of the end-user the Web switch retains transparency of Web cluster. Other “deeper” architectural details of Web cluster such as back-end application and database servers are also kept from the user by means of the Web switch design.

We developed a fuzzy-neural request distribution algorithm FNRD for dispatching user requests and selecting the target server in a Web cluster system driven by a Layer-7 Web switch [4]. In this paper we present new features of our algorithm and show its capabilities for providing quality and differentiation of service. We demonstrate through a set of new simulation experiments that proposed FNRD dispatching policy aiming to optimize request response time on a Web server outperforms other

state-of-the-art layer-7 dispatching policies, including the CAP (Client-Aware Policy) [9] and LARD (Locality Aware Request Distribution) [14] algorithms which are regarded in the literature as “points of reference” policies.

In FNRD we employ the artificial intelligence methods, namely fuzzy sets and neural networks to use effectively uncertain information. FNRD’s decisions are dynamic, i.e., they are based on current knowledge about server state and about the request. The algorithm classifies the client request to a distinguished class of requests, estimates the Web servers’ expected response times for that request and assigns the request to the server with the least expected response time. In the estimation of the expected response times FNRD uses the fuzzy estimation mechanism. The learning algorithm is used to refine information about possible response times. For this purpose we use a neural network based feedback approach. A particular feature of our approach is that our algorithm makes no *a priori* assumptions about the way the Web servers work.

A good survey of Web switch different dispatching algorithms can be found in [7]. The advent of QoS (Quality of Service) sensitive Internet applications raises a need for building Web site services supporting quality and differentiation of service [8, 10]. A step towards achieving this is to design Web sites that would provide their services taking into account the standpoint of an individual request. However, the load and sharing performance strategies that are now in common use are essentially developed focusing on administrator’s point of view, aiming at optimal utilization of site resources. Differently, the FNRD policy allocates a request to a target server to minimize expected response time of individual request. As this policy is optimal from the standpoint of an individual request it can support user perceived performance. Moreover it is an example of the development of the flexible adaptive dispatching policy. Such good FNRD’s features are achieved due to a fuzzy-neural decision model which is a new approach in the area of HTTP request distribution. However such or simpler (i.e. fuzzy) models were used in other cases of computer systems, for example, a fuzzy model for load balancing in distributed systems was studied in [11].

The remainder of this paper is organized as follows. In Section 2, we outline the fuzzy-neural Web switch. In Section 3, we present the simulation testbed and analyze experimental results. In Section 4, we give our final remarks.

2 FNRD Web Switch

FNRD Web switch is placed in front of the set of Web servers using a two-way architecture where each URL request arrives through the Web switch and the resulting resource is sent back also through the Web switch. Resources are fully replicated on all servers therefore each server can handle a request. The request is directed to the server that is expected to provide the fastest request response time from among all other servers. By minimizing the response time for each resource from a requested page we can minimize the total page latency time. The request response time is the time between opening and closing of the TCP session that is established between the Web switch and a target Web server to get a resource.

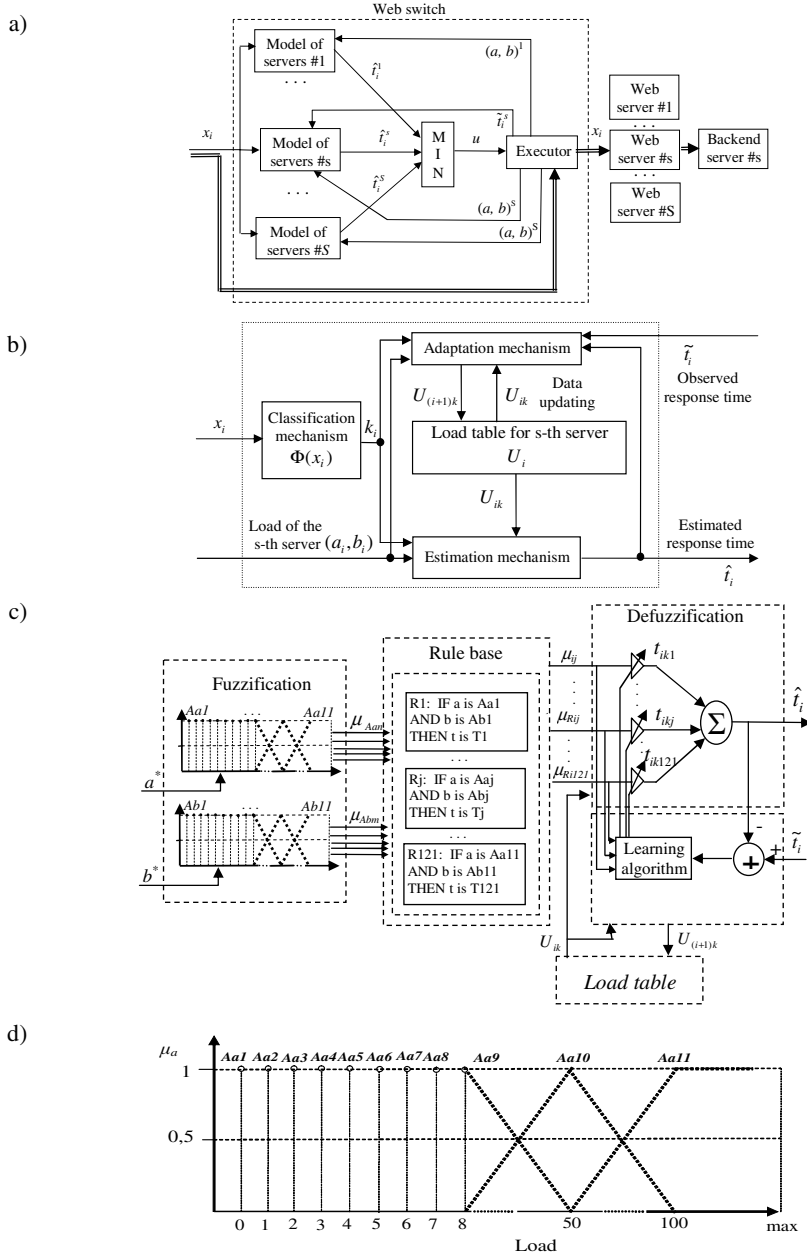


Fig. 1. Fuzzy-Neural Web switch: (a) Architecture; (b) Model of Server module; (c) Estimation mechanism; (d) Membership functions for inputs

FNRD-based switch employs a neuro-fuzzy model (Fig. 1). The inputs are: HTTP request x_i (i is a request index) and a Web server load given by a load tuple (a_i, b_i) , where the load a_i is the load of the Web server - it is estimated by the number of active TCP connections on Web server, and the load b_i is the load of the back-end database server – it is estimated by the number of TCP actively serviced requests on a database server. Both loads are indexed by i . The classification mechanism categorizes the incoming request to a certain class $k \in \{0, \dots, K-1\}$ based on the information on the size of the requested object. The estimation mechanism calculates the request response time \hat{t}_i using information U_{ik} from the knowledge base $U_i = [U_{i0}, \dots, U_{ik}, \dots, U_{i(K-1)}]^T$ called the load table. The system is based on the fuzzy-neural model the structure of which is shown in Fig. 1c. It is based on the Mamdani's model [12, 15] in which only a defuzzification module was transformed into an artificial neuron network. The input in the model described is the load tuple. Based on the semantics of both loads we define eleven membership functions $\mu_{Aa1}(a^*), \dots, \mu_{Aaj}(a^*), \dots, \mu_{Aa11}(a^*)$ and $\mu_{Ab1}(b^*), \dots, \mu_{Abj}(a^*), \dots, \mu_{Ab11}(a^*)$, respectively. The shape and distribution of the fuzzy sets underlying control variables correspond to the way in which control is expressed in the model. The model is capable of balancing between the high precision of the estimation of the request response time desired for the loads lower than (and around) the system operating point, and the computation time needed to calculate the estimator. Increasing the number of fuzzy sets generally increases the precision of the estimation but at a price of more number of rules and more computations. The operating point is the number of requests that are handled by the system simultaneously in typical situations. It was determined on the basis of empirical data as 8 requests. Consequently, we use 8 singleton membership functions for the first eight values of inputs and triangular and trapezoid membership functions distributed for higher loads as shown in Fig 1d.

The rule base is as follows: $R1$: IF $(a=Aa1)$ AND $(b=Ab1)$ THEN $(t=T1)$... Rj : IF $(a=Aam)$ AND $(b=Abn)$ THEN $(t=Tj)$... $R121$: IF $(a=Aa11)$ AND $(b=Ab11)$ THEN $(t=T121)$, where a, b - server loads (number of active connections), t - response time, $Aa1, \dots, Aa11, Ab1, \dots, Ab11$ - servers' load fuzzy sets, and $T1, \dots, T121$ - sets of output t . The rule base contains 121 rules. The fuzzy sets of outputs $T1, \dots, T121$ are singletons and define the values $t_{ik1}, \dots, t_{ikj}, \dots, t_{ikj}$. In our model they are the weights of artificial neurons used in the defuzzification module; these weights are stored in the knowledge base for each class k of the requested objects separately. For the presented rules the degree μ_{Rj} of the membership of rule Rj is equal to the membership degree

to the fuzzy set of input $\mu_{Aaj}(a^*, b^*)$. A linear neuron used in the defuzzification block calculates the response time on the server using the formula $\hat{t}_i = \sum_{j=1}^{121} t_{jki} \mu_{Rj}$.

The new weights $t_{(i+1)kj}$ are worked out in accordance with the learning rule for the ADALINE neuron, each time after completing the request, according to the formula $t_{(i+1)kj} = t_{ikj} + \eta \mu_{Rj} (\tilde{t}_i - \hat{t}_i)$, where \hat{t}_i is the estimated request response time, \tilde{t}_i is the observed request response time, and η is the learning rate index ($\eta=0.13$).

3 Simulation and Results

The model of a cluster-based Web system used in our simulation is shown in Fig. 2. We assumed that both the Web switch and local area network were fast enough and do not introduce significant delay that might influence results. The main delay in request servicing is assumed to be introduced by the servers in the Web cluster, i.e. the Web and database servers.

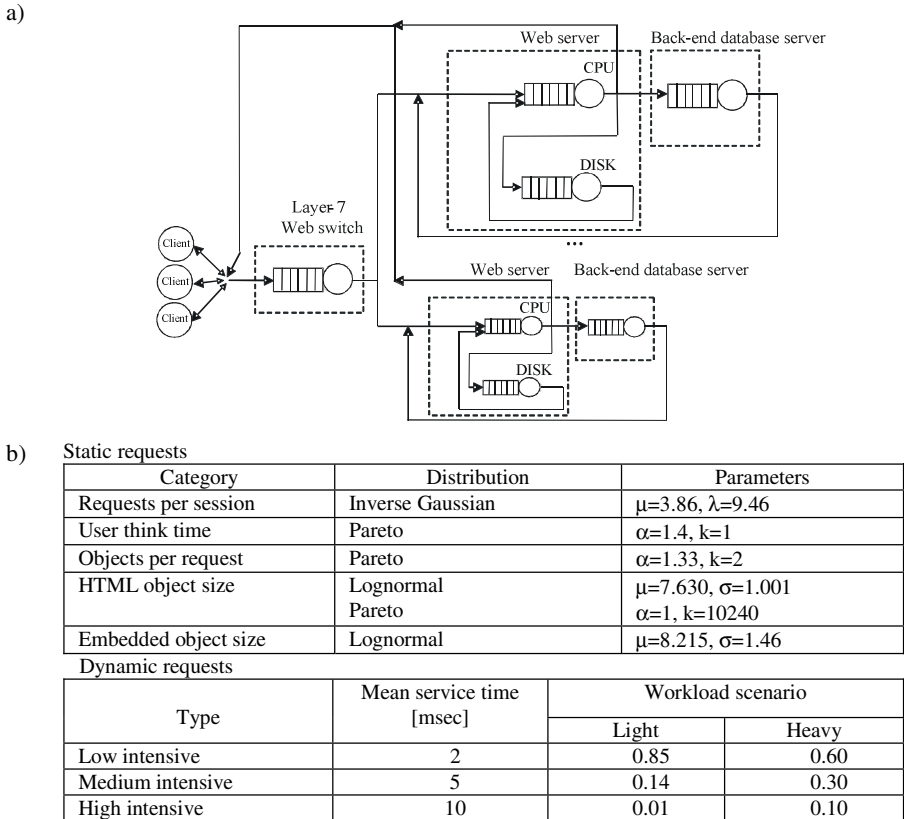


Fig. 2. (a) A simulation model; (b) Workload model parameters

Our CSIM19-based [13] simulator runs using almost the same assumptions about CPU speed, amount of memory, number of disks and other parameters as described in [2, 6, 14] where the processing costs are calculated for Pentium II 300 MHz PC with FreeBSD 2.2.5. In the simulations due to the new server and network developments we decrease all time parameters by the factor of ten. We also reduce database service times. The connection establishment and teardown costs are set at 14.5 μ s of CPU time each, while transmit processing incurs 4.0 μ s per 512 bytes. Disc costs are the following: reading a file from disk has a latency of 2.8 ms, the disk transfer time is

41.0 μ s per 4 KByte. For files larger than 44 KBytes, an additional 1.4 ms (seek plus rotational latency) is charged for every 44 Kbytes of file length in excess of 44 KBytes. The Least Recently Used (LRU) cache replacement policy is used; but files with a size of more than 500 KB are never cached. The total memory size used for caching is 85 MB.

In our previous work we evaluated FNRD via trace-driven simulation [5] (using real trace data from the 1998 FIFA World Cup Web site [1]) and via benchmarking experiments [4] (using a Web switch prototype, real Web servers and own benchmarking system). In this paper to evaluate the system, we have performed a set of new simulation experiments using CSIM19 package, a well accepted simulation tool used for building Web systems models. We evaluate the system performance using the workload model incorporating the most recent research on Web load which is heavy-tailed and self-similar [1, 3].

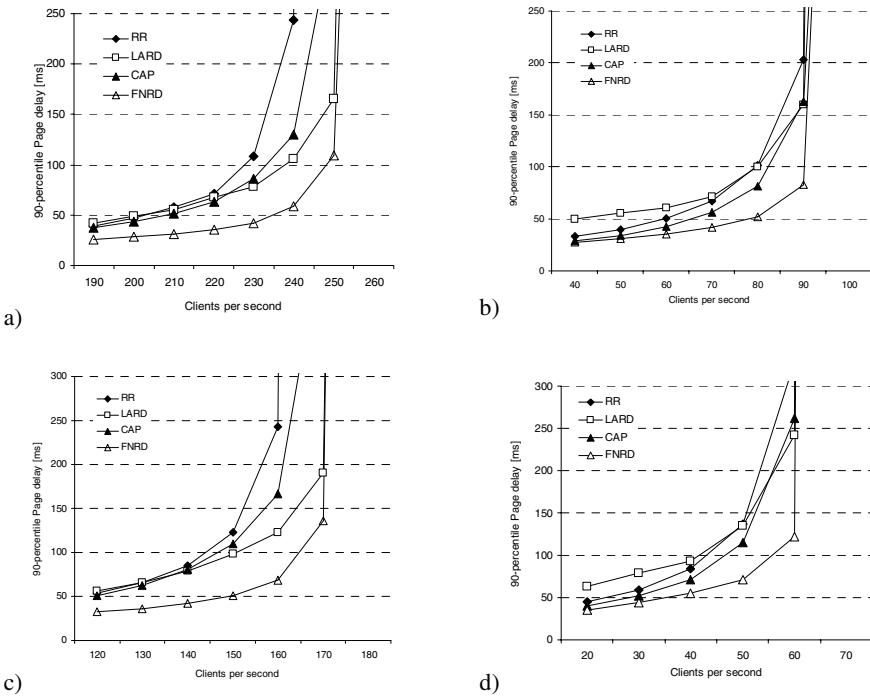


Fig. 3. 90-percentile of page delay vs. number of clients per second. Light workload: (a) 20%, (b) 60% of dynamic requests. Heavy workload: (c) 20%, (d) 60% of dynamic requests

We consider two classes of requests: *static* and *dynamic*. Static requests are serviced directly by the Web server whereas in the dynamic requests the objects are dynamically generated by the back-end servers. Fig. 2b shows the probability distributions and their parameters we use in the workload model. The size of dynamic requests is simulated based on the same size distribution as the static ones. They are

additionally classified into three classes, namely high, medium and low intensive, according to the workload size incurred while database processing. The service time on the database server for a dynamic request is modeled according to a hyper-exponential distribution with the parameters as shown in Fig. 2b. In the simulation we assumed three workload compositions consisting of 20 and 60% of dynamic requests for both *light* and *heavy* scenarios. We simulated the browsing of the Web site of the total size of 200 MB size equipped with 4 servers (each server is consisting of the Web and database server) in the cluster. We measured the page latency time being the sum of object's service times versus the number of Web clients serviced by the cluster per second. We assume that the service of every object takes the same time. The results of the simulations are presented in Fig. 3 and Fig. 4. For comparison reasons, we consider three well known dispatching policies: *Round Robin* (RR) [7], *Locality Aware Request Distribution* (LARD) [14] and the *Client-Aware Policy* (CAP) algorithm [9]. RR is content-blind baseline policy that allocates arriving requests based on a round robin discipline. LARD is known to be the best content-aware dynamic policy that is aimed to partite files among servers and increase RAM cache hit rate. CAP has been introduced as the newest competitor to LARD.

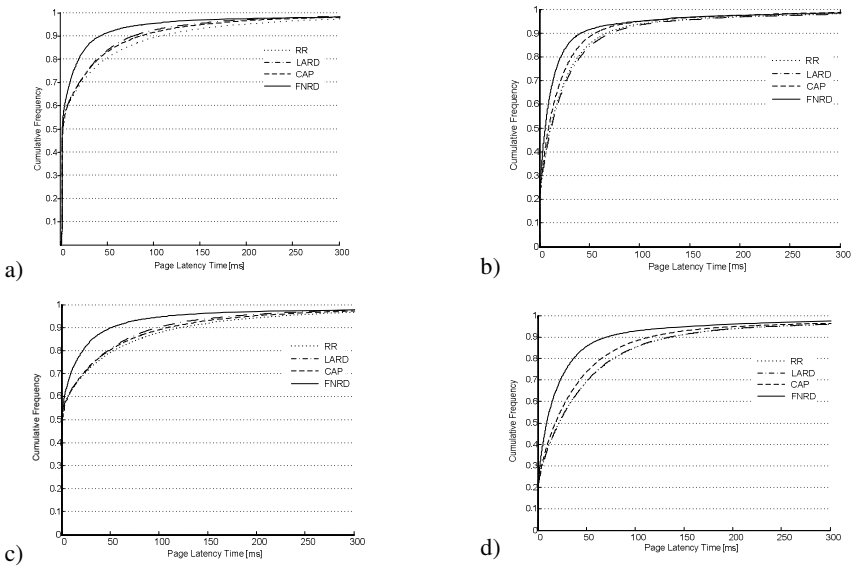


Fig. 4. Cumulative frequency of page latency time. Light workload: (a) 20%, (b) 60% of dynamic requests. Heavy workload: (c) 20%, (d) 60% of dynamic requests.

All results show that FNRD outperforms all competitors for the whole range of the load size for all scenarios. FNRD is especially effective when the Web cluster is not heavily loaded – then the page latency time for the competitive algorithms is even two times bigger than for FNRD policy. Especially, FNRD has very good performance for the heavy loads – then its page latency time is several times less than for the competitors. LARD achieves good performance but after FNRD under heavy load however

under light load it works not so well. A *vice versa* behavior can be observed in case of CAP. It performs well generally for both scenarios loaded with more than 50% of dynamic requests. Fig. 4 shows that for all loads and scenarios FNRD guarantees with very high probability that the most of pages are serviced in a time which is much less than for other algorithms.

4 Conclusions

We evaluated a fuzzy-neural HTTP distribution policy called FNRD for Web cluster through the simulation using CSIM19 package. FNRD optimizes request response time. We showed that a neuro-fuzzy approach is useful in the design of the content-aware Web switch. Our algorithm outperformed other competitive dispatching policies including the state-of-the-art content-aware algorithms CAP, LARD, as well as the very popular content-blind RR policy. Especially, FNRD achieved very good performance for the heavy loads – then its page latency time is several times less than for the competitors. Moreover, FNRD guarantees with very high probability that the most of pages are serviced in a time which is much less than for other algorithms.

References

1. Arlit M., Jin T.: A Workload Characterization Study of the 1998 Word Cup Web Site, IEEE Network, May/June (2000) 30-37
2. Aron M., Druschel P., Zwaenepoel W.: Efficient Support for P-HTTP in Cluster Based Web Servers. Proc. Usenix Ann. Techn. Conf., Monterey, CA. (1999)
3. Barford P., Crovella M.E.: A Performance Evaluation of Hyper Text Transfer Protocols. Proc. ACM SIGMETRICS '99, Atlanta, (1999) 188-197
4. Borzemski L., Zatwarnicki K.: A Fuzzy Adaptive Request Distribution Algorithm for Cluster-Based Web Systems. Proc. of 11th Conf. on Parallel, Distributed and Network-based Processing, IEEE CS Press Los Alamitos (2003) 119-126
5. Borzemski L., Zatwarnicki K.: Using Adaptive Fuzzy-Neural Control to Minimize Response Time in Cluster-Based Web Systems. LNAI, Vol. 3528. Springer-Verlag Berlin (2005) 63-68
6. Bunt R., Eager D., Oster G., Williamson C.: Achieving Load Balance and Effective Caching in Clustered Web Servers. Proc. 4th Int'l Web Caching Workshop (1999)
7. Cardellini V., Casalicchio E., Colajanni M., Yu P.S.: The State of the Art in Locally Distributed Web-Server Systems. ACM Comp. Surv. Vol. 34, No. 2 (2002) 263-311
8. Cardellini V., Casalicchio E., Colajanni M., Mambelli M.: Web Switch Support for Differentiated Services. ACM Perf. Eval. Rev., Vol. 29, No. 2 (2001) 14-19
9. Casalicchio E., Colajanni M.: A Client-Aware Dispatching Algorithm for Web Clusters Providing Multiple Services. Proc. WWW10 (2001) 535-544
10. Cheng R.G., Chang C.J.: A QoS-Provisioning Neural Fuzzy Connection Admission Controller for Multimedia Networks. IEEE Trans. on Networking, vol. 7, no. 1, Feb. (1999) 111-121
11. Kwok Y.-K., Cheung L.-S.: A New Fuzzy-Decision Based Load Balancing System for Distributed Object Computing. J. Parallel Distribut. Comput. 64 (2004) 238-253

12. Mamdani E.H.: Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis. IEEE Trans. on Computers, vol. C-26, No.12, Dec. (1977) 1182-1191
13. Mesquite Software Inc. CSIM19 User's Guide. Austin, TX. <http://www.mesquite.com>
14. Pai V.S., Aront M., Banga G., Svendsen M., Druschel P., W. Zwaenpoel, Nahum E.: Locality-Aware Request Distribution in Cluster-Based Network Servers. Proc. of 8th ACM Conf. on Arch. Support for Progr. Languages (1998)
15. Yager R.R., Filev D.: Essentials of Fuzzy Modeling and Control, John Wiley and Sons, New York (1994)

Image Watermarking Algorithm Based on the Code Division Multiple Access Technique

Fan Zhang¹, Guosheng Yang¹, Xianxing Liu¹, and Xinhong Zhang²

¹ Institute of Advanced Control and Intelligent Information Processing,
Henan University

Kaifeng 475001, P.R. China

zhangfan@vip.sohu.com,

{yangguosheng, liuxianxing}@henu.edu.cn

² Department of Computer Center, Henan University,

Kaifeng 475001, P.R. China

hnkfzxh@163.com

Abstract. A multiple watermarking algorithm based on the Code Division Multiple Access (CDMA) spread spectrum technique is presented. Multiple copyright identifiers are convolutional encoded and block interleaved, and the orthogonal Gold sequences are used to spread spectrum of the copyright messages. The CDMA encoded copyright messages are embedded into the wavelet sub-bands. As a blind watermarking algorithm, the copyright identifiers are extracted without the original image. The experimental results show that the proposed algorithm improves the detection BER, and the multiple copyright identifiers have preferable robustness and invisibility.

1 Introduction

Digital watermarking is the process that embeds data called a watermark into a multimedia object such that watermark can be detected or extracted later to make an assertion about the object. Copyright identification is an important application of digital watermarking. A user that receives an image may need to identify the source of a document or the document itself with a high degree of certainty, in order to validate this document for a specific use.

Recently, some watermarking algorithms based on the CDMA (Code Division Multiple Access) technique have been proposed. Ruanaidh [1] presented a watermarking algorithm that spread spectrum the watermark message by a m-sequence based on the DS-CDMA (Direct Sequence CDMA), and embedded the CDMA encoded message into images in the DCT (Discrete Cosine Transform) domain. Silvestre [2] performed the DFT (Discrete Fourier Transform) to the original images and form many independent bands by choosing selected DFT coefficients. Then the watermark message was spread according to the CDMA coding by two orthogonal sequences and embedded in the independent bands. Kohda [3] transformed images from RGB domain to the YIQ domain, and embedded watermark message in independent the CDMA channels that are formed

by the first 15 DCT coefficients of Y , the first 6 DCT coefficients of I and the first 3 DCT coefficients of Q . Vassaux [4] decomposed image into multiple bit planes in spatial domain, which are regarded as the independent CDMA channels, and embedded watermark message in it. Fang [5],[6] proposed a DWT (Discrete Wavelet Transform) based image watermark algorithm using the DS-CDMA techniques to resist cropping attack. Zou [7] proposed a CDMA based multiple-watermark algorithm. Hartung [8] assumed small correlation between the secret key, the image and hides data using spread spectrum in the spatial domain or compressed domain. Mobasseri [9] proposed a CDMA based digital video watermarking algorithm. Digital video is modeled as a bit-plane stream along the time axis. Using a modified m-sequence, bit-planes of specific order are pseudo randomly marked for watermarking.

In the previous research of multiple watermarking, Cox [10] applied perturbation to the first 1,000 largest DCT coefficients of the entire image. The perturbations were drawn from a normal random number generator. Hsu [11] generated the pseudo random sequences using a linear feedback shift register to modulate the multiple watermarks, and embedded the modulated watermark message into images. Wong [12] proposed a scheme to embed multiple watermarks simultaneously in the same watermark space using different secret key. Raval [13] proposed a multiple watermarking algorithm in the wavelet transform domain. Lu [14] extended the "cocktail watermark" to a blind multipurpose watermarking system that serves as both robust and fragile watermarks, capable of detecting malicious modifications if the watermark is known.

This paper presents a CDMA based multiple copyright identification watermarking algorithm. According to the multiple accessing technique of the CDMA system, multiple copyright identifiers are embedded into digital images in wavelet transform domain. The rest of this paper is organized as follows. Section 2 describes the embedding algorithm in detail. The detection algorithm is described in Section 3. Section 4 presents the experimental results to demonstrate the performance of this scheme. The conclusion is drawn in Section 5.

2 Embedding Algorithm

In the watermarking schemes, the watermarking can be considered as a communication process. The image in which the watermark is embedded is the communication channel, and the watermark is the message to be transmitted.

The CDMA is an application of spread spectrum technique. In the CDMA system, multiple users share the same frequency band at the same time. Unique channels are created and each user directly modulates their information by a unique, high bit rate code sequence that is essentially uncorrelated with that assigned to any other user. The number of users on the system at the same time is a function of the number of unique code sequences assigned, and the ratio of the code sequence bit rate to information bit rate. The properties of CDMA just satisfy the requirement of multiple watermarking algorithms.

The convolutional codes are one type of code used for channel coding. The convolutional codes are usually described by two parameters: the code rate and the constraint length. The code rate k/n , is a measure of the amount of added redundancy. A rate $R = k/n$ convolutional encoder with memory order m can be realized as a k -input, n -output linear sequential circuit with input memory m , i.e., inputs remain in the encoder for an additional m time units after entering. Typically, n and k are small integers, $k < n$. The constraint length parameter K , denotes the length of the convolutional encoder. Closely related to K is the parameter $m = K - 1$, which indicates how many encoder cycles an input bit is retained and used for encoding after it first appears at the input to the convolutional encoder. The m parameter can be thought of as the memory length of the encoder. A (n, k, m) convolutional encoder consists of k shift register with m delay elements and with n modulo-2 adders.

The convolutional decoder regenerated the information by estimating the most likely path of state transition in the trellis map. The maximum likelihood decoding means the decoder searches all the possible paths in the trellis and compares the metrics between each path and the input sequence. The path with the minimum metric is selected as the output. So the maximum likelihood decoder is the optimum decoder. In general, a convolutional code (n, k, m) has $2(m - 1)k$ possible states. At a sampling instant, there are $2k$ merging paths for each node and the one with the minimum distance is selected and called surviving path. At each instant, there are $2(m - 1)k$ surviving paths are stored in the memory. When all the input sequences are processed, the decoder will select a path with the minimum metric and output it as the decoding result.

In this paper, a $(2, 1, 2)$ convolutional code with the code rate $1/2$ is used, and the redundant is 1 bit. This convolutional code can be used for the correction of one bit error from four bits. In this scheme, the copyright identifiers are 32×32 binary images. If the copyright identifier is text or other media format, it can be expanded or compressed to 1,024 bits binary sequence, and then turn into a 32×32 binary images. In the convolutional coding, each row of the copyright identifier image, 32 bits, is input to the convolutional encoder, eventually, the convolutional coded copyright message is a 64×32 matrix.

The convolutional coded codeword is block interleaved. The block interleaving is simply done by transmitting them by column after column, and the interleaved depth is 32. The interleaved code corrects single bursts of length 32 or less, and the random errors will corrected by the convolutional coding.

In this paper, a Gold sequences set with length $L = 2^9 - 1 = 511$ is chosen to orthogonally spread spectrum of the multiple copyright identifiers. Firstly, we transform the convolutional and interleaved coded copyright identifiers to a binary sequence \tilde{a}_k , $k = 1, 2, \dots, n$; n is the number of the copyright identifiers. Then we expand sequence \tilde{a}_k according to the chip ratio Cr , and the expanded sequence is \tilde{b}_k . The choice of chip ratio Cr is decided by the size of original images, the size of copyright messages, the convolutional coding rate and the embedding condition in wavelet domain. A particular Gold code sequence \tilde{g}_k is assigned to each copyright identifier. The corresponding expanded sequence \tilde{b}_k

is modulated by Gold code sequence \tilde{g}_k , and the watermark message that we want to embed in the image is the sum of the modulated sequences,

$$\tilde{w}(t) = \sum_{k=1}^n \tilde{b}_k(t) \cdot \tilde{g}_k(t). \quad (1)$$

Watson proposed a mathematical model about the noises detection thresholds of Discrete Wavelet Transform [15]. Watson's perceptual model is human experience based; it satisfies the requirements of Human Vision System (HVS). During image compression in the wavelet domain, the wavelet coefficients are quantized using the same quantization factor in a sub-band. And if we control the watermark amplitude (distortion) under the quantization factor in a wavelet sub-band, the watermark will be invisible. Watson's quantization factor can be written as follow:

$$Q_{\lambda,\theta} = \frac{2}{A_{\lambda,\theta}} a 10^{k(\log \frac{2^\lambda f_0 g_\theta}{\gamma})^2}, \quad (2)$$

where λ and θ are the wavelet level and orientation respectively, γ is the display resolution, $A_{\lambda,\theta}$ is the amplitude of basis function, and a , k , f_0 , g_θ are constants. If the distortion is less than the quantization factors $Q_{\lambda,\theta}$, the distortion will be invisible, namely $Q_{\lambda,\theta}$ is the maximum allowable distortion or watermark amplitude of each coefficient in the wavelet domain.

In this paper, the copyright identifiers are additively embedded in original images. In order to keep the robustness and invisibility, the watermark messages are embedded in the wavelet sub-bands except the wavelet HH1 sub-band with the maximum allowable watermark amplitude. The embedded position in the selected wavelet sub-bands is decided randomly by a PN sequence. Each bit of the PN sequence corresponding to a wavelet coefficient, if the bit is 0, the amplitude corresponding to a wavelet coefficient is not modified, and if the bit is 1, the amplitude corresponding to a wavelet coefficient is modified as follows. Assume that w_1 and w_2 are the maximum and minimum of CDMA encoded watermark sequence respectively, I and I' denote the original image and the watermarked (stego) image respectively, the watermark embedding formula is as follows,

$$I'(i, j) = I(i, j) + \frac{\tilde{w}(t)}{w_1 - w_2} Q_{\lambda,\theta}. \quad (3)$$

3 Detection Algorithm

In the detection of copyright identifiers, a bi-orthogonal 9/7 DWT is used to decompose the original image. The embedded position of watermark in the selected wavelet sub-bands is decided randomly by a PN sequence using the same key. The watermark message is extracted according to the embedded amplitude that is decided by Watson's perceptual model of wavelet transform domain. Then the Gold sequences sets, which are same as using in the watermark embedding, are used to demodulate the spread spectrum watermark message. After the convolutional decoding and interleaved decoding, the multiple copyright identifiers are extracted.

Usually watermarked image may suffer some image processing or attacking, and the extracted watermark may not same as the original watermark. The correlation between the extracted copyright identifiers and the original copyright identifiers can be used as a criterion to estimate if an image is embedded some especial copyright messages.

Assume w, w' denote the original copyright identifiers image and the extracted copyright identifiers image respectively, and the size of image is $m \times n$, then the correlation coefficient between the two images is as follows,

$$Cor(w, w') = \frac{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (w(i, j) - \bar{w}) \cdot (w'(i, j) - \bar{w}')}{\sqrt{\left(\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (w(i, j) - \bar{w})^2 \right) \cdot \left(\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (w'(i, j) - \bar{w}')^2 \right)}}, \quad (4)$$

where $w(i, j)$ and $w'(i, j)$ denote the pixel of original copyright identifiers image and the extracted copyright identifiers image respectively, \bar{w} and \bar{w}' are the mean of $w(i, j)$ and $w'(i, j)$ respectively. The correlation coefficient $Cor(w, w')$ distributed in the interval $[-1, 1]$. If $Cor(w, w') > T$, T is a threshold, then we can judge the image is embedded some especial copyright messages.

4 Experimental Results

In the experiments, 512×512 gray images Lena, Baboon, Fishingboat, Peppers, Couple and Girl are used. The copyright identifiers are 32×32 binary images. Fig. 1 shows the watermarked Baboon image, The PSNR (Peak Signal to Noise Ratio) is 35.78 dB, and the watermarked Peppers, the PSNR is 36.26 dB. The two extracted copyright logo images that are extracted in the condition of no



Fig. 1. Watermarked Baboon image (PSNR is 35.78 dB), watermarked Peppers image (PSNR is 36.26 dB) and extracted copyright logo images without noises or other attacks (BER = 0)

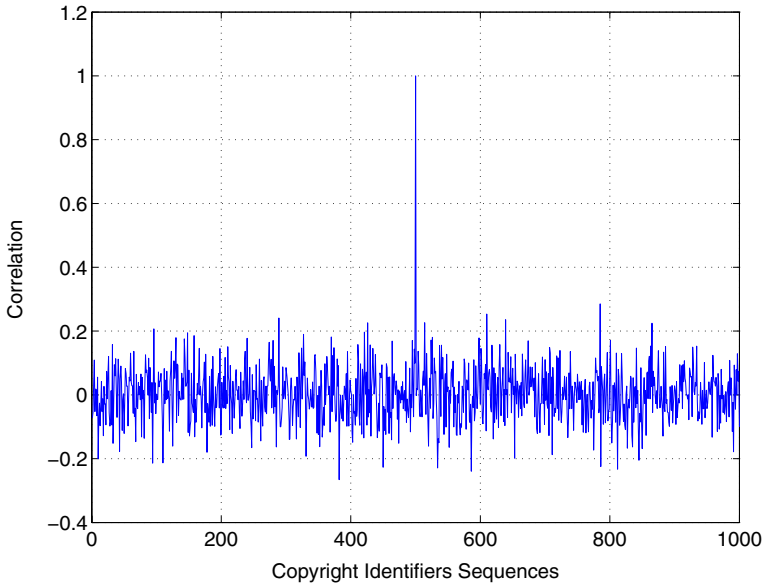


Fig. 2. Experimental result of correlation detection (Fishingboat image)

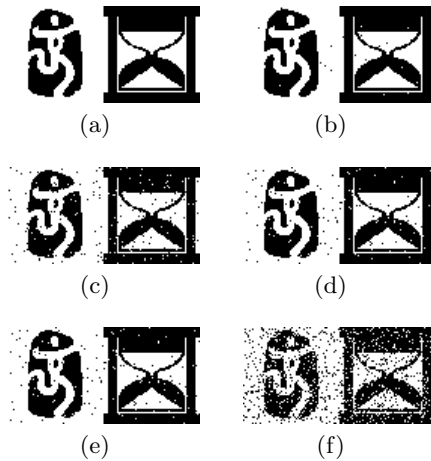


Fig. 3. Robustness performance (Lena image). (a) JPEG compression ($Q=50$), PSNR=31.72dB, BER=0. (b) JPEG compression ($Q=40$), PSNR=30.24dB, BER=0.16%. (c) JPEG compression ($Q=30$), PSNR=28.49dB, BER=2.51%. (d) Gaussian noises ($\sigma^2 = 1$), PSNR=29.32dB, BER=1.67%. (e) Gaussian noises ($\sigma^2 = 4$), PSNR=25.47dB, BER=7.62%. (f) Median filtering (4×4), PSNR=24.32dB, BER=12.38%.

noises or other attacks (Baboon and Peppers images), $BER = 0$, are also show in Fig. 1.

In order to estimate the robustness of this scheme, watermarked image is attacked by JPEG compression, Gaussian noises and median filtering etc. Fig. 2 is the experimental results of correlation detection in the condition of no noises or other attacks (Fishingboat image). The experimental results of robustness performance of Lena image are show in Fig. 3.

Because of the different experimental methods and different parameters, the comparison between the different experimental results is difficult. Comparing with the previous algorithms[4],[5],[6],[7], both of the size of watermark messages and the amplitude of watermark are bigger than above references, the PSNR is a bit smaller than them. For example, as the experimental results of [5],[6], the PSNR of the watermarked images are about 40–45dB, the PSNR of the watermarked images of [7] are about 38–40dB. While in this paper, the PSNR of the watermarked images is 35 dB. A small PSNR means a higher distortion of the watermarked images, but because of the wonderful performance of proposed scheme, the detection BER is even smaller than the previous algorithms. The proposed scheme improves the detection BER profit from the application of the convolutional coding and interleaved coding. When the JPEG compression (the quality factor is 50) is performed in Lena image, the BER of [4] is about 5%–25%the BER of [5],[6] is about 0%–0.2%, while the BER of the proposed scheme is 0 in this case.

5 Conclusion

This paper presents a CDMA based multiple copyright identification watermarking algorithm. Each of the copyright identifiers can be embedded and extracted independently without impacts to each other. As a blind watermarking algorithm, the copyright identifiers are extracted without the original image. The experimental results show that the proposed algorithm improves the detection BER, and the multiple copyright identifiers have preferable robustness and invisibility.

Acknowledgements. This work was supported by the Natural Science Foundation of Henan University under Grant No. 05YBZR009.

References

1. Ruanaidh, J., Pun, T.: Rotation, scale and t translation invariant spread spectrum digital image watermarking, *Signal Processing*, 1998, **66**(3): 303–317
2. Silvestre, G., Dowling, W.: Embedding data in digital images using cdma techniques. In: *Proceedings of 2000 IEEE International Conference on Image Processing*, Vancouver, Canada, 2000, **1**: 589–592

3. Kohda, T., Ookubo, Y., Shinokura, K.: Digital watermarking through CDMA channels using spread spectrum techniques. In: Proceedings of IEEE 6th International Symposium on Spread Spectrum Techniques and Applications, Parsippany, NJ, USA, 2000, **2**: 671–674
4. Vassaux, B., Bas, P., Chassery, J.: A new CDMA technique for digital image watermarking enhancing capacity of insertion and robustness. In: Proceedings of 2001 IEEE International Conference on Image Processing, Thessalonica, Greece, 2001, **3**: 983–986
5. Fang, Y., Wu, S., Huang, J.: DWT-Based CDMA Watermarking Resist Cropping. *ACTA Automatica Sinica*, 2004, **30**(3): 442–448
6. Fang, Y., Huang, J. Shi, Y.: Image watermarking algorithm applying CDMA. Proceedings of the 2003 International Symposium on Circuits and Systems, 2003, **2**: 948–951
7. Zou, F., Lu, Z., Ling, H.: A Multiple Watermarking Algorithm Based on CDMA. In: International Multimedia Conference archive Proceedings of the 12th annual ACM international conference on Multimedia, 2004, 424–427
8. Hartung, F., Girod, B.: Watermarking of uncompressed and compressed video. *Signal Processing*, 1998, **66**: 283–301
9. Mobasseri, B.: Exploring CDMA for watermarking of digital video. Proceedings of the SPIE 1999 Security and Watermarking of Multimedia Contents. San Jose, CA, USA, SPIE, 1999, **3657**: 96–102
10. Cox, I., Kilian, J., Shamoon, T.: spectrum watermarking for multimedia. *IEEE Trans on Image Processing*, 1997, **6**: 1673–1687
11. Hsu, C., Wu, J.: Hidden digital watermarks in images. *IEEE Transactions on Image Processing*, 1999, **8**(1): 58–68
12. Wong, P. Au, O., Yeung, Y.: A novels blind multiple watermarking techniques for images. *IEEE Transactions on Circuits and Systems*, 2003, **13**(8): 813–830
13. Raval, M., Priti, P.: Discrete wavelet transform based multiple watermarking schemes. In: 2003 Conference on Convergent Technologies for Asia-Pacific Region, 2003, **3**: 935–938
14. Lu, C., Liao, H.: Multipurpose watermarking for image authentication and protection. *IEEE Transactions on Image Processing*, 2001, **10**: 1579–1592
15. Watson, A., Yang, G.: Visibility of wavelet quantization noise. *IEEE Transactions on image Processing*. 1997, **6**(8): 1164–1174

Construction of Symbolic Representation from Human Motion Information

Yutaka Araki, Daisaku Arita, Rin-ichiro Taniguchi, Seiichi Uchida,
Ryo Kurazume, and Tsutomu Hasegawa

Department of Intelligent Systems, Kyushu University
6-1, Kasuga-koen, Kasuga, Fukuoka, 816-8580, Japan

Abstract. In general, avatar-based communication has a merit that it can represent non-verbal information. The simplest way of representing the non-verbal information is to capture the human action/motion by a motion capture system and to visualize the received motion data through the avatar. However, transferring raw motion data often makes the avatar motion unnatural or unrealistic because the body structure of the avatar is usually a bit different from that of the human beings. We think this can be solved by transferring the meaning of motion, instead of the raw motion data, and by properly visualizing the meaning depending on characteristics of the avatar's function and body structure. Here, the key issue is how to symbolize the motion meanings. Particularly, the problem is what kind of motions we should symbolize. In this paper, we introduce an algorithm to decide the symbols to be recognized referring to accumulated communication data, i.e., motion data.

1 Introduction

Non-verbal information is very important in human communication, and video-based communication seems to be the simplest way. However, it has several problems, such as use of large network bandwidth, lack of spatioperceptual inconsistency, restriction of the number of participants, etc. As a possible solution to these problems, we are developing an avatar-based communication system[1]. It has an important merit that a virtual scene can be constructed as a communication environment, which can make the communication richer and efficient. Recently, this idea is extended to robot-based communication[2], where a robot is used instead of avatar and where a virtual communication environment is established in a physical 3D space.

In general, the avatar-based communication consists of acquisition of the contents of human communication, its transmission, and presentation of the transmitted contents. To present the non-verbal information, it seems that the simplest way is to capture the human action/motion by a motion capture system and to visualize the received motion data through the avatar. However, transferring the raw motion data causes several problems:

- The difference of body structure between a human and an avatar makes the reconstructed avatar motion unrealistic or physically impossible.
- The disturbance in communication network makes the avatar motion unnatural, because raw motion data is time synchronous data.

We think this can be solved by transferring the meaning of motion, instead of the raw data of motion, and by properly visualizing the meaning depending on characteristics of the avatar's function and its body structure. In this framework, the key issue is how to represent the meaning of motion referring to observed motion data, or how to symbolize the motion meaning. Particularly, the problem is what kind of motions we should symbolized. Of course, one method is to decide the symbols according to observation by ourselves, but it requires much time when we examine a large amount of accumulated communication data. Here, instead, we introduce an algorithm to decide the symbols to be recognized referring to accumulated communication data, i.e., motion data. Our basic idea is that frequent occurring motion patterns, i.e., motion motifs (or motifs for short), usually convey meaningful information, and that we automatically extract such motifs from the accumulated motion data.

2 Motif Extraction

To extract motifs[3], we propose a three-step procedure: the first step is compressing multi-dimensional motion information into lower-dimensional one by Principal Feature Analysis (PFA)[4]; the second is labeling time slices of each dimensional motion information according to its value and generating label sequences; the third step is recursive extraction of frequently occurring label patterns from multi-dimensional label sequences as motifs based on Minimum Description Length (MDL) principle[5]. Although motion motif extraction based on the MDL was used in [6], here, we deal with extraction of multiple motifs and explicit integration of multiple features.

2.1 Reduction of Redundant Dimension by PFA

Feature space reduction of high dimensional feature data such as human motion information is a common preprocessing step used for pattern recognition, clustering, compression, etc. Principal component analysis (PCA) and independent component analysis (ICA) have been extensively used for the space reduction. These methods find a mapping function from the original feature space to a lower space. However, they mix the original feature components and the original feature components are not handled directly. It is not easy to directly extract and describe motion patterns of subsets of body parts, such as a motion pattern of arms, or a motion pattern of legs.

Therefore, Principal Feature Analysis (PFA), which automatically determine a subset of feature components representing the original feature space, is used instead. Human motion information is described as a set of measured positions of body parts¹. We represent each feature as a single vector $\mathbf{f}_i = [f_{i,1} \ f_{i,2} \ \cdots \ f_{i,n}]^T \in \mathbb{R}^n$ and all motion information as a matrix $\mathbf{M} = [\mathbf{f}_1 \ \mathbf{f}_2 \ \cdots \ \mathbf{f}_s] \in \mathbb{R}^{n \times s}$, where s is the number of features and n is the length of motion information. Here, we suppose each feature is normalized so that the average of the feature values is zero. Then, principle features are selected by three steps as follows.

¹ Each position is composed of three features, i.e., 3D spatial coordinates, (x, y, z) .

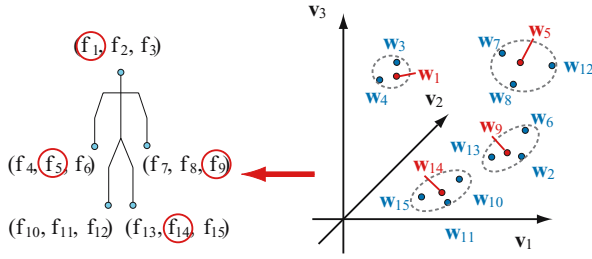


Fig. 1. Selection of principal features by PFA, where $q = 3, p = 4$. Each w_i written in red is closest to the mean of the cluster and corresponding f_i circled in red is a principal feature.

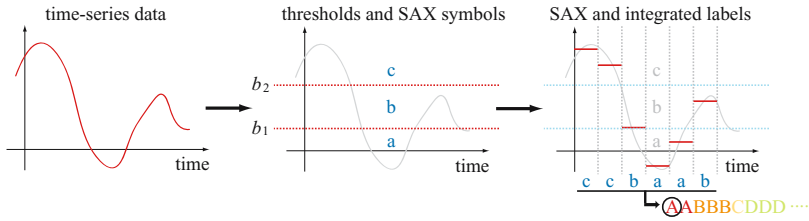


Fig. 2. Transforming a time-series data into a label sequence with SAX algorithm

First, the eigenvectors $v_i \in \mathbb{R}^s$ of M and their matrix $V = [v_1 \ v_2 \ \dots \ v_s]$ are calculated. Second, principal components $M_{pc} \in \mathbb{R}^{n \times q}$ of M is calculated by the following equations:

$$M_{pc} = MW^T \tag{1}$$

where $W = [w_1 \ w_2 \ \dots \ w_s] \in \mathbb{R}^{q \times s}$ is the first q columns of V^T . q is decided according to the variability of all the features. Then, vectors $\{w_i\}$ are clustered into $p(\geq q)$ clusters by K-Means algorithm using Euclidean distance. Here each vector w_i represents the projection of f_i to the principal component space.

Finally, in each cluster, the representing vector w_j , which is the closest to the mean of the cluster, is selected. Thus, each f_j corresponding to w_j is selected a dominant feature. Figure 1 illustrates the concept of PFA.

2.2 Transformation of Time-Series Data into Label Sequences

To reduce the computational complexity, time-series data $\{f_j\}$ ² selected by PFA is transformed into a label sequence by Symbolic Aggregate approximation (SAX)[7]. A label sequence is acquired by reducing the length of f_j and re-quantization. First, a time series data f_j of length n , i.e., n -dimensional vector f_j is regularly re-sampled into a w -dimensional vector $\bar{f}_j = [\bar{f}_{j,1}, \dots, \bar{f}_{j,w}]^T$. The i th element of \bar{f}_j is the average of its corresponding interval of f_j .

² Subscript j indicates the label of the selected feature.



Fig. 3. Recursive motif extraction: (a)an original label sequence, (b)extracting the first motif, (c)extracting the second motif, (d)final segments

$\bar{f}_{j,k}$ is re-quantified to a label-based form, SAX symbol, by thresholding. Each $\bar{f}_{j,k}$ is symbolized to $\hat{f}_{j,k}$ as follows:

$$\hat{f}_{j,k} = \text{sym}_l, \text{ iff } b_{l-1} \leq \bar{f}_{j,k} \leq b_l \quad (l = 1, \dots, N) \quad (2)$$

where b_l is a threshold and sym_l is a SAX symbol. The thresholds are decided so that generation probability of each SAX symbol is equal to others, assuming that distribution of \bar{f}_j is Gaussian. Then, each series of c SAX symbols, “ $\hat{f}_{j,k} \cdots \hat{f}_{j,k+c-1}$ ” is assigned to a single unique label $l_{j,k}$, and a label sequence $\mathbf{L}_j = [l_{j,1} \cdots l_{j,w-c+1}]$ is constructed.

Finally, in order to reduce the influence of the variation in motion speed, \mathbf{L}_j is compressed into $\hat{\mathbf{L}}_j$ by the run-length coding.

2.3 Recursive Motif Extraction Based on MDL Principle

For example, when a label sequence $\hat{\mathbf{L}}$ shown in Figure 3(a) is given, intuitively “BCD” can be a frequently occurring symbol pattern. Here, we need clear definition of what a frequently occurring pattern, or a motif, is, and we define the motif based on the MDL principle.

MDL is to find the best model which most efficiently compresses a label sequence by means that label patterns are replaced by unique meta-labels. We can consider that label patterns replaced by unique meta-labels in the best model are the motifs, because those label patterns are frequently occurred. In other words, the frequency of label patterns is evaluated based on the MDL.

To use the MDL principle, the description length of a label sequence should be defined, and it is defined based on a description model h and a label sequence $\hat{\mathbf{L}}$ as follows[6]:

$$DL = DL(\hat{\mathbf{L}}|h) + DL(h) \quad (3)$$

$$DL(\hat{\mathbf{L}}|h) = \sum_i^m \sum_j -w_{ij} \log_2 \frac{w_{ij}}{t_i} \quad (4)$$

$$DL(h) = \sum_i^m \log_2 t_i + m \log_2 \left(\sum_i^m t_i \right) \quad (5)$$

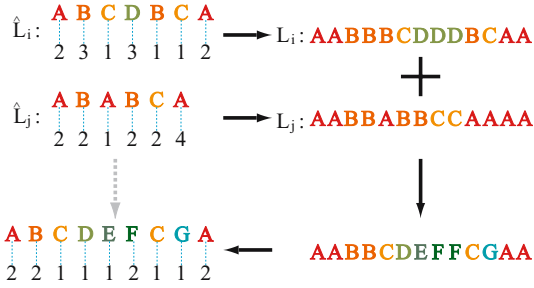


Fig. 4. Integration of \hat{L}_i and \hat{L}_j

Where the model h is a segmentation of the label sequence \hat{L} , m is the number of segments, w_{ij} is the frequency of the j -th label in the i -th segment, t_i is the length of the i -th segment.

Suppose t_L is the length of the label sequence \hat{L} , the number of possible segmentations is $O(2^{t_L})$, which is too much computation to find the best segmentation in a naive manner. Therefore, we propose a sub-optimal method to solve this problem approximately by a recursive scheme as follows.

First, the most frequent label pattern, or a motif candidate, is searched by traversing a label sequence with a fixed size search window. By changing the size of search window, we can find the best pattern, or a motif, from all the selected candidates based on equation (3). The this process is show in Figure 3(a), and the total computation of finding a motif is $O(t_L^3)$.

Second, the selected frequent pattern, or the selected motif, is replaced by a unique meta-label, and the next motif is searched in the same way as the first step. Iterate the second step until no frequent pattern whose length is more than or equal to 2 is found. All the computation cost of finding possible motifs is $O(t_M t_L^3)$, where t_M is the number of the iterations. This process is illustrated in Figure 3(b), (c). When the process finishes, the all the motifs are detected as shown in Figure 3(d).

2.4 Integration of Features

In the previous sections, we mentioned a method to extract motifs from one feature of human motion information. To analyze full-body motion information, it is necessary to integrate all the features. In this paper, simply, the integrated label sequences are generated from the all combinations of each \hat{L}_i and the motifs are extracted from each integrated label sequence based on MDL. Each \hat{L}_i is integrated using corresponding L_i . For example of integrating \hat{L}_i and \hat{L}_j , each pair (l_{it}, l_{jt}) is re-labeled to the new label to be unique of the other pairs' shown in Figure 4. Where l_{it} correspond to t th element of L_i .

Thus, when we have N features, we have $2^N - 1$ label sequences, and we extract motifs from each of the label sequence. After extracting motifs, we check similarity of the motifs and select proper motifs. However, there still remain improper motifs and we have to check them manually. Improvement of motif extraction accuracy is the future work.

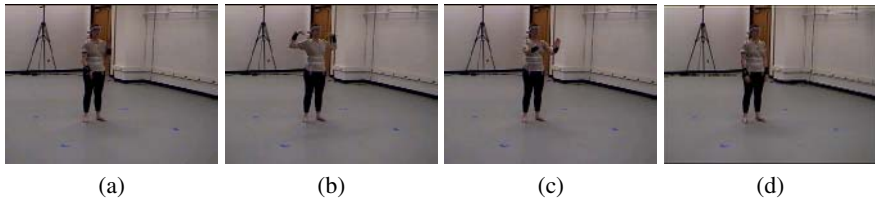


Fig. 5. The human motion scene: (a)waving hands up and down, (b) putting hands on the shoulders, (c)thrusting hands forward, (d)standing

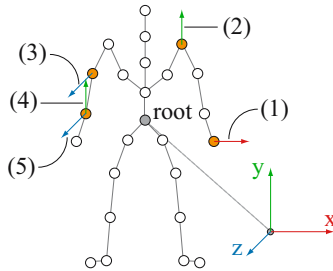


Fig. 6. Selected features: (1)left hand's x-axis, (2)left clavicle's y-axis, (3)right humerus z-axis, (4)right elbow's y-axis, (5)right elbow's z-axis

3 Experimental Results

We demonstrate the motif extraction method using motion data from Carnegie Mellon University's Graphics Lab motion-capture database³. The motion data used in this experiment is composed of 25 measured markers and 2486 frames. Each marker is composed of data of (x, y, z) -axis, i.e. the number of features is 75. The input motion data includes three basketball signals (A) waving hands up and down, (B) putting hands on the shoulders, and (C) thrusting hands forward shown in Figure 5(a), (b) and (c), each of which is repeated three times and connected by standing posture shown in Figure 5(d). In this experiment, the origin and the orientation of the coordinate system of motion information is fixed on a marker called "root" shown in Figure 6.

Five features shown in Figure 6 are selected by PFA: the left hand's x-axis, the left clavicle's y-axis, the right humerus' z-axis, the right elbow's y-axis and the right elbow's z-axis. In this experiment, the number of recursions of motif extraction is decided to eight empirically since it is enough for extracting all essential motions from the input motion information. For example, three motifs extracted from the integrated feature of (4) and (5) are shown in Figure 7. These motifs are extracted as the third, fifth and seventh motifs and correspond to motion (B), (A) and (C) respectively. The other motifs correspond to the standing posture. The third and seventh motifs are corresponding to the whole motion (B) and (C) shown in Figure 7(b),(c),(e),(f). However, the fifth motif is corresponding to a part of the motion (A) shown in Figure 7(a),(d). It is because motion

³ <http://mocap.cs.cmu.edu/>

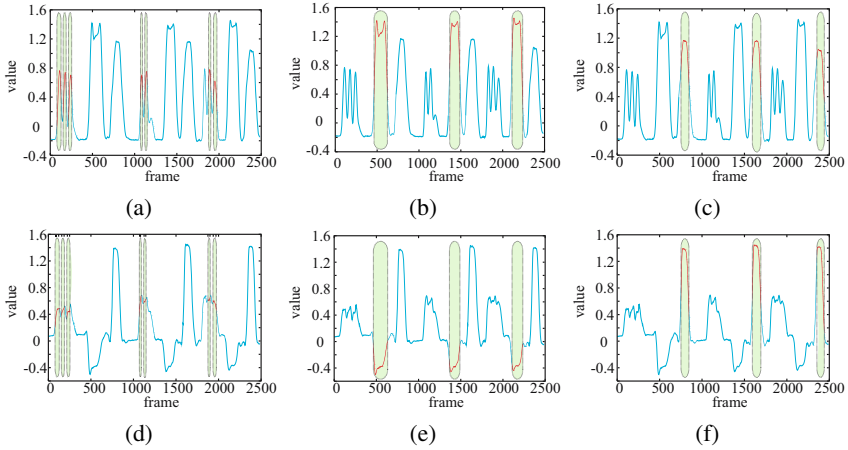


Fig. 7. The extracted 3rd(center), 5th(left) and 7th(right) motifs from integrated features: (a),(b),(c) illustrate y-position of the right elbow, (d),(e),(f) illustrate z-position of the right elbow.

(A), shaking hands up and down, is an iterative motion in comparison with motion (B) and (C) which are non-iterative. It is difficult to extract an iterative motion as a motif since an iterative motion includes sub-motifs, which are extracted as frequent patterns.

4 Discussion

We have proposed a motion motif extraction method from human motion information. The method automatically extracts all frequent motions as motifs from the whole body motion information. Our idea is that meaningful motion patterns, which are symbolized for avatar-based communication, are frequently appeared, and that they can be extracted by the proposed method. Of course, it is just preliminary study and we have to investigate the effectiveness of the idea, i.e., the following issues:

- The meaning of motion pattern sometimes differs depending on the context where the motion pattern appears. We should incorporate a context depending interpretation mechanism.
- Criteria for meaningful motion pattern other than the frequency of motion pattern should be investigated. Relationship between human posture and other persons (or objects) in the environment should be considered.

The preliminary experimental result of the motif extraction shows that the our method is effective to extract all motifs. However, there remain two problems. One is that our algorithm can not extract an iterative motion but just one cycle of the iterative motion. Another is that the computation cost is not small especially when we handle large motion databases including a variety of motion patterns, which are essentially high dimensional data.

Acknowledgment. This work has been partly supported by “Intelligent Media Technology for Supporting Natural Communication between People” project (13GS0003,

Grant-in-Aid for Creative Scientific Research, the Japan Society for the Promotion of Science), “Real-time Human Proxy for Avatar-based Distant Communication” (16700108, Grant-in-Aid for Young Scientists, the Japan Society for the Promotion of Science), and “Embodied Proactive Human Interface,” (the Ministry of Public Management, Home Affairs, Posts and Telecommunications in Japan under Strategic Information and Communications R&D Promotion Programme (SCOPE)).

Our thanks to Carnegie Mellon University’s Graphics Lab for allowing us to use their Motion Capture Database, which was supported with funding from NSF EIA-0196217.

References

1. D. Arita, H. Yoshimatsu, D. Hayama, M. Kunita, R. Taniguchi, Real-time human proxy: an avatar-based interaction system, *CD-ROM Proc. of Int. Conf. on Multimedia and Expo*, 2004.
2. A. Mori, S. Uchida, R. Kurazume, R. Taniguchi, T. Hasegawa, H. Sakoe, Early recognition of gestures, *Proc. of 11th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, pp.80–85, 2005.
3. J. Lin, E. Keogh, S. Lonardi, P. Patel, Finding motifs in time series, *Proc. of the 2nd Workshop on Temporal Data Mining*, pp.53–68, 2002.
4. I. Cohen, Q. Tian, X. S. Zhou, T. S. Huang, Feature selection using principal feature analysis, Submitted to Int. Conf. on Image Processing ’02, <http://citeseer.ist.psu.edu/cohen02feature.html>.
5. P. Grünwald, A tutorial introduction to the minimum description length principle, In *Advances in Minimum Description Length: Theory and Applications* (edited by P. Grünwald, I. J. Myung, M. Pitt), MIT Press, 2005.
6. Y. Tanaka, K. Iwamoto, K. Uehara, Discovery of time-series motif from multi-dimensional data based on MDL principle, *Machine Learning*, vol.58, pp.269–300, 2005.
7. J. Lin, E. Keogh, S. Lonardi, B. Chiu, A symbolic representation of time series with implications for streaming algorithms, *Proc. of 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp.2–11, 2003.

Toward a Universal Platform for Integrating Embodied Conversational Agent Components

Hung-Hsuan Huang¹, Tsuyoshi Masuda¹, Aleksandra Cerekovic²,
Kateryna Tarasenko¹, Igor S. Pandzic², Yukiko Nakano³, and Toyoaki Nishida¹

¹ Department of Intelligence Science and Technology, Graduate School of Informatics,
Kyoto University, Japan
{huang, masuda, ktarasenko, nishida}@ii.ist.i.kyoto-u.ac.jp

² Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
{aleksandra.cerekovic, Igor.Pandzic}@fer.hr

³ Department of Computer, Information and Communication Sciences,
Tokyo University of Agriculture & Technology, Japan
nakano@cc.tuat.ac.jp

Abstract. Embodied Conversational Agents (ECAs) are computer generated life-like characters that interact with human users in face-to-face conversations. To achieve natural multi-modal conversations, ECA systems are very sophisticated and require many building assemblies and thus are difficult for individual research groups to develop. This paper proposes a generic architecture, the Universal ECA Framework, which is currently under development and includes a blackboard-based platform, a high-level protocol to integrate general purpose ECA components and ease ECA system prototyping.

1 The Essential Components of Embodied Conversational Agents and the Issues to Integrate Them

Embodied Conversational Agents (ECAs) are computer generated life-like characters that interact with human users in face-to-face conversations. To achieve natural communications with human users, many software or hardware assemblies are required in an ECA system. By their functionalities in the information flow of the interactions with human users, they can be divided into four categories:

ECA Assemblies in the Input Phase. Non-verbal behaviors are the indispensable counterpart of verbal information in human conversations and thus embodied agents have to possess the capabilities of both of them. In addition to capturing natural language speech, non-verbal behaviors such as head movements, gaze directions, hand gestures, facial expressions, and emotional conditions are acquired by various types of sensors or visual methods in ECA researches. Further, input understanding tasks such as speech and gesture recognition are also required to be done in this phase.

ECA Assemblies in the Deliberate Phase. This is the central part of an intelligent agent to determine its behaviors in responding to the inputs from the outside environment. An inference engine with a background knowledge base and a dialogue

manager are required for conducting a discourse plan to achieve the ECA's conversational goal according to the agent's internal mental state. Talking to a conversational agent without emotions and facial expressions is weird and will be easily satiated while being like a human in the real world, personality, emotion, culture, and social role models are incorporated into ECAs to improve their believability.

ECA Assemblies in the Output Phase. Verbal output or natural language synthesis is generally done by a Text-To-Speech (TTS) engine to speak out the text output from the dialogue manager. Spontaneous non-verbal behavior outputs such as facial expressions, eye blinks, spontaneous hand gestures, and body vibrations are generated randomly or depending on the syntactical information of accompanied utterance by using the result of statistical analysis like CAST [5]. At last, a 2D/3D character animation player that renders the virtual character body and probably the virtual environment where the character resides on the screen is necessary.

A Platform for Integrating ECA Components. To integrate all the various assemblies of an ECA system described above, a platform or framework that seamlessly integrates them is a critical part. This platform has to transport all the sensor data streams, decisions, and command messages between all the components. It has been proposed that there are four essential requirements in the ECA component integration issue [4, 6]. First, the platform has to keep all of output modalities to be consistent with the agent's internal mental state. Second, all the verbal and non-verbal outputs are required to be synchronized. Third, ECAs have to be able to response to their human users in real-time. Fourth, the support for two ways of the information flow, "pull data from a component" and "push data to a component" are required in ECAs.

2 Universal Embodied Conversational Agent Framework

ECA systems are so sophisticated and their functions actually involve multiple research disciplines in very broad range such that virtually no single research group can cover all aspects of a full ECA system. Moreover, the software developed from individual research result is usually not meant to cooperate with others. There is a number of outstanding ECA systems that have been proposed previously, however, their architectures are ad hoc designed [2] and are not for a general purpose use.

Therefore, if there is a common and generic backbone framework that connects a set of general-purpose reusable and modularized ECA components which communicate with each other in a well-defined and common protocol, the rapid building and prototyping of ECA systems become possible, and the redundant efforts and resource uses of ECA researches can be prevented. This work proposes such an architecture that eases the development of ECA systems for general purposes. In our current design, it contains the following three parts, a general purpose platform (Universal ECA Platform) which is composed by a set of server programs for mediating and transporting data stream and command messages among stand-alone ECA software modules, a specification of a high-level protocol based on XML messages (UECAML) that are used in the communication between a standardized set of ECA components, and an application programming interface (UECA API) for easy development of the wrappers for the ECA software modules. These basic concepts are shown in Fig. 1.

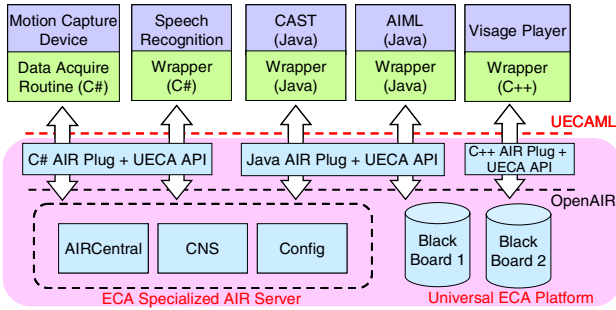


Fig. 1. The conceptual diagram of our Universal ECA Framework that includes the Universal ECA Platform server, UECA API, and a high-level protocol, UECAML

We use blackboard model as the backbone platform and OpenAIR [6] as the low-level routing and message passing protocol for the following reasons:

- Distributed architecture and XML absorb the differences of operating systems and programming languages of components and distribute the computation complexity.
- Unlike a long pipelined architecture, the single-layer topology provides the possibility to support reflexive behaviors that bypass the deliberation of the agent.
- The weak inter-connectivity of the components allows the online switching of components and thus makes online upgrading and maintaining of components.
- Components with different levels of complexity can be integrated into the ECA system as long as they understand and generate the same message types and the system can still work even some components are absent.
- Logically isolated multiple blackboards can distribute information traffic that is concentrated on only one blackboard in traditional systems.

Based on this framework, we are specifying an XML based high-level protocol for the communications between ECA components. Every XML message belongs to a message type, for example, “input.speech.text”, “output.body.gesture”, etc. Each message type has a specified set of elements and attributes, for example, “intensity”, “time_interval”, “start_time”, etc. Each component *subscribes* its interested message type(s), read them from the blackboard when they are *published* by another component, generates its output and publishes messages in other types to the blackboard. In the current stage, we are focusing on the specification on input and output phases and categorized the message types in the procedure of the I/O phases into an abstract hierarchy having three layers in the blackboard according to their abstractness. This basic idea is depicted in Fig. 2(a) and described below.

Low-level Parameter Layer in Input Phase: To absorb the possible variance even for the same modality in the lowest-level raw sensors’ data, the sensor data handling components interpret raw data into low-level parameterized representations, and then write them into the blackboard. For example, rather than the raw wave data from the voice capture subsystem, the user’s voice is interpreted into a recognized text stream by a speech recognition component, rather than the absolute current positions and angles of a sensor of the motion capture system, the numbers are transformed into

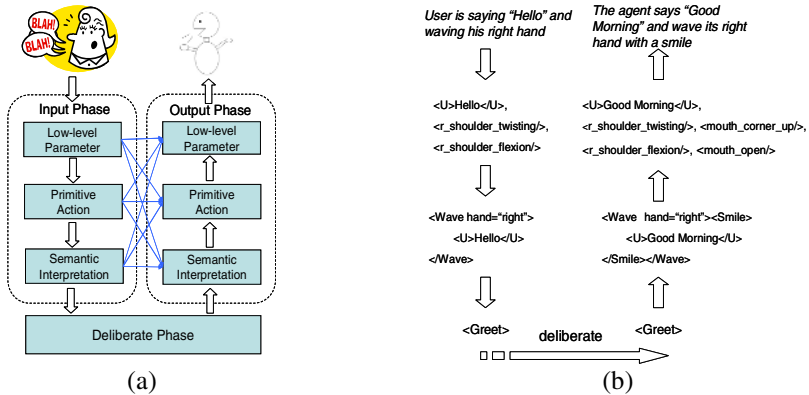


Fig. 2. (a)The hierarchy of the high-level protocol, UECAML, notes that reflex action links are shown in blue arrows (b) Example messages for a greeting response of the conversational to a human user’s greeting behavior

angles of the joints of a human body. As a result, the total output of the components in this stage is a parameterized text representation of movements of human users includes the angles of body joints, eye gaze directions, facial expression primitives, speech in text, physiological parameters and so on. Because the parameters in this stage must be specified in great detail with specific expert knowledge, we are going to specify the protocol of this layer based on appropriate and popular existing standards such as MPEG-4 FBA.

Primitive Action Layer in Input Phase: The task of this layer is to read the low-level parameters from the last layer and to interpret them into messages expressing abstract primitive actions. For example, from the change of the head orientation in horizontal and vertical directions to higher level primitive actions like “*head-shaking*” or “*head-nodding*”, from the change of bending angles of the joints of shoulder, elbow, and wrist to recognize an user action like “*waving-right-hand*” or “*raising-left-hand*”, from the angles of the joints of arms and fingers to recognize the action, “*pointing (X, Y, Z).*” Because there is virtually no limit in the range of body actions, we are going to specify just a generally useful set for specifying primitive body actions. At the same time, the message format should be flexible enough to allow new primitive actions to be included in the messages and that action will be interpreted as long as the specified component dealing with messages in this layer can recognize and understand it. A recognized primitive action is then propagated through the platform as an event in the instant when it is recognized with a timestamp and probably additional attributes such as intensity and time interval.

Semantic Interpretation Layer: The messages belong to this layer are semantic meaningful events and are interpreted by components from the primitive actions, for example, a user behavior recognizing component may interpret the primitive actions “*smiling*” and “*waving-right-hand*” done by the users to a “*greeting*” semantic explanation.

Deliberate Phase Layer: We plan to specify the messages in this layer to include the inference, knowledge representation, dialogue management, personality model, emotion model, and social role model functionalities as future works. Currently, we assume that the inputs of this black box are text streams recognized from human users' utterances which are annotated with semantic events or primitive action markups. The outputs are then utterances of the conversational agents that are going to be spoken by a TTS engine and annotated with markups to specify facial expressions, body movements, gestures, and other non-verbal behaviors.

Output Phase: In output phase, message flows are processed in a reversed order comparing to input phase, where messages from the deliberate phase are decomposed to more concrete concepts with lower abstractness by responding components. For example, when the deliberate phase decided that the agent should *greet* the user, this semantic command then may be interpreted by an action catalogue component into the “utterance (“Good Morning”)”, “smile” and “wave-right-hand” primitive actions. These two primitive actions are then further interpreted into low-level facial animation parameters and body animation parameters by a FAP / BAP database component to drive the CG character of a MPEG-4 FBA compatible player to smile and wave its right hand. A sample message flow that follows the framework of a process for an agent to greet in response to a human user's greeting behavior is shown in Fig. 2 (b).

The shared blackboard(s) mechanism allows the components to exchange information easier between different logical layers; a component can write its outputs arbitrarily into other layers and thus components with different level of sophistication can work together. Further, reflex action controlling components that bridge input phase messages directly to output phase messages are also allowed in this architecture.

Generally, blackboard architecture suffers from two major disadvantages. First, due to the distributed problem-solving methodology, it usually lacks a mechanism to centrally direct how a problem is going to be solved. This problem as well as the multi-modality consistency issue can be remedied by introducing a centralizing component to issue action confirming messages in the deliberate phase, that is, the actions sent to all output modalities will not be executed without confirmation except the reflexive behaviors. Second, the additional information traffics involving the shared blackboard cause inefficiency. The performance deterioration can be reduced by the direct information exchanges between the components while the message traffic load centralized on a single blackboard can be reduced by using multiple logically isolated blackboards at the same time. Besides, we plan to address the ECA component synchronization issue by the following ways, to require all the machines composing the system to be synchronized with each other to absolute standard time by NTP and to utilize the explicit timestamp field in each message as well as incorporating “*after the next action*”, “*begin at the same time as the next action*” specifiers for primitive actions.

3 Prototype of the Universal ECA Platform

We have implemented a Java prototype of the platform that routes and transports the communication between the ECA components those have registered in it. In addition to the reference Java AIR Plug implementation from mindmakers.org, we have

developed a C# version and are developing a C++ AIR Plug library. Based on the backbone platform, we are defining UECAML, which is currently focused on multi-modality inputs and CG character animation outputs.

As a premier evaluation, we developed two experimental ECA systems. One is a campus guide agent, it stands in front of a photo of somewhere in a campus while human users can ask it what an object in that photo is with natural language, hand pointing and head movements. As shown in Fig. 3(a), 3(b), the campus guide agent is composed with seven modules, head movement module utilizes an acceleration sensor to detect head shakes and nods, pointing module uses data from a magnetic motion capture device to judge which area the user is pointing at with his (her) right hand, a wrapped SAPI compatible Japanese recognition engine, a wrapped AIML [1] interpreter for dialogue management, gesture selector module is a wrapped CAST engine, input integrator module integrates all the tree modalities into individual input events, and a character animator player developed with visage!SDK [7]. The other one experimental system is an application for experiencing cross-culture differences of gestures and is shown in Fig. 3 (c). In this virtual environment, an avatar replays the user's hand gestures such as beckoning, and there are multiple computer controlled agents that react to those gestures. Their reaction differs depending on which country they are supposed to come from for example, Japan or Britain.

The two experimental ECA systems themselves are relatively simple; however, this work is not emphasizing on how strong the built ECAs are but is trying to ease the development and provide sufficient capabilities for general ECA researches. In the preliminary evaluation, the campus guide agent proves the platform's capability to seamlessly deal with multimodal inputs and sufficient performance for smooth real-time conversation. In the gesture experiencing application, our three-machine configuration showed satisfying performance to drive an avatar with motion capture device and ten computer controlled agents in real-time. Besides, both of these two systems can be built by incorporating software tools which are not specifically designed for these systems with little efforts, just by wrapping those tools according to the specification of universal ECA platform, and then an ECA system works. For example, the campus guide agent was built in three hours by writing two scenarios in AIML and the input integrator in addition to the other general purpose modules and pre-defined gestures. Further, in these two experimental systems, it usually requires only several dozen lines of code to wrap a software tool.

4 Future Works and Evaluation

This work is yet far from reaching its objectives. We are going to complete the definition of the standard high-level protocol to allow the integration of common ECA assemblies, improve the infrastructure to support the necessary features for the protocol, a set of client-side libraries supporting easy integration of ECA assemblies developed in various programming languages on various operating systems as well as a set of wrapped common ECA tools. The ultimate objective of this work is to pack all of these as an ECA development toolkit including a workable skeleton ECA.

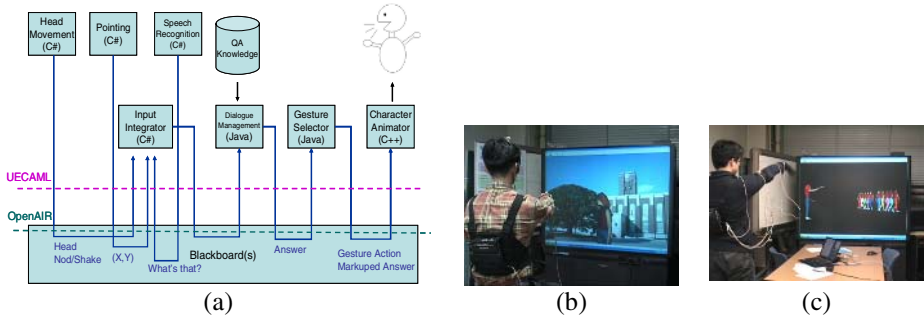


Fig. 3. (a) The campus guide agent's configuration (b) snapshot of the campus guide agent while it is performing a pointing gesture (c) An application for experiencing the cross-culture differences of hand gestures

We plan to produce a preliminary release for a field test in our proposed project at the eNTERFACE'06 [3] summer workshop on multimodal interfaces. We expect that several participants will join our team during the workshop; they will be provided with an initial release of the platform and jointly develop an ECA application during the relatively short four-week workshop period. During the practical field use of the platform, we expect to evaluate the platform in the following aspects, expressiveness of the high-level protocol, the ease of use, and the performance of the platform in the sense of responsiveness and the consistency of all modalities. We do not expect the platform to be fully satisfying the requirements in the preliminary release but are going to update the requirements, gather problem reports and other suggestion during the workshop. We will then improve the platform based on these experiments and make a public release.

References

- [1] Artificial Intelligence Markup Language (AIML), <http://www.alicebot.org/>
- [2] Cassell, J., Vilhjalmsson, H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit, in *The Proceedings of SIGGRAPH '01*, pp.477-486, 2001.
- [3] The eNTERFACE'06 workshop on multimodal interfaces, <http://enterface.tel.fer.hr>
- [4] Gratch, J., Rickel, J., Andre, E., Cassell, J., Petajan, E., and Badler, N.: Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems*, pp.54-63, 2002.
- [5] Nakano, Y., Okamoto, M., Kawahara, D., Li Q., Nishida, T.: Converting Text into Agent Animations: Assigning Gestures to Text, in *The Proceedings of The Human Language Technology Conference (HLT-NAACL04)*, 2004.
- [6] Thorisson, K., List, T., Pennock, C., and DiPirro, J.: Whiteboards: Scheduling Blackboards for Semantic Routing of Messages & Streams, AAAI-05 Workshop on Modular Construction of Human-Like Intelligence, 2005.
- [7] visageSDK, visage technologies, <http://www.visagetechologies.com/index.html>

Flexible Method for a Distance Measure Between Communicative Agents' Stored Perceptions

Agnieszka Pieczyńska

Institute of Information Science and Engineering, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
agnieszka.pieczynska@pwr.wroc.pl

Abstract. In this paper a flexible method for a distance measure between communicative agents' stored perceptions is proposed. It is assumed that each agent observes the states of external objects that are remembers them in a private temporal database in the form of base profiles. Distance measure is applied by the agent in the algorithm for the messages generation or during integration of other agents' opinions collected during communication activities in order to create objective picture of the current state's of objects. Proposed distance measure between base profiles is based on a computing of the costs of transformation one base profile into other. The objects' movements hierarchy is introduced and mathematically explained using expected value and random variable.

1 Introduction

One of the key research issues in the area of the MASs (multiagent systems) is solving complex and distributed tasks. The agents as autonomous entities cooperate with each other in order to achieve their individual goals. These activities include tasks allocation by means of applying the negotiation procedures and integration of results from the members of the multiagent population. Due to the agents' autonomy, they are characterized by some level of knowledge incompleteness what follows that the current states of some objects from external world are not directly perceived. In such cases the agents might ask others from population about their opinions. These answers might be not identical therefore some strategy for solving conflicts of inconsistent opinions is needed. On the other hand the agent not asking others might solve the problem of knowledge incompleteness using some internal mechanism for approximation the current states of the objects on the basis of previous experience that are stored in its private temporal database. This mechanism is called the algorithm for the messages generation [13,15,16]. As well as in the process of agents' opinions integration and in the algorithm for the messages generation distances between agents' perceptions must be computed, according to the following rule: more unanimous opinions smaller distance between them. In this paper, soft method for a distance measure between agents' perceptions is proposed. It is assumed, that the internal representation of the states of the external world has a relational structure and

the objects are described by means of possessing or not possessing some properties. If at the particular time point the states of some objects are for the agents not known then this area of cognitive processing is called empirical incompetence. It is assumed that distance between perceptions is understood by means of the costs of transformation one opinion into other. This work is organized as follows in section 2 the internal agent's knowledge organisation is given. In section 3 an overview of the methods for distance measure known in the literature is presented. Section 4 presents the original approach for a distance measure between structured objects. Finally in section 5 the concluding remarks are given.

2 Agent's Knowledge Representation

It is assumed that the external world is consisted of the set of objects $O = \{o_1, o_2, \dots, o_w\}$. The objects from O posses or don't posses some properties from the set $P = \{P_1, P_2, \dots, P_Z\}$.

Definition 1. One snapshot of the state of the world recognised by the agent $a \in A$ at the time point $t_n \in T$, where T is the set of time points, is embodied in agent as a base profile that is defined as [13]:

$$BP^a(t_n) = \langle O, P^+_1(t_n), P^-_1(t_n), P^\pm_1(t_n), \dots, P^+_K(t_n), P^-_K(t_n), P^\pm_K(t_n) \rangle,$$

where:

1. $P^+_i(t_n)$ denote the set of objects that at the time point t_n have been recognized by the agent $a \in A$ as possessing the property P_i .
2. $P^-_i(t_n)$ denote the set of objects that at the time point t_n have been recognized by the agent $a \in A$ as not possessing the property P_i .
3. $P^\pm_i(t_n)$ denote the set of objects from the area of a-agent's incompetence (the agent couldn't observe their state in relation to the property P_i).

Obviously, for each $i=1,2,\dots,K$, the condition $P^+_i(t_n) \cap P^-_i(t_n) = \emptyset$ holds.

Base profile is a structured object consisted of triplets of the sets of objects.

Two additional concepts related to the base profiles are: the area of incompetence related to a particular property $P_i \in P$ and the overall state of agent's knowledge:

Definition 2. The area of the agent's incompetence related to a property P_i is given by the following set of objects:

$$P^\pm_i(t_n) = O \setminus (P^+_i(t_n) \cup P^-_i(t_n))$$

Definition 3. The overall state of stored perceptions is given as a temporal database:

$$KS^a(t_c) = \{BP(t_n) : t_n \in T \text{ and } t_c \leq t_n\},$$

where t_c is a current time point.

3 Related Work

In the literature there are well-known correlation measures between structured objects that base on two kinds of coefficients: 1) similarity coefficients such as: Jacard's index, simple matching coefficient, Yule's index, Hamann's index, overlap or cosine

measure and 2) dissimilarity coefficients such as: Minkowski distance measure in two variants Manhattan and Euclidean distance measure, weighted Minkowski distance measure, Canberra metric coefficient and Hamming distance measure [8,17,18], [19]. In [11] similarity measure between structured objects is understood in terms of the minimal costs of transforming one object into other. But in this kind of similarity measure it is necessary to define the costs of transforming component elements of one object into component elements of the second object. Such methods for distance computing is popular for preference relation [5]; linear order [1], identity relation [6], [7], for strings [9], for n-trees [4] and semilattices [2,10]. In work [14] a distance measure between cognitive agent's stored perceptions has been presented. Three requirements towards this measure have been formulated with reference to the objects' movements hierarchy between the sets: $P^+_i(t_n)$, $P^-_i(t_n)$ and $P^\pm_i(t_n)$. It was assumed that the costs of the objects' movements between the sets $P^\pm_i(t_n)$ and $P^+_i(t_n)$ and the sets $P^\pm_i(t_n)$ and $P^-_i(t_n)$ should be lower than the costs of the objects' movements between the sets $P^+_i(t_n)$ and $P^-_i(t_n)$. The membership of the objects o to the set $P^+_i(t_n)$ in one base profile $PB^a(t_n)$ and to the set $P^-_i(t_k)$ of the other base profile $PB^a(t_k)$ means that the states of this objects are completely different. It was assumed that the weight of the object's movement between: 1) the sets $P^\pm_i(t_n)$ and $P^+_i(t_n)$ and the sets $P^\pm_i(t_n)$ and $P^-_i(t_n)$ is equal 0.5, 2) the sets $P^+_i(t_n)$ and $P^-_i(t_n)$ is equal 1 and 3) the sets $P^+_i(t_n)$ and $o \in P^+_i(t_k)$ or $P^-_i(t_n)$ and $P^-_i(t_k)$ is equal 0. In this paper a mathematical justification of the costs of objects movements between the sets $P^+_i(t_n)$, $P^-_i(t_n)$, $P^\pm_i(t_n)$ is presented. It is mathematically explained why the costs of the objects movements between the sets $P^+_i(t_n)$ and $P^-_i(t_n)$ should be higher than the one between the sets $P^+_i(t_n)$ (or $P^-_i(t_n)$) and $P^\pm_i(t_n)$. It is assumed that the costs of the objects' movements are measured by means of mathematical expected value and random variable.

4 Distance Between Perceptions

In order to compute a distance between two base profiles $PB^a(t_n)$ and $PB^a(t_k)$ it is necessary to define the cost of transformation one base profile into other. By transformation we understand the objects movements for each triple $(P^+_i(t_n), P^-_i(t_n), P^\pm_i(t_n))$, $P_i \in P$ that result in obtaining the sets $(P^+_i(t_k), P^-_i(t_k), P^\pm_i(t_k))$. These transformations consist on adding or deleting the objects from the sets: $P^+_i(t_n)$, $P^-_i(t_n)$, $P^\pm_i(t_n)$.

4.1 Random Variable and Expected Value

Let us assume that the state of an object $o \in O$ at the time point $t_n \in T$ in relation to the property $P_i \in P$ is a random variable O_{P_i} . The random variable according to the well-known semantic interpretation of this concept is a quantity that in consequence of some experience is equal a certain value that is only known after realization of this experience. The realization of random variable is considered as a state of an object o at the time point t_n that is perceived by the agent a :

1. Occurrence of the property P_i : $o \in P_i^+(t_n)$. If $o \in P_i^+(t_n) \Rightarrow$ for $o_{P_i} \in O_{P_i}, o_{P_i} = 1$.
2. Not occurrence of the property P_i : $o \in P_i^-(t_n)$. If $o \in P_i^-(t_n) \Rightarrow$ for $o_{P_i} \in O_{P_i}, o_{P_i} = 0$.

In order to compute a distance between two base profiles (the sum of the costs of the objects' movements) it is necessary to apply a parameter that describes a random variable i.e. expected value.

Definition 4. For the property $P_i \in P$ and random variable O_{P_i} expected value is defined as:

$$E(O_{P_i}) = \sum_{i=1}^K o_{P_i} * p(o_{P_i})$$

where:

o_{P_i} - the random variable's realization for the property P_i (the value of state of an object o remembered by the agent's a), $o_{P_i} \in \{1, 0\}$.

$p(o_{P_i})$ - the probability of occurrence particular state of an object in relation to the property P_i , $p(o_{P_i}) \in [1, 0]$

Remark. The probability of the occurrence of the property P_i in the object o can be considered from two points of view: 1) in relation to the one object (taking into account the temporal changeability of the state of this object) and 2) in relation to the states of all objects (regardless in which object the property P_i was observed). In this work the second case is considered.

As it was stated above, the object o might be in one of two states: has a property P_i or doesn't have the property P_i . If the property P_i is observed in the object o (remembered by the agent a in a base profile $PB^a(t_n)$ as $o \in P_i^+(t_n)$), then the value of the state of this object o_{P_i} is equal 1 and the probability $p(o_{P_i}=1)=1$ and $p(o_{P_i}=0)=0$. If the property P_i is not observed in the object o ($o \in P_i^-(t_n)$), then the value of the state of this object o_{P_i} is equal 0 and the probability $p(o_{P_i}=0)=1$ and $p(o_{P_i}=1)=0$. But if the state of an object o at the time point t_n is for an agent not known ($P_i^\pm(t_n)$), then $p(o_{P_i}=1) \in (0, 1)$ and $p(o_{P_i}=0) \in (0, 1)$. It means that the probability that $o_{P_i}=1$ and $o_{P_i}=0$ might be equal and depends on the overall tendencies to occurrence of particular states of objects.

In this connection expected value E of the state of an object o depends on the direction of the objects' movements within the confines of the sets: $P_i^+(t_n)$, $P_i^-(t_n)$, $P_i^\pm(t_n)$ because this direction has an influence on the realization of the random variable that is in the form of the state's of object.

For each property $P_i \in P$ two groups of objects' movements have been distinguished.

1. The first group consists of the following objects' movements:

- a) The object o is moved from $P_i^\pm(t_n)$ to $P_i^+(t_n)$.
- b) The object o is moved from $P_i^\pm(t_n)$ to $P_i^-(t_n)$.

- c) The object o is moved from $P_i^+(t_n)$ to $P_i^\pm(t_n)$.
- d) The object o is moved from $P_i^-(t_n)$ to $P_i^\pm(t_n)$.
- e) The object o is moved from $P_i^\pm(t_n)$ to $P_i^\pm(t_n)$.

2. The second group consists of the following objects' movements:

- a) The object o is moved from $P_i^+(t_n)$ to $P_i^-(t_n)$.
- b) The object o is moved from $P_i^-(t_n)$ to $P_i^+(t_n)$.

In order to compute the costs of objects' movements definition 4 should be applied.

For the first group G_i^1 : $o_{P_i} \in \{0,1\}$ and $p(o_{P_i}) \in (0, \dots, 1)$, for the second G_i^2 - $o_{P_i} \in \{0,1\}$ and $p(o_{P_i}) \in \{0,1\}$.

4.2 The Overall Cost of Objects' Movements

The overall cost of objects' movements that are necessary to transform $PB^a(t_n)$ into $PB^a(t_k)$ is equal the sum of minimal costs of objects' movements for each property $P_i \in P$ within the sets: $P_i^+(t_n)$, $P_i^-(t_n)$, $P_i^\pm(t_n)$. Computing such costs it is necessary to establish the directions of objects movements. For each object $o \in O$ and property $P_i \in P$:

1. $o \in u_i \Leftrightarrow$ object o is moved from $P_i^+(t_n)$ to $P_i^-(t_n)$ and $u_i = (P_i^+(t_n) \setminus P_i^+(t_k)) \setminus (O \setminus P_i^-(t_k))$
2. $o \in h_i \Leftrightarrow$ object o is moved from $P_i^-(t_n)$ to $P_i^+(t_n)$ and $h_i = (P_i^-(t_n) \setminus P_i^-(t_k)) \setminus (O \setminus P_i^+(t_k))$
3. $o \in s_i \Leftrightarrow$ object o is moved from $P_i^+(t_n)$ to $P_i^\pm(t_n)$ and $s_i = (P_i^+(t_n) \setminus P_i^+(t_k)) \setminus (O \setminus P_i^\pm(t_k))$
4. $o \in y_i \Leftrightarrow$ object o is moved from $P_i^-(t_n)$ to $P_i^\pm(t_n)$ and $y_i = (P_i^-(t_n) \setminus P_i^-(t_k)) \setminus (O \setminus P_i^\pm(t_k))$
5. $o \in f_i \Leftrightarrow$ object o is moved from $P_i^\pm(t_n)$ to $P_i^+(t_n)$ and $f_i = (P_i^\pm(t_n) \setminus P_i^\pm(t_k)) \setminus (O \setminus P_i^+(t_k))$
6. $o \in b_i \Leftrightarrow$ object o is moved from $P_i^\pm(t_n)$ to $P_i^-(t_n)$ and $b_i = (P_i^\pm(t_n) \setminus P_i^\pm(t_k)) \setminus (O \setminus P_i^-(t_k))$
7. $o \in g_i \Leftrightarrow$ object o is moved from $P_i^\pm(t_n)$ to $P_i^\pm(t_n)$ and $g_i = P_i^\pm(t_n) \setminus (P_i^+(t_k) \cup P_i^-(t_k))$

Definition 5. The overall cost of objects' movements that are necessary to transform one base profile $PB^a(t_n)$ into $PB^a(t_k)$ is equal a distance d between $PB^a(t_n)$ and $PB^a(t_k)$:

$$d(PB^a(t_n), PB^a(t_k)) = \sum_{i=1}^K G_i^1 * E(O_{P_i}) + G_i^2 * E(O_{P_i})$$

where:

$$G_i^1 = (\text{card}(s_i) + \text{card}(y_i) + \text{card}(f_i) + \text{card}(b_i) + \text{card}(g_i))$$

$$G_i^2 = \text{card}(u_i) + \text{card}(h_i)$$

card(x_i) is the cardinality of the set x_i for x ∈ {s, y, f, b, g, h}

4.3 Computational Example

The following set of base profiles is given. Each of them is related to a different time point t_n. All of them represent the experienced knowledge of agent a in relation to the states of objects {o₁, ..., o₆} for the property P₁.

	Property P ₁		
	P ⁺ ₁	P ⁻ ₁	P [±] ₁
PB ^a (t ₁)	o ₁ , o ₂	o ₃ , o ₄	o ₅ , o ₆
PB ^a (t ₂)	o ₁	o ₂ , o ₃ , o ₄	o ₅ , o ₆
PB ^a (t ₃)	o ₁	o ₃ , o ₄	o ₂ , o ₅ , o ₆
PB ^a (t ₄)	o ₃	o ₁	o ₂ , o ₄ , o ₅ , o ₆
PB ^a (t ₅)	o ₂ , o ₅	o ₁	o ₃ , o ₄ , o ₆

Let us assume that p(o_{P1} = 1) = p(o_{P1} = 0) = 0.5.

We obtain following distance values:

$$d(PB^a(t_1), PB^a(t_2)) = 2 * (0 * 0.5 + 1 * 0.5) + 1 * (0 * 1 + 1 * 1) = 2$$

$$d(PB^a(t_1), PB^a(t_3)) = 3 * (0 * 0.5 + 1 * 0.5) + 0 * (0 * 1 + 1 * 1) = 1.5$$

$$d(PB^a(t_1), PB^a(t_4)) = 4 * (0 * 0.5 + 1 * 0.5) + 2 * (0 * 1 + 1 * 1) = 4$$

$$d(PB^a(t_1), PB^a(t_5)) = 4 * (0 * 0.5 + 1 * 0.5) + 1 * (0 * 1 + 1 * 1) = 3$$

The smallest distance value is between PB^a(t₁) and PB^a(t₃). The states of three objects: o₁, o₃, o₄ are the same in both profiles. There is one base profile PB^a(t₂) in which also the states of three objects are the same as in PB^a(t₁). But let us note that in PB^a(t₂) object o₂ ∈ P⁻₁(t₂) and in PB^a(t₁) object o₂ ∈ P⁺₁(t₁). It means that the state of o₂ at the time points t₁ and t₂ were completely different. In the first case the agent a perceived the property P₁ in o₂, in the second one – didn't perceive. While according to PB^a(t₃), where o₂ ∈ P[±]₁(t₃) it is possible that the object o₂ at the time point t₃ possessed the property P₁ as it was at the time point t₁. For that reason PB^a(t₃) is closer to PB^a(t₁) than PB^a(t₂).

5 Conclusion

In this work, a flexible method for distance measure between relational structures was presented. It was assumed, that cognitive agents observe the states of external objects and create private temporal databases that consist of the set of base profiles related to

the time points each. The distance measure between base profiles might be applied by a single agent both in an algorithm for messages generation and in a process of semantic integration of other agents' opinions. Proposed distance measure is based on a computing the costs of transformation one structure into other. It was mathematically explained using expected value and random variable why the costs of the objects movements between the sets $P^+_i(t_n)$ and $P^-_i(t_n)$ are higher than the one between the sets $P^+_i(t_n)$ (or $P^-_i(t_n)$) and $P^\pm_i(t_n)$.

For the further work it is necessary to develop the method for approximation the probabilities of occurrence particular states of objects on the basis of analysis agents' temporal databases situated in a real environment.

References

1. Arrow K.J., Social choice and individual values, 2nd ed., Wiley J., NY, (1963).
2. Bartell B.T., Cottrell G.W., Belwe R.K., Optimizing similarity using multi-query relevance feedback, *Journal of the American Society for Information Science*, vol.49, (1998), 742-761
3. Barthelemy J.P., Dictorial consensus function on n-trees, *Mathematical Social Sciences*, vol.25, s.59-64, 1992. Boland R., Brown E., Day W.H.E., Approximating minimum-length-sequence metrics: A cautionary note, *Mathematical Social Sciences*, vol.4, (1983), 261-270
4. Barthelemy J.P., *Dictorial consensus function on n-trees*, *Mathematical Social Sciences*, vol.25, (1992), 59-64
5. Bogart K.P., *Preference structure I: distance between transitive preference relations*, *Journal of Math. Sociology*3, (1973), 455-470
6. Boland R., Brown E., Day W.H.E., *Approximating minimum-length-sequence metrics: A cautionary note*, *Mathematical Social Sciences*, vol.4, (1983), 261-270
7. Day W.H.E., Optimal algorithms for basic computations in partition lattices, Tech. Rep. CS 7819, Dept. of Computer Science, Southern Methodist University, Dallas, TX 75275, (1978)
8. Egghe, L., Michel, C., Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques, *Information Processing and Management*, vol. 39, (2003), 771-807.
9. Martinez-Hinarejos C.D., Juan A., Casacuberta F., Median Strings for k-nearest neighbour classification, *Pattern recognition Letters* 24, (2003), 173-181
10. McMorris F.R., Mulder H.M., Roberts F.S., The median procedure on median graphs, *Discrete Applied Mathematics* 84, (1998), 165-181
11. Nguyen N.T., Consensus Choice Methods and their Application to Solving Conflict in Distributed Systems, Wroclaw University of Technology Press, Poland, (2002)
12. Nwana H.S., Software agents: An Overview, *Knowledge Engineering Review*, vol.11, no.3, (1996), 205-244
13. Katarzyniak R., Pieczyńska-Kuchtiak A., A consensus based algorithm for grounding belief formulas in internally stored perceptions, *Neural Network World*, no. 5, (2002), 461-472
14. Katarzyniak R., Pieczyńska-Kuchtiak A., Distance measure between cognitive agent's stored perceptions, Proceedings of the 22nd IASTED International Conference on Modelling, Identification, and Control, Ed. M. H. Hamza. Innsbruck, Austria, February 10-13, 2003. Anaheim [i in.]: Acta Press, (2003)

15. Katarzyniak R., Pieczyńska-Kuchtiak A., Grounding and extracting modal responses in cognitive agents: AND query and states of incomplete knowledge, *International Journal of Applied Mathematics and Computer Science*, vol.14, no.2, (2004)
16. Pieczyńska-Kuchtiak, A.: Grounding a descriptive language in cognitive agents using consensus methods. In *Proceedings of ICCS 2005, Atlanta, USA, Lecture Notes in Computer Science Vol. 3516*, (2005), 671-678
17. Rousseau R., Jaccard similarity leads to the Marczewski-Steinhaus topology for information retrieval, *Information Processing and Management*, vol.34, (1998), 87-94
18. Sarker B.,R., The Resamblance coefficients in group technology: A survey and comparative study of relational metrics, *Computers and. Engineering*, vol.30, no.1, (1996) 103-116
19. Wilson D.R., Martinez T.R., Improved Heterogeneous Distance Functions, *Journal of Artificial Intelligence Research* 6, (1997)

Harmonisation of Soft Logical Inference Rules in Distributed Decision Systems

Juliusz L. Kulikowski

Institute of Biocybernetics and Biomedical Engineering PAS, Warsaw, Poland
jlkulik@ibib.waw.pl

Abstract. It is considered a problem of harmonisation of diagnostic rules used in a distributed set of diagnostic centres, the rules being based on a *k-most-similar-cases* approach. Harmonisation is reached due to an exchange of diagnostic cases among the reference sets stored in the centres. The method is based on general concepts of similarity, semi-similarity and structural compatibility measures used to evaluation of adequacy of records in remote data files to the requirements connected with supporting decision making in a given, local diagnostic centre. The procedure of local reference set extension by diagnostic cases selection and acquisition is described.

1 Introduction

Development of computer networks in the last decades opened new perspectives in computer-aided decision making. In various application areas like administration, commerce, natural environment monitoring, public health care, civil security, certain types of scientific research, etc. decisions should be made on the basis of permanently updated information resources stored in distributed databases. Two models of decision making based on distributed information resources can be taken into account: 1st distributed data – central decision making (*DD-CD*), and 2nd distributed data – distributed multi-level decision making (*DD-DD*). In the first case the necessary data are retrieved in distributed databases and transmitted to a centre where they are used for final decision making. In the second case data retrieved in distributed local databases are subjected to primary processing (decision making), the results being then transmitted to a centre where they are used to final decision making. The reason for *DD-DD* model using is not only economy in the costs of huge data volumes transmission to the centre but also a possibility of immediate using the results of the lower-level decisions for local purposes. In the *DD-DD* case more than two decision making levels are admissible, however, in this paper the two-level case will be considered.

As an example let us take into consideration the following situation. It is assumed that several medical centres specialised in diagnosis and therapy of a certain type of diseases try to collaborate in exchanging experience and improving their diagnostic methods. In the beginning specialised local centres use their proper diagnostic methods consisting of: 1st collecting, in a standard form, a set of records of observed symptoms and diagnostic parameters assessed in the patients, 2nd using a fixed scale

for the disease's progress evaluation, and 3rd using a set of logical inference rules of the "If... then..." form assigning a disease's qualification and/or progress level to the given record of symptoms and parameters observed in any individual patient [1]. A pair consisting of a record of individual symptoms and parameters and of the corresponding disease's qualification and progress level will be called a *diagnostic case*. A collection of diagnostic cases stored in any given medical centre during a long period of its activity constitutes its *information resource*, a selected part of it called a *reference set* being used to new decision making. For attention focusing it will be assumed that the *k-nearest-neighbours (kNN)* or *k-most-similar-cases (kMSC)* approaches to decision making are used [2]. All local medical centres solve thus similar pattern recognition problems; however, their reference sets as well as based on them decision rules are, in general, different. It may thus happen that two identical records of symptoms and parameters in different centres are differently classified and recognised. This means that in such case, at least in one centre, the corresponding decision rule can be improved.

It is assumed that reference sets can be extended and actualised, not only on the basis of local diagnostic cases, but also by exchange of data among the centres. For this purpose the *DD-CD* and *DD-DD* decision making models can be used. In the *DD-CD* model each local centre acquires selected diagnostic cases from other centres in order to include them into its own reference set and to use the, so extended, local reference set to decision making. In the *DD-DD* model a local centre being faced with a new disease case distributes the corresponding symptoms and parameters among several other centres calling for their recognition of the case. Their decisions are mailed back to it, a consensus between them, if necessary, is established and the new diagnostic case is included into the local reference set. In both cases a direct peer-to-peer communication or a multi-agent data exchange technology can be used [3,4]. All decision centres having free access to the information resources of other centres can try to improve their decision rules in one of the above-described ways; this process is called here a *harmonisation* of decision rules. However, it does not mean that harmonisation leads to an unification of decision rules and to a common reference set in all local centres. It only should improve, as far as possible, the quality of decisions made in local centres.

The aim of this paper is showing the way and analysis of conditions under which the improvement of decision making can be reached. Considerations are limited below to the *DD-CD* model. As not directly involving the specialists in local centres into the decision making processes initiated in outer centres the *DD-CD* model seems to be more realistic than the *DD-DD* one. Let us also remark that despite the fact that our considerations are illustrated by the example of distributed medical diagnostic centres they suit as well to any other distributed system in which patterns or objects recognition problems are to be solved.

2 Basic Model Assumptions and Notions

It is assumed that a number K of local decision centers constituting a *DD-CD* system store their diagnostic cases in the form of structured records of a general form illustrated in Fig. 1.

Identifier	Attributes	Decision	Comments
------------	------------	----------	----------

Fig. 1. General structure of a record of diagnostic case

All its fields are composite, i.e. consisting, in general, of several sub-fields containing qualitative or quantitative data; the field *Identifiers* is used to discrimination of records on the basis of formal data: current record's number, personal data of the patient, date of examination, etc. *Attributes* contain ordered strings of qualitative and/or quantitative data representing the results of medical examinations and analyses of a patient. The result of medical diagnosis based on the former data is presented in a standard form in the *Decision* field. *Comments* may contain any additional remarks, information, etc. that may be of interest in the next diagnostic or medical treatment processes.

For diagnostic cases exchange between the decision centers the contents of the *Attributes* and *Decision* fields are of particular interest. However, in different centers the records may be differently defined and interpreted. Exactly speaking, this means that:

- The sub-fields constituting the *Attributes* in different centers may be different,
- Even in the case of sub-fields to which in several centers the same semantic interpretation has been assigned their contents may be evaluated in different numerical scales.

That is why a direct inclusion of diagnostic cases imported from outer centers to a given reference set is, in general, not possible. The key to solve the problem consists in a suitable definition of *concordance* of diagnostic cases. This is here considered as a binary relation described on a non-empty set U (*universe*) of objects (diagnostic cases) satisfying some conditions that will be specified below. For this purpose there will be reminded some concepts connected with the measure of *similarity* used in various applications.

Definition 1

For a given universe U a real function

$$\sigma: U \times U \rightarrow [0, \dots, 1] \quad (1)$$

satisfying the conditions:

- a/ $\sigma(u, u) \equiv 1$ for any $u \in U$;
- b/ $\sigma(u_i, u_j) \equiv \sigma(u_j, u_i)$ for any $u_i, u_j \in U$;
- c/ $\sigma(u_i, u_j) \cdot \sigma(u_j, u_k) \geq \sigma(u_i, u_k)$ for any $u_i, u_j, u_k \in U$

is called a *similarity measure* of the elements of U .

The similarity measure can be defined on the basis of a distance measure, angular measure, Boolean test functions, as well as on some their combinations; the corresponding examples have been given in [5]. However, in practice some other functional expressions can be used for widely defined *proximity of objects* assessment. In particular, a function $\sigma(u_i, u_j)$ satisfying Definition 1 excepting the symmetry condition b/ will be called a *semi-similarity measure* and a one satisfying the Definition excepting condition c/ will be called a *neighborhood measure*.

Example 1

Let us take into consideration a non-empty finite *alphabet* U and a family Σ_U of all non-empty linearly ordered subsets of U (taken without repetition of their elements), called *strings*. A real function

$$c: \Sigma_U \times \Sigma_U \rightarrow [0..1] \tag{2}$$

for any $\mathbf{v}, \mathbf{w} \in \Sigma_U$ given by the cosine formula:

$$c(\mathbf{v}, \mathbf{w}) = \frac{n_{\mathbf{v}, \mathbf{w}}}{\sqrt{n_{\mathbf{v}} \cdot n_{\mathbf{w}}}} \tag{3}$$

where $n_{\mathbf{v}, \mathbf{w}}$ is the number of common elements in \mathbf{v} and \mathbf{w} while $n_{\mathbf{v}}, n_{\mathbf{w}}$ denote, correspondingly, the lengths of the strings \mathbf{v} and \mathbf{w} , expresses a *structural compatibility* of the strings. It can be easily proven that the above-defined function satisfies the conditions a/ and b/ of Definition 1. In particular, $c(\mathbf{v}, \mathbf{w}) = 1$ if \mathbf{v} and \mathbf{w} consist of the same elements of the alphabet (maybe, taken in different orders) and $c(\mathbf{v}, \mathbf{w}) = 0$ if \mathbf{v} and \mathbf{w} contain no common elements. The condition b/ (symmetry) follows directly from the definition of $c(\mathbf{v}, \mathbf{w})$. However, condition c/ is not satisfied and, thus, $c(\mathbf{v}, \mathbf{w})$ is a sort of neighborhood measure.

The above-defined structural compatibility measure has the following interpretation in distributed data retrieval. Let U be a vocabulary of the names of attributes (sub-fields) occurring in the records of a system of distributed databases. Let it also be assumed that the elements of U reflect semantic meaning of the attributes in an unique way. Σ_U is then a family of all possible records' structures having the form of linearly ordered strings of attributes. The measure $c(\mathbf{v}, \mathbf{w})$ characterizes a structural compatibility of a pair of records from the point of view of containing the same attributes, regardless of their order within the records. However, even a maximal structural compatibility of any two types of records does not guarantee their full *concordance* in semantic sense. The problem consists in different scales of values assigned to the attributes or, more generally, in different *ontologies* [6,7] that can be used as a basis of real objects description.

A *semantic concordance* of attributes occurring in different records can be understood as existence of one-to-one correspondences between terms expressing the attributes' values assigned to (abstract or real) objects' properties or states.

Example 2

Let us take into consideration two qualitative attributes, H_1 and H_2 , existing in two different structured records. Let the following terms be assigned to the admissible values of attributes as their names:

$$\begin{aligned} H_1: & P, Q, R, S, T \\ H_2: & I, II, III, IV, V. \end{aligned}$$

The attributes are in semantic concordance if isomorphism between the sets of objects denoted by H_1 and H_2 can be established i.e. if :

- o each object qualified as "being P " (i.e. being denoted by the term P) can also be qualified as "being I " and vice versa;
- o each object qualified as "being Q " can also be qualified as "being II " and vice versa;

etc .

On the other hand, a semantic discordance between H_1 and H_2 may arise, in general, as a result of:

1. using the same units but different intervals on a numerical scale for quantitative attributes' representation;
2. representing quantitative values of attributes evaluated by incomparable methods;
3. representing quantitative values of attributes in different and mutually incompatible discrete scales;
4. expressing the values of qualitative attributes in semantically incompatible systems of terms;
5. expressing the values of qualitative attributes in systems of differently defined fuzzy terms, etc.

In each of the above-listed cases the *semantic concordance level* of compatible pairs of attributes can be evaluated by an adequately chosen similarity measure of the scales of the corresponding attributes' values. Below, the case of qualitative attributes in semantically incompatible systems of terms (point 4) will be considered in a more detailed form.

There will be denoted by H_A and H_B two systems of terms assigned to the same attribute and by A_i, B_j two terms belonging, respectively, to H_A and H_B . Then the following relationships between A_i and B_j may arise:

$A_i \cap B_j = \emptyset$ – the given terms are totally semantically incompatible,

$A_i \cap B_j \neq \emptyset$ – the terms are partially semantically compatible what means that one of the following situations arise:

$A_i \subseteq B_j$ – all objects denoted by A_i are also denoted by B_j ,

$A_i \supseteq B_j$ – all objects denoted by B_j are also denoted by A_i ,

$A_i \neq A_i \cap B_j \neq B_j$ – some objects denoted by A_i are also denoted by B_j and vice versa, existence of objects denoted only by A_i or only by B_j being also assumed.

For semantic compatibility of H_A and H_B assessment it will be taken into consideration a $m \times n$ matrix $V_{AB} = [v_{ij}]$, called a *terms' compatibility matrix*, where m, n stand, correspondingly, for the number of terms in H_A and in H_B ; $v_{ij}, 0 \leq v_{ij} \leq 1$, denotes a *logical weight* assigned to the assertion that if A_i to an object can be assigned then B_j can be assigned to it as well. The logical weights should be chosen according to the following principles:

- a) $v_{ij} = 0$ when $A_i \cap B_j = \emptyset$ (terms are fully semantically incompatible);
- b) $v_{ij} = 1$ when $A_i \equiv B_j$ (terms are fully semantically compatible);
- c) for any fixed $i, 1 \leq i \leq n$, and $j, 1 \leq j \leq m$, the sum S_{i*} of all v_{ij} over j and the sum S_{*j} of all v_{ij} over i satisfy the inequalities:
 $0 \leq S_{i*} \leq 1, 0 \leq S_{*j} \leq 1$;
- d) $0 < v_{ij} < 1$ when $A_i \neq B_j$; v_{ij} is larger if $A_i \supset B_j$ than if $A_i \subset B_j$ and in both cases it is larger than if $A_i \cap B_j \neq \emptyset$ and neither $A_i \supset B_j$ nor $A_i \subset B_j$;
- e) if for a certain pair of terms of H_A and H_B a logical value $v_{ij} = 1$ has been established and a third system (say, H_C) is given then for any its (say, k -th) term the logical weights should satisfy the conditions: $v'_{ik} = v''_{jk}, v'_{ki} =$

v''_{kj} , where v'_{ik} , v'_{ki} are elements of a term's compatibility matrix V_{AC} while v''_{jk} , v''_{kj} are the elements of V_{BC} .

The weights v_{ij} can be established on the basis of a semantic investigation of the corresponding terms, of statistical observations or in intuitive way. Point d) introduces a sort of asymmetry between A_i and B_j caused by an intention to distinguish between a situation when one wants to include B_j into A_i as its sub-term and the reverse one. According to this two *semi-similarity measures* between the systems of terms H_A and H_B will be introduced. Let S_{AB} denote the sum of all elements of V_{AB} (of course, it is $S_{AB} = S_{BA}$). There will be defined the ratios:

$$\sigma_{AB} = S_{AB} / n, \tag{5a}$$

$$\sigma_{BA} = S_{AB} / m, \tag{5b}$$

The following properties of σ_{AB} and σ_{BA} should be remarked: a/ their values are kept between 0 and 1, b/ they are equal 0 if and only if for all pairs of terms it is $A_i \cap B_j = \emptyset$, c/ they are equal 1 if and only if for all pairs of terms it is $A_i \equiv B_j$, d/ the larger is the number n of terms of a system H_A with respect to other systems of terms corresponding to the same attribute, the lower are the measures of semi-similarity of the other systems to the given one.

Example 3

Let us assume that a qualitative attribute *Colour* in three systems is coded by the terms:

H_A : red, green, blue,

H_B : red, grass-green, sea-green, blue,

H_C : red, green, blue.

The following terms' compatibility matrices for them can be established:

V_{AB}, V_{CB}	red	grass-green	sea-green	blue
red	1.0	0	0	0
green	0	0.6	0.4	0
blue	0	0	0	1.0

V_{AC}	red	green	blue
red	1.0	0	0
green	0	1.0	0
blue	0	0	1.0

Then, according to (5a), (5b) the following semi-similarity measures for the systems of terms can be calculated:

$$\sigma_{AB} = \sigma_{CB} = \frac{3}{4},$$

$$\sigma_{BA} = \sigma_{BC} = \sigma_{AC} = \sigma_{CA} = 1.0$$

This means that transferring the *Colour* value from H_B to H_A , from H_B to H_C as well as between H_A and H_C is possible without constraints while transferring similar data from H_A to H_B or from H_C to H_B is made difficult by the fact that the term *green* with logical weight 0.6 corresponds to *grass-green* and with the weight 0.4 to

sea-green. In the case of transferring data from H_B to H_A through H_C the semi-similarity measure will be

$$\sigma'_{BA} = \sigma_{BC} \cdot \sigma_{CA} = 1.0$$

while for a reverse transfer it will be

$$\sigma'_{AB} = \sigma_{AC} \cdot \sigma_{CB} = 1.0 \cdot \frac{3}{4} = \frac{3}{4}$$

because some information is lost on the way between H_C and H_B .

3 Exchange of Diagnostic Cases

Extending of a reference set for improvement of diagnoses by additional diagnostic cases acquisition should be realised in two main steps:

1st, data files selection: choosing in the system of distributed databases a set of data files adequate to the given diagnostic problem;

2nd, diagnostic cases selection: choosing in the data files records suitable for supporting the given diagnostic task.

Let us assume that a new diagnostic task is to be solved. Formally, it is given in the form of a structured record shown in Fig. 1 where the contents of *Decision* and *Comments* are to be established according to the values of *Attributes*. For realisation of the first step the following notion will be introduced:

Definition 2

Let V and W denote the structures of records describing, respectively, a current diagnostic task under examination and a diagnostic case in an outer reference database; the corresponding strings of attributes are denoted by \mathbf{v} and \mathbf{w} . A vector

$$\theta_{VW} = [c(\mathbf{v}, \mathbf{w}), \sigma_{v_1, w_1}, \sigma_{v_2, w_2}, \dots, \sigma_{v_K, w_K}] \tag{6}$$

where $c(\mathbf{v}, \mathbf{w})$ is a structural compatibility of the strings of attributes and $\sigma_{v_1, w_1}, \sigma_{v_2, w_2}, \dots, \sigma_{v_K, w_K}$ being semi-similarity measures of the pairs of structurally compatible attributes, will be called a *characteristic of adequacy of W to V* .

Data files selection consists in:

- 1/ finding in the system of databases reference data sets (W -files) for which it is $c(\mathbf{v}, \mathbf{w}) = 1$;
- 2/ evaluation for them the partial semi-similarity measures $\sigma_{v_1, w_1}, \sigma_{v_2, w_2}, \dots, \sigma_{v_K, w_K}$ and calculation of overall semi-similarity measures:

$$\sigma_{VW} = \prod_{\kappa=1}^K \sigma_{v_{\kappa}, w_{\kappa}} \tag{7}$$

- 3/ arranging the W -files according to their non-increasing values of σ_{VW} and taking into consideration only those for which it is $\sigma_{VW} > \sigma_0$, where $\sigma_0, 0 < \sigma_0 < 1$, is a fixed semi-similarity threshold.

Data files selection being done, the diagnostic cases selection should be performed. For this purpose it is necessary, first, to define a similarity measure (see

Definition 1) $\sigma^*(A_c, A_w)$ where A_c, A_w denote, correspondingly, the strings of attribute instances in the current diagnostic task and the diagnostic case in an outer reference database. Then the diagnostic cases selection consists in:

- 1) Finding in the selected reference data sets the records satisfying a condition:

$$\sigma^*(A_c, A_w) > \sigma^*_0, \text{ for } 0 < \sigma^*_0 \leq 1, \quad (8)$$

where σ^*_0 stands for a similarity threshold.

- 2) Including the, so chosen, records (diagnostic cases containing attributes and assigned to them decisions) into the reference set used to the solution of the current diagnostic task.

So obtained, extended reference set then can be used to improved diagnostic decision making based on the *k-most-similar-objects* approach. The method can be, in particular, adapted to an exchange of reference data among medical centres collecting rare or non-typical diagnostic cases, very important for quick and correct diagnosis in other suspicious cases.

4 Conclusions

A collaboration between distributed diagnostic centres where decision rules are based on the sets of reference cases can be improved due to an exchange of suitable diagnostic cases among the centres. This procedure leads to a harmonisation of decisions reducing the rate of cases of different diagnostic decisions being made in different centres when based on similar input data. The reference diagnostic cases exchange can be reached on the basis of assessment adequacy of records in remote data files to the requirements connected with diagnostic decision making in a given diagnostic centre.

References

1. Kurzynski M., Sas J., Blinowska A.: Rule-Based Medical Decision Making via Unification Procedure of Information. Proc. 13th Int. Congress of Medical Informatics Europe, Vol. A, Copenhagen 1996,537-541.
2. Hart P.E.: The Condensed Nearest Neighbour Rule, IEEE Trans. Information Theory, Vol. 14, No 3 (1968), 515-516.
3. Dastani M., Gomez-Sanz J.J.: Programming Multi-agent Systems. The Knowledge Engineering Rev., Vol.20:2 (2005), 151-164.
4. Di Marzo Serugendo G., Gleizes M.-P., Karageorgos A.: Self-organization in Multi-agent Systems. The Knowledge Engineering Rev., Vol.20:2 (2005),165-189.
5. Kulikowski J. L. *Pattern Recognition Based on Ambiguous Indications of Experts*. Komputerowe Systemy Rozpoznawania KOSYR'2001 (pod red. M. Kurzyńskiego). Wyd. Politechniki Wrocławskiej, Wrocław 2001, ss. 15-22.
6. Borgo S., Guarino N., Masolo C., Vetere G.: Using a Large Linguistic Ontology for Internet Based Retrieval of Object-Oriented Components. Proc. of the Conf. on Software Engineering and Knowledge Engineering: 528-534, 1997.
7. Mayfield J.: Ontologies and Text Retrieval. The Knowledge Eng. Rev. 17(1): 71-75, 2002.

Assessing the Uncertainty of Communication Patterns in Distributed Intrusion Detection System*

Krzysztof Juszczyszyn and Grzegorz Kołaczek

Institute of Information Science and Engineering
Wrocław University of Technology, Wrocław, Poland
krzysztof@pwr.wroc.pl, grzesiek@pwr.wroc.pl

Abstract. A paper proposes a formal framework for communication patterns' uncertainty assessment within a distributed multiagent IDS architecture. The role of the detection of communication anomalies in IDS is discussed then it is shown how sequences of detectable patterns like fan-in, fan-out values for given network node and clustering coefficients can be used to detect network anomalies caused by security incidents (worm attack, virus spreading). It is defined how to use the proposed techniques in distributed IDS and backtrack the incidents.

1 Introduction

In order to process intrinsically distributed information, most of modern IDS systems are organized in a hierarchical architecture [4], consisting of low level nodes which collect information and management nodes which aim to detect large-scale phenomena. The task of management nodes is to reduce the amount of the data processed, identify attack situations as well as make decisions about responses [10].

In our approach it is assumed that the network system consists of the set of nodes. There are also two types of agents in our multiagent system: monitoring agents (MoA) and managing agents (MA). Monitoring agents observe the nodes, process captured information and draw conclusions that are necessary to evaluate the current state of system security within their areas of responsibility. Managing agents are responsible for gathering information from MoA agents and generating reports about global threats and ongoing attacks. Each agent MoA monitors its own area of responsibility consisting of the set of network nodes.

It is commonly known that in the case of worm attack there occur at least two kinds of anomalies: in observed traffic characteristics and in communication scheme which tends to be constant under normal conditions (see the next section). In this context the system properties observed by the agent MoA in the proposed architecture will fall into two basic (and physically different) categories: 1. Traffic measurement. 2. Communication pattern measurement. The decision about current situation is being made on the basis of them.

* This work was supported by the Polish State Committee for Scientific Research under Grant No. 3 T11C 029 29 (2005-2007).

The MoA agent's algorithm for decision making process is invoked periodically and uses observed values as input data. MoA also stores acquired values thus creating the history of system behaviour. The algorithm itself consists of the following steps (and was discussed in detail in [7]):

BEGIN

1. Detect traffic anomalies (using chosen technique).
2. Create a list of traffic anomalies.
3. Compute the communication patterns.
4. Create a list of communication anomalies.
5. If any of the anomalies' lists is not empty, perform an attack backtracking analysis which will return result in the form of attack graph.

END.

As mentioned, the managing agent MA successively obtains data which are related to particular moments of time from monitoring. Then the managing agent MA uses an algorithm for determining the global tree representing the attack propagation [7,8].

In this paper we deal with the 3rd and 4th steps of the above algorithm and show how to estimate the uncertainty of communication anomalies assessment and how to construct local attack graph on the basis of them. The method may be used independently but in the proposed architecture its results will be used together with other techniques (currently under development) in order to provide more accuracy in tracking attacks.

2 Communication Patterns

Network traffic show some quantitative and topological features that appear to be invariant and characteristic for given network [3,9]. Moreover, general rules underlying that features are the same for almost any network of remarkable size. These distinct features concern topology of network communication, considered as origin-destination flows graph, the distribution of data volumes sent between destinations and the in/out ratio of data sent between nodes/subnets and outside world.

With respect to these properties, wide range of network attacks can be detected by observation of communication patterns and comparison existing under normal state of the network to new ones which occur under attack. For example, in case of Internet worm attacks, within a network there could be scanning and attack flows which differ substantially from normal network activity [11]. Moreover, total scanning rate into the sub-network (or given set of nodes) is a function of the number of all infected nodes in the network.

Another invariant for a long time periods and different scales (subnet sizes) or traffic types (protocols) is proportion between a number of internal (Fan-in) and external (Fan-out) data flows [1]. Experiments showed that both Fan-in and Fan-out for given node and their distribution for all nodes tend to be constant under normal conditions. It was also shown that the IP graph has heavy-tailed degree distribution showing scale-free structure according to power law [3]. Under worm attack the structure of communication is heavily affected and the distribution changes. There is

also a detectible dependence between worm propagation algorithm, and communication pattern disturbance [9].

Similar relationships occur also on the level of given communication protocol, for example the topology of e-mail corporate networks exhibits a scale-free link distribution and small-world behaviour, as for known social networks. This result was recently used to propose an anti-spam tool [2].

2.1 Detectable Communication Patterns

Monitoring agents of proposed IDS system will gather information about communication within the network under state that is assumed to be secure. Then the existing communication patterns will be discovered. The system will be viewed as a graph consisting of nodes (each monitoring agent will have a set of nodes under control) and edges which appear if there exists data flow between given pair of nodes. In our approach we are interested in tracking the following communication patterns:

1. Clustering coefficient for a given node.

The *clustering coefficient* c is the probability that two nearest neighbours of vertex i are also neighbours of each other. The value of c provides a quantitative measure for cliques in communication graph. For node i clustering c_i is given by:

$$c_i = \frac{2k_i}{n_i(n_i - 1)} \tag{1}$$

where n_i is the number of its neighbours and k_i – the number of connections between them. High (close to one) c means that a node belongs to a clique in considered graph.

2. Fan-in and Fan-out ratios.

Fan-in is the number of nodes that originate data exchange with node i , while Fan-out is the number of hosts to which i initiates conversations.

According to results listed in previous section the above patterns are invariant during most time of normal system activity or change in a predictive way. But while attack appears they will change leading to alert and taking chosen countermeasures.

Each MoA agent stores data about communication in the form of M_c matrix. The values of M_c are set according to the following rules:

$$M_c(i, j) = \begin{cases} 1: & \text{node } i \text{ communicates with node } j \\ \epsilon: & \text{lack of knowledge about communication between } i \text{ and } j \\ 0: & \text{there is no communication between nodes } i \text{ and } j \end{cases}$$

Value ϵ reflects that accurate value of some fields in M_c matrix may be unknown (their actual values were for some reason not observed by MoA). This results in some uncertainty in attack investigation analysis. The level of this uncertainty will be also a part of the algorithm’s result. As suggested in sec.1 the M_c entries are updated periodically in discrete time moments t . The history is stored by the MoA and forms a basis for anomalies’ detection. Let’s denote the state of M_c in t as M_c^t .

Before dealing with communication patterns’ uncertainty we briefly introduce Subjective Logic, an useful and strong formalism for reasoning and expressing opinions about uncertain observations.

3 Subjective Logic

Subjective logic was proposed by A.Josang as a model for reasoning about trust propagation in secure information systems. It is compatible with Dempster-Shafer’s theory of evidence and binary logic [5]. Subjective logic includes standard logic operators and additionally two special operators for combining beliefs – consensus and recommendation. The basic definitions of subjective logic given in this section come from [5,6].

Subjective logic can be used to express so-called opinions (see below) about facts with assumption that we do not require the knowledge of how these facts were grounded or inferred. We may also have an opinion about some subject (source of information). When expressing belief about a statement (predicate) it is assumed that it is either true or false, but we’re not necessarily certain about it. Let’s denote *belief*, *disbelief* and *uncertainty* as b , d and u respectively. A tuple $\omega = \langle b, d, u \rangle$ where $\langle b, d, u \rangle \in [0,1]^3$ and $b + d + u = 1$ is called an *opinion*.

Opinions have always assigned membership (are expressed by certain agents) and are not inherent qualities of objects but *judgments* about them. For any opinions $\omega_p = \langle b_p, d_p, u_p \rangle$ and $\omega_q = \langle b_q, d_q, u_q \rangle$ about predicates p and q the following operators may be defined (definitions, proofs and in-depth discussion are to be found in [6]): Conjunction (result of the conjunction of opinions is also an opinion and is denoted by $\omega_{p \wedge q}$), Disjunction ($\omega_{p \vee q}$), Negation ($\omega_{\neg p}$).

Now assume two agents, A and B , where A has opinion about B . Opinion about other agent is interpreted as opinion about proposition “ B ’s opinion is reliable”. We’ll denote opinion expressed by agent B about given predicate p and agent’s A opinion about B as ω_p^B and ω_B^A respectively. Then the opinion of agent A about p is given by *discounting operator* (a.k.a *reputation operator*, denoted by \otimes): $\omega_p^A = \omega_B^A \otimes \omega_p^B$. From the other hand, the joint opinion of two agents A and B about given predicate is computed by *consensus operator* \oplus (ω_p^A and ω_p^B are opinions of A about B and B ’s about p): $\omega_p^{AB} = \omega_p^A \oplus \omega_p^B$.

Consensus operator is commutative and associative thus allowing to combine more opinions. Opinions about binary events can be projected onto a 1-dimensional probability space resulting in *probability expectation* $E(\omega_p)$ value for a given opinion:

$$E(\omega_p) = E(\langle b, d, u \rangle) = b + \frac{u}{2} \tag{2}$$

When ordering opinions the following rules (listed by priority) hold:

1. The opinion with the greatest probability expectation E is the greatest.
2. The opinion with the smallest uncertainty is the greatest.

Thus, for instance, $\langle 0.5, 0, 0.5 \rangle > \langle 0.4, 0.2, 0.4 \rangle > \langle 0.2, 0, 0.8 \rangle$.

4 Assessment of the Communication Patterns

We assume tracking three communication patterns: Fan-in (from here on denoted as $f_{in,i}^t$ for node i at time moment t), Fan-out ($f_{out,i}^t$) and clustering coefficient (c_i^t).

There are two main sources of uncertainty, when analyzing communication patterns. The first is lack of knowledge about existing communication (ϵ -edges in communication graph stored in communication matrix M_c). The second is that we actually need to know which communication pattern may be considered *normal* and which may be referred to as *anomaly*. This implies referring to the history (we assume that the attack is preceded by some period of normal system functioning) - the clear sign of the ongoing attack is rapid change of communication patterns being observed.

4.1 Fan-In, Fan-Out, Clustering Coefficient

As stated above our observations of communication patterns variables ($f_{in,i}^t, f_{out,i}^t, c_i^t$) are uncertain due to ϵ -values in M_c . As each of them may be in fact of the value 0 or 1 as well, we'll use the following formula to evaluate current values of the variables:

$$x = \frac{x_{(\epsilon=0)} + x_{(\epsilon=1)}}{2} \tag{3}$$

Where x stands for any of ($f_{in,i}^t, f_{out,i}^t, c_i^t$) and $x_{\epsilon=0}, x_{\epsilon=1}$ are their values under assumption that all ϵ values in M_c equal 0 or 1 respectively.

Now we should investigate which values of the parameter are *normal* (safe). Assume that the MoA's history consists of a number of observations of Fan-in values from some starting point up to current time t . So we have $f_{in,i}^1, f_{in,i}^2, \dots, f_{in,i}^t$. Let us now consider the Fan-in as a random variable $F_{in,i}$. Thus, ($f_{in,i}^1, f_{in,i}^2, \dots, f_{in,i}^t$) is a sample of size t of $F_{in,i}$. We also assume all of the $f_{in,i}^k$ to be independent. It is commonly known that the mean value and the variance of $F_{in,i}$ can be estimated by the following formulae:

$$\bar{F}_{in,i} = \frac{1}{m} \sum_{k=1}^t f_{in,i}^k \tag{4}$$

$$S_{in,i} = \frac{1}{t-1} \sum_{k=1}^t (f_{in,i}^k - \bar{F}_{in,i})^2 \tag{5}$$

$\bar{F}_{in,i}$ and $S_{in,i}$ are thus the estimations (based on the data being at our disposal) of mean value and the variance of $F_{in,i}$. Obviously the bigger our sample is, the better

they approximate $E(F_{in,i})$ and $Var(F_{in,i})$ respectively - from this point we assume that the observations' number is big enough to state that $E(F_{in,i})$ and $Var(F_{in,i})$ are known.

Let also $E(F_{out,i})$ and $Var(F_{out,i})$ for the Fan-in, as well as $E(C_i)$ and $Var(C_i)$ for clustering coefficient be defined in the same way.

4.2 Detecting Anomalies

As the MoA detects anomalies by formulating opinions about the node i 's state using Subjective Logic, appropriate values of a tuple $\omega_{i,abnormal_x}^{MoA} = \langle b, d, u \rangle$ (MoA's opinion that the value of parameter x at node i is abnormal) must be defined with respect to the current mean values and the variances of $F_{in,i}$, $F_{out,i}$ and $F_{c,i}$. The same as in equation (3) we assume that x stands for any of $(f_{in,i}^t, f_{out,i}^t, c_i^t)$. Let us define *uncertainty*, *disbelief* and *belief* of MoA's opinion in the following way:

$$u = \frac{\sum_x \epsilon}{\sum_x 1 + \sum_x \epsilon} \tag{6}$$

Where $\sum_x \epsilon$, $\sum_x 1$ is a total number of ϵ or 1 values in M_c^t for a particular x . From the Chebyshev's inequality we can estimate the upper bound of the probability that $|\bar{F} - x|$ is greater than kS . Where \bar{F} and S are mean value and the variance of X , while X denotes the random variable related to x (in this case one of the following: $F_{in,i}^t, F_{out,i}^t, C_i^t$). According to this estimation the MoA's disbelief about the statement that x value is normal can be defined as follows:

$$d = \min(1 - u, 1 - \frac{1}{\alpha k^2}) \tag{7}$$

Where α is a coefficient which value should be set during a process of IDS tuning to the real network conditions and parameter k is:

$$k = \begin{cases} 1 \\ \left| \frac{\bar{F} - x}{\sqrt{S}} \right| \end{cases} \quad \begin{cases} \text{if } \left| \frac{\bar{F} - x}{\sqrt{S}} \right| < 1 \\ \text{if } \left| \frac{\bar{F} - x}{\sqrt{S}} \right| \geq 1 \end{cases} \tag{8}$$

Finally the MoA's belief about the statement that x value is normal is evaluated using the subsequent formula:

$$b = 1 - u - d \tag{9}$$

5 The Backtracking Analysis – Building a Local Attack Graph

In the preceding section we show how the monitoring agents establish opinions about anomalies in given discrete moments of time. Let us define a communication anomaly occurring in time moment t and site s as a tuple $An_s^t = \langle s, \overline{\omega}_s, t \rangle$ where $\overline{\omega}_s$ is the Subjective Logic's opinion associated with the anomaly ($Ex(\overline{\omega}_s)$, computed as given by (2) must exceed some required threshold to be positively recognised as an anomaly. The precise value of this threshold will be set during simulations and tuning of the IDS). As mentioned in sec.1 the task is (if only at least one anomaly was detected) to perform the backtracking analysis and produce the attack graph as defined in [7] and [8]. The generic algorithm for generating the attack graph is as follows: denote by A the set of agents; S – the set of sites of monitored system and T – the set of discrete ordered moments of time. A monitoring agent A_i observes the nodes and analyses gathered information in order to determine in given time moment t a graph $G_i^{(t)} = (S_i, R_i^{(t)})$, where $S_i \subset S$ (the set of sites monitored by given A_i) and $R_i^{(t)}$ is a binary relation on S_i such that pair $\langle s, s' \rangle \in R_i^{(t)}$ for $s, s' \in S_i$ if and only if according to knowledge of agent A_i the attack has come directly from site s to s' . Let's define L_i^t as a list of all anomalies detected in time moment t within S_i by the agent A_i . The local attack graph building algorithm invoked by A_i when L_i^t is not empty has the form:

Given: M_c^t , L_i lists for all time moments up to current moment $t \in T$.

Result: The graph $G_i^{(t)} = (S_i, R_i^{(t)})$.

BEGIN

1. Set $R_i^{(t)} = \emptyset$ (no attack relations at the beginning).
2. Pick an element An_s^t from L_i^t .
3. For An_s^t select $s' \in S_i$ such that: $M_c^t(s', s) = 1$, there exists $An_{s'}^{t'}$ at moment $t' < t$, and $Ex(\overline{\omega}_{s'})$ at time t' is maximal (in case that there are more $An_{s'}^{t'}$). If there's no appropriate $s' \in S_i$ – go to step 2. Add $\langle s', s \rangle$ to $R_i^{(t)}$. Remove An_s^t from L_i^t .
 Recursively perform step 3 for $An_{s'}^{t'}$. Goto step 2.

END.

The partial attack graphs deduced by the monitoring agents are successively sent to the managing agent MA. The task of MA is to integrate their decisions in order to determine the global attack tree representing the propagation of the attack in the whole system. Owing to this global tree one should get to know the source of the attack as well as reason about its propagation plan.

6 Conclusions and Further Work

The proposed framework for communication anomalies detection is also compatible with the other elements of multiagent IDS architecture, as presented in sec.1. The use

of Subjective Logic's operators allow us to formally join (by means of Consensus operator) opinions about the same nodes on the level of MA. The MA may compute opinions about the credibility of monitoring agents (based on their former results) and apply them to their reports via Recommendation operator or use to take decisions about changing the MoA's areas of responsibility or delegating new ones. The above possibilities along with the algorithms for MA define important directions of forthcoming research. After testing and developing strategic algorithms for the managing agents we expect to create a full-fledged multiagent IDS environment.

The paper presented approach which will be the subject of further development and experiments under Polish State Committee for Scientific Research Grant no. 3 T11C 029 29 (2005-2007).

References

1. Allman M. et.al. A First Look at Modern Enterprise Traffic, In Proc. Internet Measurement Conference, October 2005, 217-231.
2. Boykin O., Roychowdhury V. Personal Email Networks: An Effective Anti-Spam Tool, *IEEE Computer*, **38(4)** (2005), 61-68.
3. Faloutsos M., Faloutsos P., Faloutsos C., On power-law relationships of the Internet topology. In Proc.ACM SIGCOMM '99 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, August 1999, 251-262.
4. Gorodetski V., Karsaev O., Khabalov A., Kotenko I., Popyack L., Skormin V.: Agent-based model of Computer Network Security System: A Case Study. In: Proceedings of International Workshop Mathematical Methods, Models and Architectures for Computer Network Security, Lecture Notes in Computer Science, vol. 2052, Springer Verlag, Berlin Heidelberg New York (2001), 39-50.
5. Jøsang, A.: A Logic for Uncertain Probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **9(3)** (2001) 279-311
6. Jøsang, A.: A Metric for Trusted Systems. In: Proceedings of the 21st National Security Conference, NSA (1998), 68-77
7. Juszczyzyn K, Nguyen N.T., Kolaczek G., Grzech A., Pieczynska A., Katarzyniak R. Agent-based Approach for Distributed Intrusion Detection System Design. International Conference on Computational Science 2006, Lecture Notes in Computer Science 3993 (2006) 224-231.
8. Kolaczek G., Kuchtiak-Pieczynska A., Juszczyzyn K., Grzech A., Katarzyniak R., Nguyen N.T. (2005): A Mobile Agent Approach to Intrusion Detection in Network Systems. In: Proceedings of KES 2005, Lecture Notes in Artificial Intelligence 3682 (2005) 514-519.
9. Kohler E., Liy J., Paxson V., Shenker S., Observed Structure of Addresses in IP Traffic, In Proc. SIGCOMM Internet Measurement Workshop, November 2002, 253 - 266.
10. Kotenko I. et al.: Multi-Agent Modeling and Simulation of Distributed Denial-of-Service Attacks on Computer Networks, In: Proceedings of Third International Conference Navy and Shipbuilding Nowaday. St. Petersburg, (2003), 38-47.
11. Nicol D., Liljenstam M., Liu J., Multiscale Modeling and Simulation of Worm Effects on the Internet Routing Infrastructure, In Proc. Performance Tools Conference, 2003, 1- 10.

An Algorithm for Inconsistency Resolving in Recommendation Web-Based Systems

Michał Malski

Institute of Computer Science, The State Higher Vocational School in Nysa, Grodzka 19 Street,
48-300 Nysa, Poland
malskim@pwsz.nysa.pl

Abstract. This paper presents a consensus-based algorithm for resolving inconsistency in Web-based recommendation systems. In such systems it is possible that a user have his own usage path (the way of usage of the system) and system can recommend individual usage path for new user. The problem of determining a new path for a new user is the subject of this paper. An algorithm for solving this problem will be described below and some results of the experiments with this algorithm will be presented.

1 Introduction

Nowadays recommendation systems are very popular in many domains of management and computer science. In earlier paper [6] methods for solving conflicts in recommendation systems are described. In that paper an example of operation of recommendation for Tax Declaration Making system has been analyzed and discussed. A conflict solution problem has been formulated and justified practically.

In this paper an algorithm for solving conflicts in recommendation systems, which is consistent with the above-mentioned conception is presented. Firstly, in Section 2 we will review the problem of solving conflicts in recommendation systems and present the foundation of the conception of solution for this problem. Next we will present an algorithm for consensus calculation, which minimizes of sum of distance function proposed for Tax Declaration Making system. In the next step we will show results of experiments executed using program in which this algorithm is implemented.

In many books, articles and papers we can find out that recommendation takes important place in different domains of life. Recommendation is an important aspect of marketing of enterprises, which help in offering products. The strategy of formulating the sequence of products in offers, which are prepared for different groups of potential clients is very important element of marketing policy of a firm. We can give many examples of such situation, where recommendation systems can be useful for enterprises. For example, we can use recommendation system for creating offers by ISP. Actually, nobody wants to provide only physical access to the Internet. It happens because the monthly cost of connection to global network for individual users is too small. So ISPs provide also services like Voice over IP, Telephone over IP, Video over IP or hosting services like eDisk, disk space for data or websites, email accounts, domains and sub-domains. Besides, such providers try to show the advantages of

connection to the big and fast local network. They share access to the local game servers, local peer-to-peer file sharing servers (local hubs) etc. As we can see there are many products (services) in offers of ISP. Basing on information about a potential client we can create his profile and classify him to the group of user, who own similar profiles. If we know which of offered services are using by each member of this group, we can prepare precise offer for a new member. It is then recommendation process and as we can see, it is a fundamental aspect of marketing one-to-one.

In [1]-[5] we can find more examples of employment of recommendation systems. Paper [1] explains the design of the system architecture for adaptive web page recommendation service and the results of experiment, evaluating recommendations provided to a group of test users. In [2] authors show us a client side personal recommendation system for news websites, which collect information using in recommendation process during its usage of online news websites. Authors of [3] propose an algorithm to classify users into groups and recommend product items based on these classified groups in system, which can help online merchants to make business decision respect concrete customer. An intelligent assistant designed to help people in making decisions, specifically a restaurant has been described in [4]. This paper explains base of recommendations in very good way. Paper [5] presents classification algorithm basing on melody style (as one of the music features to represent user's music preference) and three recommendation methods based on this classification.

In the algorithm for resolving inconsistency in Web-based recommendation system will be reviewed. The detailed description of this conception is included in [6]. In the next chapter only the most important information needed to present the mentioned algorithm will be mentioned.

2 Web-Based Recommendation System

Recommendation system described in my earlier paper [6] is an example of recommendation for Web-based Tax Declaration Making system. In the environment of functioning such system we can single out groups of users, which must to fill in the same tax forms. So they must fill in the same number of fields into these forms.

Basing on answers to the questions about sources of revenues and expenses of user in the previous year system can add him to the one of groups mentioned above. Such process consists of two actions: creating profile of user and classification. Each one of "old" members have his own (the best in his opinion) way of usage of this system (his own sequence of fields to fill in). This sequence we can be called usage path. So we have usage paths of all "old" members of group and we want to give proposition of usage path for new member. It is not easy task, because usage paths of members of the same group could not be identical and in effect we have some kind of conflict.

The solution presented in [6] relies on using consensus methods for solving conflict described above. With reference to these methods U is denoted the universe of all potential profiles, where the profile represents a usage path of user. In the example of Tax Declaration Making system by usage path we denote a sequence of act indexes, which represents the order of using the system. So, if there are 5 acts $\{1,2,\dots,5\}$ then a usage path might be for example: $\langle 5,2,4,1,3 \rangle$. For given usage paths:

c_1, c_2, \dots, c_n of all n members of the same group we can determine usage path c^* for new user using following criterion [7], [8]:

$$\sum_{i=1}^n d(c^*, c_i) = \min_{c \in U} \sum_{i=1}^n d(c, c_i).$$

According to this concept, the definition of distance function d is proposed in [6]. Likewise in this paper by distance between 2 sequences of act indexes we take the minimal weight of act indexes shifts needed for transforming one sequence into other. So, for given 2 sequences: $c_1 = \langle 5, 2, 4, 1, 3 \rangle$ and $c_2 = \langle 3, 1, 4, 2, 5 \rangle$ the weight for shifting index 1 is equal 2; for index 2 – 2; for index 3 – 4; for index 4 – 0; and for index 5 – 4. Thus the distance is equal $d(c_1, c_2) = 2 + 2 + 4 + 0 + 4 = 12$.

The problem of creating usage path for a new member becomes a problem of such choice of sequence from U , which guarantees the minimal sum of distances between chosen sequence and given sequences.

3 Inconsistency Resolving Algorithm for Recommendation System

Let us now describe formal algorithm for determination of consensus of given sequences presented above:

Input: Set of acts $A = \{1, 2, \dots, n\}$, set of usage paths $C = \{c_1, c_2, \dots, c_m\}$, where sequence $c_i = \langle c_i^1, c_i^2, \dots, c_i^n \rangle$, $c_i^j \in A$ and c_i^j is unique in c_i .

Output: Sequence c^* for which $\sum_{i=1}^m d(c^*, c_i) = \min_{c \in U} \sum_{i=1}^m d(c, c_i)$.

BEGIN

1. create matrix M of size $n \times n$, which all elements are equal 0
2. for each one of elements of C
for each one of elements of A
add one to element $m_{i,j}$ of matrix M if position of act number i in the element of C is equal j
3. create matrix D (matrix of distances) of size $n \times n$ of elements calculated by formula:

$$d_{i,j} = \sum_{m=1}^m [m_{i,m} \cdot |m - j|]$$

4. n times do:
 - find first minimal element of matrix D , which value is greater or equal 0 and remember its number of row as i_{\min} and number of column as j_{\min}
 - calculate $c_{j_{\min}}^* = i_{\min}$
 - change to -1 values of all elements of D in the row with number i_{\min} and in the column with number j_{\min}
5. recreate matrix D such like in 3

6. calculate sum of the distances between c^* (appointed in 4) an each of elements of C as following:

$$\sum_{i=1}^m d(c^*, c_i) = \sum_{j=1}^n d_{(c^j)^*, j}$$

END.

Now we can show an example of solving described problem using presented algorithm. First, we can create a matrix M , elements of which will inform us how many times given act has taken a stand on given position in a set of given sequences. For example, if an element $m_{i,j}$ of matrix M is equal 10, then we know that act i has taken a stand on position j ten times in a set of given sequences. If we have n acts then matrix M will be a square matrix of size $n \times n$.

In the situation, when 5 acts $\{1,2,\dots,5\}$ and example of set of nine sequences:

$$\begin{aligned} c_1 &= \langle 2,1,3,5,4 \rangle, c_2 = \langle 1,2,3,4,5 \rangle, c_3 = \langle 3,4,2,1,5 \rangle, \\ c_4 &= \langle 4,5,3,2,1 \rangle, c_5 = \langle 5,3,1,2,4 \rangle, c_6 = \langle 1,2,5,4,3 \rangle, \\ c_7 &= \langle 4,5,1,3,2 \rangle, c_8 = \langle 3,1,4,2,5 \rangle \text{ and } c_9 = \langle 3,2,4,5,1 \rangle \end{aligned}$$

are given, the matrix M has size 5×5 and

$$M = \begin{bmatrix} 2 & 2 & 2 & 1 & 2 \\ 1 & 3 & 1 & 3 & 1 \\ 3 & 1 & 3 & 1 & 1 \\ 2 & 1 & 2 & 2 & 2 \\ 1 & 2 & 1 & 2 & 3 \end{bmatrix}.$$

Now, we can prepare matrix D , which the rows will represent number of act and columns will represent number of act position in sequence, such like in matrix M . Each element of this matrix is calculated basing on matrix M , like it is shown below.

$$d_{i,j} = \sum_{m=1}^n [m_{i,m} \cdot |m - j|].$$

It is the effect of reviewed in previous chapter proposition of distance function, what we can present basing on following example:

$$d_{1,1} = 2 * 0 + 2 * 1 + 2 * 2 + 1 * 3 + 2 * 4 = 17,$$

because if we decide to put act 1 on first position into c^* , we have 2 sequences with act 1 on first position and distance between positions of act 1 is equal 0, 2 sequence with act 1 on second position and distance between positions of act 1 is equal 1, 2 sequence with act 1 on third position and distance between positions of act 1 is equal 2, 1 sequence with act 1 on fourth position and distance between positions of act 1 is equal 3 and 2 sequence with act 1 on fifth position and distance between positions of act 1 is equal 4.

For matrix M from example above we have

$$D = \begin{bmatrix} 17 & 12 & 11 & 14 & 19 \\ 18 & 11 & 10 & 11 & 18 \\ 14 & 11 & 10 & 15 & 22 \\ 19 & 14 & 11 & 12 & 17 \\ 22 & 15 & 12 & 11 & 14 \end{bmatrix}.$$

Matrix D can be called the matrix of distances. Having this matrix it is possible to calculate sum of distances between given sequence and a set of sequences, because for given $c = \langle c^1, c^2, \dots, c^n \rangle$ and set of m sequences we have

$$\sum_{i=1}^m d(c, c_i) = d_{c^1,1} + d_{c^2,2} + \dots + d_{c^n,n} = \sum_{j=1}^n d_{c^j,j}.$$

For described above example with 9 sequences we have calculated matrices M and D , now we can calculate sum of distance between some sequence and set of these 9 sequences. Let $c = \langle 1,5,4,2,3 \rangle$, then

$$\sum_{i=1}^9 d(c, c_i) = d_{1,1} + d_{5,2} + d_{4,3} + d_{2,4} + d_{3,5} = 17 + 15 + 11 + 11 + 22 = 76.$$

As we can see the problem of minimization of sum of distance between a usage path for new member and set of usage paths of "old members" (problem of choice c^*) becomes the problem of choice n elements of matrix D so that the sum of these elements will minimal. Important thing is fact, that we must chose one (no less or more) element from each row and each column of matrix D , because as a result we want to have sequence of all acts so act can not be repeated and two or more acts can not be on the same position.

We suggest to search out first minimal element of matrix D and remember number of row and number of column of this element. Next we must put act with the same number as number of row of minimal element into c^* on position with the same number as number of column of minimal element. Finally, we must eliminate row and column with this minimal element before next searching. We repeat these three actions n times and in effect we can determine c^* .

For matrix D illustrated above, we chose elements in such sequence: $d_{2,3}$, $d_{3,2}$, $d_{5,4}$, $d_{1,1}$ and $d_{4,5}$ an basin on this we create $c^* = \langle 1,3,2,5,4 \rangle$. Next, it is possible to calculate the sum of distance between c^* and set of given 9 sequences:

$$\sum_{i=1}^9 d(c^*, c_i) = d_{2,3} + d_{3,2} + d_{5,4} + d_{1,1} + d_{4,5} = 10 + 11 + 11 + 17 + 17 = 66.$$

In the next chapter we show results of experiments executed using program, which is working according to described algorithm. Due to these results, evaluation of correctness of operations of this algorithm is possible.

4 Experimental Results

The program, which uses described algorithm has been made in order to evaluate correctness of operation of this algorithm. This program generates optional number of

lottery sequences of 5 acts. Into next step program calculates an elements of matrices M and D . Basing on matrix D program generates sequence c^* and sum of distance between c^* and generated set of sequences such like it is described in previous chapter. For eliminating row and column with minimal element before next searching out a minimal element of matrix D program replace values of elements in this column and in this row by value -1 . In the next searching these values are ignored, because distance cannot have negative value. Next, program generate all possible sequences of these 5 acts and calculate sum of distance between each of these sequences and set of sequences generated on beginning of work. It is needed to find out optimal solution c_{op}^* . On the end program calculate relative error in percents, like below:

$$rel.error = \frac{\sum_{i=1}^n d(c^*, c_i) - \sum_{i=1}^n d(c_{op}^*, c_i)}{\sum_{i=1}^n d(c_{op}^*, c_i)} \cdot 100\% .$$

The results of experiments executed using this program we have in table 1.

Table 1. Results of program functioning

n	Number of experiment	$\sum_{i=1}^n d(c^*, c_i)$	$\sum_{i=1}^n d(c_{op}^*, c_i)$	rel.error
10	1	68	68	0
10	2	60	58	3,448
10	3	60	52	15,386
10	4	68	60	13,333
10	5	72	66	9,091
10	6	64	64	0
10	7	72	70	2,857
20	1	130	130	0
20	2	152	142	7,042
20	3	142	140	1,429
20	4	140	140	0
20	5	142	134	5,97
20	6	142	140	1,429
20	7	144	138	4,348
50	1	380	372	2,151
50	2	328	328	0
50	3	374	372	0,538
50	4	362	344	5,233
50	5	382	376	1,596
50	6	366	366	0
50	7	384	372	3,226
50	8	372	372	0
100	1	780	764	2,094
100	2	736	712	3,371
100	3	738	738	0

Table 1. (continued)

100	4	742	730	1,644
100	5	742	740	0,27
100	6	756	752	0,532
100	7	758	758	0
200	1	1566	1562	0,256
200	2	1548	1526	1,442
200	3	1530	1522	0,526
200	4	1554	1532	1,436
200	5	1556	1556	0
200	6	1540	1538	0,13
200	7	1568	1540	1,818

The table above shows us that the algorithm proposed in chapter 3 is not optimal. For example described in chapter 3 $c^* = \langle 1,3,2,5,4 \rangle$ has been generated but we can find out sequence, which has smallest sum of distance to the set of given sequences. It is $c_{op}^* = \langle 3,1,2,4,5 \rangle$ and $\sum_{i=1}^n d(c_{op}^*, c_i) = 62$. The relative errors are consequence of eliminating row and column with this minimal element. As it happens, that such choice of 5 minimal elements of matrix D not always cause that the sum of distance between c^* and the set of given sequences is minimal too. However, it is possible to be satisfied, because the relative error is small. For 10 lottery generated sequences averaged relative error is equal 6,30 percents, for 20 sequences it is equal 2,89 percents, for 50 sequences – 1,82 percents, for 100 sequences – 1,13 percents and for 200 – 0,8. In the conclusion we can say that the value of averaged relative error is going down when the number of lottery generated sequences is going up, what is showed on figure below.

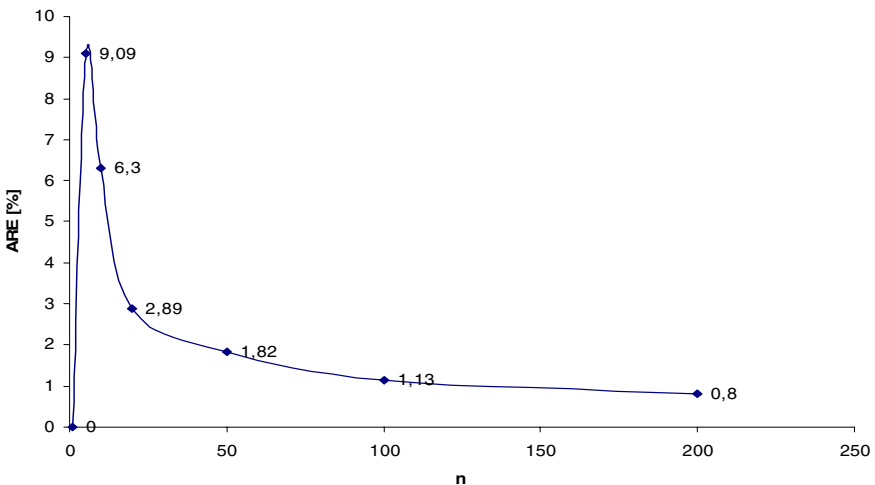


Fig. 1. Relation between number of generated sequences (n) and averaged relative error (ARE)

5 Conclusions

The recommendation algorithms are very popular since there was a need to dynamically fitting software to the user (his requirements, profiles, preferences etc.). Here we have presented one of such algorithms, which base on the one of consensus methods. Basing on experimental results executed using this algorithm the relative error of this algorithm was calculated and presented here. The results of experiments are congenial to optimal solutions so we can say that this algorithm is heuristic. In the future work it will be showed that described problem is NP-complete and the future work will concentrate on heuristic methods to modify the given algorithm to reduce the value of averaged relative error.

References

1. Balabanović M.: An adaptive Web page recommendation service. In: Proceedings of the first International Conference on Autonomous Agents (ICAA'97) (1997) 378 – 385
2. Bomhardt C.: NewsRec, a SVM-driven Personal Recommendation System for News Websites. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'04) (2004) 545-548
3. Chiu C.-F., Shih T.K., Wang Y.-H.: An Integrated Analysis Strategy and Mobile Agent Framework for Recommendation System in EC over Internet. *Tamkang Journal of Science and Engineering* 5(3) (2002) 159-174
4. Göker M.H., Thompson C.A.: The Adaptive Place Advisor: A Conversational Recommendation System. In: Proceedings of the 8th German Workshop on Case Based Reasoning, Lammerbuckel, Germany (2000) 187-198
5. Kuo F.-F., Shan M.K.: A Personalized Music Filtering System Based on Melody Style Classification. *Second IEEE International Conference on Data Mining (ICDM'02)* (2002) 649-652
6. Malski, M.: Resolving inconsistencies in recommendation Web-based systems. *Proceedings of the 11th System Modelling Control Conference (SMC'05)* (2005) 189-194
7. Nguyen, N.T.: *Methods for resolving conflicts in distributed systems*. Monograph, Wroclaw University of Technology Press (2002)
8. Nguyen, N.T.: Consensus System for Solving Conflicts in Distributed Systems. *Journal of Information Sciences* 147 (2002) 91-122

Distributed Class Code and Data Propagation with Java

Dariusz Król and Grzegorz Stanisław Kukla

Institute of Applied Informatics, Wrocław University of Technology,
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
dariusz.krol@pwr.wroc.pl, grzegorz_kukla@o2.pl

Abstract. This paper addresses the problem of distributed class code and data propagation with Java. Traditional approach based on problem-oriented structures and on predefined task language is not suitable for universal grid programming. The main contribution is the development of an automatic framework for efficient propagation of class package and data. We examine two problems suitable for code and data distribution: large n-merge sorting and document indexing. Thanks to the use reflection mechanism, we show that Java is adequate for defining new tasks on grid elements without any language extension. Relation between number of component nodes of the structure and total processing time has been checked. Furthermore the framework is fault-tolerant when some nodes fail.

1 Introduction

The introduction of reflection [8] revolutionized the field of programming methods by achieving dynamic possibilities than any other method at the time. Since then, much research has focus on design, implementation, and analysis of reflection and their generalization [12].

One of the main goals of this research has been to achieve good performance of traditional algorithms building a low cost high performance computing system [6]. There are many examples of such projects, i.e. Einstein@Home, LHC@Home, and Climateprediction.net.

The drawback to traditional grid computing is that it is not guaranteed to converge to successive end, nor does it has any guarantee on the quality of its output. Unfortunately, a program cannot run on a grid if it has not been designed for it. The division of code and location of tasks have to be coded individually. Some progress has been made assuming strict Java coding, for which algorithm is known. To minimize the impact of such obstacles, the reflection API can be used.

Java technology is very promising because the Java language has been designed with architecture independence in mind. Java programs have the ability to propagate on any computer through reflection and serialization using the RMI package. Hence using Java for grid programming seems to be a good idea.

The paper is organized as follows. Section 2 describes the main characteristics of link between code and data propagation and programming issues. The implementation details are introduced in Section 3 and empirical evaluation is sketched in Section 4. We conclude with a discussion and some hints about future work in Section 5.

2 Link Between Propagation and Programming Issues

The concept of code propagation in form of computer virus is relatively old. From the beginning of the eighties till now researchers are modeling viral code replication and propagation behavior. Creating such models is beneficial to better understand the threat posed by new virus attack and to develop improved models for disinfection and cleanup [10].

Another propagation approach is needed for applications of wireless sensor networks (WSNs) [4]. Many of these applications require dynamic remote reprogramming of sensor nodes through the wireless channel. In typical example, the base station contains a code image that needs to be propagated to the sensor nodes. Propagation should not take more than a minute more than the time required for transmission.

In heterogeneous networks in order to coordinate the propagation process the concept of keylets are introduced [11]. Keylets are mobile code used solely to direct the propagation of keys which are necessary to reveal and execute mobile agent code on any platform. A keylet executing at one platform will need to distribute specific fragments to other platforms. Thus propagation consists of two operations: suspension of execution until needed fragment becomes available, and distribution of fragments to a new platform.

The Java programming environment includes features that are suitable for high-performance distributed computing [9]: multi-threading, remote method invocation, sockets, servlets, etc. and parallel programming [7]. Usually in this environment there are implemented the following problems: square root, matrix multiply, Mandelbrot sets, Laplace solver, etc. The development of corresponding programs is hard task. Besides coding the algorithm details, the programmer must also take care of data allocation, communication and synchronization handling.

In paper [3] the authors present a static analysis, which estimates exception propagation paths of Java programs. This propagation information can guide programmers to uncaught exceptions. Due to paths mining, exceptions will be handled more specifically and extra handlers by tracing exception propagation can be implemented. This idea is very useful also in distributed environment when attempting to detect an error.

Our approach is similar to these cited above implementations. We propose a Java based framework that provides support for distributed high performance computing (DHPC). We underline two common characteristics: multiple modules needed to cooperate in solving problems are dynamically instantiated on different heterogeneous nodes. Currently, there is a lack of frameworks for code and data propagation with dynamic integration in a network.

3 Implementation Details

Reflection is defined as the activity performed by an agent when doing computations about itself [2] or is the act of reasoning itself about a computational system [5]. A reflective approach permits easy separation of management code from application code. Due to such consideration our and many others implementations have been developed.

Current research in so called reflective middleware is focused on improving customizations and more transparent use. From the other side using reflection in Java almost always requires run-time type casting, leading to potential dynamic failure. The programmer must write code to handle exceptions when run-time check fails. Furthermore, reflection breaks user-defined abstractions. But, reflection is indispensable for our system. Although other communication mechanisms such as CORBA could have been employed as well, we designed first prototype exclusively based upon Java. Grid generated from the source program is 100% Java; it consists of a set of Java programs running on a set of Java Virtual Machines.

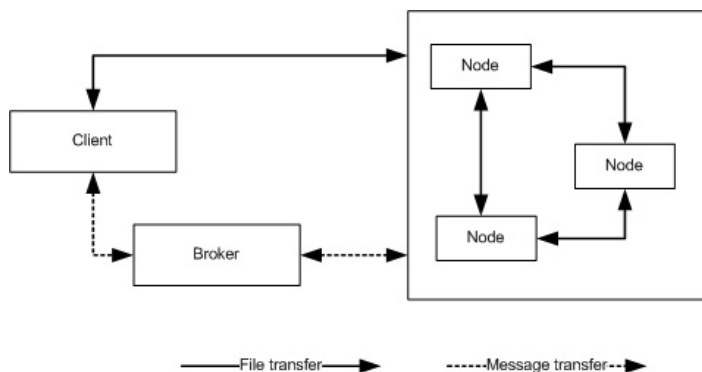


Fig. 1. Communication schema of universal grid structure

The communication schema of the framework is shown in Fig. 1. Universal grid consists of node application, broker application and a client library. Node and broker applications are stand-alone Java executable ready to deploy. Client library provides a set of classes allowing a programmer to develop client application which can communicate with the other applications mentioned. The task of a programmer is to implement the algorithm executed by the structure. Communication, data exchange, task dispensing, error handling is provided by the framework.

Client, the source of original data does not take part in computation; he takes part in task division, passes classes' code and data to the nodes and receives output data. Client is the primary source of all actions performed by the system. It commissions the grid structure to execute tasks. It is also the primary source of all data. Client implementation comes to programmer duties.

Broker is a coordination centre for all grid elements. He is responsible for optimization and task dislocation due to possible error occurrences. Broker receives task execution requests sent by the client. It is responsible for dispensing these requests to executive elements of the structure - nodes. Broker does not take part in computing. It just passes a request from the client to the proper node. Broker also manages information about all resources of the structure. It holds data about every file and machine involved in computing process. Broker makes this information accessible to the other grid structure elements in order to make communication between them possible.

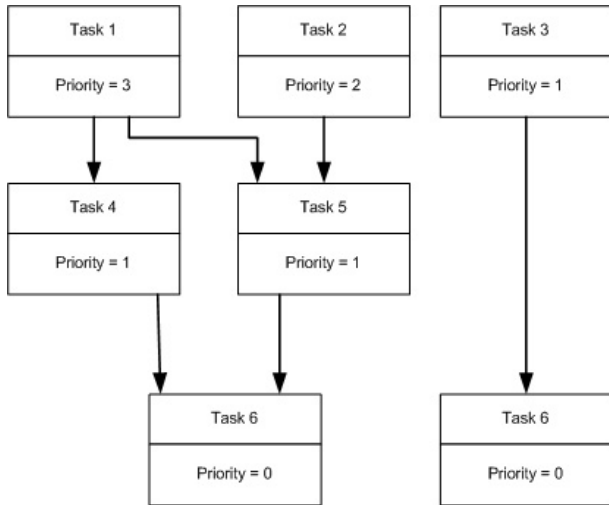


Fig. 2. Task priority computation algorithm used by broker

Fig. 2 presents the task priority computation algorithm used by broker. A priority of a certain task depends on how many other tasks are waiting for it to finish. The more tasks depend on certain task, the higher priority it has. They may change, because they are calculated every time a node requests a task from the broker. Broker uses tasks priorities to choose the proper task for the certain node.

Node computes delegated task and passes data to other node or back to client. Nodes can cooperate themselves. Node is the executive element of the structure. It executes tasks requested by the client. Data exchange between client and nodes is served by file transfer protocol. All of the nodes and client itself are able to share data. A result of task executed by one node can be an input to the task executed by the other. Nodes also can share data files.

In order to write a working application using a skeleton based Fig. 3, the programmer must perform the following steps.

1. First, the programmer has to implement tasks classes; extend the task state abstract class to fit his needs; implement a set of code blocks used to carry

out a task itself; make a JAR package consisting of the class files, name of this file will be used as a task name.

- Then, the programmer must extend the abstract client class; this class provides all the functionality needed to communicate with the broker; the task of the programmer is to implement an algorithm carried out by the grid structure.

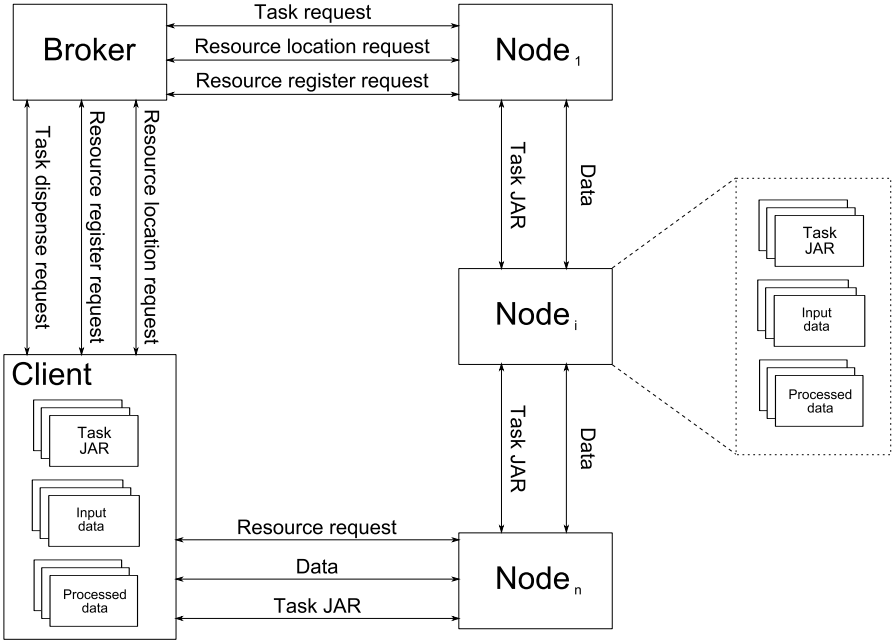


Fig. 3. Distributed architecture of framework elements

The proposed method is fault-tolerant because it does not require all of the elements of the structure to be connected and available all the time. First of all, every executive element is a source of the data. Once obtained file can be shared to the other elements. It lowers a possibility of lost of data. Also task dispensing protects the system from deadlocks. If a node that had been given a task is unavailable for a while, the broker assigns that task to the other node. Applied data exchange protocol minimizes the possibility of data corruption.

4 Empirical Evaluation

In order to assess project performance, we have done a set of experiments on a Windows XP cluster operated at our department. The cluster used for experiments hosts 16 nodes (1 GHz Pentium). The client, broker and nodes are interconnected by a local Fast Ethernet network. The experiments have been performed using JDK version 1.4.

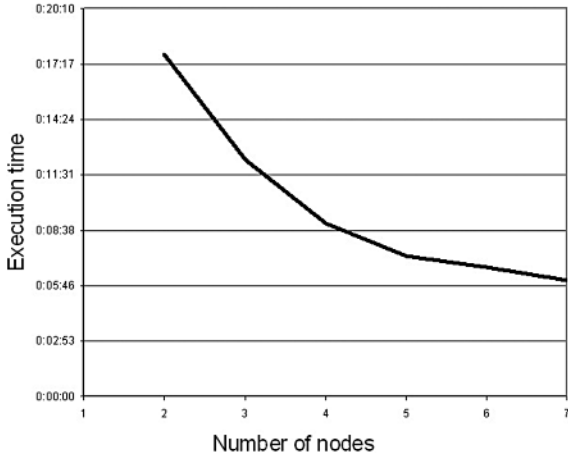


Fig. 4. Performance index for merge sorting

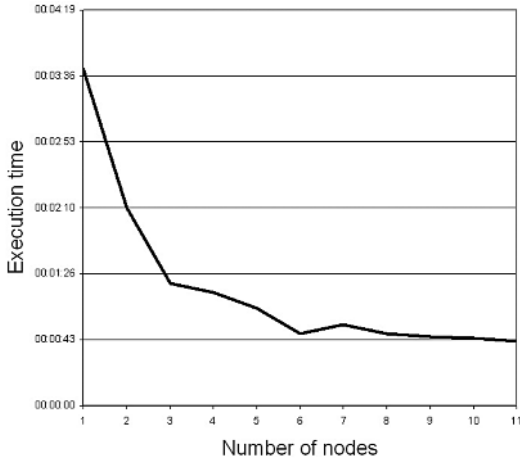


Fig. 5. Performance index for document indexing

We have implemented two test programs in order to estimate the cost of using this environment: large n-merge sorting and document indexing.

The first experiment consisted in distributed merge sorting a large set of data. Test data source includes 100 000 random generated elements.

The second one carried out distributed indexing of a selected web site. The input data consisted of 478 HTML pages from New York Times website of entire size 27 MB. The task consisted in extracting a set of indexing terms from these pages.

First of all we measured the application absolute execution time. The execution times show an additional decrement from 2 to 6 nodes onwards. Therefore it takes a shorter time to execute Java code in our grid. Second, we considered the

overhead introduced by distribution. This is visible only after 6-7 nodes added to the grid. Performance results are shown in Fig. 4-5.

During the second experiment the 7th node was intentionally shut down (in case occurrence of failure) to check the reaction of the system. The system passed the exam, see Fig. 5.

The speed-ups obtained with the algorithms show that using this method is a good solution that does not induce noticeable operating cost.

5 Conclusions and Future Work

The paper presents the prototype tool supporting universal grid computing of Java programs. The tool implements two algorithms targeting various types of computations in distributed network.

Two factors have motivated this tool. First, grid computing is a powerful and computationally expensive technique that can not be considered without the efficient aid to automation the process. Second, the Java language integrates reflection and object-oriented features that can be the basis for code and data propagation. We performed experiments that demonstrate that good scalability and efficiency figures can be achieved.

Another observation is that, that this structure has a dynamic limit to compute algorithms in an efficient way. The more nodes the better efficiency but costs also grow.

This approach is not convenient for implementing any algorithm. The main obstacle lies in the fact that the broker passes a class code to a node in an automated way and does not know how many nodes are already implemented. We need additional manager to code replication.

A key and novel idea is to combine code analysis and reflection mechanism for implementing low cost supercomputer. Java grid framework has moreover been developed in such a way that it can easily be extended with other algorithms.

First evaluation results allowed us to draw promising conclusions. Further research will be performed through other case studies to use high cost algorithms and provide practical hints how to build supercomputer using automatic code propagation.

References

1. Aldinucci M., Danelutto M., Teti P.: An advanced environment supporting structured parallel programming in Java. *Future Generation Computer Systems* **19** (2003) 611–626
2. Cazzola, W.: Remote method invocation as a first-class citizen. *Distrib. Comput.* **16** (2003) 287–306
3. Chang, B.M., Jo, J.W., Her S.H.: Visualization of Exception Propagation for Java using Static Analysis. In: *Proceedings of the Second IEEE International Workshop on Source Code Analysis and manipulation* (2002) 1–10

4. Deng, J., Han, R., Mishra, S.: Secure Code Distribution in Dynamically Programmable Wireless Sensor Networks. Technical Report CU-CS-1000-05. University of Colorado at Boulder (2005)
5. Gybels, K., Wuyts, R., Ducasse, S., Hondt, M.: Inter-language reflection: A conceptual model and its implementation. *Computer Languages, Systems and Structures* **32** (2006) 109–124
6. Haeuser, J. et al.: A test suite for high-performance parallel Java. *Advances in Engineering Software* **31** (2000) 687–696
7. Launay, P., Pazat, J.L.: Easing parallel programming for clusters with Java. *Future Generation Computer Systems* **18** (2001) 253–263
8. Laure, E.: OpusJava: A Java framework for distributed high performance computing. *Future Generation Computer Systems* **18** (2001) 235–251
9. Matsuoka, S., Itou, S.: Towards performance evaluation on high-performance computing on multiple Java platforms. *Future Generation Computer Systems* **18** (2001) 281–291
10. Serazzi, G., Zanero, S.: Computer Virus Propagation Models. IEIIT-CNR Institute (2001) 1–25
11. Tan, H.K., Moreau, L.: Mobile code for key propagation. *Electronic Notes in Theoretical Computer Science* **63** (2001) 1–22
12. Weirich, S., Huang, L.: A design for type-directed programming in Java. *Electronic Notes in Theoretical Computer Science* **138** (2005) 171–136

Conflicts of Ontologies – Classification and Consensus-Based Methods for Resolving

Ngoc Thanh Nguyen

Institute of Information Science and Engineering, Wroclaw University of Technology, Poland
thanh@pwr.wroc.pl

Abstract. Ontology can be treated as the background of an information system. If integration of some systems has to be performed, their ontologies must also be integrated. In this process it is often needed to resolve conflicts between ontologies. In this paper conflicts of ontologies are classified into classes and for each of them a method for conflict solving is proposed.

1 Introduction

Ontology has been known to be a very useful tool in defining the “spirit” or a “background” of an information system. In database systems this background is the database scheme, which consists of such elements as a set of attributes with their domains; a set of dependencies between the attributes and a set of relationships between tables. The data or knowledge which appears in the database has to comply with this scheme.

Most often ontology is defined by the following elements [4], [6]:

- C – a set of concepts (classes);
- I – set of instances of concepts;
- R – set of binary relations defined on C ;
- Z – set of axioms, which are formulas of the first order logic and can be interpreted as integrity constraints or relationships between instances and concepts, and which can not be expressed by the relations in set R .

In general, a conflict between 2 ontologies takes place when they reflect the same real world for different systems. Figure 1 presents such situation where for the same real world there are 4 ontologies defined for 4 systems. These ontologies may differ from each other. However, this definition is not always accurate because the difference of ontologies may appear on different levels and in different parameters. Besides, the areas of the real world occupied by these systems may be different, or only partly overlap. Remer [10] has considered Knowledge Integration in two aspects: Integration of different knowledge bases and integration of different representations of the knowledge but on different levels of representation. Our work is focused on the first aspect. In this work we deal with conflict resolution of ontologies which contain different states of knowledge about the same real world. We perform a classification of ontology conflicts and propose consensus-based methods for their integrating.

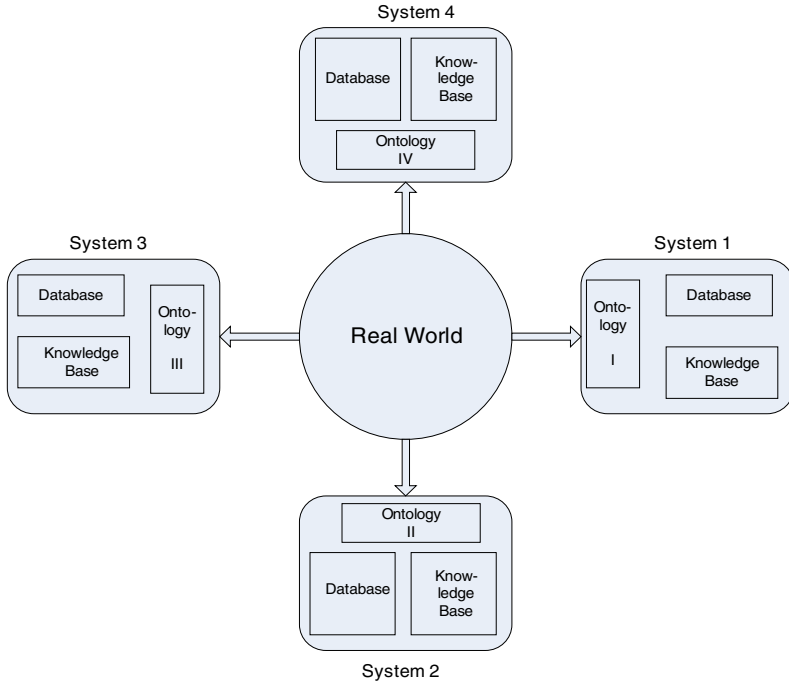


Fig. 1. A possible conflict situation of ontologies

2 Problems of Ontologies Integration

Recently, there has been increased interest in ontologies and in various tasks related to processing them, such as ontology mismatch, ontology merging or ontology integration [2], [5], [9]. In general, the problem of ontology integration can be formulated as follows: *For given ontologies O_1, \dots, O_n one should determine one ontology which best represents them.* Ontology integration is useful when there is a need to make a fusion (or merge) of systems in which ontologies O_1, \dots, O_n are used. The final system needs the ontology which arises in result of integration of these ontologies. The process of ontology integration is illustrated on Figure 2.

Notice that not always in ontology integration processes it is needed to solve conflicts of ontologies. Conflicts of some ontologies may arise only when they refer to the same real world. As stated in the Introduction, the subject of our work is focused on the aspect that the same real world object is differently reflected in ontologies. Thus during the integration process there will be the problem of choice of the proper description of the object.

In this work we consider 3 levels of ontology conflicts: *instance level*, *concept level* and *relation level*. For each level the conflict will be defined and solved with using consensus methods. Consensus methods [3], [7] have been shown to be useful in conflict solving in distributed environments, among others, in knowledge inconsistency processing [8].

Solving conflicts of ontologies is a particular case of inconsistency processing tasks. These tasks are very popular in many systems or environments [1]. Consensus methods are useful not only in case when there are different versions of knowledge or data about the same real world objects, but also in case when these versions refer to different real world parts. This is realized by postulates defined for consensus choice. In work [7] we defined a set of postulates, which in general are not consistent with each other. However, inconsistency of postulates means that particular postulates may be used in different conflict situations.

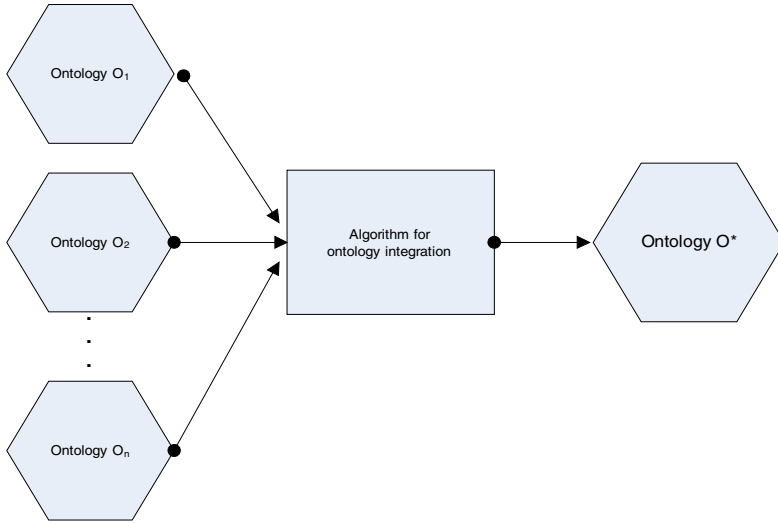


Fig. 2. Ontology integration process

3 Conflicts Between Ontologies

As stated in Introduction, by an ontology we understand a quadruple:

$$(C, I, R, Z).$$

We assume a real world (A, V) where A is a finite set of attributes and V – the domain of A , that is V is a set of attribute values, and $V = \bigcup_{a \in A} V_a$ (V_a is the domain of attribute a). We consider domain ontologies referring to the real world (A, V) , such ontologies are called (A, V) -based. In this paper we accept the following assumptions:

1. A concept is defined as a triple:

$$concept = (c, A^c, V^c)$$

where c is the unique name of the concept, $A^c \subseteq A$ is a set of attributes describing the concept and $V^c \subseteq V$ is the attributes' domain: $V^c = \bigcup_{a \in A^c} V_a$. Pair (A^c, V^c) is called the *structure* of concept c . It is obvious that all concepts belonging to the same ontology are different with each other. However, notice that within an ontology there may be 2 concepts with the same structure. Such situation may take place, for example, for

concepts “*person*” and “*body*”. In this case the relations from set \mathbf{R} will be very useful.

2. An instance of a concept c is described by the attributes from set A^c with values from set V^c . Thus an instance of a concept c is defined as a pair:

$$instance = (id, v)$$

where id is the unique identifier of the instance in world (\mathbf{A}, \mathbf{V}) and v is the value of the instance, which is a tuple of type A^c and can be presented as a function:

$$v: A^c \rightarrow V^c$$

such that $v(a) \in V_a$ for $a \in A^c$. All instances of the same concept in an ontology are different with each other.

By $Ins(O, c)$ we denote the set of instances belonging to concept c in ontology O . We have

$$I = \bigcup_{c \in C} Ins(O, c)$$

3. Between a pair of concepts there may be defined one or more relations.

4. Let O_1 and O_2 are (\mathbf{A}, \mathbf{V}) -based ontologies. A concept (c, A^c, V^c) belonging to O_1 is identical with concept $(c', A^{c'}, V^{c'})$ belonging to O_2 if and only if $c = c'$.

5. Two instances (id, v) and (id', v') are identical if and only if they have the same identifier, that is $id = id'$.

Conflicts between ontologies may be considered on the following levels:

- Conflicts on instance level: The same instance belonging to the same concept has different values in different ontologies.
- Conflicts on concept level: The same concept has different structures in different ontologies.
- Conflicts on relation level: The relations for the same concepts are different in different ontologies.

Conflicts mentioned in above way are very general and inaccurate. In the following sub-sections we will provide with the concrete definitions of them.

3.1 Conflicts on Instance Level

On this level we assume that 2 ontologies differ from each other only in values of instances. That means they may have the same concepts and relations between them.

Definition 1. Let O_1 and O_2 be (\mathbf{A}, \mathbf{V}) -based ontologies. Let concept (c, A^c, V^c) belongs to both ontologies and let the same instance i belongs to concept c in each ontology, that is $(i, v_1) \in Ins(O_1, c)$ and $(i, v_2) \in Ins(O_2, c)$. We say that a conflict takes place if $v_1 \neq v_2$.

As an example let us consider ontologies of two information systems of two universities. Consider concept “*student*” which has the following structure:

$$(\{Name, Age, Address, Specialization, Results\}, V)$$

where V is the set of attributes’ domain. It is then obvious that these two ontologies are in conflict on instance level if referring to a student who studies in both universities the values of the instances representing him will be different (for example referring to attributes *Specialization* and *Results*). A conflict situation may be presented as follows:

<i>Ontology</i>	<i>Name</i>	<i>Age</i>	<i>Address</i>	<i>Specialization</i>	<i>Results</i>
O_1	Nowak	20	Wroclaw	Information Systems	4.5
O_2	Nowak	20	Wroclaw	Discrete Mathematics	4

For solving conflicts of ontologies on instance level consensus methods seem to be very useful. Consensus methods in general deal with determining for a set of different versions of data (so called a *conflict profile*) such a version which at best represents the given versions. Different criteria, structures of data and algorithms have been worked out [7]. For this kind of conflict the consensus problem can be defined as follows:

Given a set of values $X = \{v_1, \dots, v_n\}$ where v_i is a tuple of type A , that is:

$$v_i: A^c \rightarrow V^c$$

for $i=1, \dots, n; A^c \subseteq A$ and $V^c = \bigcup_{a \in A^c} V_a$ one should find a tuple v of type A , such that one or more selected postulates for consensus are satisfied.

For this problem a consensus system, which enables describing multi-valued and multi-attribute conflicts has been defined and analyzed [7]. The structures of tuples representing the contents of conflicts are defined as distance functions between these tuples. Two distance functions (ρ and σ) have been defined. The consensus and the postulates for its choice are defined and analyzed. For defined structures and particular postulates algorithms for consensus determination have been worked out. One of proposed postulates (called *Superiority of knowledge*) is suitable to the situation described in the above example. Referring to this situation it means that consensus should contain both information that Nowak studies “Information Systems”, as well as “Discrete Mathematics”. Another very popular postulate requires minimizing the following sum

$$\sum_{i=1}^n d(v, v_i) = \min_{v' \in T(A^c)} \sum_{i=1}^n d(v', v_i)$$

where $T(A^c)$ is the set of all tuples of type A^c .

3.2 Conflicts on Concept Level

On this level we assume that 2 ontologies differ from each other in the structure of the same concept. That means they contain the same concept but its structure is different in each ontology.

Definition 2. Let O_1 and O_2 be (A, V) -based ontologies. Let concept (c_1, A^{c1}, V^{c1}) belongs to O_1 and let concept (c_2, A^{c2}, V^{c2}) belongs to O_2 . We say that a conflict takes place in concept level if $c_1 = c_2$ but $A^{c1} \neq A^{c2}$ or $V^{c1} \neq V^{c2}$.

Definition 2 specifies such situations in which two ontologies define the same concept in different ways. For example concept *person* in one system may be defined by attributes: *Name, Age, Address, Sex, Job*, while in other system it is defined by attributes: *Id, Name, Address, Date_of_birth, Taxpayer's identification number, Occupation*.

The problem is the following:

For given a set of pairs

$X = \{(A^i, V^i) : (A^i, V^i) \text{ is the structure of concept } c \text{ in ontology } O_i \text{ for } i=1, \dots, n\}$
 it is needed to determine a pair (A^*, V^*) which best represents the given pairs.

The word “best” means one or more postulates for satisfying by pair (A^*, V^*) . In general, we would like to determine such set A^* of final attributes that all attributes which appear in sets A^i ($i=1, \dots, n$) are taken into account in this set. However, we cannot simply make the sum of A^i . We should investigate the following relations between attributes:

- *Equivalence*: Two attributes are equivalent if they have the same domain and semantics, for example, attributes *Occupation* and *Job* are equivalent. We denote this relation by symbol “ \Leftrightarrow ”, for example:

$$Occupation \Leftrightarrow Job.$$

- *Generalization*: An attribute a is more general than another attribute b if the information included in fact (i, a, x) , that is attribute a referring to instance i has value x , may be implied from the information included in fact (i, b, y) for some x and y . For example, attribute *Age* is more general than *Day_of_birth* because if the day of birth of somebody is known, then his (her) age will be also known. This relation is denoted by symbol “ \Rightarrow ”, for example:

$$Age \Rightarrow Day_of_birth.$$

- *Contradiction*: An attribute a is contradictory with attribute b if their domains are a 2-element set, for example $\{0,1\}$, and for the same instance i if $(i, a, 0)$ then there must be $(i, b, 1)$, and if $(i, b, 0)$ then there must be $(i, a, 1)$. For example, attribute *Is_free* referring to books in a library is contradictory with attribute *Is_lent* where their domain is set $\{Yes, No\}$. This relation is denoted by symbol “ \Downarrow ”, for example:

$$Is_free \Downarrow Is_lent.$$

These informal notions require more precise definitions. However we will not deal with this in this work because of its limited place.

From these definitions the following properties of these relations can be formulated as follows:

- a) If $a \Leftrightarrow b$ then $a \Rightarrow b$ and $b \Rightarrow a$, and vice versa.
- b) If $a \Rightarrow b$ and $b \Rightarrow c$ then $a \Rightarrow c$.
- c) If $a \Leftrightarrow b$ and $a \Downarrow a'$ and $b \Downarrow b'$ then $a' \Leftrightarrow b'$.
- d) If $a \Rightarrow b$ and $a \Downarrow a'$ and $b \Downarrow b'$ then $b' \Rightarrow a'$.

Let us return to the problem formulated above. If the final set A^* of attributes should be full, that is their semantics should contains the semantics of all attributes belonging to sets A_i for $i = 1, \dots, n$ then we can use the following algorithm for integration:

Input: Pairs (A^i, V^i) for $1, \dots, n$, where A^i is a set of attributes and V^i is their domain

Output: Pair (A^*, V^*) which at best represents given pairs.

BEGIN

1. $A^* = \bigcup_{i=1}^n A^i$

2. For each pair (a,b) of attributes from A^* do
 Begin

If $a \Leftrightarrow b$ then $A^* := A^* \setminus \{a\}$;

If $a \Rightarrow b$ then $A^* := A^* \setminus \{a\}$;

If $a \Downarrow b$ then $A^* := A^* \setminus \{b\}$

End.

$$3. \quad V^* = \bigcup_{a \in A^*} V_a$$

END.

In many cases it is not required that the semantics of A^* should contain the semantics of all attributes belonging to sets A_i , but it should be the most representative in such sense that A^* should contain those attributes which appear frequently in sets A_i . In other words, the following condition should be satisfied:

$$\sum_{i=1}^n d(A^*, A_i) = \min_{A \subseteq A} \sum_{i=1}^n d(A, A_i).$$

This condition guarantees that set A^* will contain these attributes which are most representative for attributes occurring in set A_i . It is used in situation when the integration of ontologies does not require containing all elements of the component ontologies, but only those which are most used. For this choice the following aspects should be taken into account:

- It is needed to define distance function d between sets of attributes. In work [7] a distance function between sets of values has been defined and can be used for this aim.

- In set A^* determined by the above criterion it is also needed to move some attributes as it is realized in Step 2 of the above presented algorithm.

- The same attribute may occur in different ontologies with different domain. For example the domain of attribute *Age* in an ontology is interval $(0,18)$ while in other ontology is set $[18, +\infty)$. In this case the final domain of this attribute is set as the sum of both domains.

3.3 Conflicts on Relation Level

We consider situation in which between the same concepts c and c' different ontologies assign different relations. As an example consider concepts *Man* and *Woman*, in one ontology they are in relation *Marriage*, in other ontology they are in relation *Brother-Sister*. Notice that within the same ontology 2 concepts may be in more than one relation. Let us denote by $R_i(c, c')$ the set of relations between concepts c and c' within ontology O_i for $i=1, \dots, n$. We have the following integration problem:

For given a set $X = \{R_i(c, c') : i=1, \dots, n\}$ it is needed to determine set $R(c, c')$ of final relations between c and c' , which at best represents the given sets.

For this problem the solution seems to be simple: One can determine $R(c, c')$ as the sum of $R_i(c, c')$, or select to $R(c, c')$ only those from $R_i(c, c')$, which appear most frequently. However, the solution will not so simple if we take into account the relationships between relations. For example, if relation r is between c and c' , and relation r' is between c' and c'' , then there must be relation r'' is between c and c'' . Such kind of relationships may make the choice of set $R(c, c')$ hard. However, we do not deal with it in this paper.

Besides resolving relation conflicts may in practice appear more complicated then considered in this section - relations are not independent from the concepts so we may expect mutual interaction between results.

4 Conclusions

This paper presents a classification of ontology conflicts and consensus-based method for their resolution. Apart from the three levels of conflicts it is worth also to investigate conflicts on the level of axioms. This problem is then a subject of the future work.

References

1. Balzer, R.: Tolerating Inconsistency. In: Proceedings of the 13th International Conference on Software Engineering, IEEE Press (1991) 158-165.
2. Crow, L. & Shadbolt, N.: Extracting focused knowledge from the semantic web. *International Journal of Human-Computer Studies* **54** (2001) 155-184
3. Day, W.H.E.: Consensus methods as tools for data analysis. In: H.H. Bock (ed.): *Classification and related Methods of Data Analysis*, Proceedings of IFCS'87, North-Holland (1987) 317-324.
4. Fensel, D.: *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag (2001)
5. Fernandez-Breis, J.T., Martinez-Bejar, R.: A cooperative framework for integrating ontologies. *Int. J. Human-Computer Studies* **56** (2002) 665-720.
6. Gruber, T.R.: *A translation approach to portable ontology specifications*. Knowledge System Laboratory, Academic Press Stanford University (1993).
7. Nguyen, N.T.: Consensus System for Solving Conflicts in Distributed Systems. *Journal of Information Sciences* **147** (2002) 91-122.
8. Nguyen N.T.: Processing Inconsistency of Knowledge on Semantic Level. *Journal of Universal Computer Science* **11**(2) (2005) 285-302.
9. Pinto, H.S., Martins, J.P.: A Methodology for Ontology Integration. In: Proceedings of the First International Conference on Knowledge Capture. ACM Press (2001) 131-138.
10. Reimer, U.: Knowledge Integration for Building Organizational Memories. In Proceedings of the 11th Banff Knowledge Acquisition for Knowledge Based Systems Workshop. Canada, Vol. 2 (1998) KM-6.1-KM-6.20.

Estimation of FAQ Knowledge Bases by Introducing Measurements

Jun Harada, Masao Fuketa, El-Sayed Atlam, Toru Sumitomo, Wataru Hiraishi,
and Jun-ichi Aoe

Dept. of Information Science and Intelligent Systems
The University of Tokushima, Tokushima-Shi, 770-8506, Japan
atlam@ccr.tokushima-u.ac.jp

Abstracts. Question and answering (QA) systems in the CRM scheme require both the quality relating user's satisfaction and the amount of questions to be managed, that is to say, it depends on the cost. This paper presents an estimation method of the FAQ service by introducing the following measurements: 1) user's disrepute for products which defined by four types of classifying questions; 2) kindness for solutions replied which defined by four types of classifying answers; 3) comprehension for answers which defined by semantic expressions of questions and answers; 4) sufficiency and quality for the whole FAQ service that introduced by the 1), 2) and 3). This approach is evaluated by the FAQ data with 4,538 questions and 5,356 answers and the real time simulation to estimate user's sufficiency is computed. From this evaluation, it is verified that the presented approach is useful and effectiveness.

1 Introduction

With wide spread of products (computers, cars, handy phones, etc.) equipped with advanced technologies, a CRM (Customer Relationship Management) scheme becomes a very important task in order to settle the troubles and to keep facilities. The purpose of the CRM scheme is to achieve user's satisfaction by managing their questions and claims. Moreover, it enables us to keep high quality service for products. Question and answering (QA) systems in the CRM scheme require both the quality relating user's satisfaction and the amount of questions to be managed, that is to say, it depends on the cost. There are many QA researches for large text databases, but they are not relation to CRM schemes [8][7][14]. Most useful for the CRM researches includes answering opinion questions by [11], good and bad expression understanding by [6], and estimating sentence types by [13].

Understanding approaches for the affective expressions [10] must be introduced, not text classification approaches by [9] [12]. This paper presents a measurement of quality of the FAQ service by introducing the following measurements:

1) Measurement of user's disrepute for products which defined by four types of classifying questions (IMPOSSIBLE, SIDE EFFECT, INSUFFICIENT and UNCLEAR), and the degree for each type is defined. 2) Measurement of kindness for

solutions replied which defined by four types of classifying answers (ACTION, CONFIRMATION, EXPLANATION, and NO PROBLEM), and the degree for each type is defined. 3) Measurement of comprehension for answers which defined by semantic expressions of questions and answers. 4) Measurements of sufficiency and quality for the whole FAQ service that introduced by the 1), 2) and 3). This degree is defined by an integer and it becomes very easy to estimate the FAQ service.

2 Semantic Expression of FAQ Knowledge Bases

2.1 Direct and Indirect Intentions

A questioner is classified into three kinds of types: interrogative, imperative and declarative sentences. Consider each sentence requesting a drink as follows:

- 1) Interrogative sentence "Isn't a drink given?"
- 2) Imperative sentence "Give me a drink"
- 3) Declarative sentence "I want a drink"

Although the above examples have direct intention requesting a drink, many questions have indirect intention. This is called an indirect speech act. Consider the following question (1) with indirect intention: "Doesn't a throat become it dry?", "How is juice?" is one of the right answers if a questioner's intention is "I want a drink." The intention understanding depends on the dialogue situation and the semantic expression with a situation attribute must be formalized. The above indirect intention of question (1) = "Doesn't a throat become it dry?" can be represented by [Moisture is insufficient in the situation C], and the more formal description of a question semantic expression is denoted by [[[C], [SITUATION]]; [[moisture], [OBJECT]]; [[insufficient], [CLAM]]], where [] specifies semantic representation and [A] of [[A], [B]] is the attribute value for attribute [B]. In order to define an answer semantic expression corresponding to expected answer (a) for question (q), the question semantic expression is transformed by replacing [[insufficient],[CLAIM]] into [[supply], [SOLUTION]]. The transformed semantic expression is the answer semantic expression and it is represented as follows: [[[C], [SITUATION]]; [[moisture], [OBJECT]]; [[supply], [SOLUTION]]].

2.2 Transformation of Semantic Expressions on FAQ Knowledge

In the FAQ dialogue, a questioner (a user or a customer) expects that a respondent (a company person) provides useful answers resolving his/her claim. Therefore, no questioner gives his/her juice to the respondent as the above section 2.1. This section discusses a formal definition for FAQ dialogue systems by defining a Q-Class attribute. The Q-Class attribute means the degree of questioner's disrepute and it is defined by four kinds of classes. For examples, "Can not print out" means Q-CLASS is [IMPOSSIBLE], "printer is noisy" means Q-CLASS is [SIDE EFFECT], "Printing character is unclear" [INSUFFICIENT] and "The red lamp of the printer has been lighted up" means Q-CLASS is [UNCLEAR]. We can define the degree of user's disrepute by using Q-CLASS.

3 Estimation Measurements of FAQ Knowledge Bases

3.1 Degree of Disrepute for Questions

In the question understanding process, affective information (user's tone, sentence style and so on.) are considered and Q-CLASS defines in the question semantic expression. This section defines the degree of the user's disrepute from questions as follows:

1) DISREPUTE ([[IMPOSSIBLE],[Q-CLASS]])=4

Value [IMPOSSIBLE] means the function which should be committed essentially does not work, so the degree of user's disrepute is the highest level. This point is defined by 4,

2) DISREPUTE ([[SIDE EFFECT],[Q-CLASS]])=3

Value [SIDE EFFECT] means there is a bad phenomenon unrelated to the original function, so the degree of user's disrepute is in the second level. This point is defined by 3,

3) DISREPUTE ([[INSUFFICIENT],[Q-CLASS]])=2

Value [INSUFFICIENT] means a function is lower than the expected performance, so the degree of user's disrepute is in the third level. This point is defined by 2,

3) DISREPUTE ([[UNCLEAR],[Q-CLASS]])=1

Value [UNCLEAR] means the operating method and the results are unclear, so the degree of user's disrepute is the lowest level. This point is defined by 1.

4 Experimental Observations

In the simulation, the natural language analyzer with retrieving a variety of dictionaries have been utilized by using many techniques [1] [2] [3] [4].

4.1 Experimental Data and Their Properties

For FAQ data, 4,538 questions and 5,356 answers have been prepared for six kinds of products (computers, telephones/facsimiles, digital cameras, AV equipments, home electronics and cars). The 1,513 questions and answers have been collected from FAQ web pages and FAQ documents of products where Japanese has been translated into English. The remaining 3,025 questions have been produced by ten expert persons (10 Ph.D. Students). Let $N(Q)$ be the number of questions obtained from FAQ data, let $N(Q+)$ be the number of questions produced by ten expert persons, let $N(A)$ be the number of answers, let $AVE(N(A)/N(Q))$ be the average number of answers to one question, let $MIN(N(A)/N(Q))$ be the minimum number of answers to one question, and let $MAX(N(A)/N(Q))$ be the maximum number of answers to one question. Table 1 shows information about the FAQ data. In Table 1, $MIN(N(A)/N(Q))$ of the computer field is the highest value 6.7 while those of other fields are smaller from 3.0 to 3.2. The reason is that technical terms and operations of that field are more difficult than other fields.

Table 1. Information about FAQ data

Products	N(Q)	N(Q+)	N(A)	AVE(N(A)/N(Q))	MIN(N(A)/N(Q))	MAX(N(A)/N(Q))
<Computers>	155	310	1032	6.7	1	41
<Telephones/Facsimiles>	213	425	739	3.5	1	16
<Digital cameras>	136	274	425	3.1	1	10
<AV equipments>	567	1133	1702	3.0	1	16
<Home electronics>	249	497	839	3.4	1	19
<Cars>	193	386	619	3.2	1	9
Total	1,513	3,025	5,356	3.5	1	41

4.2 Experimental Results by Semantic Expressions

4.2.1 Situations and Objects of Semantic Expressions

Table 2 shows the number of same and different attributes (situations, objects) between question and answer semantic expressions in order to observe values to be computed in the degree of COMPREHENSION. While the rate of different situations is smaller than that of different objects, it is a reasonable result because FAQ data with many different situations means poor in the degree COMPREHENSION. That is to say, for the degree of COMPREHENSION, the rate of different situations should have priority over that of different objects. The rate of different objects is considered for the next estimation of COMPREHENSION, and the definition in section 3 depends on this observations Field <Computers> has the highest rate 38% in the fields and that FAQ knowledge base is poor in different situations, but that rate of different objects is the fifth. The rate of different objects is from 38% to 79% and the rate for field <Telephones/Facsimiles> has the highest. The degree of COMPREHENSION will be shown in Table 3.

Table 2. Information about the number of words

Products	N(WQ)	MIN (N(WQ))	MAX (N(WQ))	N(WA)	MIN (N(WA))	MAX (N(WA))
<Computers>	7.9	3	40	14.2	2	47
<Telephones/Facsimiles>	6.8	3	23	10.8	4	39
<Digital cameras>	5.8	3	25	10.7	2	29
<AV equipments>	6.7	2	27	12.1	3	45
<Home electronics>	5.9	1	18	13.8	2	48
<Cars>	6.1	3	21	12.6	3	37
Total	6.5	1	40	12.7	2	48

4.2.2 Question Semantic Expressions

Classified results by question semantic expressions in the FAQ data are discussed. First of all, Table 3 shows the classified results for questions.

From Table 3, the following observations are obtained: 1) The total rate 90% of IMPOSSIBLE and UNCLEAR for field <Computers> are larger than the total rate 46% for field <Home Electronics>. 2) The total rate 54% of SIDE EFFECT and INSUFFICIENT for field <Home electronics> are larger than the total rate 10% for field <Computers>.

Table 3. Classified results by question semantic expressions

Q-CLASS	<Computers>		<Telephones /Facsimiles>		<Digital cameras>		<AV equipments>		<Home electronics>		<Cars>	
(IM)	282	61%	488	76%	246	60%	1125	66%	261	35%	303	52%
(SE)	33	7%	57	9%	72	18%	299	18%	315	42%	201	35%
(IN)	15	3%	12	2%	45	11%	72	4%	87	12%	24	4%
(UN)	135	29%	81	13%	48	12%	204	12%	83	11%	51	9%
TOTAL	465	100%	638	100%	411	100%	1700	100%	746	100%	579	100%

(IM)=IMPOSSIBLE;(SE)=SIDE-EFFECT; (IN)=INSUFFICIENT; (UN)=UNCLEAR

Operations associating with field <Computers> are generally difficult for users, so the IMPOSSIBLE and UNCLEAR questions increase. In other words, users can not determine whether their claims for field <Computers> are concrete classes SIDE EFFECT and INSUFFICIENT, or not. Consequently, many questions of field <Computers> belong to IMPOSSIBLE and UNCLEAR. However, products for fields <Home electronics> must be safe and complete because it is easy to discover the dissatisfaction of the home electronics used in everyday life. Therefore users are affective for SIDE EFFECT and INSUFFICIENT. This observation can be reflected on other fields, for example, fields <Home Electronics> and <Cars> have the similar tendency.

5 Conclusions

This paper has been presented an estimation method of the FAQ service by introducing the following measurements: 1) user’s disrepute for products which defined by four types of classifying questions; 2) kindness for solutions replied which defined by four types of classifying answers; 3) comprehension for answers which defined by semantic expressions of questions and answers; 4) sufficiency and quality for the whole FAQ service that introduced by the 1), 2) and 3). The presented approaches have been evaluated by the FAQ data with 4,538 questions and 5,356 answers.

References

- [1] Aho, A. V. & Corasick, M. J. Efficient string matching: An aid to bibliographic search. *Communications of the ACM*, 18(6), (1975), 333-340.
- [2] Aoe J., Morimoto K., Shishibori M. and Park, K-H. A Trie Compaction Algorithm for a Large Set of Keys. *IEEE Tra. on Knowledge and Data Engineering*, 8(3), (1996), 476-491.
- [3] Atlam E.-S., El-Marhomy G., Morita K., Fuketa M. and Aoe J. Automatic Building of New Field Association Word Candidates Using Search Engine. *Information Processing & Management*, 42 (4), (2006), 951-962.
- [4] Atlam E.-S, Morita K., Fuketa M. and Aoe J. Documents similarity measurement using field association terms. *Information Processing & Management*, 39, (2003), 809-824.
- [5] Fuketa, M., Lee, S., Tsuji, T., Okada, M., & Aoe, J. A document classification method by using field association words. *Journal of Information Sciences*, 126(1), (2000), 57-70.

- [6] Fuketa M., Kadoya Y., Atlam E.-S., Kunikata T., Morita K., Kashiji S., and Aoe J. A Method of Extracting and Evaluating Good and Bad Reputations for Natural Language Expressions. *Information Technology & Decision Making*, 4 (2), (2005), 177-196.
- [7] Fukumoto, J., Kato, T., and Masui, F. Question Answering Challenge (QCA-1) Question answering evaluation at NTCIR Workshop 3. in Working Notes of the Third NTCIR Workshop Meeting, Part IV: Question Answering Challenge (QAC1), (2002), 1-10.
- [8] Kiyota, Y., Kurohashi, S., & Kido, F.. Dialog Navigator: A Question Answering System based on Large Text Knowledge Base. *Natural Language Processing J.*, 10 (4), (2003), 1-30
- [9] Kwon, O., and Lee, J. Text categorization based on k-nearest neighbor approach for Web site classification. *Information Processing and Management*, 39(1), (2003), 25-44.
- [10] Kadoya Y., Morita K., Fuketa M., Oono M., Atlam E.-S., Sumitomo T. and Aoe J. A Sentence Classification Technique by Using Intention Association Expressions. *Computer Mathematics*, 82 (7), (2005), 777-792.
- [11] Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., and Fukushima, T.. Collecting Evaluative Expressions for Opinion Extraction. *Journal of Natural Language Processing*, 12 (3), (2005), 203-222 (in Japanese).
- [12] Lam, W., Ruiz, M., Srinivasan, P., Automatic Text Categorization and Its Application to Text Retrieval. *IEEE Tras. on Know. and Data Engineering*, 11(6), (1999), 865-879.
- [13] Tokunaga H., Atlam E.-S., Fuketa M., Morita K., Tsuda K. and Aoe J., Estimating sentence types in computer related new product bulletins using a decision tree. *Information Sciences*, 168 (1-4), (2004),185-200.

Efficient Stream Delivery over Unstructured Overlay Network by Reverse-Query Propagation

Yoshikatsu Fujita^{1,2}, Yasufumi Saruwatari², Jun Yoshida¹, and Kazuhiko Tsuda²

¹ Matsushita Electric Industrial Co., Ltd., Corporate eNet Business Division
2-13-10, Kyobashi, Chuo-ku, Tokyo 104-0031, Japan
{fujita.yoshikatsu, yoshida.jun}@jp.panasonic.com
<http://home.hi-ho.ne.jp/>

² Graduate School of Business Sciences, University of Tsukuba, Tokyo
3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
{saru, tsuda}@gssm.otsuka.tsukuba.ac.jp
<http://www.gssm.otsuka.tsukuba.ac.jp>

Abstract. We propose reverse-query mechanism to deliver broadband contents over unstructured overlay network. Due to the nature of scale-free network, newly defined reverse-query message will cover more than 80% of all the clients, with relaying reverse-query message under probability as low as 10%, being effective to reduce the total traffic generated by query propagation. This platform is built on percolation theory to propagate the message and contents, which each agent on P2P client relays reverse-query to randomly selected peers. This can be applied for quasi-broadcast platform on P2P network, and for business application, to flyer delivery over the Internet targeting common attributes such as residential area.

1 Introduction

In Japan, domestic network infrastructure expands rapidly thanks to governmental e-Japan strategy. Today, number of broadband users sums up to 20 million and overall penetration rate for households counts for more than 40%. In addition, digital broadcast which began in December 2003, and standardization activity for server type broadcasting, expects breakthrough for digital contents market demands. With the emergence of such broadband contents delivery infrastructure, people can enjoy any contents which satisfy their own lifestyle among a huge amount of contents archives.

However, when it comes to the matter of how we should deliver broadband contents over the Internet, the total throughput will be determined by any bottleneck somewhere between contents provider to consumers. For example, even a user purchases 100Mbps optical fiber service, one's requested contents comes from distant server only 100Kbps because of narrow path somewhere over the Internet. Most of proxy servers distributed over the Internet are aimed for static contents like homepage objects, and are not tuned for broadband stream contents.

For instance, once many users try to pull large video streams at the same time, it is clear that backbone network is easily falls into overflow. This requires a new contents delivery technology which supports a huge simultaneous access transaction for

broadband objects. For this purpose, to deliver broadband contents over the Internet, CDN (Contents Delivery Network) architecture [1], [2] and contents distribution algorithm for replication [3], [4] are actively studied. But such CDN solutions for large scale contents delivery faces difficulty because the number of acceptable simultaneous access is almost determined by hardware specification of cache servers, and this falls into optimal cache server distribution problem with considering dynamic request load balance under the exact forecast of contents popularity and hardware availability. This is also regarded as a big issue for realizing broadcast type traffic over the Internet. Existing technique to handle telephone call over the telephone network is specific for point to point traffic, and it is not applicable for clearing simultaneous access to a contents sever, that makes it difficult to deliver broadband contents to many clients.

This study aimed for overcoming this problem by building our proposed overlay network over the Internet, and to manage generated traffic under our control, to be new quasi-broadcast platform. We develop new contents delivery network for broadband contents based on the fact that many link status on the Internet follows power law distribution [5]. For example, one of the most popular pure P2P network Gnutella has been analyzed that its nodes' outgoing degree can be expressed as $P(k) \sim k^{-\tau}$ ($\tau \geq 0$) [6]. However, such an unstructured network is not manageable in nature, and makes it difficult to apply for contents delivery for its fundamental network architecture. In this paper, we employ percolation theory [7] that is mainly studied in physics, to model the "percolating information" for contents delivery over pure P2P network. In other words, the query message released by client seeking for requested contents is not merely used for this purpose, but we define a new "reverse-query" message to find any client who needs a certain contents and try to apply the percolation theory to manage the generated traffic. This will lead to reduce the explosive P2P query traffic while maintaining fairly high clients cover rate over our proposed overlay network. An outlook of our model is shown in Fig.1.

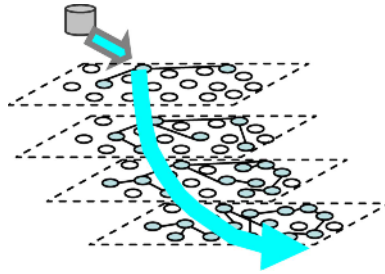


Fig. 1. Percolation on the P2P Network

2 Related Work

2.1 P2P Network

Recently, file sharing application over P2P network has been pervasive and it plays an important role as contents distribution infrastructure.

P2P network holds such technical problems in nature as:

- Interoperability between peers
- Scalability and reliability without center server
- How to keep anonymity and privacy
- Adhoc join and leave behavior

When we apply P2P network for contents delivery, under the pure P2P architecture which has no center server, it is major concern how to find available resources and required contents from all over the network. In this field of study, such projects as CAN[8], Chord [9], Pastry [10], Tapestry [11] are trying to employ Distributed Hash Table (DHT)[12].

However, in order to apply for commercial services, there exists more problems in this DHT solution.

- When peers join/leave the network, they have to takeover management table to some other peers, results in heavy overhead.
- Network structure becomes complicated.

In addition, traffic generated by those P2P nodes has been increasing more and more, whose amount of load gives serious impact to today's ISP backbone network.

2.2 Power-Law Nature

In this study, we propose a new network architecture for broadband contents delivery which can solve problems shown above and applicable for quasi-broadcast platform over the internet.

To estimate the amount of total traffic for retrieval query sent by each client, there is a report from [13]. And for modeling method of network with a power law link distribution, there is an algorithm by [5] with preferential attachment model. However, this method focuses mainly on the growth of network itself and analyzing dynamic traffic behavior has been still left for further study. Also, validation of those theoretical network structure and real P2P network on the Internet is not fully analyzed yet.

3 Reverse Query Mechanism

We have already introduced a fundamental idea to propagate a query message over the unstructured peer-to-peer network called "Reverse Query Mechanism" [14].

In that paper, we have defined new "do-you-need" query and shown that contents can be delivered just by relaying this query from server to peer to peer. Details are explained in [14].

3.1 Analysis

3.1.1 Percolation on Generalized Random Graph

The percolation behavior on generalized random graph can be led as follows [15].

Suppose the degree distribution of each node to be $p(k)$, then the general function of this distribution can be defined as:

$$G_0(x) = \sum_{k=0}^{\infty} p(k)x^k \tag{1}$$

Suppose the vertex distribution of connected component on generalized random graph holds general function $H_0(x)$, and general function for the size of connected component from a certain branch to be $H_1(x)$. Then the average size of connected component $\langle C_0 \rangle$ to be

$$\begin{aligned} \langle C_0 \rangle &= H_0'(1) = 1 + G_0'(1)H_1'(1) \\ &= 1 + \frac{G_0'(1)}{1 - G_1'(1)} \end{aligned} \tag{2}$$

The state transition will take place when right-hand side becomes 0,

$$\begin{aligned} G_1'(1) = 1 &\Leftrightarrow \sum_k k(k-2)p(k) = 0 \\ &\Leftrightarrow \frac{\langle k^2 \rangle}{\langle k \rangle} = 2 \end{aligned} \tag{3}$$

Here, the percolation threshold can be

$$q_c = \frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1} \tag{4}$$

If we assume our delivery platform to be generalized random graph, we can say that almost all node can receive the transferred messages when each node relays received messages more than q_c .

3.1.2 Percolation on Power Law Overlay Network

The percolation behavior on the power law overlay network can be led as follows.

(1) From Origin to the First Neighbor

We will consider the number of vertex that is 1 hop away from the origin.

Let us employ the expression k_m^n , which shows the degree of vertex to be m hops away from origin (lower right) and nth vertex out of the set of m vertexes (upper right). Fig.2 shows an example.

Then, remember the assumption that every node on the network has degree according to power law. This assumption is observed from real P2P network [6].

When we assume the origin holds connection with k_0 vertexes, then the number of vertex that can be reachable within 1 hop from origin is

$$k_1^1 - 1 \tag{5}$$

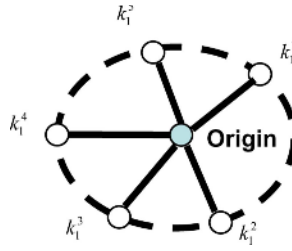


Fig. 2. First neighbors

As there are k_0 vertexes as a whole, then the total number of reachable vertexes is

$$\sum_{i=1}^{k_0} (k_1^i - 1) \tag{6}$$

(2) Duplicated First Neighbor

Then, we will reduce the duplicated number in the case of graph of Fig.6.

In the Fig.3, we need to subtract the paths k_i to k_j and k_j to k_i , because both k_i and k_j are counted as just 1 hop away from the origin.

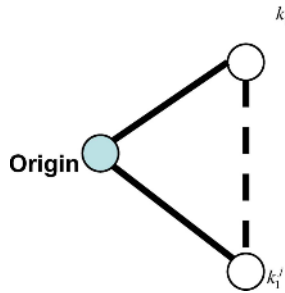


Fig. 3. Duplicated path

The expected value for the number of vertex, that k_i is 2 hops away from the origin is

$$\sum_i \frac{k_1^i - 1}{n - 2} (k_0 - 1) \tag{7}$$

Let the total number of vertexes D , which is reachable within 1 hop from origin can be expressed from (6) and (7)

$$D = \sum_{i=1}^{k_0} (k_1^i - 1) - \sum_i \frac{k_1^i - 1}{n - 2} (k_0 - 1) \tag{8}$$

As this number is overestimated, it is enough to employ this equation to calculate the reachability from the origin.

(3) Distant Neighbor

The number of vertex V_m that m hops away from the origin can be expressed as

$$V_m = \sum_i \frac{k_2^i - 1}{n - m} (D + k_0^1) \tag{9}$$

By using recursive equation, the number of vertex n hops away from origin can be obtained in the same way.

When we want to deliver a certain message to more than 80 % of nodes on the overlay network by applying (4), it is enough to deliver the message as:

$$\alpha^{V_m} \bullet (1 - p)^{V_m} \bullet p \geq 0.8 \tag{10}$$

In the next chapter, we have investigated the dynamics of α upon mathematical simulation.

4 Evaluation

In order to evaluate our proposed model, we prepared a random network with a power law link distribution generated by Pajek [16], and implemented our algorithm on R environment. This overlay network is generated based on generalized BA model, which presumes that every vertex has at least some baseline probability of gaining an edge, to generate edges by mixture of preferential attachment and uniform attachment [17]. For generating condition, we set the total node number $N = 1000$, $M_0 = 3$, $TTL = 25$ and average degree = 2.7. In order to evaluate the results, we counted the number of generated messages (= reverse-query) and cover rate (= how much of nodes receives the reverse-query) upon relay probability implanted in the reverse query message.

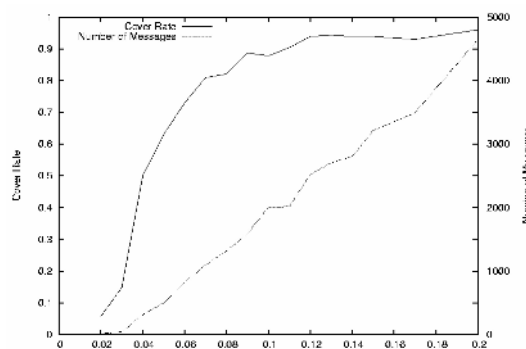


Fig. 4. Simulation Results

Fig.4 shows a typical example of message propagation over this generated overlay network. The propagating message will cover almost 80% of nodes upon relay probability 0.08, and the total traffic increases linearly. After several hops of relaying messages, the copy of message body will just go out from the P2P network by using up the TTL, and this will lead to reduce the explosive P2P query traffic while maintaining fairly high clients cover rate over our proposed mechanism.

5 Conclusion

In this paper, we employ percolation theory to model the “message propagation” for contents delivery over the pure-P2P network. We defined a new “reverse-query” message to find any client who needs a certain content and try to apply the percolation theory to prevail the message over the network. We analyzed validity of the model through dynamics of simulation. This concludes that our proposal is effective for contents delivery over the Internet.

For commercial application example, we can attach a small video clip in the reverse-query message, which is possible to use this contents delivery mechanism as propagating electric flyer all over the clients, just like a quasi-broadcast platform.

References

- [1] D.Karger, E.Lehman, T.Leighton, R.Pnigrahy, M.Levine and D.Lewin: “Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web”, Proceedings of the 29th annual ACM Symposium on Theory of Computing, ACM Press, pp. 654-663 (1997)
- [2] M.Abrams, C.R.Standridge, G.Abdulla, S.Williams and E.A. Fox: “Caching Proxies: Limitations and Potentials”, Proceedings of 4th International World Wide Web Conference, pp. 119-133 (1995)
- [3] Y.Chen, R.H.Katz and J.D.Kubiatowicz: “Dynamic Replica Placement for Scalable Content Delivery”, IPTPS 2002, pp. 306-318 (2002)
- [4] Y.Li and M.T.Liu: “Optimization of Performance Gain in Content Distribution Networks with Server Replicas”, SAINT2003 Proceedings, pp. 182-189 (2003)
- [5] R.Albert and A.L.Barabasi: “Statistical Mechanics of Complex Networks”, Reviews of Modern Physics, Vol. 74, pp. 47-97 (2002)
- [6] M.Faloutsos, P.Faloutsos and C.Faloutsos: “On Power-law Relationships of the Internet Topology”, ACM SIGCOMM, pp. 251-262 (1999)
- [7] D.Stauffer and A.Aharony: “Introduction to Percolation Theory”, Taylor and Francis, London (1994)
- [8] S.Ratnasamy, P.Francis, M.Handley, R.Karp and S. Schenker: “A Scalable Content-addressable Network”, Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, ACM Press, pp.161-172 (2001)
- [9] I.Stoica, R.Morris, D.Karger, M.F.Kaashoek and H. Balakrishnan: “Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications”, Proceedings of ACM SIGCOMM, ACM Press, pp. 149-160 (2001)

- [10] A.Rowstron and P.Druschel: "Pastry: Scalable, Decentralized Object Location, and Routing for Large-Scale Peer-to-Peer Systems", Lecture Notes in Computer Science, Vol. 2218, pp. 329-350 (2001)
- [11] K.Hildrum, J.D.Kubiatowicz, S.Rao and B.Y.Zhao: "Distributed Object Location in a Dynamic Network", Proceedings of the 14th ACM Symposium on Parallel Algorithms and Architectures, pp. 41-52 (2002)
- [12] H.Sunaga, T.Hoshiai, S.Kamei and S.Kimura: "Technical Trends in P2P-Based Communications", IEICE TRANS. COMMUN, Vol. E87-B, No.10, pp.2831-2846 (2004)
- [13] L.A.Adamic, R.M.Lukose, A.R.Puniyani, and B.A.Huberman: "Search in power-law networks", Physical Review E, Vol. 64, No. 046135 (2001)
- [14] Yoshikatsu Fujita, Jun Yoshida, Kazuhiko Tsuda: "Reverse-Query Mechanism for Contents Delivery Management in Distributed Agent Network", KES2005, Part 4, pp.758-764 (2005)
- [15] N.Masuda and N.Konno: "Science of Complex Network", Sangyo Tosho (2005) (In Japanese)
- [16] Pajek: <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>
- [17] D.M.Pennock, G.W.Flake, S.Lawrence, E.J.Glover and C.L. Giles: "Winners Don't Take All: Characterizing the Competition for Links on the Web", Proceedings of the National Academy of Sciences, Vol. 99, No.8, pp. 5207-5211 (2002)

A Method for Development of Adequate Requirement Specification in the Plant Control Software Domain

Masakazu Takahashi^{1,2}, Yoshinori Fukue³, Satoru Takahashi⁴, and Takashi Kawasaki⁵

¹ Shimane University, 1060 Nishikawatsu, Matsue, Shimane, Japan
masakazu@cis.shimane-u.ac.jp

² Galaxy Express Corporation, 18-16-1 Hamamatsucho, Minato-ku, Tokyo, Japan
hanzawa@galaxy-express.co.jp

³ Tokushima University, 2-Ijyonan-mishima, Tokushima, Tokushima, Japan
fukue@is.tokushima.ac.jp

⁴ Tsukuba University, Otsuka 3-2-9 Bunkyo-ku, Tokyo, Japan
satoru@gssm.otsuka.tsukuba.ac.jp

⁵ New Energy and Technology Development Organization, Muza Kawasaki, Central Tower
19F, 1310 Omiya-cho, Saiwai-ku, Kawasaki, Japan
kawasakitks@nedo.go.jp

Abstract. This paper proposes a method for development of adequate requirement specifications in the Plant Control Software (PCSW). Before we propose this method, we have analyzed this domain and developed the components as parameter-customized-style in order to facilitate the customization. In the proposed method, PCSW requirement specification is developed from information that is used to customize components. We applied it to five development cases, and achieved 91[%] of Requirement Coverage and 94 [%] of the Requirement Conformity Rate. This result indicates that proposed method have sufficient capabilities to develop exhaustive and adequate PCSW requirement specification.

1 Introduction

A plant is an industrial mechanical facility for manufacturing chemical products or processing materials. In these days, the plant operation becomes more difficult, and requires assistance with software. Here, we define the software of assisting plant operation for “Plant Control SoftWare (PCSW)”.

A PCSW is developed usually based on the purposes. The client and the manufacturer had requirements review meetings and decide the requirement specification before starting the actual software development. However, in the practical PCSW development, some revisions sometimes have happened, because clients often requested.

These revisions caused the deterioration in PCSW quality, the extended development term, and the raised cost. In order to develop a requirement specification that fully reflects client’s requirements, it is important to present the clearer image of PCSW requirements by operating the PCSW prototype in front of the client, in addition to the requirement review meeting.

This paper proposes a method to develop adequate PCSW requirement specification efficiently by; “developing software components”, “composing PCSW prototypes from the components”, and “reflecting the results of PCSW behavior checks to the requirement specification.”

2 Evaluation of Existing Method

We show existing methods (concrete methods and past studies) for developing requirement specification, and describe some disadvantages that they are applied to PCSW development.

At first, we show some concrete methods. Structured Method [2], [4] has some disadvantages such as “there are many items to be determined” and “support tools are necessary to apply”. Object Oriented Method [3], [7] has disadvantages such as “it is difficult to list up all requirements” and “it is difficult to keep consistencies between requirements”. The algebraic specification method [6] has some disadvantages such as “it requires advanced mathematical knowledge” and “it is difficult to image the behavior of the target software.” The prototyping method [8] has some disadvantages such as “the development of prototype increases the development period and cost” and “it is hard to ensure that every raised issue is reflected to the requirement specification.

Next, we show some past studies. Uchitel [9] proposed the method that develops requirement specification from operation scenario. But we can not list up all scenarios when the development have started. Kabei [5] proposed the method that integrated Functional and Object Oriented Methodology. Because this method is close to conventional method, there are same disadvantages that concrete methods have. Braberma [1] proposed the method that is similar to formal notation like as algebraic language. We don’t need such a detailed requirement description.

As the result of evaluation, each method has some disadvantages, and it is difficult to apply either of them to the developments of PCSW requirement specifications directly.

3 Proposed method for Developing Requirement Specification

Section 3.1 summarizes the results of the domain analysis, and section 3.2 gives the outline of the proposed method.

3.1 The Result of Domain Analysis for PCSW

We carried out domain analysis in three steps: “analyze plants”, “analyze PCSW functions”, and “develop software components corresponding to the functions.”

At first, we carried out some analyses on plants. It is impossible to figure out all of devices that plants consist of. We need not to simulate the detailed behaviors, and classified the main devices into “Sequence Control Device (SCD)” and “Feedback Control Device (FCD)”. SCD controls the motions according to the pre-determined operation sequence. FCD controls the motions by dynamically calculating the operation volumes using state volumes and the target volumes.

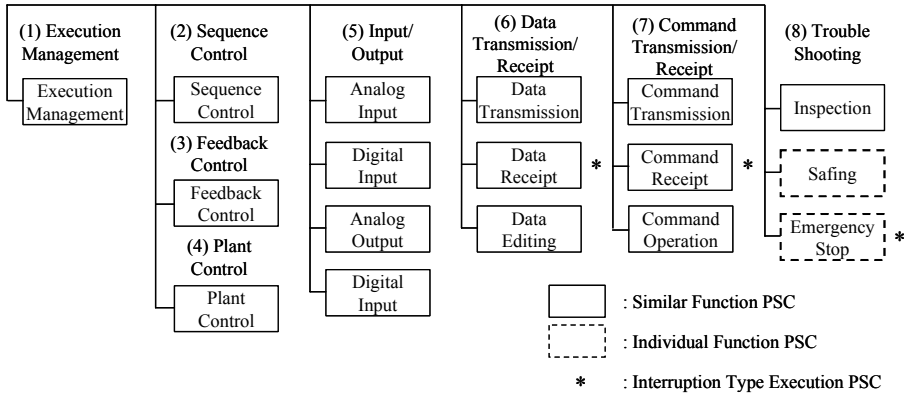


Fig. 1. List of Developed PSCs

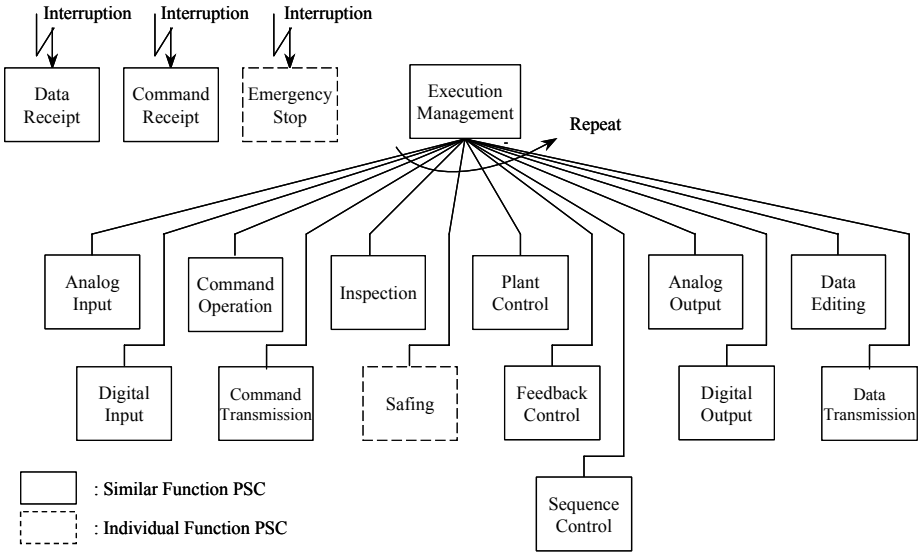


Fig. 2. Activation Sequence of PSCs

At second, we analyzed the functions of SCD and FCD, and found that the functions were categorized into “similar function” and “individual functions.” At last, we developed software components corresponding to pre-analyzed functions. Figure 1 shows the PCSW Software Components (PSCs). The squares with solid lines represent PSC for the similar functions, and the squares with broken lines represent PSC for the individual functions. The similar functional PSCs take customization parameters for achieving the targeted functions. The individual function PSCs are developed depending on the plant’s requirements. And we also considered the activation

sequences of the PSCs based on the sequence diagrams for the operational directives and the data transmission/receipt in the typical states (Figure 2).

3.2 Developing Requirement Specifications Using Prototype

Figure 3 gives the outline of the proposed method for developing requirement specifications. In step 1, we collect information about a PCSW and organize the required information for composing the prototype. In step2, we compose the prototype by setting the information to the developed PSCs as the parameters. In step3, we collaborate with the client to check the behavior of the prototype. If there are some lacks or conformities in a function, we return to the step 1 and re-organize the required information. The prototype fits customer’s requirements, and PCSW requirement specification is developed from the design information of the prototype as step 4 (Figure 4).

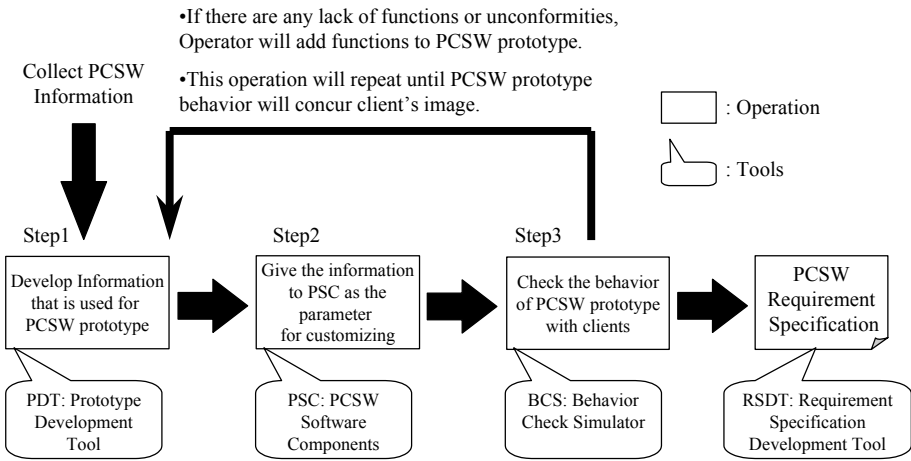


Fig. 3. Development Procedure in Proposed Method

4 Applications and Evaluations of Proposed Method

Section 4.1 explains the result of applying the method, and section 4.2 describes the evaluation of the method.

4.1 Results of Applying Proposed Method

We have applied the proposed method to five developments. The plant A, D and E are material production facilities, the plant B is the motor control device, and the plant C is temperature control device.

Table 1 shows the number of requirements that we developed and modified after completion of PCSW requirement specification. Here, the modified requirements mean that they are completely new or redundant requirements that we have not assured. If the rate of developed requirement is high percent, it is thought that requirement specification covers requirements.

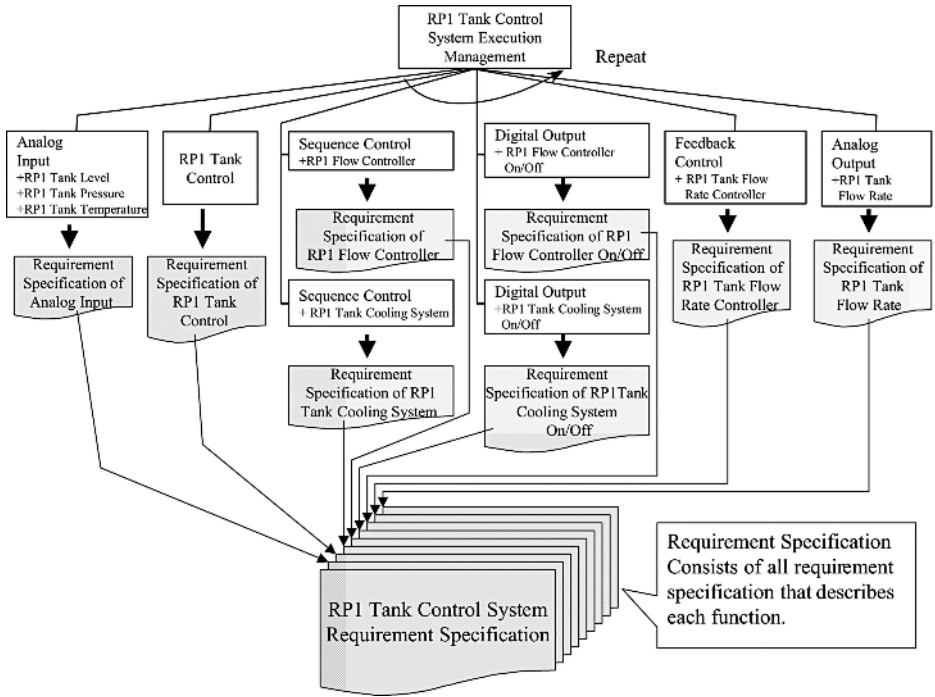


Fig. 4. PCSW requirement specification developed from design information

Table 2 shows the number of requirements when we have developed and when we have not changed through development. If the rate of non-changed requirements is high percent, it will be thought that developed requirements are adequate.

Here, requirements that were modified or revised depending on the plant facility changes are eliminated, because these are not concerned with coverage and adequacy.

Table 1. Result of Requirement Coverage

Plant Name	Number of Requirements that were described in requirement specification using Proposed method (NRP)	Number of requirements that were modified after completion of developing requirement specification (NRA)	Requirement Coverage (RC) [%]
A	92	8	92
B	37	4	90
C	68	5	93
D	87	10	90
E	82	7	92

Table 2. Result of Requirement Adequate Rate

Plant Name	Number of requirements that were described in requirement specification using proposed method (NRP)	Number of requirements that were not changed after completion of developing requirement specification (NRC)	Requirement Adequate Rate (RAR) [%]
A	92	87	95
B	37	34	92
C	68	65	96
D	87	82	95
E	82	75	92

4.2 Evaluations of Applying Proposed Method

We evaluated the coverage and the adequacy of developed requirement specifications.

1) Evaluation of coverage of requirements

We defined Requirement Coverage as an evaluation of specification:

$$RC = 100 * NRP / (NRP + NRM) \tag{1}$$

RC: Requirement Coverage

NRP: number of requirements that were described in requirement specification

NRM: number of requirements that were modified after completion of requirement specification

As shown in Table 5, the RCs of 92, 90, 93, 90, and 92 [%] were achieved respectively, and the average RC of 91 [%] was achieved. In addition, added requirements were mainly related to control method and malfunction. This result indicates that PCSW requirement specification that is developed by using proposal method covers actual user requirements sufficiently.

2) Evaluation of adequacy of requirements

We defined Requirement Adequate Rate as an evaluation of specification:

$$RAR = 100 * NRC / (NRP + NRC) \tag{2}$$

RAR: Requirement Adequate Rate

NRC: number of requirements that were not changed through development

As shown in Table 6, the RARs of 95, 92, 96, 95 and 92 [%] were achieved respectively, and the average RRR of 94 [%] was achieved. In addition, revised (not adequate) requirements were mainly related to data editing, transmission and receipt. This result indicates that requirements that are developed by using proposal method are adequate sufficiently.

The results of 1) and 2) indicate that developed requirement specification using proposed method covers actual requirements sufficiently and each requirements are adequate enough. Consequently, we consider that we are able to develop PCSW requirement specification using proposed method.

5 Conclusion

This paper has proposed the method for development of adequate requirement specification using component-based prototype in PCSW domain. As a result of applying proposed method, we have achieved the RC of 91 [%], the RAR of 94[%] in average. These results indicate that the proposed method is able to satisfy adequate PCSW requirement specification that covers actual user requirements.

Acknowledgement

This research is conducted as a part of "Basic Technology for Next Generation Transportation System Design", which is a delegation from New Energy and Industrial Technology Development Organization (NEDO).

References

- [1] V. Braberman, N. Kicillof and A. Olivero, TA Scenario-Matching Approach to the Description and Model Checking of Real-Time Properties, IEEE Transactions on Software Engineering, Vol.31, No.12, pp.1028-1041, DECEMBER 2005.
- [2] T. Demarco, Structured Analysis and System Specification, Prentice-Hall, 1979.
- [3] B. Douglass, Real-Time UML: Developing Efficient Objects for Embedded Systems, Second Edition, Addison Wesley Longman, 2000.
- [4] D. Hatley and I. Pirbhai, Strategies for Real-Time System Specification, Dorset House Publishing, 1988.
- [5] J. Kabeli and P. Shoval: Comprehension and quality of analysis specifications - a comparison of FOOM and OPM methodologies, Information and Software Technology, Vol.47, Issue 4, PP.271-290, 2005.
- [6] H. Kobayashi, Y. Kawata, M. Maekawa, A. Kawasaki, A. Yabu and K. Onogawa: Modeling External Objects of Process Control Systems in Executable Specifications, Journal of Information Processing Society of Japan, Vol.35, No. 7, pp. 1402-1409, July 1994 (in Japanese)
- [7] I. Jacobson, Object-Oriented Software Engineering - A Use Case Driven Approach -, the ACM Press, 1992.
- [8] J. Martin: "Rapid Application Development", Macmillian Publishing Company, 1991
- [9] S. Uchitel, J. Kramer, and J. Magee: Synthesis of Behavioral Models from Scenarios, IEEE Transactions on Software Engineering, Vol.29, No.2, PP.99-115, 2003.

Express Emoticons Choice Method for Smooth Communication of e-Business

Nobuo Suzuki and Kazuhiko Tsuda

Graduate School of Business Sciences, University of Tsukuba
Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
{nobuo, tsuda}@gssm.otuka.tsukuba.ac.jp

Abstract. For the business communication by email with cellular phones, it has an important weak point. That is to hard to tell to be utterance speed and the pitch of sounds involved in the sentences, because it communicate by letters only. Emoticons are often used to make up for this weak point. This paper describes techniques to predict emotions of sentences in Japanese emails and give an emoticon to end of a sentence automatically. This is achieved by learning information of emotions with emoticons used and analyzing the text of email with cellular phone by collecting and analyzing our corpus of emails. We also examined consistency evaluation with real email sentences input by cellular phones and emoticons automatically generated by this technique. We could get correct answer rate of 87.7%.

Keywords: Prediction of emotions, Morpheme analysis, Emoticons.

1 Introduction

We often use email for our business communications. In such situation, it is difficult to express emotions because we usually use letters only. Therefore it is common to express emotions by using emoticons. Recently, the cellular phones connected to the Internet become the daily tools and most familiar input tools to the Internet. We use emoticons with cellular phones more frequently than normal PCs. So, we collected and analyzed Japanese email sentences input from cellular phones. Fig.1 shows examples of sentences with emoticons input from cellular phones.

- | |
|---|
| <ul style="list-style-type: none">• How about this? o(^-^)• The other person was in Osaka, and I was in Saitama (>_<), but she said he didn't know when the baggage reach there. (*_*)• Could you tell me from which exit of Shinjuku station I can get to Kousei-Nenkin-Kaikan!!!? And tell me how long will it take? m(__)m |
|---|

Fig. 1. Examples of sentences with emoticons

In this paper, we analyze emotions of emoticons used in sentences by the results of collecting and classifying inputs by cellular phones. Then we describe a technique to

predict the emotions of email text in cellular phones and give an emoticon to the end of a sentence automatically.

Fig.2 shows our method for generating best emoticons to express emotions.

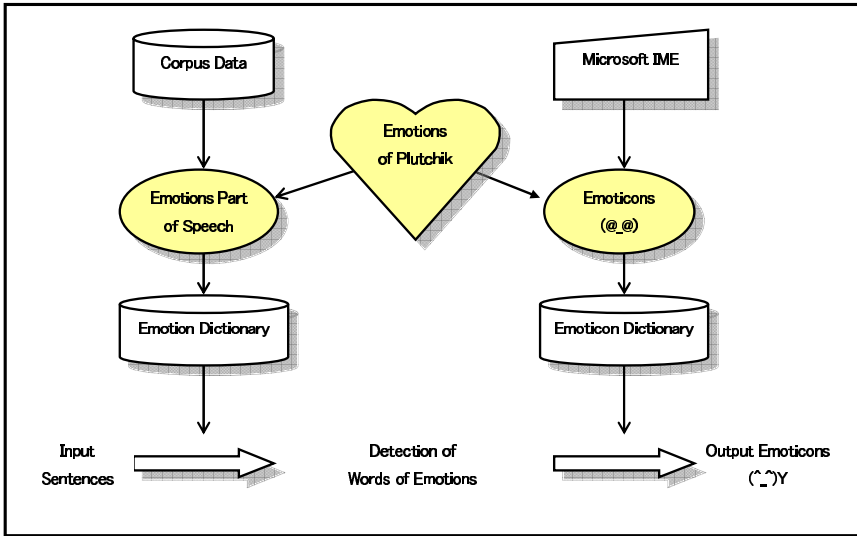


Fig. 2. The processing model of this method

2 Related Works

Various classification methods of emotions based on this study were suggested so far [1]. Woodworth said he could classify emotions to six basic emotions, and Schlossberg expanded these six emotions and suggested a three-dimensional model of emotions. Plutchik also made a criterion of basic emotions clear and defined eight emotions. This definition suggests the opposite meaning of emotions and a three-dimensional strength model based on corpus, so it is finer model than others.

A lot of automatic distinction methods of emotions in the various media with computers have been studied. Kanoh applied information of emotions to expressions in robots [2], and Matsumoto suggested emotions recognition technique by images and sounds [3]. Keila also examined emotions understanding method of emails as technique for customer's problem discovery [4].

Kort analyzed the emotion transition for learning situation [7]. They proposed an interesting four phase emotions transition model. This is important point for the smooth communication.

In addition, Nakamura suggested the technique with neural network for emotions distinction of emoticons in dialogues [5]. This is one of the methods for understanding the meaning of emoticons.

In these works, they didn't examine about automatic emoticons grants technique for the smooth communication.

3 Classification Emotions

At first, it is important to define classifications of human emotions itself that decides what kind of emotions emoticons express. Many classification methods of emotions were suggested so far. We use a classification method based on eight basic emotions of Plutchik.R which can express various emotions by the following reasons^[1]. Fig. 3 shows the eight basic emotions and its opposite meanings.

- (1) It expresses the strength of emotions.
- (2) It expresses the opposite meaning of emotions.
- (3) It defines actions with emotions and relation of the actions and the emotions.
- (4) It defines the mixture of basic emotions and can explain various emotions.

Table 1 shows the relation of feelings with Plutchik method and words of emotions. We express emotions of emoticons by these words of emotions and additional ones described in next clause.



Fig. 3. The processing model of this method

Table 1. Emotions classification of Plutchik

Basic Emotions	(Strong) ← Strength → (Weak)
Acceptance	Love, Goodwill, Trust, Generosity, Acceptance
Fear	Grief, , Worry, Sorrow, Discouragement, Sadness
Disgust	Hatred, Hate, Antipathy, Disgust
Anger	Anger, Rage, Fury, Indignation, Hostility
Anticipation	Anticipation, Expectation, Caution, Curiosity, Attention
Joy	Pride, Joy, Satisfaction, Pleasure, Peace

4 Definition of Emoticons That Show Emotions

We define emoticons corresponding to the words of emotions that were showed before. The emoticons to use were picked up representative emoticons equivalent to

each emotion words by one or more questionnaire from 172 emoticons defined in Microsoft IME 2003.

When we simply relate IME to Plutchik's words of emotions, it is concerned about falling off emoticons and words of emotions with high frequency using. Therefore, we pull out and add words of emotions which are used in emails of cellular phones and doesn't appear in Plutchik's words of emotions from our collection of email data. We also add words which are in Plutchik's words of emotions and short in IME. We show additional words of emotions in Table 2.

We picked up pairs of emoticons and words of emotions as above. Table 3 shows parts of these. We store them in our system as "Emoticon Dictionary".

Table 2. The additional words of emotions

Irritating
Ridiculous
Apology
Sleepy
Tired
Greeting
Normal farewell
Sad farewell
Request
Fear
Love
Acceptance

Table 3. The part of "Emoticon Dictionary"

Emoticons	Words of emotions
(^_^)/	Greeting
(>_<)	Disgust
(^^♪)	Joy
(T_T)	Sadness
(••?)	Doubt
(+o+)	Perplexity
m(_)_m	Apology or Request
<(^^^)>	Pride
(^_^;)	Irritating
(@_@)	Surprise
(•o•)	Unrest
(-_-)zzz	Sleepy
(=.=)	Tired

5 Morpheme Analysis for Emotions Distinctions

We use morpheme analysis in this technique to distinguish what kind of emotions emails of cellular phones have. By using morpheme analysis, we can expect that this

technique can handle it more precisely than the method of searching all strings. We use ChaSen^[2] for morpheme analysis.

We chose the parts of speech to express emotions from morphemes defined in ChaSen. We call this "Emotions parts of speech". Table 4 shows a list of these.

Table 4. A list of Emotions parts of speech (in Japanese)

Emotions part of speech	Examples of words	Number of words in the dictionary of ChaSen
Noun, Changed connection of "Sa"	愛着 (Attachment), ひと安心 (Settled)	12,041
Noun, Stem of adjective verb	安易 (Easygoing), だめ (No good)	3,313
Noun, Stem of adjective for "Nai"	申し訳 (Excuse), 仕方 (No choice)	42
Adjective, Adjective+Step "i"	哀しい (Sad), 楽しい (Fun)	654
Adjective, Unchanged type	かつこいい (Cool), きもちいい (Comfortable)	8
The number of words in total		16,058

Next, we extracted words for each emotion part of speech from 2,218 Japanese sentences input by cellular phones which we actually collected. We decided what kind of emotions these words expressed by questionnaires and built "Emotion Dictionary" such as Table 5.

Table 5. The part of "Emotion Dictionary"

Parts of Speech	Emotion words	Emotions
Noun, Changed connection of "Sa"	お願い (Request)	Request
	お祝い (Celebration)	Pleasure
Noun, Stem of adjective verb	不安 (Worry)	Perplexity
	不利 (Disadvantageous)	Sadness
Noun, Stem of adjective for "Nai"	仕方 (No choice)	Sadness
	申し訳 (Excuse)	Apology
Adjective, Step "i"	よろしく (Best regards)	Greeting
	あいくるしい (Lovely)	Love
Adjective, Unchanged type	かつこいい (Cool)	Pride

6 Emoticon Automatic Grant Technique

This chapter describes the technique to give an emoticon to express emotions for a sentence in an email of cellular phones by using Emoticon dictionary and Emotion Dictionary showed in last chapter. This technique is carried out by the following procedures.

- (1) Input one sentence.
- (2) Get the emotion part of speech at the end of the sentence by morpheme analysis. (Because it is often that cellular phone email sentences have an emoticon in the last of a sentence, we also grant an emoticon to the end of a sentence.)
- (3) Get an emotion word from Emotion dictionary using an emotion part of speech and real words as keys.
- (4) Get an emoticon for that emotion word from Emoticon Dictionary.
- (5) When emotions part of speech that we get is "Noun - Stem of adjective for Nai", check whether there are "auxiliary verb - special Nai" just after that. If it gets one, it defines an emoticon that shows the opposite meaning of emotions at (4). (It can find opposite emotions by emotion classification method of Plutchik.)
- (6) Output an emoticon which is converted from the punctuation mark at end of an input sentence.

7 Evaluation Experiment

We compared the real emoticons with emoticons acquired by this technique for 65 sentences with emoticons input by cellular phones. When the emoticon did not completely accord, we assumed it was correct answer if correct semantically. Table 6 shows examples of output sentences and Table 7 shows our results.

Table 6. Examples of output sentences

Input sentences	Output emoticons	Decision
Anyone has same job know this, please reply to me. m(_m	m(_m	Good
How about cookies or cakes for his birthday? (^_)	(^_)	Good
I received an application, but I was worried about the expence. (*_*)	(^_)	Good
I think the jobs that a high school student can work are the most reliable one. (^_;	(^o^)/	Good
I was happy but ... (^_)	(T_T)	No Good
If you are worried about it, please examine it by books of cats. V(^~^)	(~_~;)	No Good

Table 7. Evaluation results

	Number of sentences	Ratio
Correct	57	87.7%
Wrong	8	12.3%
Total	65	100.0%

As a result of evaluation, the correct answer ratio was 87.7% and our method is almost effective enough. It was 12% wrong answer ratio in our evaluation. We describe some reasons of them below.

(1) It cannot understand a conjunctive particle.

It is ambiguous meaning such as “けどね” with information only for one sentence, and difficult to distinct emotions if it has emoticons even human being. For example, “楽しかったけどね(^-^)” (“I was happy, but...” in English.)

(2) Morphemes out of the range

For example, case of the sentence “もし心配なら猫の本とか見て調べてみて下さいV(^-^)V” (“If you are worried about it, please examine it by books of cats.” In English), it chose an emoticon for “心配”, but the correct choice is one for “調べてみて下さい”(Its morpheme is Verb – five steps, “Ra” line special). We are able to handle this problem by extension of morphemes to intend for.

8 Conclusion

In this paper, we defines the emotions classification with emoticons and proposed the technique to grant an emoticon which express emotions of it at end of email sentence by input from cellular phones.

We were able to confirm the correct answer rate of 87.7% as a result of evaluation experiment and the effectiveness of this method.

We plan to study to understand of emotions by context for ambiguous expressions in future. We think it is important evaluation point that using this method in real world.

References

1. Fukui Y.: Psychology of emotions, Kawashima Publish (1990)
2. Gotoh M, Kanoh M., Kato S., Kunitachi T., Itoh H.: Face Generation Using Emotional Regions for Sensibility Robot, Transactions of the Japan Society for Artificial Intelligence (2006) Vol.21 No.1 pp.55-62
3. Matsumoto S., Yamaguchi T., Komatani K., Ogata T., Okuno H.: Emotion recognition by integration face image information and sound information for using in robots, Proceedings of 22th The Robotics Society of Japan Conventions (2004) 3D14
4. Keila P.S., Skillicorn D.B: Detecting unusual and deceptive communication in email, External technical report, School of Computing, Queen’s University (2005)
5. Nakamura J., Ikeda T., Inui N., Kotani Y.: Learning Face Mark for Natural Language Dialogue System, Natural Language Processing Study Report of Information Processing Society (2003) No.154-24
6. Matsumoto H.: A morpheme analysis system “ChaSen”, Information Processing (2000) Vol.41 No.11
7. Kort B., Reilly R., Picard R.: An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy – Building a Learning Companion, Proceedings of International Conference on Advanced Learning Technologies (2001)

A New Approach for Improving Field Association Term Dictionary Using Passage Retrieval

Kazuhiro Morita, El-Sayed Atlam, Elmarhomy Ghada, Masao Fuketa,
and Jun-ichi Aoe

Department of Information Science and Intelligent Systems
University of Tokushima, Tokushima 770-8506, Japan
atlam@ccr.tokushima-u.ac.jp

Abstract. Large collections of full-text document are now commonly used in automated information retrieval. Readers generally identify the subject of a text when they notice specific terms, called *Field Association (FA) terms*, in that text. Previous researches showed that evidence from passage can improve retrieval results by dividing documents into coherent units with each unit corresponding to a subtopic. Moreover, many current researchers are extracting *FA terms* candidates from the whole documents to build *FA term* dictionary automatically. This paper proposes a method for automatically building new *FA term* dictionary from documents after using passage retrieval. A WWW search engine is used to extract *FA terms* candidates from passage document corpora. Then, new *FA terms* candidates in each field are automatically compared with previously determined *FA terms* dictionary. Finally, new *FA terms* from extracted term candidates are appended automatically to the existence *FA terms* dictionary. From experimental results the new technique using passage documents can automatically append about 15% of *FA terms* from terms candidates to the existence *FA term* dictionary over the old method. Moreover, *Recall* and *Precision* significantly improved by 20% and 32% over the traditional method. The proposed methods are applied to 38,372 articles from the large tagged corpus.

1 Introduction

In recent years there is a tremendous growth of large amount of online text information available on the Internet, digital libraries, medical diagnostic systems, remote education, news sources and electronic commerce. There is an extreme need to search and organize huge amounts of relevant information in text documents. In text retrieval, it is important to recognize wanted portion of text information depending on user query. Moreover, to retrieve relevant texts is important rather than to investigate entire documents especially when the text document is large. There is research and development related to automatic classification of document files [8]. Particularly, in document classification, the technique is based on vector-space [11] and probabilistic [7] methods. By using these methods it is possible to retrieve and classify texts in response to arbitrary databases without referring to systematic

classified information. However, due to multiple topics and document fields, the content to be searched usually exists in only part of the file [5][14][16].

Atlam et al. [4] proposed a method for automatic building of new field association term candidates using search engine from the whole documents. The drawback of this method that it used whole documents so, the extract *FA terms* is not more accurate. A new approach detects effectiveness *FA term* by using passage document which divided text into coherent units with each unit corresponding to subtopic. This approach presents a new system for automatically building *FA terms* candidates using search engine from these passage documents. This new system uses a *WWW* search engine [17] to extract *FA terms* candidates from passage document data for each field instead of using whole documents. These candidates terms are automatically compared with *FA terms* already exist in a *FA terms* dictionary. Then, new *FA terms* can be added to a dictionary of *FA terms*. From experimental results the new technique automatically appending 15% new *FA terms* over Atlam's method to *FA term* dictionary.

2 Document Field Association Terms and Passage Retrieval

2.1 Document Field Tree

A *document field tree* structure ranks relationships between document fields [1][4][18]. The field tree in Figure 1, based on Imidas'99 [6], contains 14 *super-fields*, 443 *sub-fields* and 393 *terminal fields*. *Root names* are omitted unless there is conflict between *super-fields* and *sub-fields*. In such cases, only *terminal fields* are described and *FA terms* and paths are manually assigned. For example, path <*SPORTS*/Ball Games/Tennis> describes the document field <*SPORTS*> as *Super-field* of < Ball Games > of *sub-field* <Tennis>.

2.2 Document Field Association Terms

A *Single Field Association (FA)* is a minimum unit (*word*) in the term. A *Compound FA Term* that consists of two or more single *FA Term* is regarded as being single if it loses field information when divided (e.g. *nuclear weapon*, *global warming*) [2][3][23]. A computer that is taught selected *FA Terms* saves those terms in the field tree as a knowledge base.

2.3 FA Term Levels

A document field can be ranked as: *super-fields*, *sub-fields* or *terminal fields*. *FA Terms* are grouped according to how well they indicate specific fields. *FA Terms* have different rank to associate with document fields, so five *Field levels* can be used to classify *FA Terms* according to document fields, as in Table 1.

3 Passage Retrieval

3.1 Passage Retrieval

Passage retrieval techniques have been extensively used in standard *IR* settings, and have proven effective for document retrieval when documents are long or when there

Table 1. Single FA Terms with Paths and Field Levels

<i>FA Term</i>	<i>Field Association Path</i>	<i>Field accuracy Level</i>
<i>Seismic</i>	<Civil Engineering/Structure Engineering/ Earthquake>	1 = <i>Perfect-FA Terms</i>
<i>Single</i>	<SPORTS/Ball Games/Tennis> <SPORTS/Ball Games/ Badminton>	2 = <i>Imperfect- FA Terms</i>
<i>Game or Contest</i>	<SPORTS>	3 = <i>Super-FA Terms</i>
<i>Victory & Defeat</i>	<SPORTS> <POLITICS & LAW/ Election> <HOBBY & ENTERTAINMENT/Games/ Chess>	4 = <i>Multiple-FA Terms</i>
<i>the or use</i>	---	5 = <i>Non-Specific FA Terms</i>

(Field Level 1) Perfect-FA Terms are associated with one *sub-field* (e.g. *Seismic* and *earthquake ground motion* are associated with *sub-field* <Earthquake> only).

(Field Level 2) Imperfect FA terms are associated with more than one *sub-field* in one *super-field* (e.g. *goal* and *goalkeeper* are associated with *sub-fields* <Soccer> and <Hockey> of *super-field* <SPORTS>).

(Field Level 3) Super-FA Terms are associated with one *super-field* (e.g. *team* and *player* are associated with *super-field* <SPORTS>).

(Field Level 4) Multiple-FA Terms are associated with more than one *sub-field* of more than one *super-field* (e.g. *winner* is associated with *super-field* <SPORTS> and *sub-field* <POLITICS/Election>).

(Field Level 5) Non-Specific FA Terms do not specify *sub-field* or *super-field* and also include stop words (e.g. *articles*, *prepositions*, *pronouns*).

are topic changes within a document, thus making it an appealing candidate for the present work. Second, from an *IR* system user's standpoint, it may be more desirable that the relevant section of a document is presented to the user than the entire document. Passages can be defined based on the document structure [3][10][13][20]. This entails using author-provided marking (e.g. period, indentation, empty line, etc.) as passage boundaries. Examples of such passages include paragraphs, sections, or sentences. Passages can also be defined according to subject or content of the text. The main idea is to divide documents into coherent units with each unit corresponding to a subtopic. A well-known algorithm for deriving such passages is Text Tiling [9][10]. Other algorithms have been reported in [12][19][21]. The third type of passage is window, which consists of a fixed number of words or bytes. Passages in this category may or may not take logical structure of the document into account.

4 Improvement of Automatic Building FA Terms

4.1 Passage Field Association Terms

This paper presents a technique for extracting *FA terms* from passage documents instead of whole document, passage retrieval divide documents into coherent units

with each unit corresponding to a subtopic. The method presented is based upon reducing whole text to produce coherent passages which described the text and extracting *FA Terms* from the selective documents. The advantage of new method to extract discriminating words to produce more accuracy and effectiveness *FA terms*.

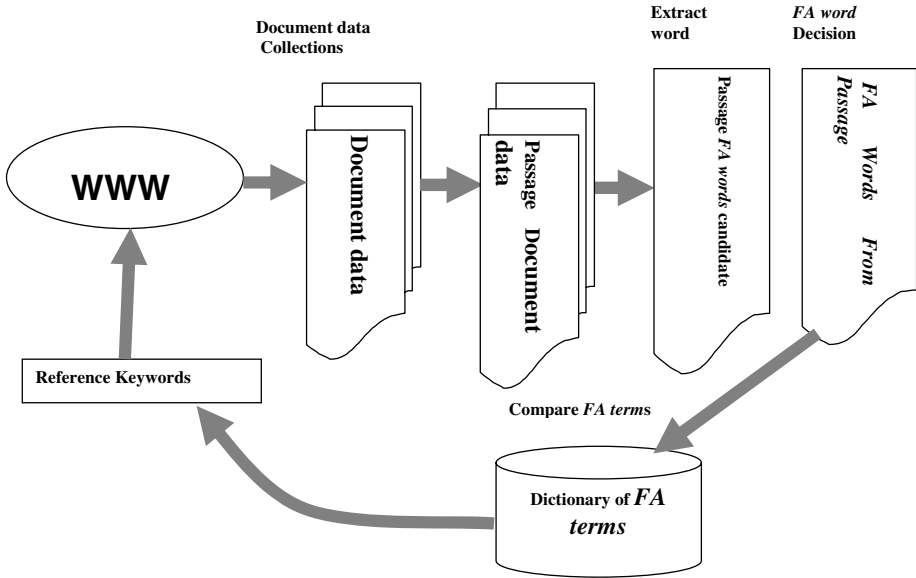


Fig. 2. System Outline

4.2 System Outline

To retrieve document data, a word from an existing *FA term* dictionary is used as a reference keyword for a *WWW* search engine (see Fig. 2). Word frequencies are collected for every document field in the field tree. *FA Terms* are decided from words which are associated with specific fields. At first passage documents are extracted from whole text, then extracted new *FA term* candidates from Passage retrieval are appended to the dictionary of *FA terms* by comparison with *FA terms* already listed in that dictionary. By using this system a New *FA terms* dictionary is produced with high accuracy.

5 Experiment Evaluation

5.1 Data

Word collection data are used for experimental data and there are 38,372 term candidates and 4.52MB document data from a data set of 20 Newsgroup from CNN Web Site (1996-2003). Concentration ratio is changed from 0.5 ~ 0.9 to decide the levels of *FA term* candidates.

5.2 Experiment Simulation Results

Figure 3 shows the number of extracted *FA terms* on level 1 become higher after using passage documents because after using passage documents the number of extracting level 1 become performing the systems to more accurate extracting *FA terms* related to the field. Based on different concentration ratios, the number of extracted *FA terms* of Level 1 creases as concentration ratios decrease. Because of, when the concentration ratio is high, *FA terms* of Level 2, Level 3 and Level 4 are extracted accurate but, when the ratio is low, *FA terms* of Level 2, Level 3 and Level 4 become Level 1 *FA terms*.

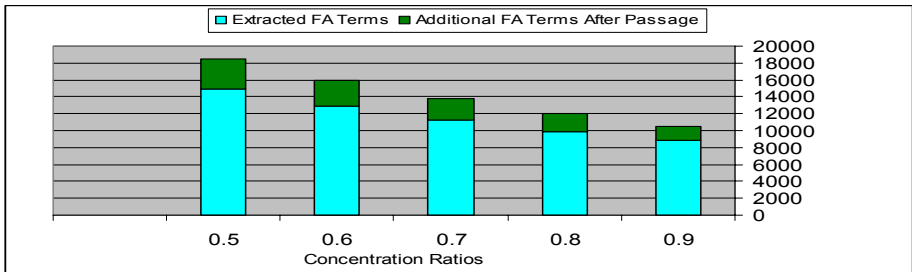


Fig. 3. Level 1 *FA terms* and additional one after using passage documents with concentration ratios

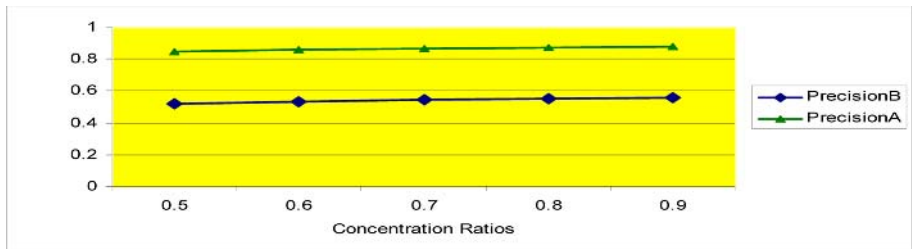


Fig. 4. Precision Before and After Using Passage documents with Concentration Ratios

From Figure 4, we notice that the Precision (PrecisionA) after using extracted *FA terms* from passage documents make *FA terms* dictionary are higher than using the whole documents (PrecisionB) because of when new *FA terms* candidates in each field are automatically compared with previously determined *FA terms* dictionary, the new appended *FA words* become restricted and accurate from passage documents so, Precision become high. Moreover, the number of relevant *FA terms* is increasing with the increasing of concentration ratio of threshold, so, the accurate terms are extracted at the increasing concentration ratio of threshold. Therefore, concentration ratio of threshold at 0.9 is effective enough for that propose than other threshold values. Moreover, the Precision after using *FA terms* from passage documents higher than using *FA terms* from whole documents by around 32%.

In conclusion, new approach using extract *FA terms* from passage documents for improving appending *FA terms* dictionary is also performing well effective in Recall and Precision than traditional method.

6 Conclusion

With increasing popularity of the Internet and tremendous amount of on-line text, automatic document classification is important for organizing huge amounts of data. Readers can know the subject of many document fields by reading only some specific *FA terms*. Moreover, document fields can be decided efficiently if there are many *FA terms* and their *frequencies* rate become high. Previous researches showed that evidence from passage can improve retrieval results by dividing documents into coherent units with each unit corresponding to a subtopic. Moreover, many researchers have been extracted automatically *FA terms* from the whole Documents. This paper proposed a method for automatically building new *FA term* from documents after using passage retrieval. A *WWW* search engine is used to extract *FA terms* candidates from passage document corpora. Then, new *FA terms* candidates in each field are automatically compared with previously determined *FA terms* dictionary. Finally, new *FA terms* from extracted term candidates from passage documents are appended automatically to the existence *FA Terms*. New Method appended and improved around 15% of new *FA terms* from extracted term candidates to an existence *FA terms* dictionary, which means around 15% of appended new terms over than Alam's method. Moreover, Precision and Recall are achieves 88% and 92% respectively using new method. Future work could focus on using automatic building of *FA terms* with similarity measurements.

References

- [1] Aoe, J., Morita, K. & Mochizuki, H. An Efficient Retrieval Algorithm of Collocate Information Using Tree Structure. *Transaction of the IPSJ*, 39 (9), (1989), 2563-2571.
- [2] Atlam, E.-S., Morita, K., Fuketa, M. & Aoe, J.. A New Method for Selecting English Compound Terms and its Knowledge Representation. *Information Processing & Management Journal*, 38 (6), (2002), 807-821.
- [3] Atlam, E.-S., Fuketa, M., Morita, K. & Aoe, J. Document Similarity measurement using Field association terms. *Information Processing & Management Journal*, 39(6), (2003), 809-824.
- [4] Atlam, E.-S., G Elmarhomy, Fuketa, M., Morita, K. & Aoe, J.. Automatic Building of New Field Association Word Candidates Using Search Engine. *Information Processing & Management Journal*, 42 (4), (2006), 951-962.
- [5] Breiman, L., Friedman, J.H., Olshen, R. A. & Stone, C.J. *Classification and Regression Trees*. Chapman & Hall, (1984)..
- [6] Callen, J. P. Passage and level evidence in document retrieval. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1994), 302-310.
- [7] Dozawa, T. *Innovative Multi Information Dictionary Imidas'99*, Annual Series, Zueisha Publication Co., Japan (1999) (In Japanese).

- [8] Fuhr, N. (1989). Models for retrieval with probabilistic indexing, *Information Processing and Retrieval* 25 (1), 55-72.
- [9] Fukumoto, F., Suzuki, Y. Automatic Clustering of Articles using Dictionary definitions. In proceeding of the 16th International Conference on Computational Linguistic (COLING'96), (1996), 406-411.
- [10] Hearst, M.A., & Plaunt, C. Subtopic structuring for full-length document access. In R. Korfhage, E. Rasmussen, & P. Willet (Eds.), *Proceedings of the 16th annual international ACM-SIGIR conference on research and development in information retrieval*, Pittsburgh, PA New York: ACM, (1993), 59-68.
- [11] Hearst, M. A.. TextTiling, a quantitative approach to discourse segmentation. Technical Report 93/24 Sequoia 2000 Technical Report, University of California, Berkeley, (2000).
- [12] Iwayama, M. & Tokunaga, T. Probabilistic Passage Categorization and Its Application. *Journal of Natural language Processing*. 6 (3) (1999), 181-198.
- [13] Jiang, J., Zhai, C.X. (2004). UIUC in HARD 2004-Passage Retrieval Using HMMs, University of Illinois at Urbana-Champaign, TREC 2004.
- [14] Jones, K. S.. Automatic summarizing: factors and directions, Computer Laboratory, University of Cambridge, (1998).
- [15] Kaszkiel, M. & Zobel, J. Passage retrieval revised In Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in information Retrieval, (1997), 178-185.
- [16] Kawabe, K. & Matsumoto, Y.. Acquisition of normal lexical knowledge based on basic level category. *Information Processing Society of Japan, SIG note*, NL125-9, (1998), 87-92. (in Japanese).
- [17] Melucci, M.. Passage Retrieval and a Probabilistic technique". *Information Processing and Management*.34 (1), (1998), 43-68.
- [18] Ohkubo, M., Sugizaki, M., Inoue, T. & Tanaka, K.. Extracting Information Demand by Analyzing a WWW Search Login". *Trans. of Information Processing Society of Japan*, 39(7), (1998), 2250-2258.
- [19] Salton, G., & McGill, M.J. *Introduction of Modern Information Retrieval*. New York: McGraw-Hill, (1983).
- [20] Salton, G., Allan, J. and Singhal, A.K. Automatic text decomposition and structuring. *Information Processing and Management*, 32 (2), (1996), 127-138,.
- [21] Salton, G. (1989). *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.
- [22] Salton, G., Allan, J. and Buckley, C. Approaches to passage retrieval in full text information systems. In Proceedings of the 16th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, (1993), 49-58.
- [23] Tsuji, T., Nigazawa, H., Okada, M. and Aoe, J. Early Field Recognition by Using Field Association Words". In the Proceeding of the 18th International Conference on Computer Processing of Oriental Language, 2, (1999), 301-304.
- [24] Tsuji, T., Fuketa, M., Morita, K. & Aoe, J.. An Efficient Method of Determining Field Association Terms of Compound Words. *Journal of Natural Language Processing*. 7(2), (2000), 3-26.

Analysis of Stock Price Return Using Textual Data and Numerical Data Through Text Mining

Satoru Takahashi^{1,2}, Masakazu Takahashi³, Hiroshi Takahashi⁴, and Kazuhiko Tsuda²

¹ Mitsui Asset Trust and Banking Co., Ltd., 3-23-1 Shiba, Minato-ku, Tokyo, Japan
Satoru_1_Takahashi@mitsuitrust-fg.co.jp

² Graduate School of Systems Management, The University of Tsukuba, 3-29-1 Otsuka, Bunkyo-ku, Tokyo, Japan
{satoru, tsuda}@gssm.otsuka.tsukuba.ac.jp

³ Shimane University, 1060 Nishikawatsu-cho, Matsue-shi, Shimane, Japan
smasakazu@cis.shimane-u.ac.jp

⁴ Okayama University, 1-1, Tsushima-Naka, 1-Chome, Okayama, Japan
htaka@e.okayama-u.ac.jp

Abstract. In finance task domain, it is indispensable to get and analyze information as quickly as possible. Analyst's reports are one of the important information in asset management, and these include a large amount of text information. However, it is very difficult to handle text information of analyst's reports, few research and development have been conducted. In [5] and [6] we explored the feasibility to extract valuable knowledge for asset management through text mining using analyst's reports as text data. And we found the effectiveness of keyword information. In this paper we make further research of analyst's reports. From empirical study on the practical data, we have confirmed the effectiveness of using keyword information and numerical information together: (1) the effectiveness of keyword information is different by the direction of change of earning estimate; (2) the keyword of "Upward (or Downward) surprise in forecast" has strong effect to stock price return.

1 Introduction

The environment surrounding asset management has dramatically been changing. The rapid progress of telecommunication technologies, such as the Internet, has eliminated the time lag on distributing vast amount of investment information. Although this kind of information contains valuable information for asset management, it is impossible to manually handle all information. It is an important task in asset management to make full use of such information more efficiently and quickly.

In many studies of finance, the reaction against only numerical information that based on financial statement or analyst's reports has been analyzed [2],[3],[4]. Analyst's reports, which are one of the important information sources for investors, contain both numerical and text data, which describe the state-of-the-art business conditions of firms. For instances, the text information about such as "business

reconstruction" or "business restructuring" are not numerical data but have strong impacts to markets. So, the use of text information of analyst's reports is indispensable to analyze financial market.

One of the problems to analyze text information of analyst's reports is difficulty of handling text information. In order to solve this problem, text mining, which can analyze large quantities of text data systematically, is very effective method. Text mining aims at obtaining valuable knowledge from enormous amount of text data by analyzing the tendencies and correlation of the contents based on the change histories of the texts and the distribution of the keywords in the text data.

Although there is still few research which uses text mining for analysis of finance, some researches are reported [1],[7]. These analyses are against usual information like web news, and the analysis of the expert information on security analysts does not exist. So far, we have analyzed the relationship between text information of analyst's reports and stock prices [5],[6]. In [5] and [6], we extracted keyword information from title of analyst's reports, and we studied how stock prices move in existence of extracted information. We found that analyst's reports contained valuable information that influenced to the stock prices. And we identified the same influence on manufacturing and non-manufacturing industries. We also found that multiple analysts reacted to the same information. As a result of eliminating such duplication of information, we succeeded to extract more valuable information from analyst's reports. In this paper, we analyze analyst's reports by combining text information with numerical information to extract more valuable information.

The paper is organized as follows. We explain the outline of our research and the data we used, in section 2 and 3. We illustrate the result in section 4, and then we conclude our discussion in Section 5.

2 Outline of Research

This section describes the outline of our research. In this paper, we use the information of existence of keyword in the title of analyst's reports. We analyze how the information of whether keywords, such as "upward revision of estimate" or "reconstruction of an organization", exist or not in the title of analyst's reports has affected stock prices. We show the design of the whole analysis in Figure 2.1:

- 1) obtain analyst's reports via WWW or e-mails,
- 2) construct Keyword DB by a knowledge representation,
- 3) classify analyst's reports using the existence of a keyword in the title,
- 4) for each keyword, measure stock price return on the basis of the time that the analyst's reports is released,
- 5) analyze the effect of a keyword by comparing the pattern of stock price return of Group A and Group B.

And then, we measure the influence of the report in detail. We use numerical information based on financial statement database to evaluate whether we can get more

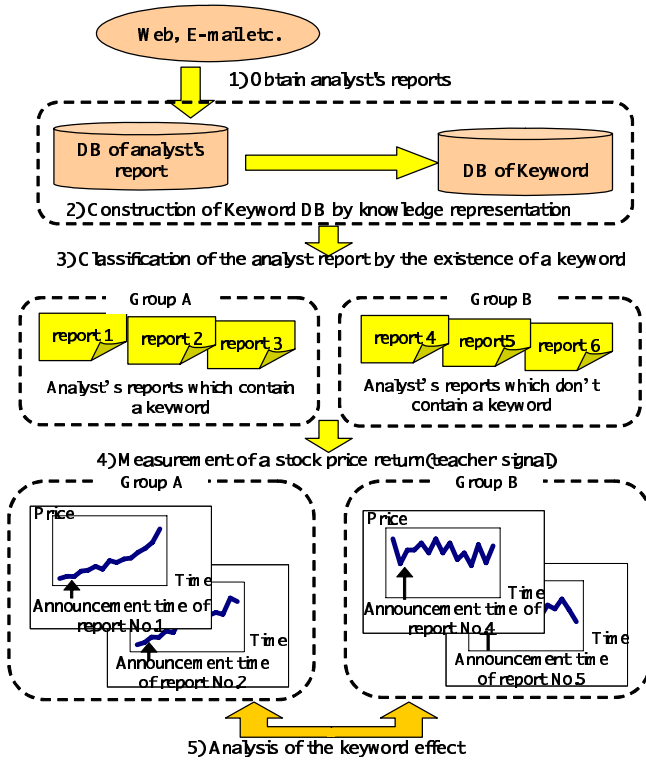


Fig. 2.1. Outline of the model

valuable information or not. We classify company into two groups using numerical information, and then, we conduct analysis shown in Figure 2.1 to both of groups. If the result is different in both groups, it turns out that using numerical information makes it possible to validate text information of analyst's report (Figure 2.2).

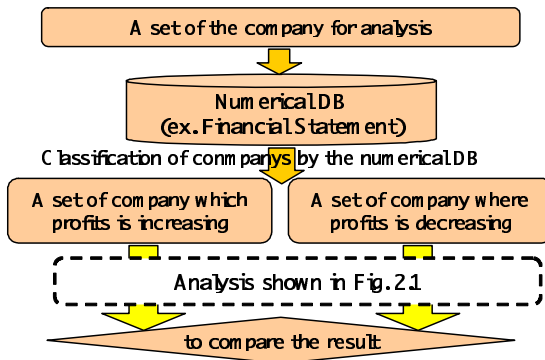


Fig. 2.2. Outline of the detail research

3 Data Used for Analysis

3.1 Type of Company and Analyst's Reports

We use all stock data listed in the first section of the Tokyo Stock Exchange in the period from January 1st, 2001 to March 31st, 2003. There are 1,619 firms listed during the period. Using Thomson Financial Web service, we have obtained 77,256 analyst's reports related to the listed firms. And we use the I/B/E/S Consensus Estimates as numerical data. The obtained analyst's reports have characteristic that the analyst coverage in the larger companies is greatly different from the ones of the smaller companies.

3.2 Kind and Type of Keywords

We have extracted keywords from the title of the obtained reports. As shown in Table 3.1, we classify the extracted keywords into twelve groups after the difference adjusting. We also classify the keywords into three news types: Good, Bad, and Neutral News.

Table 3.1. Classification of keywords

NO	Notation adjusted keyword	News type
1	Increase in profits	Good News
2	Upward surprise in forecast	Good News
3	Downward surprise in forecast	Bad News
4	No surprise in forecast	Neutral
5	Business restructuring	Good News
6	Upward earnings revision	Good News
7	Downward earnings revision	Bad News
8	Rating "Sell"	Bad News
9	Rating "Buy"	Good News
10	Rating unchanged	Neutral
11	Upgrade of rating	Good News
12	Downgrade of rating	Bad News

3.3 Teacher's Signal

We use stock price returns as the target concepts to measure the influences of the keywords. The influence is measured as a difference of stock price returns before and after the analyst's reports has been published. We classify the stocks into two groups: Group A, which contains keywords in the report title and Group B, which does not contain them. Then, we use statistical test of the differences of the average stock price returns between Groups A and B. We employ Welch's test to measure the differences of mean values, since there exists heteroskedasticity in stock price returns between the groups.

3.4 Numerical Information

As numerical information, we use change of monthly consensus earning estimate for next fiscal year (CESFY1). CESFY1 is the mean value of analyst's forecast of next fiscal year earning. So, CESFY1 can be considered as next fiscal year earning

estimate that the investors of the market expect on the average. CESFY1 has important information to stock price; especially the change of CESFY1 has a great influence on a stock price. We use change of monthly CESFY1. We classify data into two groups by using whether monthly CESFY1 changes upward or downward revision. We define upward or downward revision of monthly CESFY1 as follows:

- Upward revision : $(EPS_t - EPS_{t-1}) / (abs(EPS_t) - abs(EPS_{t-1})) > 0$
- Downward revision : $(EPS_t - EPS_{t-1}) / (abs(EPS_t) - abs(EPS_{t-1})) < 0$

where EPS_t : CESFY1 per share in month t

4 Analysis of Stock Price Return Using Textual Data and Numerical Data

In [5] and [6], we confirmed that analyst’s reports had important influence to stock price and the effect of analyst’s reports did not depend on a type of industry, but it depended on a size of firm. In this subsection, we analyze whether this effect will change or not if we use numerical data with text data. We use the change of mean value of analyst’s forecast of next fiscal yare earning per share. We classify data into two groups by using whether monthly earning forecast changes upward or downward revision. We define upward or downward revision of monthly earning estimate in subsection 3.4.

First we show the result of using all keywords (Fig 4.1).

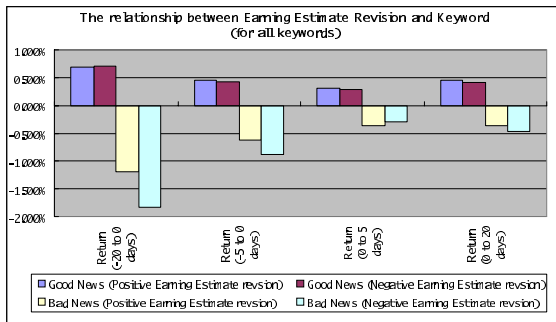


Fig. 4.1. Analysis of keyword and earning estimate revision on all keyword

As shown in Fig 4.1, we can find the stock price reacts in positive direction to Good News and in negative direction to Bad News, regardless of whether the earning estimate revision is upward or downward revision.

Next, we use only two keywords “Upward earning revision” and “Downward earning revision”. If an analyst uses the keyword of “Upward (or Downward) earning revision” in his/her report when CESFY1 changes upward (or downward) revision, it

means that his/her earning forecast of a certain company is almost the same as market valuation. We show the result of analyses in Figure 4.2. We can see the same result as Fig 4.1. So when an analyst's opinion is the same as a market, the effect of keyword of “Upward (or Downward) earning revision” hardly is different from that of other keywords.

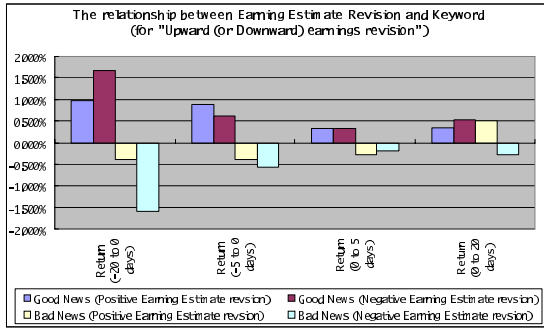


Fig. 4.2. Analysis of keyword and earning estimate revision on “Upward (or Downward) earning revision”

Then we use other two keywords that are thought to have a strong effect to direction of earning estimate revision, “Upward surprise in forecast” and “Downward surprise in forecast”. If an analyst uses the keyword of “Upward (or Downward) surprise in forecast” in his/her report, it means that his/her earning forecast of a certain company is larger (or smaller) than he assumed previously. We can see a very interesting result in Fig 4.3.

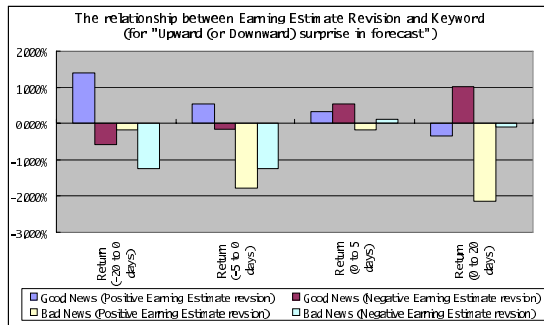


Fig. 4.3. Analysis of keyword and earning estimate revision on “Upward (or Downward) earning revision”

When the earning estimate changes upward direction and an analyst uses Good News in his/her analyst’s reports, the stock price reacts to negative direction. On the other hand, even when the earning estimate changes downward direction, if an analyst uses Good News in his/her analyst’s reports, the stock price reacts to positive

direction very strongly. And even if Bad News is found, an effect of keywords is lower when the earning estimate changes downward direction. In other words, we find that an impact of text information of analyst's reports is larger when analysts express their opinion opposite to the direction of average earning estimate revision.

5 Conclusion

In this paper, we have analyzed the effect of text information of analyst's reports through text mining. In [5] and [6], we confirmed that analyst's reports had important influence to stock price. So we analyze the validity of combining text information with numerical information. We have identified that to combine text information with numerical information make it possible to extract more valuable knowledge from analyst's reports.

In our future work, to explain mechanism of stock prices decision in detail, we combine text information with numerical information in higher level. And we hope to construct more precision model.

References

1. Antweiler, W., and Frank, M. Z.: Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *The Journal of Finance*, 2004, Vol. 59, 1259-1294.
2. Chopra, V. K.: Why So Much Error in Analysts' Earnings Forecasts?, *Financial Analysts Journal*, 1998, November/December, 35-42.
3. Clement, M., B.: Analyst Forecast Accuracy: Do Ability, Resources, and Portfolio Complexity Matter?, *Journal of Accounting and Economics*, 1999, No.27,285-303.
4. Dreman, D., and Berry, M.: Analyst Forecasting Error and Their Implications for Security Analysis, *Finance Analysts Journal*, 1995, Vol.51, No.3, 30-41.
5. Takahashi, S., Masakazu, T., Takahashi, H., Tsuda, K.: Learning Value-added Information of Asset management from Analyst Reports through Text Mining, *Lecture Notes in Computer Science: 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 2005, Proceedings, Part IV, Springer-Verlag Heidelberg*, 2005.
6. Takahashi, S., Takahashi, H., Tsuda, K., Terano, T.: Analyzing Asset Management Knowledge from Analyst's Reports through Text Mining, *International IPSI-2004*, 2004.11.
7. Wuthrich B., Cho V., Leung S., Permunetilleke D., Sankaran K., Zhang J., Lam W.: Daily Prediction of Major Stock Indices from Textual WWW Data, *KDDM'98 Conference NY, AAAI Press (1998) 364-368*

A New Approach for Automatic Building Field Association Words Using Selective Passage Retrieval

El-Sayed Atlam, Elmarhomy Ghada, Kazuhiro Morita, and Jun-ichi Aoe

Department of Information Science and Intelligent Systems
University of Tokushima Tokushima, 770-8506, Japan
atlam@ccr.tokushima-u.ac.jp

Abstract. Large collections of full-text document are now commonly used in automated information retrieval. When the stored document texts are long, the retrieval of complete documents may not be in the users' best interest and extract *Filed Association (FA) words* is not accurate. In such circumstances, efficient and effective retrieval *FA words* may be obtained by using passage retrieval strategies designed to retrieve text excerpts of varying size in response to statements of user interest.

New approaches are described in this study for implementing selective passage retrieval systems, and identifying text passage response to particular user needs. Moreover an automated system is using for extract accurate *FA words* from that passage and evaluate the usefulness of the proposed method. From the experimental results, when passage retrieval are accessible leading to the retrieval of additional extracted relevant *FA word* with corresponding improvements in Recall and Precision. Therefore, *Recall* and *Precision* improved by 30% than using whole texts and traditional methods.

1 Introduction

Both Internet information and users are rapidly growing. Hence, using a search engine to retrieve useful information has become important [12]. In operational retrieval environments, it is now possible to process the full text of all sorted documents. Many long, book-size documents are stored, often containing a mix of different topics covered in more or less detail. In these circumstances, it is not useful to maintain the integrity of complete undivided documents. Instead individual text passage should be identified that are more responsive to particular user needs than the full documents texts, as Salton, Allan, and Buckley (1993)[14] indicated, "in such cases, efficient and effective retrieval results may be obtained by using passage-retrieval strategies designed to retrieve text excerpts of varying size in response to statements of user interest". Moreover, most information retrieval systems return a ranked list of whole documents as answers to a query. However, when documents are long and have multiple topics, retrieval at passage-level, i.e., returning relevant passages, rather than whole documents, may be more useful to the user as the user does not need to read through a whole document to find the most relevant part. Passage retrieval also

enables an IR system to re-score documents based on their relevant passages, and exploit feedback more accurately based on passages rather than whole documents.

In this paper, *field* means basic and common knowledge that can be used in human communication [10][16]. It is natural for people to identify the field of document when they notice specific words. We shall refer to these specific words as *Field Association (FA)* words; especially they are words that allow us to recognize intuitively a field of text or documents. For example, people who encounter the words “*election*” or “*seismic*” can recognize the document *super-field* <Politics> or *sub-field* <Earthquake>.

FA word is a minimum word [17] which can not be divided without losing semantic meaning. Based on specific *FA word* information, topics of documents (e.g. *Politics*, *Earthquake*) can be recognized. For example, the word “*election*” can indicate *super-field* <POLITICS> or word the “*seismic*” can indicate *sub-field* <EARTHQUAKE>.

Several advantages are apparent when individual text passages [6][8] are independently accessible. First, the efficiency of text utilization may be improved because the users no longer faced with large mass of retrieved material, but can instead concentrate on the most immediately relevant text passage. Moreover, extracted accurate and restricted *FA word* to the field will be efficient from that text passage. Second, the effectiveness of the retrieval activities may also be enhanced because relevant short texts are generally more easily retrievable than longer one. The longer items covering a wide diversity of different subjects may not closely resemble narrowly-formulated, specific user queries. When text excerpts are accessible, extracted *FA word* is often higher for the text excerpts than the corresponding full texts, leading to the retrieval of additional relevant *FA words* with corresponding improvements in recall and precision.

Most text items are naturally sub-dividable into recognizable units, such as text sections, paragraphs, and sentences. This leads to the notion of assembling text excerpts of varying size covering just the right amount of information to satisfy the user population. In the remainder of this study, effective methods are introduced for identifying relevant text excerpts in response to user interest statements as passage, and extracting the accurate *FA words* from that relevant text passages.

2 Passage Retrieval

2.1 FA Word Using Passage Retrieval

Passage retrieval in this study is based on the use of text paragraphs rather than sentences for the construction of text passages. In that case, relationships are computed between individual text paragraphs, based on the number of common text components in the respective paragraphs. Certain paragraphs are then chosen for abstracting purposes, replacing the originally available text [15]. In this study the use of complete paragraphs is generalized to include text passage of varying length, covering the subject at varying levels of detail, and responding to varying kinds of user needs. A top-down approach is used whereby large text excerpts are chosen first,

that are successively broken down into smaller pieces covering increasingly specific user needs. This makes it possible to retrieve full texts, text sections, text paragraphs, or sets of adjacent sentences depending on particular user requirements.

2.2 Retrieval of Text Excerpts

The global/local text matching strategy described in the previous section is capable of retrieving relevant documents with a high degree of accuracy. However, restricting the retrieval to full document texts presents two main problems. Most obviously, the users will be overloaded rapidly when large document text are involved. Second, a potential loss in retrieval effectiveness (recall) occurs because the global query similarity of the long, discursive documents that cover a number of different topics will be low, implying that many long documents will be rejected, even when they contain relevant passages.

The user overload can be reduced and the retrieval effectiveness enhanced by making it possible to retrieve text passage [9][11] instead of full document only, whenever the query similarity of a text excerpt is larger than the similarity of the complete document. This suggests that a hierarchical text decomposition system be used which successively considers for retrieval text sections, text paragraphs, and sets of adjacent text sentences. In each case, the text excerpt with the highest query similarity may be presented to the user first, while providing options for obtaining larger and smaller text pieces. Because of the local context checking requirement used for retrieval purposes, text excerpts such as paragraphs and sentences are already individually accessible and additional resources needed for passage retrieval purpose are relatively modest.

3 Document *Field Association Terms* and Levels

3.1 Document Field Tree

A *document field tree* structure ranks relationships between document fields [1][5]. The field tree, based on Imidas'99 [7], contains 14 *super-fields*, 443 *sub-fields* and 393 *terminal fields*. *Root names* are omitted unless there is conflict between *super-fields* and *sub-fields*. In such cases, only *terminal fields* are described and *FA terms* and paths are manually assigned. For example, path <SPORTS/Ball Games/Tennis> describes the document field <SPORTS> as *Super-field* of <Ball Games > of *sub-field* <Tennis>.

3.2 Document *Field Association Terms*

A *Single Field Association (FA) term* is a minimum unit (*word*) which can not be divided without losing semantic meaning (e.g. *computer*, *player*). A *Compound FA term* that consists of two or more single *FA term* is regarded as being single if it loses field information when divided (e.g. *nuclear weapon*, *global warming*) [2][3][4].

A computer that is taught selected *FA terms* saves those terms in the field tree as a knowledge base.

An *FA term* can be a word (e.g. *game*) or a phrase (e.g. *victory* and *defeat*) that indicates subject matter category in the classification scheme. The basic concept underlying *FA terms* involves choosing a limited set of words that best match a given document, so *FA terms* describe a set of discriminating words. *FA terms* are not always the same as words that specifically identify subject fields. *FA terms* appear in a document, but subject words may not appear in that document, so *FA terms* may be better for discriminating between documents than subject words. Many *FA terms* are not subject words (e.g. *case* or *use*). There are few semantic differences among *FA terms* and the choice of *FA terms* used in a document is mainly a matter of style. *FA terms* generally have high *inverse document frequencies (idf)* and important role in passage retrieval.

3.3 Single FA Term Levels

A document field can be ranked as: *super-fields*, *sub-fields* or *terminal fields*. *FA terms* are grouped according to how well they indicate specific fields. *FA terms* have different rank to associate with document fields, so five *Field levels* can be used to classify *FA terms* according to document fields.

4 Experiment Evaluation

4.1 Data

Around 6.52 MB document data from a data set of 20 Newsgroup from CNN Web Site (1996-2004) and word collection data are used for experimental data and there are 38,372 word candidates after passage retrieval. Concentration ratio in this experimental is changed from 0.5 ~ 0.9 to decide the levels of *FA word* candidates.

4.2 Experiment Simulation Results

Figure 1 shows the number of extracted *FA words* on each level **before** (Level *ib*, $i = 1, 2, 3, 4$) and **after** (Level *ia*, $i = 1, 2, 3, 4$) using passage retrieval based on different concentration ratios. From Figure 2, we notice that the number of extracted *FA words* of Level *1a*, Level *2a*, Level *3a* and Level *4a* are higher than the number of extracted *FA words* of Level *1b*, Level *2b*, Level *3b* and Level *4b*, which means that the number of accurate extracted *FA words* increase with passage retrieval and the effectiveness of using passage retrieval and our new approach. Moreover, the number of extracted *FA words* of Level 1 is higher than the number of extracted *FA words* of Level 2, Level 3 and Level 4. Also, the number of extracted *FA words* of Level 1 increases as concentration ratios decrease. Conversely, the number of extracted *FA words* on Level 2, Level 3 and Level 4 decreases with decrease in concentration ratios. Therefore, when the ratio is high, *FA words* of Level 2, Level 3 and Level 4 are

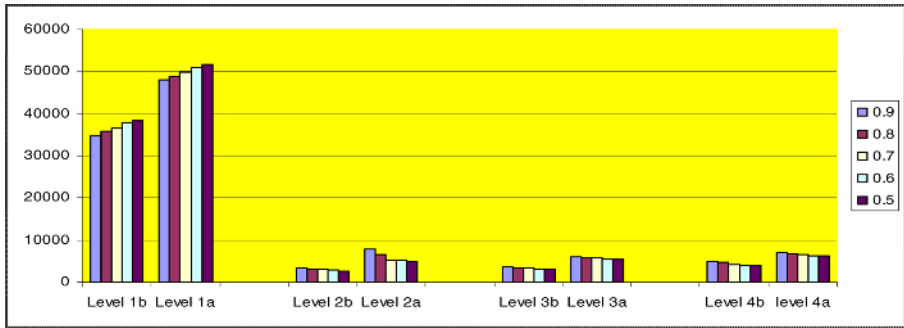


Fig. 1. Number of Extracted FA Words and Levels in each Document Data before and after using Passage Retrieval

extracted and when the ratio is low, FA words of Level 2, Level 3 and Level 4 become FA words of Level 1. This study uses only extracted FA words of Level 1. P and R of Level 2, Level 3 and Level 4 are difficult to measure because FA words on those levels exist in more than one field.

4.3 Sample of Extracting FAW Using Passage Retrieval

The Section and paragraph retrieval strategies outlined earlier are used to retrieve the most closely matching text sections and text paragraphs in answer to available query articles. For example suppose we have the following whole text as in Figure 2 (a):

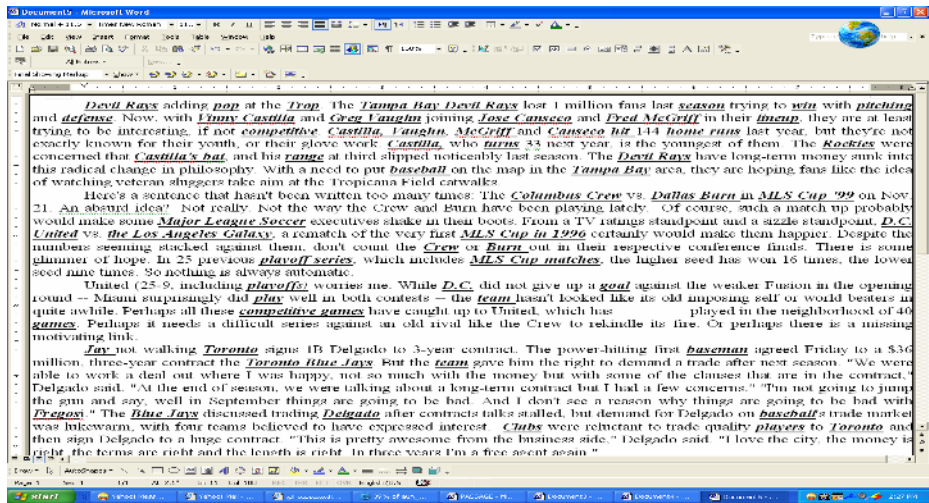


Fig. 2(a). Sample of the Whole Document

and Queries in this text are: *Fred McGriff, Home Run, Baseball and MLS Cup*, the most closely matching paragraphs for this query are the first and second paragraphs as follows as in Figure 2 (b):

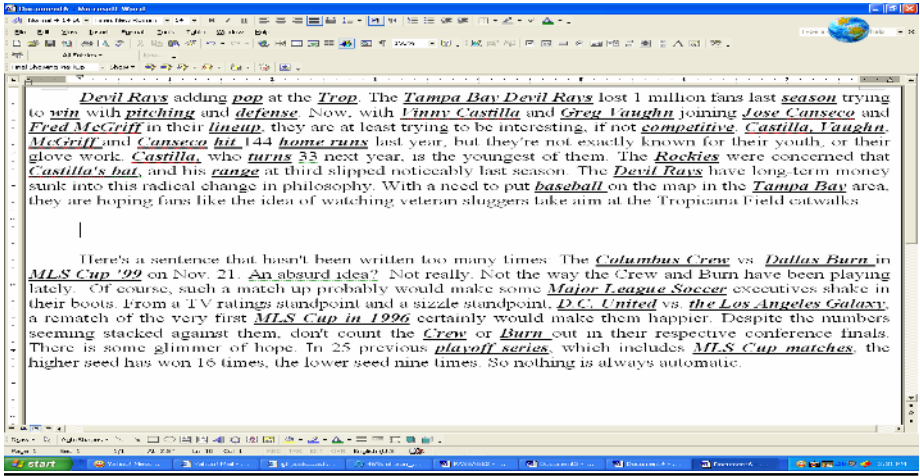


Fig. 2(b). Excerpt Document Related to the Query Figure 2, Example of extracting FA words using passage retrieval

From Figure 2 (b), we notice that the first and second paragraphs have many accurate extracted FA word related to the filed Baseball and the number of relevant FA word compared to total extracted word are higher than using the whole text which will make improvement in the system evaluation as in the following section.

4.4 System Evaluation

Precision and *Recall* used to evaluate this system [13] is defined as follows:

$$Precision (P) = \frac{\text{Number of FA word extracted by system}}{\text{Total Number of FA words extracted by system}}$$

$$Recall (R) = \frac{\text{Number of FA word extracted by system}}{\text{Total Number of FA words extracted manually}}$$

Figure 3 shows the effective results of using passage retrieval on Recall and Precision. From Figure 3, we notice that the *Recall* and *Precision (RecallA and PrecisionA)* after using passage retrieval are about 30% higher than before using passage retrieval (*RecallB and PrecisionB*) because when text excerpts (passage retrieval) are accessible leading to the retrieval of additional extracted relevant FA word with corresponding improvements in *Recall* and *Precision*. Moreover, the number of FA words is increasing with the decreasing of concentration ratio of threshold, so, the reliability is low with decreasing the concentration ratio of

threshold. Therefore, concentration ratio of threshold at 0.9 is effective enough for that propose than other threshold values. New approach using passage retrieval is also performing well effective in Recall and Precision than traditional method.

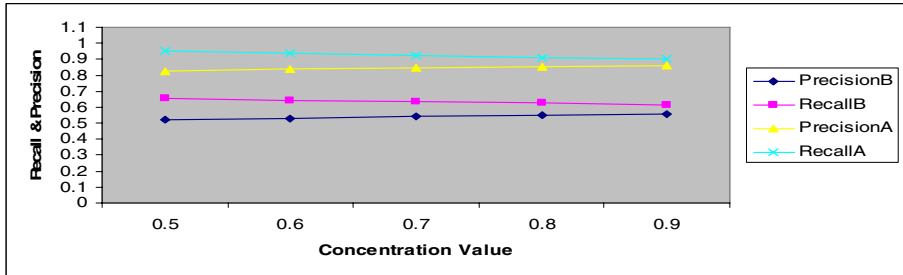


Fig. 3. Recall and Precision Before and After Using Passage Retrieval with Concentration Ratios

5 Conclusion

With increasing popularity of the Internet and tremendous amount of on-line text, automatic document classification is important for organizing huge amounts of data. Readers can know the subject of many document fields by reading only some specific *FA words*. Moreover, document fields can be decided efficiently if there are many *FA words* and their *frequencies* rate become high. In this study, effective methods are introduced for identifying relevant text excerpts in response to user interest statements as passage, and extracting the accurate *FA words* from that relevant text passages. Therefore, *Recall* and *Precision* using new approach improved by 30% than using whole texts and traditional methods. Moreover, Recall and Precision become high than traditional method with effective concentration ratio 0.9 for that propose. Future work could focus on using automatic building of *FA words* classification with attributes.

References

- [1] Aoe, J., Morita K., and Mochizuki. H. An Efficient Retrieval Algorithm of Collocate Information Using Tree Structure". *Transaction of the IPSJ*, 39 (9), (1989), 2563-2571.
- [2] Atlam, E.-S., Morita K., Fuketa, M. and Aoe, J. A New Method For Selecting English Compound Terms and its Knowledge Representation. *Information Processing & Management Journal*, 38, (2000), 807-821.
- [3] Atlam, E.-S., Fuketa M., Morita, K. and Aoe, J. Document Similarity measurement using Field association terms". *Information Processing & Management Journal*, Vol. 39, pp. 809-824, 2003.
- [4] E.-S. Atlam, G. Elmarhomy, M. Fuketa, K., Morita and Jun-ich Aoe. "Automatic building of new Field Association word candidates using search. *Information Processing & Management Journal*, 42(4), (2006), 951-962.

- [5] Breiman, L., Friedman, J.H., Olshen R. A. and Stone C.J.. *Classification and Regression Trees*. Chapman & Hall, (1984).
- [6] Callen J. P. Passage and level evidence in document retrieval. *In Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1994), 302-310,.
- [7] Dozawa T.. Innovative Multi Information Dictionary Imidas'99, Annual Series, Zueisha Publication Co., Japan (1999) (In Japanese).
- [8] Iwayama M.and Tokunaga T.. Probabilistic Passage Categorization and Its Application". *Journal of Natural language Processing*. 6 (3), (1999), 181-198.
- [9] Kaszkiel M. and Zobel J. Passage retrieval revised. *In Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in information Retrieval*, (1997), 178-185.
- [10] Kawabe K.and Matsumoto Y. Acquisition of normal lexical knowledge based on basic level category. *Information Processing Society of Japan, SIG note*, NL125-9, (1998), 87-92.
- [11] Melucii M. Passage Retrieval and a Probabilistic technique. *Information Processing and Management*. 34(1), (1998), 43-68.
- [12] Risvik K. M.and Michelsen R. Search Engines and Web Dynamics. *Computer Networks*, 39, (2002), 289-302.
- [13] Salton G., & McGill M.J. *Introduction of Modern Information Retrieval*. New York: McGraw-Hill, (1983).
- [14] Salton G., Allan J., Buckley C. Approaches to Passage Retrieval in Full Text Information Systems. *presented at the Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)* (1993),.
- [15] Salton G., *Automatic text Processing- The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley Publishing Company, Reading, MA, (1989).
- [16] Tsuji T., Nigazawa H., Okada M., and Aoe J. Early Field Recognition by Using Field Association Words. *In the Proceeding of the 18th International Conference on Computer Processing of Oriental Language*, 2, (1999), 301-304.
- [17] Tsuji T., Fuketa M., Morita K., and Aoe J. An Efficient Method of Determining Field Association Terms of Compound Words. *Journal of Natural Language Processing*. 7 (2), (2000), 3-26.

Building New Field Association Term Candidates Automatically by Search Engine

Masao Fuketa, El-Sayed Atlam, Elmarhomy Ghada, Kazuhiro Morita,
and Jun-ichi Aoe

Department of Information Science and Intelligent Systems
University of Tokushima, Tokushima 770-8506, Japan
atlam@ccr.tokushima-u.ac.jp

Abstract. With increasing popularity of the Internet and tremendous amount of on-line text, automatic document classification is important for organizing huge amounts of data. Readers can know the subject of many document fields by reading only some specific *Field Association (FA) words*. Document fields can be decided efficiently if there are many *FA words* and if the *frequency* rate is high. This paper proposes a method for automatically building new *FA words*. A *WWW* search engine is used to extract *FA word* candidates from document corpora. New *FA word* candidates in each field are automatically compared with previously determined *FA words*. Then new *FA words* are appended to an *FA word* dictionary. From the experiential results, our new system can automatically appended around 44% of new *FA words* to the existence *FA word* Dictionary. Moreover, the concentration ratio 0.9 is also effective for extracting relevant *FA words* that needed for the system design to build *FA words* automatically.

1 Introduction

With recent growth of the internet systems, many document retrieval systems have been developed and a variety of facilities are increasingly need in each area such as keyword retrieval [9], similar file retrieval [8], passage retrieval [6][9][10][12] and so on. In this paper, *field* means basic and common knowledge that can be used in human communication [11][15]. Generally, people know the subject *super-field* <SPORTS> or *sub-field* <Baseball> of a document based on specific words in that document. For example, people who encounter the words “*election*” or “*home run*” can recognize the document *super-field* <Politics> or *sub-field* <Baseball>.

Field Association (FA) word is a minimum *word* which can not be divided without losing semantic meaning. Based on specific FA word information, topics of documents (e.g. *Sports*, *Baseball*) can be recognized. For example, the word “*election*” can indicate *super-field* <POLITICS> or word the “*home run*” can indicate *sub-field* <Baseball>. Some keywords appear frequently in texts and relate strongly to text topics. Ohkubo et al. [13] proposed a method to estimate information that users might need for analyzing login data on a *WWW* search engine. By that method, word

groups connected with search words change due to time series variation; for example, new sport personalities and new words related to fashion can appear periodically (e.g. every year). The research documented in this paper focuses on finding new *FA words* automatically. To do so, a large quantity of document data for building *FA word* candidates is necessary. This paper presents a system for automatically building *FA words* by adding new *FA word* candidates which change over time to an *FA word* dictionary. This new system uses a *WWW* search engine to extract *FA word* candidates from document data for each field. These candidates are automatically compared with *FA words* already in a dictionary of *FA words*. Then, new *FA words* can be added to a dictionary of *FA words*.

2 Field Association Words, Field Trees and Levels

2.1 Document Field Trees

Document field trees represent ranked relationships connecting document fields [1][5]. A leaf node in a document field tree is a terminal document field and other nodes are medium document fields. In this study, a field tree based on Imidas'99 [7] contains 14 super-fields, 50 medium fields and 393 terminal fields (sub-fields). *Root names* are omitted when there is no conflict between super-field and terminal field and in such cases only terminal fields are indicated. *FA words* and paths are manually ranked. For example, path <SPORTS/Water Sports/ Swimming> describes *super-field* <SPORTS> having *sub-field* <Water Sports>, and document field <Swimming>.

2.2 *FA word* Levels

According to the scope of a field [2][3][4][15], fields can be classified as: (1) terminal fields, (2) medium fields or (3) multiple fields (i.e. set of terminal or medium fields). Moreover, *FA words* having different scopes and five levels are used to associate *FA words* to document fields.

3 Automatic Building of *FA Words*

3.1 System Outline

To retrieve document data, a word from an existing *FA word* dictionary is used as a reference keyword for a *WWW* search engine (see Fig. 1). Word frequencies are collected for every document field in the field tree. *FA words* are decided from words which are associated with specific fields. New *FA word* candidates are appended to the dictionary of *FA words* by comparison with *FA words* already listed in that dictionary.

To decide *FA word* candidates requires a large collection of documents and a *WWW* search engine quickly retrieves document data. Generally, a search engine text document has tables and images, but this system requires only document data.

Useless line feed and blanks are also cut from document data. Figure 4 shows some filtered document data.

4 Experiment Evaluation

4.1 Data

Word collection data are used for experimental data and there are 59,161 words and 4.52MB document data from a data set of 20 Newsgroup from CNN Web Site (1996-2003). Concentration ratio is changed from 0.5 ~ 0.9 to decide the levels of *FA word* candidates.

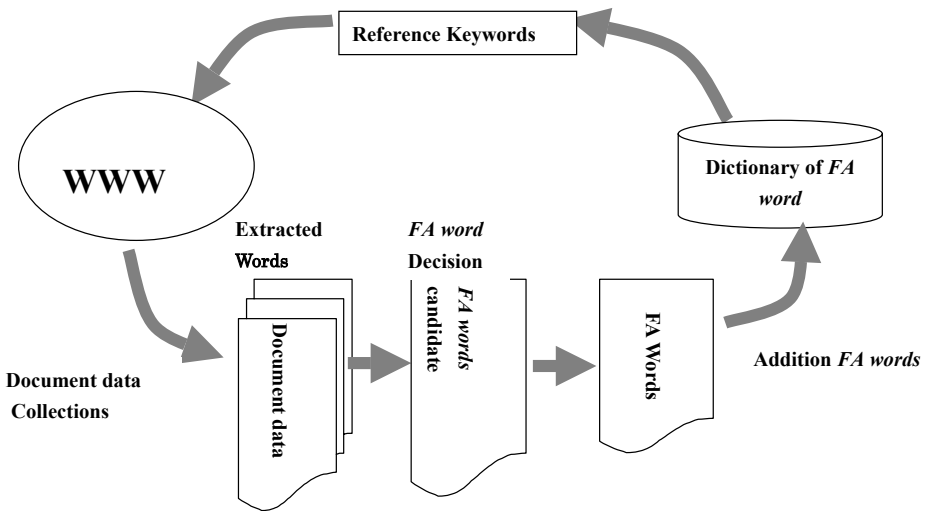


Figure 1 System Outline

Fig. 1. System Outline

4.2 System Evaluation

Precision and *Recall* used to evaluate this system are defined:

$$Precision (P) = \frac{\text{Number of Relevant } FA \text{ word extracted by system}}{\text{Total Number of } FA \text{ words extracted by system}}$$

$$Recall (R) = \frac{\text{Number of Relevant } FA \text{ word extracted by system}}{\text{Total Number of Relevant } FA \text{ words extracted Manually}}$$

If the number of extracted *FA words* equals the number of decided fields, the *FA words* are considered to be relevant.

4.3 Experiment Simulation Results

Figure 2 shows the number of extracted *FA words* on each level. Based on different concentration ratios, the number of extracted *FA words* of Level 1 is higher than the number of extracted *FA words* of Level 2, Level 3 and Level 4. Moreover, the number of extracted *FA words* of Level 1 increases as concentration ratios decrease.

Conversely, the number of extracted *FA words* on Level 2, Level 3 and Level 4 decreases with decrease in concentration ratios. Therefore, when the ratio is high, *FA words* of Level 2, Level 3 and Level 4 are extracted and when the ratio is low, *FA words* of Level 2, Level 3 and Level 4 become *FA words* of Level 1. This study uses only extracted *FA words* of Level 1. *P* and *R* of Level 2, Level 3 and Level 4 are difficult to measure because *FA words* on those levels in more than one field.

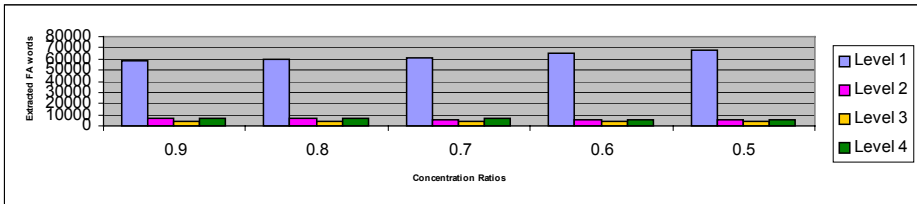


Fig. 2. Number of Extracted *FA Words* and Levels in each Field of Document

Table 1 shows that our systems selected 38,372 of *FA words* candidates from 59,161 words of document data and automatic appended from them 16,761 new *FA words* to existence *FA word* dictionary. From Table 1, we notice that the automatically appended number of *FA word* using Level1, Level2, Level3 and Level4 are: 16,761(%44). This means that our new system can automatically appended around 44% of new *FA words* to the existence *FA word* Dictionary.

Figure 3 (a) shows number of extracted *FA words* and number of relevant *FA words*. In Figure 3 (a), the number of extracted *FA words* increases with decrease of the concentration ratios and the number of extracted *FA words* which are relevant *FA words* increases too.

In Figure 4, changing the concentration ratio causes no significant change in *R*. *R* of concentration ratio 0.9 is 0.935 but *R* of concentration ratio 0.5 slightly increases to 0.956. This mean that, the number of extracted *FA words* increases at concentration ratio 0.9 but relevant *FA words* which could not be extracted at concentration ratio 0.9 could be extracted at concentration ratio 0.5.

P values decrease with decrease in concentration ratios. *P* at concentration ratio 0.9 is 0.744 but *P* at concentration ratio 0.5 slightly decreases to 0.68. So, change in *P* is bigger than change in *R*, meaning that the number of relevant *FA words* increases as the number of extracted *FA words* increases.

Table 1. Automatic Appended new *FA Words*

	Level 1	Level 2	Level 3	Level 4	Total Number
Total Number of Automatic Extracted <i>FA Words</i>	8,106 (% 48)	2,379 (%14)	2,488 (15%)	3,788 (%23)	16,761 (% 44)

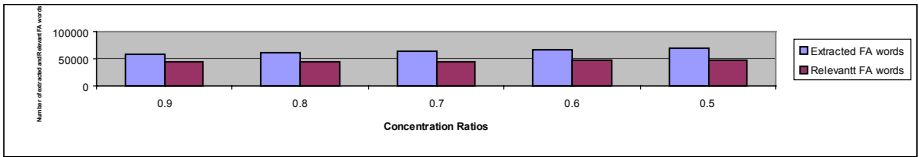


Fig. 3. Relevant *FA words* with Changing Concentration Ratios

Figure 4 shows the efficiency of document data set on *R* and *P*. *R* is almost constant as concentration ratios change, however *P* decreases with decrease of concentration ratios. Also, the number of relevant *FA words* increases with decrease of concentration ratios. Reliability is low with decrease in concentration ratios. This means that relevant *FA words* increases with decrease of concentration ratios but not accurate one. However, accurate relevant *FA words* are needed for the system design to build *FA words* automatically, so we chose the highest concentration ratio (0.9) which gives us the best simulation results and very effective for that propose.

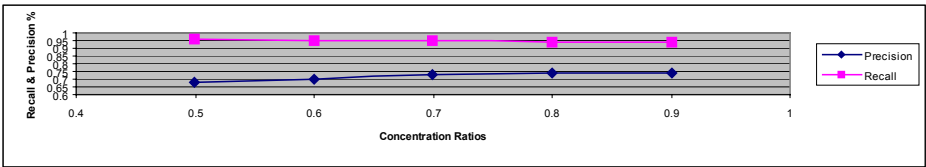


Fig. 4. Recall and Precision with Changing of Concentration Ratios

5 Conclusion

With increasing popularity of the Internet and tremendous amount of on-line text, automatic document classification is important for organizing huge amounts of data. Moreover, document fields can be decided efficiently if there are many *FA words* and if the *frequency* rate is high. A *WWW* search engine retrieved keywords from document corpora. Words were extracted from those document corpora to get *FA word* candidates. Furthermore, new *FA word* candidates were appended to an *FA word* dictionary by comparing these words automatically with *FA words* already in

that dictionary. From the experiential results, our new system can automatically appended around 44% of new *FA* words to the existence *FA word* Dictionary. Also, the concentration ratio 0.9 is also effective for that propose. Future work could focus on using automatic building of *FA words* with *Passage* retrieval.

References

- [1] Aoe, J., Morita, K. & Mochizuki, H. An Efficient Retrieval Algorithm of Collocate Information Using Tree Structure. *Transaction of The IPSJ*, 39 (9), (1989) 2563-2571.
- [2] Atlam, E.-S., Elmarhomy, G., Morita, K., Fuketa, M. & Aoe, J. A New Algorithm for Construction Specific Field Terms Using Co-occurrence Words Information, 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Wellington, New Zealand, Part 1, . (2004), 530-540.
- [3] Atlam, E.-S. & Aoe, J. A new algorithm for automatic extracting *FA word* candidates from document corpora. *The Interim Report of Tokushima University*, (2004), 25-27.
- [4] Atlam, E.-S., Morita, K., Fuketa, M. & Aoe, J. A New Method for Selecting English Compound Terms and its Knowledge Representation. *Information Processing & Management Journal*, 38 (6), (2002). 807-821.
- [5] Atlam, E.-S., Fuketa, M., Morita, K., & Aoe, J. Document Similarity measurement using Field association terms. *Information Processing & Management*, 39(6), (2003), 809-824.
- [6] Callen, J. P. Passage and level evidence in document retrieval In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (1994), 302-310.
- [7] Dozawa, T. Innovative Multi Information Dictionary Imidas'99, Annual Series, Zueisha Publication Co., Japan, (1999) (In Japanese).
- [8] Fuhr, N. Models for retrieval with probabilistic indexing, *Information Processing and Retrieval* 25 (1), (1989). 55-72.
- [9] Fukumoto, F., Suzuki, Y. Automatic Clustering of Articles using Dictionary definitions. In *proceeding of the 16th International Conference on Computational Linguistic (COLING'96)*, (1996). 406-411.
- [10] Iwayama, M. & Tokunaga, T. Probabilistic Passage Categorization and Its Application. *Journal of Natural language Processing*. 6 (3) ,(1999), 181-198.
- [11] Kawabe, K. & Matsumoto, Y. Acquisition of normal lexical knowledge based on basic level category. *Information Processing Society of Japan, SIG note*, NL125-9, (1998), 87-92.
- [12] Melucii, M.. Passage Retrieval and a Probabilistic technique". *Information Processing and Management*.34 (1), (1998), 43-68.
- [13] Ohkubo, M., Sugizaki, M., Inoue, T. & Tanaka, K. Extracting Information Demand by Analyzing a WWW Search Login". *Trans. of Information Processing Society of Japan*, 39(7), (1998). 2250-2258.
- [14] Salton, G., & McGill, M.J.. *Introduction of Modern Information Retrieval*. New York: McGraw-Hill. (1983).
- [15] Tsuji, T., Fuketa, M., Morita, K. & Aoe, J. An Efficient Method of Determining *FA* Terms of Compound Words. *Journal of Natural Language Processing*. 7(2), (2000), 3-26.

Efficient Distortion Reduction of Mixed Noise Filters by Neuro-fuzzy Processing

M. Emin Yüksel and Alper Baştürk

Digital Signal and Image Processing Lab., Dept. of Electrical and Electronics Eng.,
Erciyes University, Kayseri, 38039, Turkey

Abstract. A simple method for reducing undesirable distortion effects of mixed noise filters for digital images is presented. The method is based on a simple 2-input 1-output neuro-fuzzy network. The internal parameters of the neuro-fuzzy network are adaptively optimized by training. The training is easily accomplished by using simple artificial images generated on a computer. The method can be used with any type of mixed noise filters since its operation is completely independent of the filter. The proposed method is applied to two representative mixed noise filters from the literature under different noise conditions and image properties. Results indicate that the proposed method may efficiently be used with any type of mixed noise filters to effectively reduce their distortion effects.

1 Introduction

Contamination by noise is a frequently encountered problem in acquisition, transmission and processing of digital images. Preservation of image details while removing the noise is usually not possible during the restoration process, but both are essential for subsequent image processing tasks (eg. edge detection, image segmentation, object recognition, analysis and evaluation) [1].

In many applications, image noise is modeled with either an impulsive, a uniform, or a Gaussian distribution. In practice, however, this assumption is not adequate to model the effects of the noise in digital images. Hence, noise is modeled by mixing the impulse and the Gaussian noise with various density and variances in this paper.

Impulse noise is a special type of noise that can be caused by atmospheric disturbances, strong electromagnetic fields, transmission errors, etc. It is characterized by short, abrupt alterations in gray levels. Many noise removal methods can not sufficiently remove the impulse noise because they incorrectly assume the noise pixels as edges to be preserved. For this reason, a separate class of nonlinear filters have been developed specifically for the removal of impulse noise; many are extensions of the standard median filter [3,4], some use rank statistics [5,6,7] and others use adaptive neuro-fuzzy inference systems [8,9,10,11,12]. When applied to images corrupted with the Gaussian or the uniform noise, however, such filters are not effective.

Conventional communication system models usually assume that the noise in the transmission channels is Gaussian. Gaussian noise can be analytically described and has a characteristic bell shape. The mean filter shows a relative achievement against the Gaussian and the mixed noise. The Wiener filter alleviates some of the difficulties inherent in inverse filtering by attempting to model the error in the restored image through the use of statistical methods [2].

The adaptive neuro-fuzzy inference system (ANFIS) can be considered as a class of adaptive networks which are functionally equivalent to fuzzy inference systems (FISs). The main aim of the ANFIS is to optimize the parameters of the equivalent FIS by applying a training algorithm using input-output data sets.

In this work, performances of mixed noise filters (the mean and the Wiener filter in this paper) are improved by means of using an adaptive neuro-fuzzy inference system. Performances of these filters are tested with and without the adaptive neuro-fuzzy inference system. The results obtained show that the use of the proposed ANFIS based method causes considerable performance improvements. The rest of the paper is arranged as follows: the structure of the ANFIS is discussed in Section 2. The proposed method and its training are explained in Section 3. Filtering experiments and their results demonstrating the comparative performances of the proposed method are reported in section 4. Finally, discussion and conclusions are presented in Section 5.

2 Adaptive Neuro-fuzzy Inference System

The FIS is a popular computing framework based on the concepts of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning [13]. Among many FIS models, the Sugeno fuzzy model, which was first introduced in [14] and [15], is the most widely applied one for its high interpretability, computational efficiency and suitability for optimization using adaptive techniques. The Sugeno fuzzy model provides a systematic approach to generate fuzzy rules from a set of input output data pairs. The ANFIS is a FIS implemented in the framework of an adaptive fuzzy neural network. The ANFIS combines the learning capability of the artificial neural networks (ANNs) and ambiguity modeling ability of the FISs. Thus, the ANFIS combines the benefits of ANNs and FISs in a single model. Fast and accurate learning, excellent explanation facilities in the form of semantically meaningful fuzzy rules, the ability to accommodate both data and existing expert knowledge about the problem, and good generalization capability features have made neuro-fuzzy systems popular in the last few years [8, 9, 10, 11, 12, 16].

A simplified architecture of the ANFIS is shown in Fig.-1, in which a circle indicates a fixed node, whereas a square indicates an adaptive node. It was assumed that the ANFIS has two inputs x and y and one output f and implements a first-order Sugeno fuzzy model. For this model, a typical rule set with two fuzzy if-then rules can be expressed as :

$$\begin{aligned} \text{Rule1 : If } x \text{ is } A_1 \text{ and } y \text{ is } B_1 \text{ then } f_1 &= p_1x + q_1y + r_1 \\ \text{Rule2 : If } x \text{ is } A_2 \text{ and } y \text{ is } B_2 \text{ then } f_2 &= p_2x + q_2y + r_2 \end{aligned} \quad (1)$$

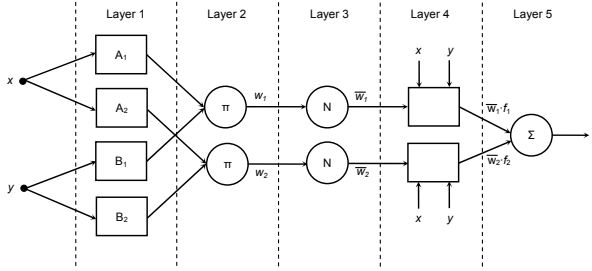


Fig. 1. Architecture of the ANFIS (For simplicity in expressions, it is assumed that there are two fuzzy if-then rules. In normal situation, rules contain all possible combinations of the inputs and the membership functions.)

where \$A_i\$ and \$B_i\$ are the fuzzy sets in the antecedent, and \$p_i, q_i\$ and \$r_i\$ are the design parameters that are determined during the training process. As in Fig.-1, the ANFIS consists of five layers:

Layer 1 : Each node in the first layer employs a node function given by:

$$O_i^1 = \begin{cases} \mu_{A_i}(x), & i = 1, 2 \\ \mu_{B_{i-2}}(y), & i = 3, 4 \end{cases} \tag{2}$$

where \$\mu_{A_i}(x)\$ and \$\mu_{B_{i-2}}(y)\$ can adopt any fuzzy membership function (MF). Here, \$O_i^n\$ denotes the \$i\$th output of the \$n\$th layer. In this paper, *the generalized bell* type MFs are used as follows:

$$\text{bell}(x; a, b, c) = \frac{1}{1 + \left| \frac{x - c}{a} \right|^{2b}} \tag{3}$$

where \$\{a_i, b_i, c_i\}\$ are the parameters that change the shapes of the MFs. Parameters in this layer are referred to as *the premise parameters*.

Layer 2 : Each node in this layer calculates the firing strength of a rule via multiplication:

$$O_i^2 = w_i = \mu_{A_i}(x) \mu_{B_i}(y), \text{ for } i = 1, 2 \tag{4}$$

Layer 3 : The \$i\$th node in this layer calculates the ratio of firing strength of the \$i\$th rule to the sum of firing strengths of all rules:

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2}, \text{ for } i = 1, 2 \tag{5}$$

where \$\bar{w}_i\$ are referred to as *the normalized firing strengths*.

Layer 4 : In this layer, each node has the following function:

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i), \text{ for } i = 1, 2 \tag{6}$$

where \$\bar{w}_i\$ is the output of the layer 3, and \$\{p_i, q_i, r_i\}\$ is the parameter set. Parameters in this layer are referred to as *the consequent parameters*.

Layer 5 : The single node in this layer computes the overall output as the summation of all incoming signals, which is expressed as:

$$O^5 = f = \sum_{i=1}^2 \bar{w}_i f_i = \frac{w_1 f_1 + w_2 f_2}{w_1 + w_2} \tag{7}$$

It is clear that the ANFIS has two sets of adjustable parameters, namely the premise and consequent parameters. During the learning process, the premise parameters in the layer 1 and the consequent parameters in the layer 4 are tuned until the desired response of the FIS is achieved. In this paper, the hybrid learning algorithm [17, 13], which combines the least square method (LSM) and the gradient descent algorithm, is used to rapidly train and adapt the FIS. When the premise parameter values of MFs are fixed, the output of the ANFIS can be written as a linear combination of the consequent parameters:

$$f = (\bar{w}_1 x) p_1 + (\bar{w}_1 y) q_1 + (\bar{w}_1) r_1 + (\bar{w}_2 x) p_2 + (\bar{w}_2 y) q_2 + (\bar{w}_2) r_2 \tag{8}$$

Hybrid learning algorithm has a two-step process. First, while holding the premise parameters fixed, the functional signals are propagated forward to layer 4, where the consequent parameters are optimized by the LSM. Then, the consequent parameters are held fixed while the error signals, the derivative of the error measure with respect to each node output, are propagated from the output end to the input end, and the premise parameters are updated by the gradient descent algorithm.

3 The Proposed Method and Its Training

The proposed method is shown in Fig.-2a. The fundamental element in the proposed method is a first-order Sugeno type ANFIS with two inputs and one output. Five generalized bell type MFs (in Eq. (3)) are used for both inputs of the ANFIS. The first input of the ANFIS is the output of a mixed noise filter (the output of the mean filter or the output of the Wiener filter), and the second input is the noisy image itself. The rule base of the ANFIS in the proposed method has 25 rules (5^2 , 5 denotes number of the MFs, and 2 denotes number

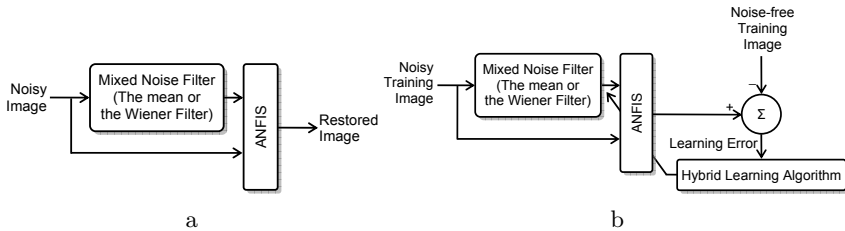


Fig. 2. a. The proposed method, b. Training setup of the proposed method

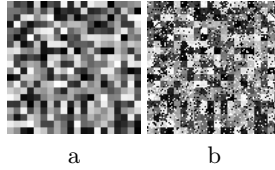


Fig. 3. The training images. a. Noise-free training image. b. Noisy training image. (Impulse noise has a density of 20%, Gaussian noise has a mean of 0 and a variance of 64).

of the inputs). Rules contain all possible combinations of the inputs and the membership functions.

Training setup for optimizing the internal parameters of the ANFIS is shown in Fig.-2b Throughout the training, internal parameters of the ANFIS are optimized by the hybrid learning algorithm [17, 13] so as to minimize the learning error which occurs due to difference between the output of the ANFIS and the ideal output. This process is called training. In this setup, inputs of the ANFIS are suitable input data set which is obtained from the noisy training image given in Fig.-3b and the one of the obtained outputs by the mean or the Wiener filter for the noisy training image. Besides, desired output of the ANFIS is noise-free training image shown in Fig.-3a. Both training images are 100-by-100 sized 8-bit gray level images. The noise-free training image is a simple synthetic image which is generated by a PC. Each box in this image is 5-by-5 sized and consists of 25 elements with same gray value chosen randomly from the interval [0-255]. The noisy training image is obtained by corrupting the noise-free training image with mixed noise including impulse noise having a density of 20% and Gaussian noise having a mean of 0 and a variance of 64.

4 Results

The proposed method is implemented and detailed simulations are performed with popular 256-by-256 sized 8-bit gray level test images in order to determine its performance. Experimental images are obtained by corrupting every test image by mixed noise which consists of impulse noise having a density from 1% to 50% and Gaussian noise having a mean of 0 and variances of from 1 to 255. Noisy test images are restored with the proposed method and the performance of the method is evaluated by means of *peak signal-noise ratio* (PSNR) criterion. Obtained PSNR performance values by the methods are given in Fig.-4, 5 and Table-1 (in the figures and table, IN and GN denote the impulse noise and the Gaussian noise, dens. and var. denote density and variance). Fig.-4 shows PSNR performance graphs of the methods. In Fig.-4a and b, while the variance of the Gaussian noise is constant (192), noise density of the impulse noise varies from 1% to 50%. Similarly, in Fig.-4c and d, while the variance of the Gaussian noise varies from 1 to 255, noise density of the impulse noise is constant (30%). Fig.-5 shows the visual performances of the methods for noisy *Goldhill*

Table 1. Obtained PSNR performances by the methods. BF denotes before, -A denotes normal filtering without ANFIS, +A denotes proposed filtering with ANFIS, a. For mean filter, b. For 3x3 window sized Wiener filter.

IN dens. → GN var. ↓	10%		20%		30%		40%		50%	
	BF	+A	BF	+A	BF	+A	BF	+A	BF	+A
51	15.598	19.397	12.476	18.129	10.803	17.195	9.487	16.278	8.545	15.554
102	15.373	19.334	12.484	18.167	10.749	17.174	9.527	16.305	8.547	15.489
153	15.249	19.291	12.428	18.177	10.726	17.150	9.558	16.347	8.575	15.570
204	15.156	19.282	12.304	18.050	10.714	17.155	9.451	16.201	8.507	15.474
255	14.976	19.213	12.282	18.068	10.595	17.036	9.468	16.234	8.520	15.468
a										
IN dens. → GN var. ↓	10%		20%		30%		40%		50%	
	BF	+A	BF	+A	BF	+A	BF	+A	BF	+A
51	15.386	18.528	12.530	17.228	10.777	16.327	9.522	15.497	8.557	14.899
102	15.408	18.607	12.512	17.229	10.730	16.262	9.512	15.532	8.538	14.887
153	15.176	18.496	12.457	17.235	10.699	16.289	9.522	15.611	8.526	14.869
204	15.063	18.442	12.342	17.155	10.698	16.280	9.425	15.477	8.498	14.808
255	15.087	18.539	12.332	17.112	10.671	16.247	9.426	15.430	8.526	14.932
b										

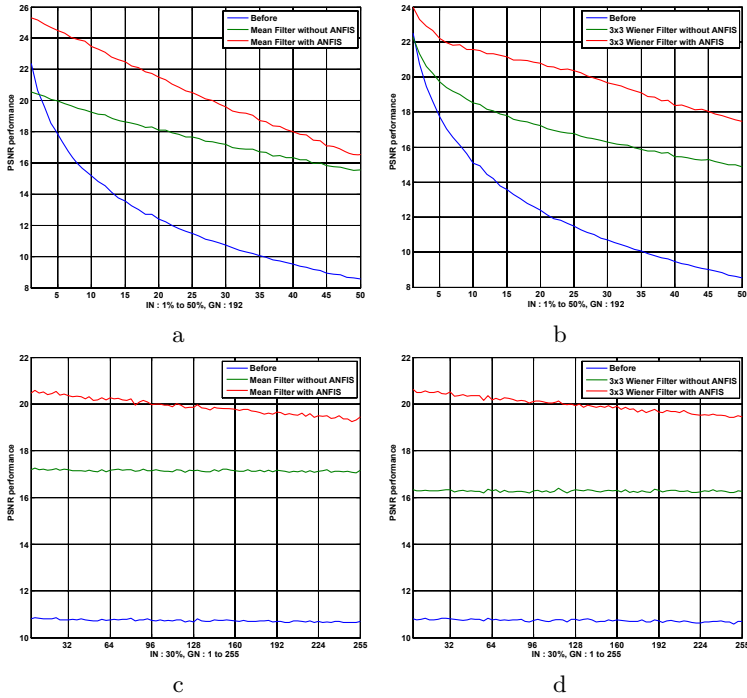


Fig. 4. Performance graphs of the methods for noise removal. Before:Blue, Without ANFIS:Green, With ANFIS:Red. In a and b GN have constant variance (192) and IN have varying noise density (1% to 50%). In c and d IN have constant noise density (30%) and GN have varying variance (1 to 255).

image. Table-1.a and 1.b show the simulation results obtained by the mean and the Wiener filters for both varying impulse noise density (from 10% to 50% by 10% steps) and varying Gaussian noise variance (from 51 to 255 by 51 steps), respectively.

5 Discussion and Conclusions

It can be seen from the figures and the table that the proposed method provides performance improvement and better restoration capability to the mean and the Wiener filters with a simple computational structure. As it can be seen from the Fig.-4 and Table-1, the PSNR values of the proposed method is better than the PSNR values of the mean and the Wiener filters for all impulse noise densities and all Gaussian noise variances. The effectiveness of the proposed method in processing the images can easily be evaluated by looking at the Fig.-5. Computational cost of the proposed method depends on the number of the fuzzy rules. In order to reduce the computational cost, fuzzy structure in the proposed filter use only 25 rules, and in order to simplify computational requirements,

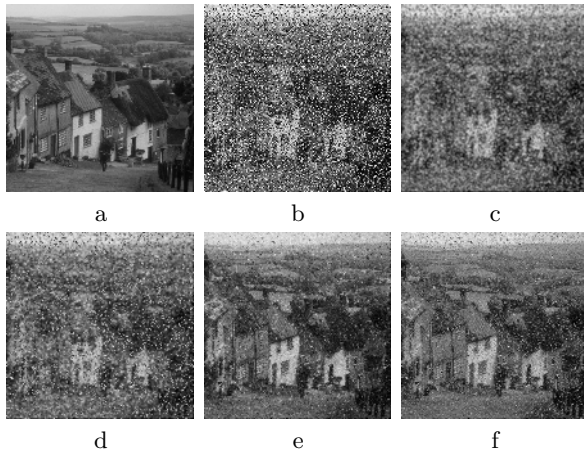


Fig. 5. Visual performance comparisons of the methods for *Goldhill* test image, a. Noise-free image, b. Noisy image (corrupted by mixed noise including impulse noise having a density of 30% and Gaussian noise having a mean of 0 and a variance of 128), c. Output of the mean filter, d. Output of the 3x3 window sized Wiener filter, e. Output of the proposed method using with the mean filter, f. Output of the proposed method using with the 3x3 window sized Wiener filter.

proposed method uses the simple bell and linear type MFs at the inputs and output of the ANFIS structure, respectively. It is obvious that the proposed method requires training the ANFIS, but it guarantees better restoration results.

It is concluded that the proposed method supplies more pleasing restoration results aspect of visual perception and it can be used with the mean and the Wiener filters as a simple but powerful tool for efficient removal of mixed noise in digital images.

References

- [1] L. Breveglieri, V. Piuri: Digital median filters. *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*. **31** (2002) 191–206.
- [2] S. E. Umbaugh: *Computer Vision and Image Processing*. Prentice-Hall International Inc, Upper Saddle River, NJ. (1998).
- [3] H. Lin, Jr. A. N. Willson: Median filters with adaptive length. *IEEE Trans. on Circuits Syst.* **35** (1998) 675–690.
- [4] T. Sun, Y. Neuvo: Detail-preserving median based filters in image processing. *Pattern Recognition Letters*. **15(4)** (1994) 341–347.
- [5] A. C. Bovik, T. S. Huang, D. C. Munson, Jr.: A generalization of median filtering using linear combinations of order statistics. *IEEE Trans. on Acoust., Speech, Signal Processing*. **ASSP-31(6)** (1983) 1342–1349.
- [6] R. C. Hardie, K. E. Barner: Rank conditioned rank selection filters for signal restoration. *IEEE Trans. on Image Processing*. **3** (1994) 192–206.
- [7] G. Pok, J. Liu, A. S. Nair: Selective removal of impulse noise based on homogeneity level information. *IEEE Trans. on Image Processing*. **12** (2003) 85–92.

- [8] M. E. Yüksel, A. Baştürk: Efficient removal of impulse noise from highly corrupted digital images by a simple neuro-fuzzy operator. *Int. J. Electron. Commun.* **57(3)** (2003) 214–219.
- [9] M. E. Yüksel, E. Beşdok: A simple neuro-fuzzy impulse detector for efficient blur reduction of impulse noise removal operators for digital images. *IEEE Transactions on Fuzzy Systems.* **12(6)** (2004) 854–865.
- [10] M. E. Yüksel, A. Baştürk, E. Beşdok: Detail-preserving restoration of impulse noise corrupted images by a switching median filter guided by a simple neuro-fuzzy network. *EURASIP Journal on Applied Signal Processing.* **2004(16)** (2004) 2451–2461.
- [11] M. E. Yüksel, A. Baştürk: A simple generalized neuro-fuzzy operator for efficient removal of impulse noise from highly corrupted digital images. *Int. J. Electron. Commun.* **59(1)** (2005) 1–7.
- [12] F. Russo: Noise removal from image data using recursive neurofuzzy filters. *IEEE Trans. on Instrumentation Measurement.* **49(2)** (2000) 307–314.
- [13] J.-S. R. Jang, C. T. Sun, E. Mizutani: *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence.* Prentice-Hall, Upper Saddle River, NJ. (1997).
- [14] T. Takagi, M. Sugeno: Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics.* **15** (1985) 116–132.
- [15] M. Sugeno, G. T. Kang: Structure identification of fuzzy model. *Fuzzy Sets and Systems.* **28** (1988) 15–33.
- [16] J. Kim, N. Kasabov: HyFIS: Adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems. *Neural Networks.* **12** (1999) 1301–1319.
- [17] J.-S. R. Jang: ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. on Systems, Man, and Cybernetics.* **23** (1993) 665–685.

Texture Segmentation with Local Fuzzy Patterns and Neuro-fuzzy Decision Support

L. Caponetti, C. Castiello, A.M. Fanelli, and P. Górecki

Dipartimento di Informatica, Università degli Studi di Bari,
Via E. Orabona, 4 - 70126 Bari Italy
{laura, castiello, fanelli, przemislaw}@di.uniba.it

Abstract. In this paper we propose a split and merge texture segmentation method. The presented approach is characterised by the introduction of a novel operator, the Local Fuzzy Pattern for texture discrimination, and the employment of a neuro-fuzzy decision support strategy, which supervises the overall split and merge procedure. The effectiveness of the proposed approach is evaluated on a set of artificial and natural texture images.

1 Introduction

Texture analysis plays an important role in image interpretation and refers to characterisation of images based on local spatial variation of pixel intensities [6]. A texture definition was not definitively assessed, other than some kind of “fuzzy” determinations, like the one suggested by Sklansky [7]: “a region in a image has constant texture if a set of local statistics or other local properties are constant, slowly varying, or approximately periodic”. One of the goals of texture analysis consists in texture segmentation, which is the task of dividing an image into regions containing the same textural properties. Consistent research efforts have been recently devoted to investigate feature extraction methods which could stand as a preliminary stage for segmentation tasks. These methods fall into the following general categories [10]: statistical, geometrical, model-based and signal processing methods. Based on the extracted features, a number of heuristics may be devised to produce region segmentation: one approach is to apply soft computing and clustering techniques [8,3,4].

In this paper, we propose a texture segmentation method based on the employment of the Local Fuzzy Pattern operator and a strategy of neuro-fuzzy decision support. In particular, our approach for texture segmentation aims at dividing the original image into smaller regions, in order to gather a set of high order image statistics from each of them. Afterwards, homogeneity among regions is detected, employing feature similarity measures. This kind of analysis is useful to put into action an agglomerative merging process, intended to produce the final texture segmentation. The peculiarity of our approach is basically twofold. Firstly, we introduce a particular operator for texture discrimination, the Local Fuzzy Pattern, which stands as a fuzzified extension of the Local Binary Pattern operator. Moreover, the overall split-and-merge process, at the basis of the segmentation procedure, is supervised by a neuro-fuzzy decision support strategy.

This allows to overcome the threshold-based dimension representing the most common condition faced by researchers in this field. The effectiveness of the proposed approach is evaluated on a set of artificial and natural texture images obtained from the Brodatz and VisTex databases [1,11].

The paper is organised as follows. In the next session we detail the texture feature analysis process, based on the employment of the Local Fuzzy Pattern operator. In section 3 we describe the segmentation process, performed by the successive application of the image decomposition and region merging steps. Some details concerning the adopted neuro-fuzzy strategy will be also given. Section 4 presents some experimental results and closes the paper with some conclusive remarks.

2 Texture Analysis

In this section we introduce a novel operator, the Local Fuzzy Pattern (*LFP*), for texture discrimination. It represents a fuzzified version of the Local Binary Pattern (*LBP*) operator, originally proposed in [5], which is reviewed in the following. Moreover, a similarity measure to determine the homogeneity degree between two given texture regions is illustrated.

2.1 The Local Binary Pattern Operator

The *LBP* operator was developed as a gray-scale invariant pattern measure based on the thresholding of gray level differences. In this approach, the 3×3 neighbourhood of each pixel in a the image is considered. The gray level values x_i ($i = 1, \dots, 8$) of pixels around the central pixel x_c are thresholded by the value of the centre pixel x_c . The obtained values b_i are given by:

$$b_i = \begin{cases} 1 & \text{if } x_i - x_c \geq 0, \\ 0 & \text{otherwise.} \end{cases} \quad i = 1, \dots, 8. \quad (1)$$

The concatenation $B = [b_1 \dots b_i \dots b_8]$ of the obtained digits corresponds to one of the possible 256 patterns (*LBP* binary numbers) employed to describe the centre pixel. In fig. 1 an example is shown: the *LBP* binary number is obtained by concatenating the b_i digits which are read row by row. The corresponding *LBP* number is simply derived by converting B into a decimal number. This can be done by firstly multiplying the neighbourhood window by the weight window, as shown in fig. 1, and then adding the obtained values. The *LBP* operator describes spatial structure of a local texture and it is invariant to any monotonic gray-scale transformation, but it does not address the contrast of the texture. Therefore, a contrast measure C is additionally calculated as the difference between the average gray level of those pixels having $b_i = 1$ and those having $b_i = 0$ (see fig. 1).

The *LBP/C* distribution may be approximated by a two-dimensional histogram of size 256×8 , where 8 is the number of bins for contrast measure. The number of bin is chosen as a trade-off between the discriminative power and

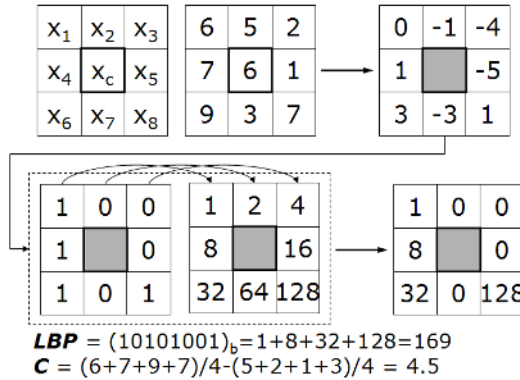


Fig. 1. Computation of Local Binary Pattern and contrast measures

the stability of the texture description [5]. To determine the texture homogeneity between two different regions, the corresponding *LBP/C* distributions are compared using a non-parametric pseudo-metric, called *G*-statistic [9], which is applied over a pair of histograms.

2.2 The Local Fuzzy Pattern Operator

Whereas signed gray level differences measure both the spatial organisation (pattern) and the contrast (amount) of local image texture, *LBP* information focuses only on spatial structure and discards contrast. That is, the *LBP* number embeds a crisp comparison between pixels, losing the information related to the actual amount of difference between pixel values. Moreover, it can be easily argued that the *LBP* operator is also sensitive to noise. In fact, if we consider a uniform neighbourhood with gray level values similar to the one of the centre pixel, the *LBP* distribution will be affected even by a minimal addition of random noise. To overcome these drawbacks, we introduce a Local Fuzzy Pattern (*LFP*) operator. This is an extension of the *LBP* operator by the introduction of fuzzy, rather than crisp, thresholding in the comparison of two gray levels. The gray level difference d_i between x_i and x_c , is characterized using two fuzzy sets $\{N, P\}$, representing negative and positive differences.

The piecewise linear membership functions of the fuzzy sets are symmetrical and the amount of fuzziness of *N* and *P* is determined by the parameter e . The functions are shown in fig. 2. Indicating by n_i and p_i the membership degree values for a given d_i , each pixel x_i is represented as a pair (n_i, p_i) . The *LBP* operator associates each pixel centre with only one binary number *B*, namely to a single *LBP* number. Instead, the *LFP* operator associates to each pixel centre the entire set of 256 numbers, with different fulfilment degrees. If we consider the k -th pattern B_k , composed of digits b_i^k , the fulfilment degree f_k of the pattern is calculated as follows:

$$f_k = \prod_{i \in I^k} p_i \prod_{j \in J^k} n_j \quad k = 1, \dots, 256. \tag{2}$$

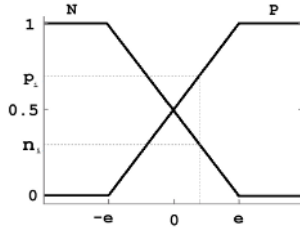


Fig. 2. Membership functions of the fuzzy sets $\{N, P\}$

where $I^k = \{i = 1, \dots, 8 | b_i^k = 1\}$ and $J^k = \{j = 1, \dots, 8 | b_j^k = 0\}$. It should be noted that $\sum f_k = 1$. Similarly, the fuzzy contrast measure FC is:

$$FC = \sum_{i=1}^8 x_i p_i / \sum_{i=1}^8 p_i - \sum_{i=1}^8 x_i n_i / \sum_{i=1}^8 n_i. \tag{3}$$

As an example, table 1 reports the membership function values for the fuzzy sets P and N (with $e = 2$), evaluated for the pixels configuration illustrated in fig. 1 and the corresponding list of fulfilment degrees evaluated for the LBP numbers.

Finally, each computed f_k is accumulated in the k -th bin of a histogram to obtain the LFP/FC texture distribution and the G -statistic test is used to quantitatively compare a pair of histograms.

3 Segmentation Algorithm

To obtain the segmentation of a textured image, two subsequent steps are performed. Initially, the image is recursively decomposed into square blocks of different size, until all regions satisfy homogeneity criterion. Successively, similar adjacent regions are iteratively merged to produce the final segmented image. In order to establish the stop criterion both for the region decomposition and for the merging iterative procedures, a neuro-fuzzy strategy is applied. In this way, the limitations of classical threshold-based approaches can be overcome.

In particular, the adopted neuro-fuzzy strategy aims at generating a base of fuzzy rules in the IF-THEN form. These rules serve to drive both the splitting and the merging processes on the basis of the analysis of the LFP/FC distributions obtained for each region. Actually, in the case of the splitting process the antecedent part of each rule expresses a condition on the configuration of distribution distances (in the way they are evaluated by the G -statistic). In the case of the merging process, the antecedent part of each rule expresses a condition on the increasing rate of the distance values, registered at every iteration. In both the splitting and merging cases, the consequent part of each rule provides the final decision about the prosecution or termination of the iterative process. The fuzzy rule base is obtained via the connectionist learning of a particular neural

Table 1. Evaluation of membership functions for pixels shown in figure 1 and the corresponding list of fulfilment degrees

indexes	values							
i	1	2	3	4	5	6	7	8
x_i	6	5	2	7	1	9	3	7
d_i	0	-1	-4	1	-5	3	-3	-1
p_i	0.5	0.25	0	0.75	0	1	0	0.75
n_i	0.5	0.75	1	0.25	1	0	1	0.25
binary number	fulfilment degree							
$B_1 = 00000000$	$f_1 = 1 * (0.5 * 0.75 * 1 * 0.25 * 1 * 0 * 1 * 0.25) = 0$							
...	...							
$B_{213} = 11010100$	$f_{213} = (0.5 * 0.25 * 0.75 * 1) * (1 * 1 * 1 * 0.25) = 0.023438$							
$B_{214} = 11010101$	$f_{214} = (0.5 * 0.25 * 0.75 * 1 * 0.75) * (1 * 1 * 1) = 0.070313$							
...	...							
$B_{256} = 11111111$	$f_{256} = (0.5 * 0.25 * 0 * 0.75 * 0 * 1 * 0 * 0.75) * 1 = 0$							

network (the neuro-fuzzy network), whose structure and parameters reflect those embedded in the fuzzy inference system. The neuro-fuzzy network is trained on the basis of input data derived from a set of pre-labelled texture images. The learning scheme of the network is articulated in two successive steps, intended to firstly initialise a knowledge structure and then to refine the obtained fuzzy rule base. We do not provide the mathematical details concerning the formalisation of the learning algorithms (whose discussion does not concern the scope of this article), addressing the reader to some other previous works of ours [2].

In the following, we are going to detail the adopted quadtree decomposition and agglomerative merging processes.

3.1 Quadtree Decomposition

Quadtree decomposition is a technique involving subdivision of an image into homogeneous rectangular blocks [10]. In our investigation, we aim at expressing region homogeneity in terms of texture similarity. Initially, the image is segmented into a number of large square blocks of size S_{max} . Then, to evaluate texture homogeneity inside the blocks, each of them is divided into four sub-blocks and the corresponding *LFP/FC* distributions are calculated. On the basis of the *G*-statistic comparison results, it could be possible to determine if the four sub-blocks constitute a single texture region or it is necessary to replicate the quadtree decomposition procedure over each of them. In order to avoid the selection of threshold values, the uniformity test is left to the assessment provided by a neuro-fuzzy strategy. In particular, the fuzzy rule base originated through the learning of the neuro-fuzzy network determines the stopping criterion for the decomposition process. The input of the network (corresponding to the antecedent part of the rules) is represented by configurations of six pairwise distances between the *LFP/FC* distributions of the four sub-blocks obtained during the quadtree decomposition process. The output of the network (corresponding to the consequent part of the

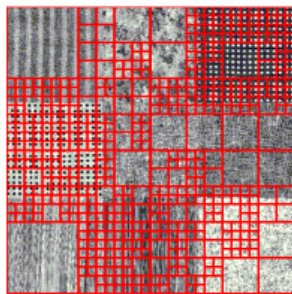


Fig. 3. An illustration of the quadtree decomposition performed over a sample image

rules) is simply constituted by the final decision concerning the termination or prosecution of the decomposition process. If a decision is made to split the block, each sub-block is recursively tested for homogeneity until the minimum block size S_{min} is reached (S_{min} should contain enough pixels for the LFP/FC distribution to be meaningful and stable). Details concerning the choice for the S_{max} and S_{min} values are provided in the final section of the paper. Fig. 3 shows an example of quadtree decomposition.

3.2 Agglomerative Merging

After dividing the image into regions with uniform texture, the merging step is executed to join similar adjacent regions. During each iteration of the algorithm, two adjacent regions with maximum similarity are merged into one region and their LFP/FC distributions are summed together.

The merging process is regulated in its iteration by the evaluation provided by the neuro-fuzzy strategy. Analogously to the previous decomposition process, a set of fuzzy rule is properly derived through neural learning to compile a base of knowledge useful for the decision-making processes. In this case, the neuro-fuzzy network is trained to automatically determine whether the agglomerative merging has to be stopped or it is necessary to perform the next iteration of the procedure. This kind of evaluation is produced by considering the loss of region homogeneity occurring after each step of the merging process. Again, we must refer to the distances between the LFP/FC distributions (evaluated for adjacent regions). In particular, the increasing rate of the distance values, registered at every iteration, represents the input of the network (correspondingly, the antecedent part of the fuzzy rules). The network output (the consequent part of the fuzzy rule) consists in the final decision concerning the termination or prosecution of the merging process.

4 Experimental Results and Conclusive Remarks

In order to assess the feasibility of the proposed approach, we are going to illustrate some experimental results. In particular, natural scene images from the

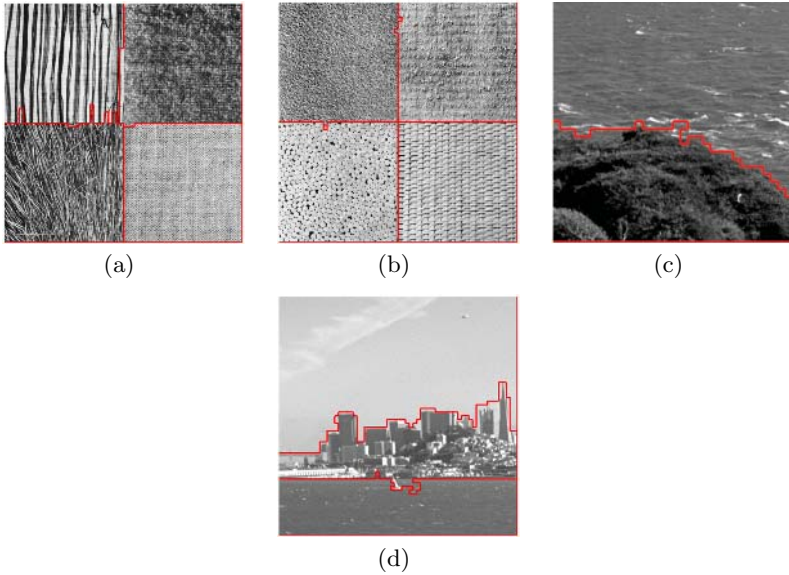


Fig. 4. Sample texture images used in experiments of composite Brodatz textures (a,b) and Vistex natural scenes (c,d)

VisTex dataset [11] and the mosaic images constructed by combining different Brodatz primitive textures [1] have been employed to perform the testing. The size of all the images is 512×512 pixels. For the experimental sessions, the fuzzification parameter ϵ was set equal to 10 and the quadtree decomposition was performed with $S_{max} = 128$ and $S_{min} = 8$. As considering the learning of the neuro-fuzzy network, the used training set was obtained by creating mosaics of different Brodatz images and their corresponding ground truth segmentation.

The segmentation results obtained using the *LFP* operator have been compared with those deriving from the application of the original *LBP* operator. Similar performances have been registered when employing both the operators for the segmentation of the images in the employed dataset. Sample results of the segmentation obtained with the proposed approach are presented in fig. 4. It can be noticed from fig. 4(a)-(b) that synthetic texture images are correctly segmented. Also, the segmentation of natural scenes in fig. 4(c)-(d) agrees with human perception. By definition, the *LFP* operator is more robust to noise than *LBP*. In order to assess this peculiarity, we performed an experimental session over images distorted by additive Gaussian white noise with zero mean and a variance equal to 0.01. As an example, in fig. 5 (where noisy images are considered) it is shown that the application of the *LFP* operator produced better results during the segmentation of natural images, with a more accurate preservation of texture boundaries.

To conclude, the experimental session demonstrated that the novel *LFP* operator is capable of segmenting the image into uniform texture regions, while

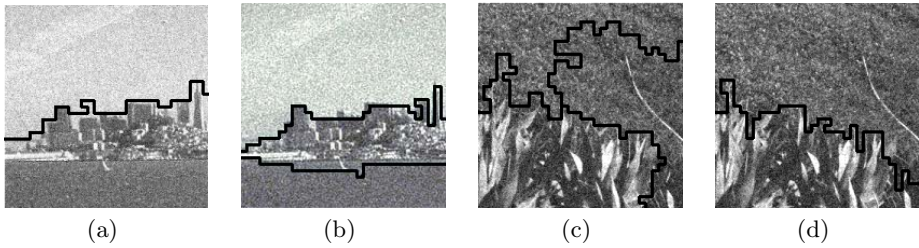


Fig. 5. Segmentation of noisy natural images using the *LBP* (a,c) and the *LFP* (b,d) operators

preserving fine texture boundaries, performing better than *LBP* for images distorted by noise. As concerning future work, we plan to complete the segmentation method adding a pixelwise classification step, thus improving localisation of texture boundaries.

References

1. Brodatz, P.: Textures: A photographic Album for Artists and Designers (1966)
2. Castellano, G., Castiello, C., Fanelli, A.M., Mencar, C.: Knowledge discovering by a neuro-fuzzy modeling framework. *Fuzzy Sets Syst.* **149**(1) (2005) 187–207
3. Goltsev, A.: An Assembly Neural Network for Texture Segmentation. *Neural Networks* **9**(4) (1996) 643–653
4. Karmakar, G., Dooley, L., Murshed, M.: Fuzzy rule for image segmentation incorporating texture features. *Proc. Int. Conf. Image Processing* **1** (2002) 797–800
5. Ojala, T., Pietikäinen, M.: Unsupervised texture segmentation using feature distributions. *Pattern Recognition* **32** (1999) 477–486
6. Pietikäinen, M.K.: *Texture Analysis in Machine Vision*. World Scientific (2000)
7. Sklansky, J.: Image segmentation and feature extraction. *IEEE Trans. Syst. Man Cybern.* **8** (1978) 237–247
8. Si, W.L., He, X.: Textured image segmentation using autoregressive model and artificial neural network. *Pattern Recognition* **28**(12) (1995) 1807–1817
9. Sokal, R.R., Rohlf, F.J.: *Introduction to Biostatistics*. Freeman and Co (1987)
10. Tuceryan, M., Jain, A.K.: *Texture Analysis*. In: *The Handbook of Pattern Recognition and Computer Vision*, World Scientific, (1998)
11. MIT Vision Texture (VisTex) database, <http://www-white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html>

Lineal Image Compression Based on Lukasiewicz's Operators*

N.M. Hussein Hassan and A. Barriga

Instituto de Microelectrónica de Sevilla, CNM-CSIC
Avda. Reina Mercedes s/n,edif CICA, E-41012 - Sevilla, Spain
{nashaat, barriga}@imse.cnm.es

Abstract. We proposed the use of Lukasiewicz's operators for lineal image compression. These operators have been applied to the approximation of piecewise linear functions. In this sense we showed two basic piecewise lineal static image compression techniques in which the use of this operators lets to reduce hardware resources.

1 Introduction

The development of the theoretical concepts of multi-valued logics began in the decade of the 20s by J. Lukasiewicz, who established the generalization of the classic logic to multi-valued logic. Later, at the end of the 1950s, C.C. Chang formalized the multi-valued algebra based on Lukasiewicz's logic. In this paper we are interested in the representation of n -variables Lukasiewicz's formulas by means of piecewise linear functions [1]. In this sense the concept of rational McNaughton's function defined as a continuous piecewise linear function in which every piece has rational coefficients constitutes universal approximators. We are going to show a hardware realization of these systems and we will apply them to functions interpolation. The application of linear interpolation consists of a special case of the approximation for splines of degree 1.

The linear interpolation has application in different areas of static image processing and video. A usual example corresponds to the interpolation that is performed on the image for a zooming operation. In this case a usual and simply technique consists to replicate a pixel. For the increase of precision a solution consists of realizing the linear interpolation between neighboring pixels. In our case we are going to apply the interpolation to static images compression, although the technique that we propose can spread directly to another type of applications that need linear interpolation.

Hardware realization of the basic Lukasiewicz's operators will be shown in section 2. We will illustrate an example of piecewise linear approximation using the previously design operators in section 3. In section 4 we will apply piecewise linear approximation to apply for a simple image compression technique. Finally, in section 5, another

* This work was partially supported by projects TEC2005-04359/MIC from the Spanish Ministry of Education and Science as well as TIC2006-635 from the Andalusian regional Government.

piecewise-based image compression algorithm will be discussing as function approximation usign Lukasiewicz's operators.

2 Basic Operators

A multi-valued Lukasiewicz algebra is a structure $A = (A, \oplus, \otimes, \neg, 0, 1)$ that satisfies the following properties:

- $x \oplus (y \oplus z) = (x \oplus y) \oplus z$
- $x \oplus y = y \oplus x$
- $x \oplus 0 = x$
- $x \oplus 1 = 1$
- $\neg 0 = 1 \quad y \quad \neg 1 = 0$
- $\neg(\neg x \oplus \neg y) = x \otimes y$
- $x \oplus (\neg x \otimes y) = y \oplus (\neg y \otimes x)$

The multi-valued algebra coincides with the Boolean algebra if idempotency happens ($x \oplus x = x$). The operators definition are:

- $x \oplus y = \min(1, x + y)$ (1)
- $x \otimes y = \max(0, x + y - 1)$ (2)
- $\neg x = 1 - x$ (3)

On the other hand, the following connectives are useful:

- $x \vee y = \max(x, y) = (x \otimes \neg y) \oplus y$ (4)
- $x \wedge y = \min(x, y) = (x \oplus \neg y) \otimes y$ (5)

A continuous piecewise linear function, in which every piece has integer coefficients, is associated with a Lukasiewicz's formulae [2].

An $f : [0,1]^n \rightarrow [0,1]$ function is an n variables McNaughton's function if the following conditions are fulfilled:

- f is a continuous function
- f is piecewise linear, this is, polynomials exist p_1, \dots, p_k such that for each $p_i(x) = a_i \bullet x + b_i$ for each $x \in [0,1]^n$ and $i \in \{1, \dots, k\}$ so that $f(x) = p_j(x)$.
- For each $i \in \{1, \dots, k\}$ coefficient a_i, b_i are integer.

The classes of functions determined by Lukasiewicz's logic formulae coincide with the class of McNaughton's functions [1]. An important and useful definition is that of rational McNaughton's function. A rational McNaughton's function is defined as a piecewise linear function in which every piece has rational coefficients. This definition is important because it means that any rational Lukasiewicz's formulae can be implemented as a rational McNaughton's function. Therefore these functions constitute universal approximators.

The operators that we are going to consider are defined in equations (1-5). These operators constitute the basic elements for piecewise linear function circuits as we will see in next sections. The design of the basic operators circuits can be realized according to two strategies: based on neural networks circuits and based on combinational logic circuits. In the first implementation strategy we will follow [3] where is demonstrated that it is possible to represent the neural network corresponding to a combination of propositions of Lukasiewicz's calculus using an activation function $\psi(x) = 1 \wedge (x \vee 0)$. The second design strategy is based on combinational logic circuits [4]. Figure 1 shows the design of min, bounded-product and bounded-sum operator respectively.

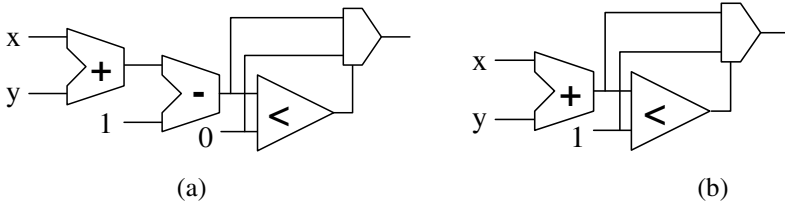


Fig. 1. a) Bounded-product circuit. b) Bounded-sum circuit

3 Function Approximation

In agreement with the definition of rational McNaughton's function it becomes that the use of Lukasiewicz's operators allows us to realize the approximation of piecewise linear functions. An approximation to the above mentioned problem is established in [5] where it is stated that a piecewise linear functions $f(x)$, and a set of its distinct components $(\{g_1, \dots, g_n\})$, can be described by means of the expression $f(x) = \bigvee_{j \in J} \bigwedge_{i \in S_j} g_i(x) \quad \forall x$, where the elements of the family $\{S_j\}_{j \in J}$ are incomparable (with respect to \subseteq) subsets of $\{1, \dots, n\}$. In order to illustrate the piecewise linear function approximation by means of Lukasiewicz's operators we are going to consider an example with one input variable as shown in Figure 2.

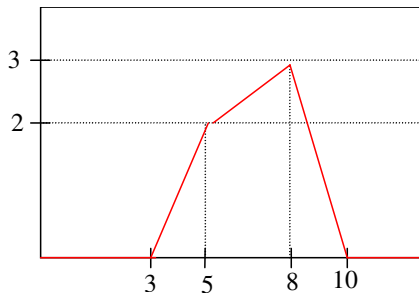


Fig. 2. Example of piecewise linear function

A direct implementation of this function would use straight line segments given by $g_i = m_i x + n_i \quad (i=1, \dots, 5)$. Another type of realization can be inferred applying Lukasiewicz's algebra. In this case, it is possible to rewrite the function using Lukasiewicz's operators. We can verify that the function $g_i = m_i x + n_i$ can be expressed as $g_i = (1 + n_i) \otimes m_i x$. Based on this expression the previous function can be written as $f_1(x) = (g_2 \wedge g_3 \wedge g_4) \vee 0$ where:

$$f_1(x) = \begin{cases} g_1 = 0 & x < 3 \\ g_2 = -2 \otimes x & 3 < x < 5 \\ g_3 = \frac{4}{3} \otimes \frac{1}{3} x & 5 < x < 8 \\ g_4 = 16 \otimes -\frac{3}{2} x & 8 < x < 10 \\ g_5 = 0 & x > 10 \end{cases}$$

We have implement function $f_1(x)$ on a Xilinx's FPGA considering an 8-bit precision for input and output variables. The four most significant bits correspond to the integer part and the four least significant bits to the decimal one. Comparing this implementation with another one using the neural network based Lukasiewicz operators given in [2] we have reduced the area from 162 slices to 25 slices and the delay from 54.9 nsec. to 21.4 nsec.

4 Basic Algorithm for Image Compression

A straightforward application of McNaughton's functions is related to image compression. We will show some example of such kind of applications. The description of a basic image compression algorithm is shown in figure 3 for a color 2D image. The information is stored in a three dimensions array (*img*). Function *Lprod* is Lukasiewicz's bounded-product. Variable *d* is the compression index of the image. The algorithm is based on a linear interpolation of the pixels of a column.

```

for k=1:3
    for i=1:length(x)
        for j=1:(length(y)/d)-1
            if (y(d*j+1)-y(d*j-1))==0
                A=0;
            else
                A=(img(i,d*j+1,k)-img(i,d*j-1,k))/(y(d*j+1)-y(d*j-1));
            end
            B=img(i,d*j-1,k)-A*y(d*j-1);
            img_out(i,j,k)=Lprod(B+1,A*y(d*j));
        end
    end
end
end

```

Fig. 3. Basic algorithm for static image compression

The example in figure 3 realizes the column compression. The compression strategy that we have used is very simple and appears in the figure 4a. It is based in considering being every row (column) as a piecewise linear function in which every pixel stores an integer value in the range from 0 to 255 (8 bits codification). For the case of a compression index of 2 ($d=2$) it selects one of every two pixels interpolating its value by the straight line that joins the neighbouring pixels. As seen in figure 4 the crosses stand for the original pixels and circles stand for approximated ones.

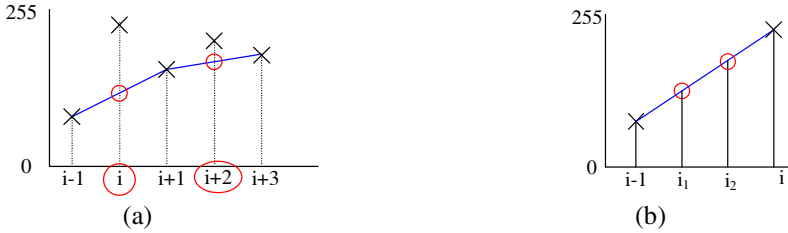


Fig. 4. (a) Compression algorithm ($d=2$). (b) Decompression algorithm ($d=3$).

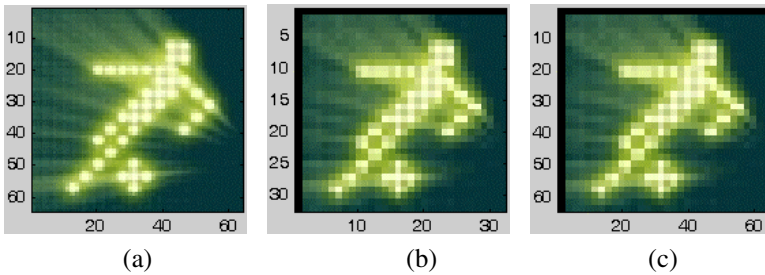


Fig. 5. Example of image compression and decompression: (a) original image, (b) compressed image, (c) decompressed image

On the other hand the decompression algorithm also applies the piecewise linear approximation criterion (figure 4b). In this case new pixels are inserted between every two by means of linear approximation. Figure 5 shows the obtained result. It is possible to observe the effects of the approximation both in the compressed image (figure 5b) and in the decompressed image (figure 5c). Since the interpolation mechanism is based on a piecewise linear approximation it is a lossy compression technique as shown in figure 5c.

5 BNK Algorithm

The algorithm that we are going to consider in this section is a piecewise linear compression technique that we have named BNK as it authors initials Bhaskaran-Natarajan-Konstantinides [6], [7]. In this algorithm the input image is consider as samples of a one dimensional waveform. The waveform can be compressed by a piecewise linear approximation within a prescribed error tolerance. The basic idea

consists of realizing the compression of a piecewise linear function so that the error of the new compressed function is bounded to a certain predefined value ϵ . Based on this fact the problem is defined as follows: given a piecewise linear function $F : [0,1] \rightarrow \mathfrak{R}^k$ specified as the interpolation of N points and an error $\epsilon \in \mathfrak{R}^k$, to construct the piecewise linear function G so that for everything $x \in [1, N]$, $\|F(x) - G(x)\|_{\infty} \leq \epsilon$ and G consists of the fewest number of segments that defines this function.

In agreement with this for every point (x, y) that defines function F the algorithm constructs the points $(x, y+\epsilon)$ and $(x, y-\epsilon)$. This creates an error tunnel as shows figure 6.

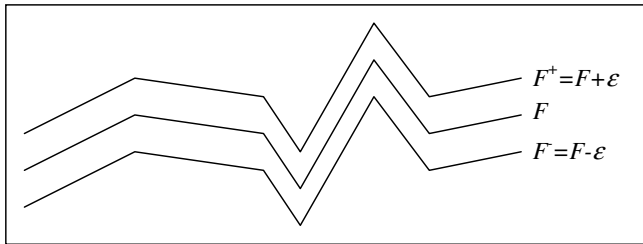


Fig. 6. Error tunnel of function F

The function described in figure 6 is a scalar function that gives place to a two dimensions tunnel. In general function F is a vector of k dimensions with an error tunnel of $k+1$ dimension. In the particular case of a static color image $k=3$. From this error tunnel the compression algorithm constructs a new piecewise linear function (G) that joins both ends of the tunnel with the fewest number of pieces. A two-dimensional image can be considered to be a collection of one-dimensional independent lines. Nevertheless it suits to exploit the two-dimensional correlation of the image [6], [7]. In fact one of the modifications that we propose to this algorithm corresponds to the scan mechanism of the image. In order to take advantage of the existing correlation between contiguous pixels we apply a zig-zag scan of the two-dimensional image. Applying zig-zag we turn a two-dimensional image into a one-dimension vector in which adjacent elements have a high correlation.

The original BNK compression scheme as proposed in [6], [7], is based on three stages: 1) error tunnel generator for adding and subtracting the target error (ϵ) to the original function ($F^{\pm} = F \pm \epsilon$), 2) waveform compression usign piecewise linear function with breakpoints limited by error tunneling, and 3) Huffman coding in order to encode the output function G using shorter bit patterns for more common characters and longer bit pattern for less common characters. Our strategy differs from the above mentioned scheme in which as previous step we fulfil a scan of the image in zigzag to which we apply the error tunnel as shown in figure 7. The stage of final codification is realized storing the new function G with the end values of every section together with the width of the above mentioned section. This gives place to a simplification of the needed hardware since there is replaced the Huffman codifier of

the original algorithm by a simple counter (we store the values of the breakpoints and the number of pixels between two adjacent breakpoints). Nevertheless the price that is paid is a reduction of the compression rate.

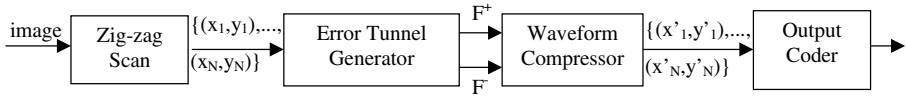


Fig. 7. Compression scheme

The results obtained for the image of figure 5 give a compression of 14 % for a 6% error and in case of 10% an error tunnel the obtained compression has been 55%. The circuit was implemented on a Xilinx Spartan3 FPGA. It required 230 slices on the device.

6 Conclusions

We have presented the implementation of Lukasiewicz operators and have used them for the approximation of piecewise linear functions. The linear interpolation has been applied to the problem of static images compression. Two techniques for image compressing have been showed: a basic technique based in selecting a pixel between several ones and another based in error tunnel.

References

1. A. Di Nola, A. Lettieri, "On normal forms in Lukasiewicz logic," *Archive for Mathematical Logic*, Springer-Verlag Heidelberg, vol 43, no. 6, pp. 795-823, Aug. 2004.
2. P. Amato, A. Di Nola, B. Gerla, "Neural Networks and Rational Lukasiewicz Logic", *Proceedings. NAFIPS. Annual Meeting of the North American*. pp. 506-510, 2002.
3. J.L. Castro, E. Trillas: "The logic of neural networks", *Mathware and Soft Computing*, vol 5, pp. 23-27, 1998.
4. N.M. Hussein Hassan, A. Barriga, S. Sánchez-Solano, "Piecewise Linear Function Interpolation Using Lukasiewicz's Operators", *Int. Symposium on Innovations in Intelligent System and Applications (INISTA'2005)*, Istanbul, June 2005
5. S. Ovchinnikov: "Max-min representation of piecewise linear functions", *Contributions to Algebra and Geometry*, vol 43, pp. 297-302, 2002
6. V. Bhaskaran, B.K. Natarajan, K. Konstantinides: "Lossy Compression of Images Using Piecewise-Linear Approximation" *Hewlett-Packard Technical Report HP-93-10*, Feb. 1993.
7. V. Bhaskaran, B.K. Natarajan, K. Konstantinides: "Optimal piecewise-linear compression of images", *IEEE Data Compression Conference (DCC'93)*, pp. 168-177, Utah, Apr. 1993.

Modelling Coarseness in Texture Images by Means of Fuzzy Sets

J. Chamorro-Martínez, E. Galán-Perales, D. Sánchez, and J.M. Soto-Hidalgo*

Department of Computer Science and Artificial Intelligence
University of Granada

C/ Periodista Daniel Saucedo Aranda s/n, 18071 Granada, Spain
{jesus, elena, daniel, soto}@decsai.ugr.es

Abstract. In this paper we model the concept of "coarseness", typically used in texture image descriptions, by means of fuzzy sets. Specifically, we relate representative measures of this kind of texture with its presence degree. To obtain these "presence degrees", we collect assessments from polls filled by human subjects, performing an aggregation of these assessments by means of OWA operators. Using this collected data, and some statistics as reference set, the membership function corresponding to the fuzzy set "coarseness" is modelled.

Keywords: Image features, texture features, fuzzy texture, visual coarseness.

1 Introduction

Visual texture is one of the most difficult features to be characterized in images due to the imprecision of the concept itself. In fact, there is not an accurate definition for the concept of texture but some widespread intuitive ideas. In this way, texture is described by some authors as local changes in the intensity patterns or gray tones. Other authors consider texture as a set of basic items called *texels* (or texture primitives), arranged in a certain way [1,2]. Moreover, it is usual for humans to describe visual textures according to some "textural concepts" like *coarseness*, *orientation*, *regularity* [3]. To describe such concepts, linguistic labels are used (e.g. coarse or fine can be used to describe coarseness).

The own imprecision of the concept of texture suggests to use representation models that incorporate the uncertainty. Nevertheless, the majority of the approaches found in the literature are crisp proposals [4] where uncertainty is not properly taken into account. To deal with the imprecision relative to visual texture, there are some approaches which introduce the use of fuzzy logic [5]. However, in many of them, fuzzy logic is often applied just during the process but the output do not usually model the imprecision (being often a crisp one).

In this paper we focus our study on coarseness, one of the textural properties most used in the literature which allows to distinguish between fine and coarse

* This work has been partially supported by "Instituto de Salud Carlos III under FIS G03/185 project, "Imagen Médica Molecular y Multimodalidad".



Fig. 1. Some examples of images with different degrees of fineness

textures. In fact, it is usual that the texture concept is associated to the presence of fineness. A *fine* texture can be considered as small texture primitives with big gray tone differences between neighbour primitives (e.g. the image in figure 1(A)). On the contrary, if texture primitives are bigger and formed by several pixels, it is a *coarse* texture (e.g. the image in figure 1(A)).

In this paper, we propose to model fineness by means of fuzzy sets to deal with the problem of imprecision found in texture characterization. To do this, two questions will be faced: what reference set should be used for the fuzzy set, and how to obtain the related membership functions. To solve the first question, a vector of measures will be automatically computed from the texture image. To answer the second question, functional relationship between a certain measure and the presence degree of a textural concept related to it will be learnt.

The rest of the paper is organized as follows. In section 2 we introduce our methodology to obtain the fuzzy sets related to fineness textural concept. In section 3 we show the results of applying the model and the main conclusions and future work are summarized in section 4.

2 Texture Modelling. Application to Fineness

In this paper, we propose to model a *textural concept* as a fuzzy set. From now on, we shall denote \mathcal{T} the textural concept we want to model, in our case $\mathcal{T} = \textit{fineness}$. As reference set, a vector of K measures obtained by carrying out an analysis of the texture image is used. These measures should give reliable information about the presence degree of the textural concept *fineness* under study. Thus, the model of the textural concept will be given by a fuzzy subset built on the domain of the chosen measure.

Furthermore, a membership function that models the textural concept fineness for the fuzzy set is needed. In this paper we propose to obtain this function by learning a functional relationship between a certain measure and the presence degree of the textural concept fineness.

To learn this relationship, we will use a set $\mathcal{I} = \{I_1, \dots, I_N\}$ of N images that fully represent the textural concept \mathcal{T} to be learnt. Given the concept \mathcal{T} , a set of measures $\mathcal{P} = \{P_1, \dots, P_K\}$ will be considered, with $P_k \in \mathcal{P}$ being a measure of the presence of \mathcal{T} in an image (e.g. in the case of $\mathcal{T} = \textit{fineness}$ we could define $\mathcal{P} = \{\textit{EdgeDensity}, \textit{Variance}, \textit{Range}\}$). Thus, for each image $I_i \in \mathcal{I}$, we will obtain (a) a vector of measures $\mathbf{M}^i = [m_1^i, \dots, m_K^i]$, where m_k^i is a value for the measure $P_k \in \mathcal{P}$ applied to the image I_i (section 2.1), and (b) an assessment v^i of the presence degree of the textural concept \mathcal{T} under study. To get this assessment, a poll will be performed (section 2.2). Once we have a multiset of valid pairs $\Psi = \{(\mathbf{M}^1, v^1), \dots, (\mathbf{M}^N, v^N)\}$, we shall estimate the membership function (section 2.4).

2.1 Fineness Measures

Given the textural concept $\mathcal{T} = \textit{fineness}$ a set of measures $\mathcal{P} = \{P_1, \dots, P_K\}$ will be considered. Different measures that characterize the presence of fine texture are found over the literature [3]. In this paper we have chosen simple measures which imply a low computational cost. We have specifically used three well known measures:

- Range, measured as the difference between the minimum and the maximum values in the image. We will note \textit{Range}^i the range related to the image I_i
- Variance of the image gray tones. We will note \textit{Var}^i the variance related to the image I_i
- Edge density, measured as the percentage of points which are an edge in the image. In this paper, we have used the Canny Edge Detector [6]. We will note $\textit{EdgeDens}^i$ the edge density related to the image I_i

In section 2.3 we will study the goodness of a measure $P_k \in \mathcal{P}$ to model a textural concept \mathcal{T} . This study will be performed by analyzing the ability of P_k to discriminate the assessments given by a set of subjects.

2.2 Assessment Collection

In this section we will describe how to get, from the image set $\mathcal{I} = \{I_1, \dots, I_N\}$, a vector $\Gamma = [v^1, \dots, v^N]$ of the assessments of the presence degrees related to the textural concept $\mathcal{T} = \textit{fineness}$.

Thus, firstly a criterion for choosing the image set \mathcal{I} is needed. After that, a poll which allows to get assessments of the presence degree of the textural concept \mathcal{T} will be designed. These assessments will be obtained for each image in \mathcal{I} . Finally, to get just one assessment that summarizes the information given by human subjects, an aggregation of the different assessments will be performed.

The texture image set. Firstly, the image set $\mathcal{I} = \{I_1, \dots, I_N\}$ that fully represents the textural concept to be learnt must be chosen. In this paper, we have selected a set \mathcal{I} of $N = 80$ images representative of the concept of *fineness*. Figure 1 shows some images extracted from the set \mathcal{I} . The selection was done to

cover the different presence degrees of fineness with a representative number of images. Furthermore, we have selected images in which, as far as possible, just one degree of fineness is perceived.

The poll. Given the image set \mathcal{I} , the next step is to obtain assessments about the perception of \mathcal{T} from a set of subjects. From now on we shall denote $\Theta^i = [o_1^i, \dots, o_L^i]$ the vector of assessments obtained from L subjects for image I_i . To get Θ^i we will ask subjects to assign images to classes, so that each class has associated a presence degree. The number of classes is fixed and an example image which represents the presence degree is associated to each class.

In particular, 20 subjects have participated in the poll and 9 classes C_1, C_2, \dots, C_9 have been considered. The first nine images in figure 1 show the nine representative images for each class $C_k, k = 1, \dots, 9$ used in this poll. It should be noticed that the images are decreasingly ordered according to the presence degree of the fineness concept. The first class C_1 (figure 1(A)), represents a presence degree of 1 while the ninth class C_9 (figure 1(I)), represents a presence degree of 0. The rest of the classes (figure 1(B)-(H), (J)) represent degrees in the interval (0,1).

Finally, a vector of 20 assessments $\Theta^i = [o_1^i, \dots, o_{20}^i]$ is obtained for each image $I_i \in \mathcal{I}$. The degree o_j^i associated to the assessment given by the subject S_j to the image I_i is computed as $o_j^i = (9 - k) * 0.125$, where k is the index of the class C_k to which the image is assigned by the subject S_j .

Assessment aggregation. Our aim at this point is to obtain, for each image in the set \mathcal{I} , one assessment that summarizes the assessments given by the different subjects about the presence degree of fineness.

To aggregate opinions we have chosen the use of OWA operator guided by a quantifier [7]. With these operators it can be represented the existing interval between logic *AND*, which allows for the representation of the quantifier *for all*, and logic *OR*, which allows for the representation of the quantifier *exists*.

Yager proposed in [7] the use of monotonically nondecreasing linguistic quantifiers. In particular, we can use quantifiers of the form:

$$Q(r) = \begin{cases} 0 & \text{if } r < a, \\ \frac{r-a}{b-a} & \text{if } a \leq r \leq b, \\ 1 & \text{if } r > b \end{cases} \tag{1}$$

being $a, b, r \in [0, 1]$. Depending on the values associated to the pair (a, b) , the quantifier interpretation would be different. Thus, in [8] the quantifier *most* is associated to the pair (0.3, 0.8), the quantifier *at least half* (0,0.5) and the quantifier *as many as possible* (0.5,1). In our proposal, the quantifier *most* is chosen.

Once the quantifier Q has been chosen, the weighting vector of the OWA operator can be obtained following Yager as $w_j = Q(j/L) - Q((j - 1)/L), j = 1, 2, \dots, L$. According to this, given a quantifier Q , for each image $I_i \in \mathcal{I}$ the vector Θ^i obtained from L subjects will be aggregated into one assessment v^i as follows:

$$v^i = w_1\hat{o}_1^i + w_2\hat{o}_2^i + \dots + w_L\hat{o}_L^i \tag{2}$$

where $[\hat{o}_1^i, \dots, \hat{o}_L^i]$ is a vector obtained by ranking in nonincreasing order the values of the vector Θ^i .

2.3 Goodness of a Measure P_k

Given a measure $P_k \in P$ we need to solve two questions: (a) is P_k a good measure to appropriately represent \mathcal{T} ? and (b) what is the ability of P_k to discriminate different presence degrees for \mathcal{T} , i.e. how many classes can P_k actually discriminate?

To face the first question, in this paper we propose to perform an ANOVA test to find out if P_k can discriminate at least two classes with a significance level of 0.05. In the test, the measure values obtained by applying P_k to each $I_i \in \mathcal{I}$ will be the dependent variable, and the class to which the image I_i has been (mostly) assigned will be the independent variable.

To face the second question, we propose to apply a set of multiple comparison tests to find out how many classes can P_k discriminate. Concretely, we will use the algorithm 1 which, starting with an initial partition, will iteratively join clusters until the partition has a number of classes that can be discriminated. In our proposal, the initial partition $Part^0 = C_1, C_2, \dots, C_{nc}$ will be the 9 classes, where $C_r \in Part^0$ contains the images assigned to the class by the majority of the subjects (section 2.2), as δ we propose to use the Mahalanobis distance by considering the mean and the variance of the involved classes, as ϕ the multiple comparison tests: Scheffé, Bonferroni, Duncan, Tukey’s least significant difference, Tukey’s honestly significant difference will be considered, and finally the number of positive tests to accept distinguishability will be fixed to $NT = 3$ between a pair of clusters.

From now on we shall denote $\widetilde{Part}_k = C_1, C_2, \dots, C_{\widetilde{nc}}$, as the classes which can be discriminated by P_k .

In the case of fineness, from the set $\mathcal{P} = \{Range, Variance, EdgeDensity\}$ considered in this paper, only the measure *EdgeDensity* passes the ANOVA test and the number of classes obtained from the algorithm 1 for this measure is 3.

2.4 Obtaining the Membership Function

In this section we will deal with the problem of obtaining the membership function for the textural concept $\mathcal{T} = fineness$. To simplify the notation, as it is usual in the scope of fuzzy sets, we will use the same notation for the textural concept, for the fuzzy set which represents it and for the membership function that defines it. In our proposal, only the measure *Edge density* will be used to obtain the membership function that models fineness as the rest of the measures of P did not pass the ANOVA test. In this way, the fineness will be modelled by a membership function defined on \mathbb{R} i.e.,

$$\mathcal{T} : \mathbb{R} \rightarrow [0, 1] \tag{3}$$

Algorithm 1. Obtaining the distinguishable clusters

Input:

$Part^0 = C_1, C_2, \dots, C_{nc}$: Initial Partition

δ : Clusters distinguishable function

ϕ : Set of multiple comparison tests

NT : Number of positive tests to accept distinguishability

1.- Initialization

$k = 0$

$dist = false$

2.- While $dist = false$

Apply the multiple comparison tests ϕ to $Part^k$

If for each pair $C_i, C_j \in Part^k$ more than NT of the multiple comparison tests ϕ show distinguishability

$dist = true$

Else

Search for the pair of clusters C_r, C_{r+1} , verifying

$\delta(C_r, C_{r+1}) = \min\{\delta(C_i, C_{i+1}), C_i, C_{i+1} \in Part^k\}$

Join C_r and C_{r+1} on a cluster $C_n = C_r \cup C_{r+1}$

$Part^k = Part^{k-1} - C_r - C_{r+1} + C_n$

$k = k + 1$

3.- Output: $\widetilde{Part}_k = C_1, C_2, \dots, C_{nc-k}$

In this paper we propose to define \mathcal{T} as a linear spline, i.e.,

$$\mathcal{T}(x) = \begin{cases} 0 & x \leq x_1 \\ T^1(x) & x \in (x_1, x_2] \\ T^2(x) & x \in (x_2, x_3] \\ \vdots & \vdots \\ 1 & x > x_{\widetilde{nc}} \end{cases} \quad (4)$$

with $x_r, r = 1 \dots \widetilde{nc}$, calculated as the mean of the measure values in the class $C_r \in \widetilde{Part}_k$, $T^r(x)$ is the straight line defined between the points (x_r, y_r) and (x_{r+1}, y_{r+1}) with y_r being the mean value of the presence degree of the textural concept \mathcal{T} for the images grouped into the cluster C_r .

In our case, as we obtained three different classes, the above mentioned function is defined as follows:

$$\mathcal{T}(x) = \begin{cases} 0 & x \leq 0.17 \\ 3.89 * x - 0.68 & x \in (0.17, 0.30] \\ 7.90 * x - 1.89 & x \in (0.30, 0.37] \\ 1 & x > 0.37 \end{cases} \quad (5)$$

3 Results

Let's consider figure 2(A) corresponding to a mosaic made by several images, each one with a different increasing fineness presence degree. For each image in



Fig. 2. Results for a mosaic image: (A) original mosaic image (B) edge map of the original images (C) presence degree of fineness textural concept obtained with the proposed model

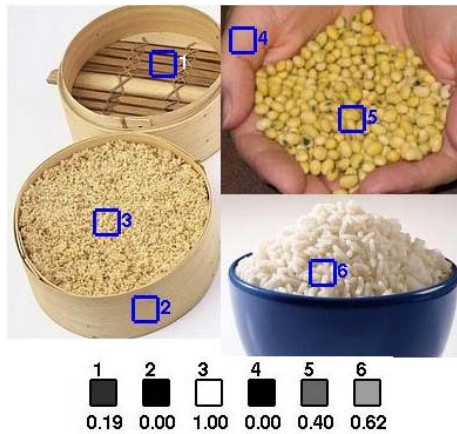


Fig. 3. Some examples of fineness degrees computed for six real subimages

the mosaic, we apply the membership function defined in equation 5. The image 2(B) shows the edge map obtained from the original mosaic. This map is used to get the edge density which is the reference set of our fuzzy set. The fineness presence degree for each subimage is shown in figure 2(C) where a white grey level means maximum membership degree, while a black one corresponds to 0 membership degree (the numeric value is also shown on each subimage). It can be noticed that our model captures the evolution of the fineness degrees.

We have also compared our model output with the assessments obtained from subjects. To get them we have aggregated the assessments of 20 subjects following the steps explained in section 2.2. The average error obtained is 0.089, which shows the goodness of our approach to represent the subjectivity found in fineness perception.

Figure 3 shows several real images where some windows have been selected corresponding to subimages with different fineness degree. We apply the model to each subimage, and the fineness degree obtained is shown at the bottom of

figure 3. As it can be noticed, our model assesses high degrees to fine texture areas and low degrees to coarse texture areas.

4 Conclusions and Future Works

In this paper we have proposed a methodology to represent the fineness concept by means of fuzzy sets. To define the membership function associated to the fuzzy set, the functional relationship between a certain measure (automatically computed over the image) and the presence degree of fineness has been learnt.

In order to obtain the perception degree of a certain textural concept, a group of 20 human subjects have been polled and their assessments have been aggregated by means of OWA operators. After that and by using as reference set a group of very simple statistical measures, we have obtained a fuzzy set which models the human perception of fineness. The results given by our approach show a high level of connection with the assessments given by subjects.

As future work, more complex statistical measures, new ways of defining the membership function and the use of linguistic labels for this fuzzy set in image retrieval by content will be analyzed.

References

1. Tuceryan, M., Jain, A.: Texture Analysis. In: *The Handbook of Pattern Recognition and Computer Vision*. 2 edn. C.H. Chen and L.F. Pau and P.S.P. Wang (1998) 207–248
2. Shapiro, L.G., Stockman, G.: Image Segmentation. In: *Computer Vision*. Prentice-Hall (2001) 297–301
3. Abbadeni, N., Ziou, D., Wang, S.: Perceptual textural features corresponding to human visual perception. In: *Proc. of the Thirteenth Vision Interface Conference, Montreal, Quebec (Canada) (2000)* 365–372
4. Reed, T.R., Buf, J.H.D.: A review of recent texture segmentation and feature extraction techniques. *CVGIP: Image Understanding* **57**(3) (1993) 359–372
5. Shackelford, A.: A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas. *IEEE Transactions on Geoscience and Remote Sensing* **41**(9) (2003) 1920–1932
6. Canny, J.: A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**(6) (1986) 679–698
7. Yager, R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man and Cybernetics* **18**(1) (1988) 183–190
8. Herrera, F., Herrera-Viedma, E.: Linguistic decision analysis: steps for solving decision problems under linguistic information. *Fuzzy Sets and Systems* **115**(1) (2000) 67–82

Fuzzy Motion Adaptive Algorithm for Video De-interlacing

P. Brox¹, I. Baturone¹, S. Sánchez-Solano¹, J. Gutiérrez-Ríos²,
and F. Fernández-Hernández²

¹ Instituto de Microelectrónica de Sevilla (IMSE-CNM-CSIC)
Avda. Reina Mercedes S/N. Edificio CICA. 41012 Sevilla-Spain
`brox@imse.cnm.es`

² Dpto. Tecnología Fotónica. Universidad Politécnica de Madrid
Campus de Montegancedo S/N. 28660 Boadilla del Monte-Madrid-Spain
`jgr@fi.upm.es`

Abstract. A motion adaptive algorithm for video de-interlacing is presented in this paper. It is based on a fuzzy inference system, which performs an interpolation between two linear techniques as a function of the motion level. Fuzzy systems with different number of 'if-then' rules have been analyzed and compared in terms of complexity as well as efficiency in de-interlacing benchmark video sequences.

Keywords: Video De-interlacing, Motion Adaptive, Fuzzy Inference Systems, Supervised Learning Algorithms.

1 Introduction

The main video transmission formats of TV signals (NTSC, PAL, SECAM) use an interlaced signal, where only half of the lines which compose a frame are transmitted. Therefore, the bandwidth required by the broadcast is effectively halved since the human visual system is less sensitive to flickering of details than to large areas flicker [1]. However, the need of progressive scanning is growing nowadays due to the advent of high-definition television, videophone, projectors, DVDs, and video on PCs. This increasing need has encouraged the development of algorithms that perform a spatio-temporal sampling to calculate the non-transmitted lines.

Among the de-interlacing algorithms two categories can be distinguished: motion-compensated algorithms that use a motion vector to interpolate the missing lines, and non-motion-compensated algorithms [2]. The first ones generally perform better than the second ones especially for sequences with a high level of motion. Unfortunately, the motion-compensated algorithms involve the high computational cost related to motion vector calculation. The different de-interlacing algorithms can be classified by considering if they always interpolate the same pixels (linear techniques) [3], [4], or if these pixels are selected accordingly to the characteristics of the image (non-linear techniques) [5], [6]. Non-linear techniques can be divided in turn into two groups: those which try

to adapt the interpolation strategy to the presence of motion [5]; and other ones that perform an edge-dependent interpolation [6].

To implement correctly the motion adaptive algorithm, it is fundamental to detect motion accurately. Basically motion detectors evaluate the difference between luminance values of pixels from two consecutive fields. However, this measurement is not usually very reliable due to the presence of edges, vertical details, and noise corrupting the TV signal. The robustness of these detectors can be increased by using more than one detector, and combining them with the logical operator 'and' [1]. In this way, only if all of them detect motion the motion signal is activated. Other authors resort to the use of a multilevel signal, rather than a binary one, to indicate the probability of motion. Several algorithms based on fuzzy-logic have also been proposed to perform an adaptive interpolation with the level of motion. They exploit the capacity of fuzzy techniques to perform interpolations where the information is uncertain and, hence, the decision is not trivial [7]. The technique proposed in [7] provides good results but it uses a complex set of rules, which requires a considerable computational cost.

A novel motion adaptive algorithm for video de-interlacing is proposed in this paper. It uses a fuzzy logic-based system to determine the interpolation between the pixels from the transmitted lines accordingly to the level of motion. The algorithm is described in detail in Section 2. Its performance when de-interlacing several image sequences is analyzed in Section 3. Finally, concluding remarks are included in Section 4.

2 Algorithm Description

The fuzzy motion adaptive algorithm is based on the following heuristic knowledge: if the pixel to interpolate belongs to an area where there is no motion, the best result is achieved by performing an interpolation among pixels from the previous field (temporal interpolation); however, in the case that the pixel corresponds to an area with a high level of motion then the best solution is to realize an interpolation among pixels from the current field (spatial interpolation). The most basic interpolations have been selected among spatial and temporal linear techniques: pixel insertion from the previous field as temporal interpolation (I_T) and the average value of the pixel from the upper and lower lines as spatial technique (I_S). The level of motion is evaluated by processing the bi-dimensional convolution signal given by the following expression:

$$mot(x, y, t) = \Sigma_{i=1}^3 (\Sigma_{j=1}^3 H_{ij} C_{ij}) \quad (1)$$

where H_{ij} and C_{ij} are the elements of the following matrices:

$$C = \frac{1}{16} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} \quad (2)$$

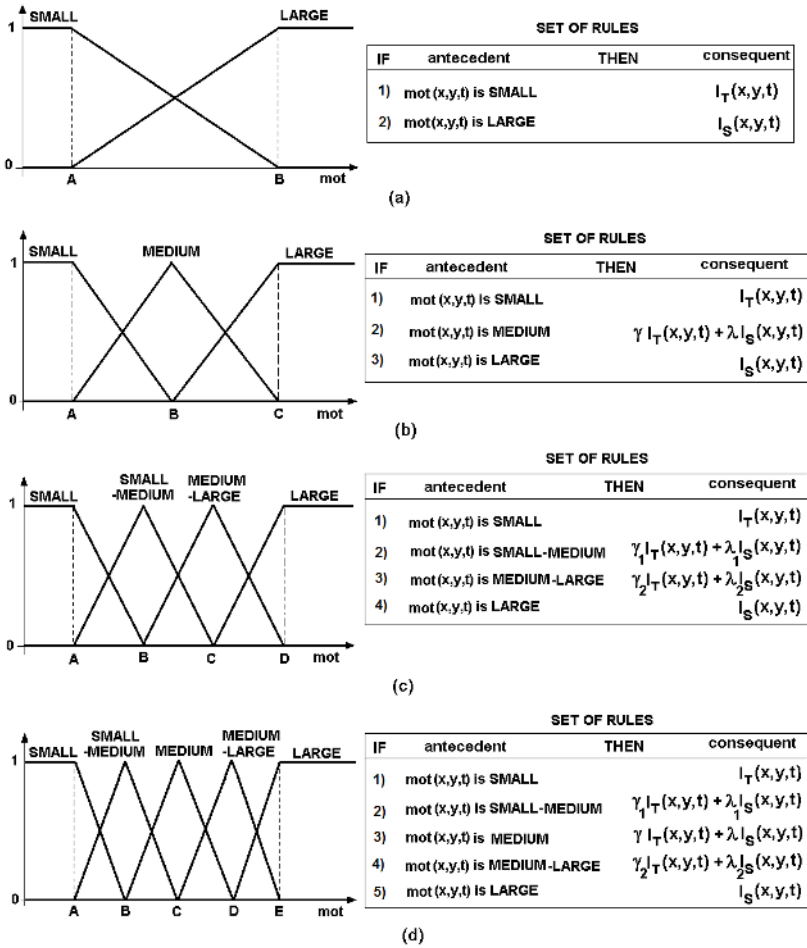


Fig. 1. Membership functions used by the different fuzzy inference systems

$$H = \begin{pmatrix} H(x-1, y-1, t-1) & H(x-1, y, t) & H(x-1, y+1, t-1) \\ H(x, y-1, t-1) & H(x, y, t) & H(x, y+1, t-1) \\ H(x+1, y-1, t-1) & H(x+1, y, t) & H(x+1, y+1, t-1) \end{pmatrix} \quad (3)$$

The notation (x,y,t) means that the pixel has the spatial coordinates (x,y) and corresponds to the instant (t) in the video sequence. Observing the size of the matrices H and C, a bi-dimensional convolution window of size 3x3 has been chosen. The idea of using bi-dimensional convolution techniques was introduced in [8]. Its main advantage is the inclusion of neighbors (with different weighting factors, as shown in expression (2)) to estimate the motion. In this case, the selected window allows to consider a spatio-temporal neighborhood. This could reduce the eventual errors introduced by the presence of noise, edges or vertical details.

In the motion adaptive technique originally introduced in [5], the level of motion was evaluated by comparing the signal value corresponding to the luminance difference between two consecutive fields with a constant threshold value. The aim of the work presented in this paper is to improve the original motion adaptive technique by fuzzifying the levels of motion so as to perform a gradual instead of abrupt change between spatial and temporal interpolation. Therefore, in those areas where the level of motion is medium and, hence, the decision is not trivial, a non-linear interpolation between I_S and I_T is realized.

2.1 Fuzzy Inference System Description

The heuristic knowledge used by the motion adaptive techniques is modeled by employing a fuzzy inference system. Firstly, a system with two rules is used, where the concepts 'SMALL motion' and 'LARGE motion' are represented by the fuzzy sets of the Figure 1(a). Nevertheless, the interpolation capacity of fuzzy logic could be further exploited by considering the possibility of extending the number of fuzzy sets. In this sense, it is possible to define a new fuzzy set represented by the MEDIUM label shown in Figure 1(b). The set of rules is enlarged with a new rule that, when activated, performs a linear combination between the techniques I_S and I_T .

The level of motion in a field could not only be considered as SMALL, MEDIUM or LARGE, but more situations can be distinguished. For example four (SMALL, SMALL-MEDIUM, MEDIUM-LARGE, LARGE) or five labels (SMALL, SMALL-MEDIUM, MEDIUM, MEDIUM-LARGE, LARGE), represented in the Figure 1(c) and 1(d), could be employed. This translates into using four or five fuzzy 'if-then' rules, respectively. The problem when trying to implement these rules is that heuristic knowledge does not provide enough information to fix the constant values gamma and lambda of the rules' consequents neither to determine the values A, B, C, D, and E that describe the five possible linguistic labels. In order to fix these values, our approach has been to use supervised learning algorithms, as detailed in the following sections.

2.2 Supervised Learning Algorithm

The above described fuzzy systems have been implemented with the development environment Xfuzzy 3 [9]. This environment eases the design of fuzzy logic-based inference systems by including different CAD tools for the description, identification, simplification, verification, tuning and synthesis of the systems. In particular, the tool named *xfl* aids in the tuning stage, which is usually one of the most complex tasks in the design of fuzzy systems. This tool allows to apply different supervised learning algorithms where the desired behavior of the system is described by a set of training patterns. In our problem of video de-interlacing, the fuzzy systems have been tuned by using a set of progressive frames to generate the training patterns. The selected supervised learning algorithm (Marquardt-Levenberg in our case) tries to minimize a function error which evaluates the difference between the current behavior and the desired one (determined by

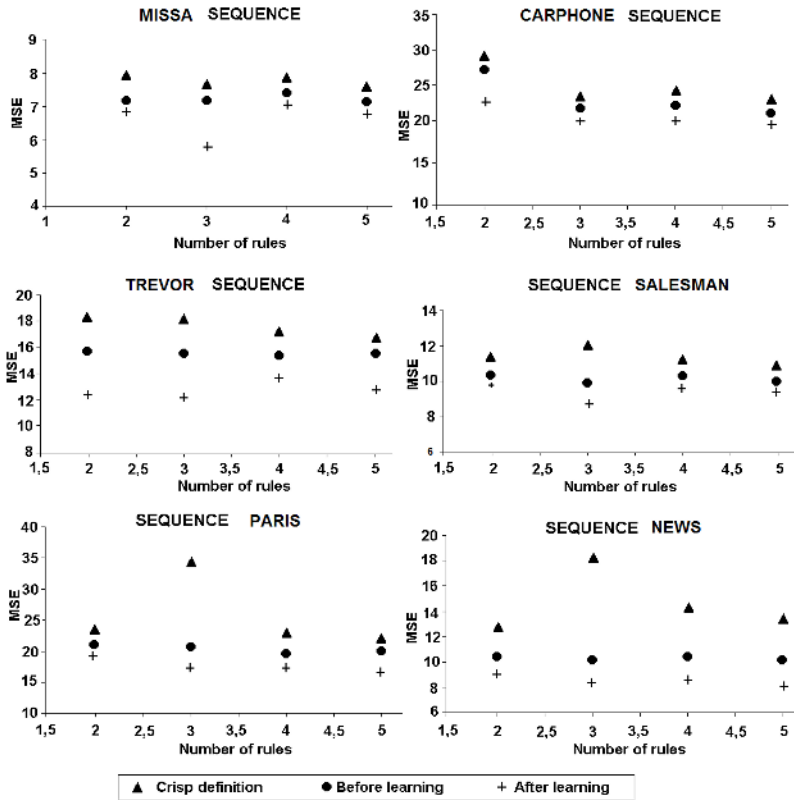


Fig. 2. MSE results obtained by the different fuzzy inference systems de-interlacing several video sequences

the input/output patterns). The tool *xfl* allows the user to select the parameters of the fuzzy system to be involved in the tuning process. The utility of this stage in the design process of fuzzy systems for de-interlacing video sequences is explained in detail in Section 3.

3 Simulation Results

In order to analyze the performance of the proposed fuzzy systems several benchmark video sequences have been considered. Since they are originally in a progressive format, a set of progressive images from all of these sequences have been selected to generate the training data for the supervised learning process. Afterwards, they have been de-interlaced artificially by eliminating lines from the frames.

Figure 2 shows the mean squared error (MSE) value obtained when the artificially interlaced video sequences are de-interlaced. This value corresponds to the average MSE value of the de-interlaced fields (approximately fifty fields of

Table 1. Average PSNR (values in dBs) when de-interlacing several video sequences

Sequence	Missa	Salesman	Carphone	Paris	Trevor	News
Format	CIF	CIF	QCIF	CIF	CIF	QCIF
LD	36.44	29.75	28.25	23.61	31.05	25.18
I_S	40.47	33.53	32.61	36.67	35.04	29.25
I_T	38.36	36.17	30.64	29.86	34.36	33.13
2fields-VT	40.25	36.54	34.08	30.73	36.61	35.46
3fields-VT	40.52	36.95	34.54	31.37	37.16	35.67
Technique in [7]	40.01	37.62	32.27	33.12	35.38	34.73
2-rule Proposal	40.18	38.29	34.78	35.28	36.69	37.51
3-rule Proposal	40.51	38.44	34.83	35.78	37.49	38.68
4-rule Proposal	39.65	38.23	34.94	35.5	36.77	38.65
5-rule Proposal	39.67	38.21	34.94	35.93	37.16	39.15

each video sequence have been simulated). The graphics in Figure 2 illustrate the results obtained by an algorithm which uses a crisp definition of the concepts SMALL, SMALL-MEDIUM, MEDIUM, MEDIUM-LARGE and LARGE. They also show the results obtained when those concepts are defined as fuzzy sets and are processed by fuzzy systems with different number of rules (before and after learning). Comparing the three series of results, a first conclusion is that the use of fuzzy instead of crisp concepts provides lower errors. A second conclusion is that performance improves when the membership functions as well as the consequents are modified by the supervised learning algorithm. Finally, analyzing the number of rules and the MSE value, it can be observed that the system with three rules always obtains better results than the system with two ones. In the other side, the systems with four and five rules either do not provide significant improvement or even introduce a slightly higher number of errors.

The proposed fuzzy logic-based technique has been compared with: (a) basic linear techniques such as line doubling (LD), line average (I_S) as spatial technique and pixel insertion from the previous field (I_T) as temporal technique; (b) with linear spatio-temporal techniques [3], [4], which are currently used in commercial chips; and (c) with the fuzzy logic-based motion adaptive algorithm reported in [7]. Analyzing the results shown in Table 1, it can be seen how the highest results of PSNR, and hence the lowest errors, correspond to the proposed fuzzy systems (the results in Table 1 are obtained with the systems after learning). The superior performance of our proposal can be also seen by analyzing the de-interlaced images in Figure 3.

Finally, an analysis of the computational cost involved in the implementation of each one of the proposed systems has been realized. All the algorithms have been executed on the same platform (a PC with a Pentium IV processor and the



Fig. 3. (a) Progressive frame of 'Carphone' sequence. De-interlaced image applying: (b) line doubling, (c) line average, (d) field insertion, (e) 2-field VT filtering, (f) 3-field VT filtering, (g) fuzzy motion adaptive in [7], (h) proposal with 2 rules and (i) proposal with 3 rules.

Table 2. Time invested in de-interlacing fifty fields of a video sequence

Algorithm	DL	I _S	I _T	VT 2fields	VT 3fields	Technique [7]	Proposal 2-3-4-5 rules
Time(s)	2.03	2.05	3.28	10.62	14.65	143.03	29.23-30.95-31.76-32.65

operating system MSWindow XP) so as to measure the time taken by each one in processing one sequence. The results are shown in Table 2. It can be seen how the linear techniques are the fastest ones but their results are widely improved by our proposal.

4 Conclusions

A fuzzy motion adaptive technique is presented in this paper. It performs a combination between two linear techniques depending on the level of motion. The proposal is inspired in the original motion adaptive idea, but it uses fuzzy definitions instead of crisp ones to describe the level of motion and employ a

supervised learning technique to adjust the parameters of the fuzzy systems. After analyzing several systems with different complexity it can be concluded that a fuzzy inference system with three 'if-then' rules provides a good trade-off between performance and computational cost.

Acknowledgement. This work has been partially funded by the project TEC2005-04359/MIC from the Spanish Ministry of Education and Science, and TIC2006-635 from the Andalusian Regional Government. The first author is supported by the Spanish Ministry of Education under the program F.P.U for Phd. students.

References

1. G. De Haan. Video Processing. University Press, Eindhoven, 2004.
2. G. De Haan and E. B. Bellers. De-interlacing: An overview. *Proc. of the IEEE*, Vol. 86, Pág.1839-1857, 1998.
3. Genesis Microchip, Inc., Preliminary data sheet of Genesis gmVLD8, 8 bit digital video line doubler, version 1, 1996.
4. M. Weston. Interpolating lines of video signals. US-patent 4, Pág.789-893, 1998.
5. A. M. Bock. Motion adaptive standards conversion between formats of similar field rates. *Signal Processing: Image Communication*, Vol. 6, no. 3, Pág.275-280, 1994.
6. T. Doyle and M. Looymans. Progressive scan conversion using edge information. *Signal Processing of HDTV*. Ed. Elsevier Science Publishers, Vol. II, Pág.711-721, 1990.
7. D. Van de Ville, B. Rogge, W. Philips and I. Lemahieu. De-interlacing using fuzzy-based motion detection. *Proc. 3rd Int. Conf. on Knowledge-Based Intelligent Information Engineering Systems*, Pág.263-267, 1999.
8. J. Gutiérrez-Ríos, F. Fernández-Hernández, J. C. Crespo and G. Triviño. Motion adaptive fuzzy video de-interlacing method based on convolution techniques. *Proc. of Information Processing and Management of Uncertainty in Knowledge-Bsed Systems*, 2004.
9. F. J. Moreno-Velo, I. Baturone, S.Sánchez-Solano and A. Barriga. Rapid design of complex fuzzy systems with XFUZZY. *Proc. IEEE Int. Conf. on Fuzzy Systems*, Págs.342-347, 2003.
10. F. J. Moreno-Velo, I. Baturone, R. Senhadji and S. Sánchez-Solano. Tuning complex fuzzy systems by supervised learning algorithms. *Proc. IEEE Int. Conf. on Fuzzy Systems*, Págs. 226-231, 2003.

Web Site Off-Line Structure Reconfiguration: A Web User Browsing Analysis

Sebastián A. Ríos¹, Juan D. Velásquez^{2,3}, Hiroshi Yasuda¹,
and Terumasa Aoki¹

¹ Applied Information Engineering, Laboratory, University of Tokyo, Japan
{srios, yasuda, aoki}@mpeg.rcast.u-tokyo.ac.jp

² Center for Collaborative Research, University of Tokyo, Japan

³ on leave from Department of Industrial Engineering,
University of Chile, Chile
jvelasqu@dii.uchile.cl

Abstract. The correct web site text content must be help to the visitors to find what they are looking for. However, the reality is quite different, many times the web page text content is ambiguous, without meaning and worst, it don't have relation with the topic that is shown as the main theme. One reason to this problem is the lack of contents with concept meaning in the web page, i.e., the utilization of words and sentences that show concepts, which finally is the visitor goal. In this paper, we introduce a new approach for improving the web site text content by extracting Concept-Based Knowledge from data originated in the web site itself. By using the concepts, a web page can be rewrite for showing more relevant information to the eventual visitor. This approach was tested in a real web site, showing its effectiveness

1 Introduction

The Web has become a immense collection of documents which contain valuable information in almost every imaginable topics. However, the problem of searching in this huge ocean of data is neither easy nor fast. These problems get worse with the fast growing of the Web and forces to a new way of designing and developing web sites [3]. Moreover, improving the web site usability, structure and content to keep the visitors interested on it is a challenging task [8].

In order to improve the web site structure and content, the web mining techniques have shown a reasonable effectiveness [4]. Specify, the Web Content Mining (WCM) and Web Usage Mining (WUM) techniques have being used for analyzing the web page content (mainly text) and the visitor browsing behavior respectively [6,15]. From the patterns extracted by using these techniques, recommendations about hyperlinks and content modifications are obtained.

This paper aims to provide a suitable content recommendation by applying a WCM technique for discovering Content-Based Knowledge, which is used for reducing the vocabulary problem present in a web page, i.e., badly utilization of irrelevant and redundant words for creating web page text content without meaning for the visitor.

This paper is organized as follows. In Section 2, a short review about related research work is done. Section 3 shows the concept discovery task in text contents. The proposed methodology is introduced in Section 4, and in Section 5 its application to a real-world case is presented. Finally, Section 6 presents the main conclusions and future work.

2 Related Work

Web site personalization according to Eirinaki et al. [3] is “*any action that adapts the information or services provided by a Web site to the needs of a user or set of users, taking advantage of the knowledge gained from the users’ navigational behavior and individual interest, in combination with the content and the structure of the Web site*”. Besides, web personalizations’ objective according to Mulvenna et al. [7] is to “*provide users with the information they want or need, without expecting from them to ask for it explicitly*”.

Many different approaches have been developed in order to perform a Web site personalization in the best way. The majority of the efforts correspond to those which only take into consideration the data of usage [6,9,13]; however, other researchers improved the personalization process incorporating the knowledge that is underlied in the textual content [1,9,10,15] or structure [10] the site.

As a natural evolution to those approaches, the need for a solution that take into account the semantical information of the web site have been developed lately. Eirinaki et al. have developed the Semantic Web Personalization System (SEWeP) [3]. This work consist in combining web usage with content Knowledge. Then, they developed an enhanced version of the web logs registers which are called C-Logs (concept logs). These C-Logs consist of web sites’ semantic information that is added to the traditional usage logs in the way of keywords. Afterwards, these C-Logs are used in the mining process to obtain better and broader recommendations.

On other hand, Knowledge Discovery in Text (KDT) concerns to the application of Knowledge Discovery in Databases (KDD) techniques over free text. Loh et al. use KDT process for developing a Concept-Based Knowledge Discovery process for texts [5]. In that work, the KDT techniques are applied over concepts rather than on attribute values, terms or keywords labeling texts. Then statistical analysis are performed to obtain interesting patterns. One of Lohs’ objectives was to allow the user to search ideas, ideologies, trends and intentions presents on text.

3 Concept-Discovery in Text for Personalization

3.1 Concept Representation Model

The concept representation is not an easy task, inclusive the meaning of the word concept also is quite ambiguous. A concept from dictionary is an “idea, opinion or thought”. We can understand from this definition that the concepts

have relation with ambiguity and also with the subjectivity that is commonly used by humans for performing different tasks, like, for instance, browsing a web site to obtain information about some topic.

Concepts are represented by a coherent combination of words [3,5]. However, in order to express an idea or event, we need to understand that not all words represent a concept in the same level, degree or context. For example, in Spanish the word “cancelar” (cancel) means “to stop doing something” however, it also means “to pay a bill”. In this example we have two different concepts represented by one word depending on the context. Another example are the phrases “i bought a dog” and “i didn’t buy a dog” the concept represented in one case is buying a pet but in the other phrase is the negation of that. For the explained reason, we need to represent the words with a weight that show the degree of relation that a term has to represent a specific concept.

From several approaches to represent concepts, we chose the Vector Space Model [12], which consist in transforming the text of the web site in a vectors of words. Each document is represented by a N dimensional vector, where N is the number of different words the whole site.

Chakrabarti [2] said that is better to use weighted vector instead of a binary model because the precision is higher. This is why we use a simple weight calculated using TF*IDF for each term and then we normalize the weight vector for each web page.

However, this weight only give some hints about the relative importance of the terms on the specific document. We still need to define how these terms represent a concept.

3.2 Definition and Identification of Concepts

For defining concepts, we used a dictionary of synonyms and antonyms for Spanish. It is important to differentiate between words in the text of the Web pages and terms. We use terms to refer those words that are used to represent a concept.

We based our work in the one proposed by Loh [5], his idea is to use a *fuzzy reasoning* model to decide wether a concept is expressed by a web document or not. Computing the possibility for a concept to be on a text based on the weights obtained before and the membership values from the terms that represent a concept. The existence of Necessary Conditions (NC) and Sufficient Conditions (SC) allows to perform such task. If a SC is present then the presence of a concept is mandatory ($TERM \Rightarrow CONCEPT$). While NC are the consequence of the presence of a concept ($CONCEPT \Rightarrow NC$). In the case of this work we will use only the NC. It means that if a term that defines a concept appears in a web page text, then there is a high possibility this concept belong to the document. According to this we generate a list of terms that define a concept based on the definition extracted from the dictionary. In this first proposal we set up a simple binary weigh system: 1 if the term is a synonym or 0 if antonym or any other word that it is not related with the concept. This system is very simple and does not take into account quasi synonyms or context of terms.

When having these two vectors, one for the words of the Web pages and other for the concepts, we need to apply our fuzzy reasoning model. We used Eq. (1) from [5] for such purpose.

$$[Concepts \times Words] = [Concepts \times Terms] \circ [Terms \times Words] \quad (1)$$

In the Eq.(1) “ \circ ” represent a combination between two fuzzy relations and the symbols “[\times]” represent a fuzzy relation (can be a matrix).

After running the process that compares all the word vectors with the terms vectors we write a file for each web page that contains the concepts for the page.

4 Concept-Based Session Analysis

4.1 Important Concept-Based Web Pages Vector

Assuming that the degree of importance in some web page content is correlated with the time spent on it by the visitors, we can state that those pages where a visitor spends more time are those more interesting to him. To represent this idea Velásquez et al. in [14] defined the ι -Most important pages vector, however, now we need to slightly change that definition to incorporate the notion of concepts. Thus we redefined the ι -Most important pages vector as follows:

Def. 1 (ι -Most important concept-based pages vector). *We define a vector of two components $\vartheta_\iota = [(\kappa_1, \tau_1), \dots, (\kappa_\iota, \tau_\iota)]$ where the pair $(\kappa_\iota, \tau_\iota)$ represents the ι^{th} most important page based on the time spent. Then κ is the vector of concepts that represent a page and τ is a scalar value to represent the percentage of time spent in it within a visitor session.*

For applying successfully the Def. 1 we need to develop a similarity measure that allows to compare vectors. In a previous work, the *Important Visited Pages Similarity* (IVS) [11,14] was introduced. However, this similarity is not suitable to be used in the present work because it do not use the ι -Most important concept-based pages vector. Therefore, IVP similarity uses a combination of the relative time spent in two web pages and the textual content of those pages.

Now, we need a new similarity that allows combine the relative time spent with the concepts web pages visited. Fortunately, if we use the new definition ι -Most important concept-based pages vector, the changes on IVS are minimal. Then it is possible to modify IVS in order to include the concepts, as we show in Eq.(2). We called this expression *Important Concepts-Based Visited Pages Similarity* (ICVS).

$$ICVS(S^i, S^j) = \sum_{p=1}^{\iota} \min\left(\frac{S_\tau^i(p)}{S_\tau^j(p)}, \frac{S_\tau^j(p)}{S_\tau^i(p)}\right) * PD(S_\kappa^i(p), S_\kappa^j(p)) \quad (2)$$

The expression shown in Eq.(2) compares the ι -most important concept-based pages vectors into the sessions of two different visitors S^i and S^j . On the other

hand, the function $PD()$ is introduced in Eq.(3). This is the way in which we combine the content of the site with the visitors browsing preferences. The term $S_\tau^i(p)$ represent the time spent on page p for the visitor i . Similarly, $S_\kappa^i(p)$ are the concepts that represents the page p for the visitor i .

$$PD(p_i, p_j) = \frac{\sum_{k=1}^W p_{ki} p_{kj}}{\sum_{k=1}^W (p_{ki})^2 \sum_{k=1}^W (p_{kj})^2} \tag{3}$$

The *Page Distance* introduced in Eq.(3) is the dot product between two vectors p_i and p_j .

When two visitors sessions are similar in browsing time $\min(\frac{S_\tau^i(p)}{S_\tau^j(p)}, \frac{S_\tau^j(p)}{S_\tau^i(p)}) \approx 1$ and if the sessions are similar in content $PD(S_\kappa^i(p), S_\kappa^i(p)) \approx 1$ furthermore $IVS(S^i, S^j) \approx 1$. In the opposite case, if the text contents are dissimilar then $PD(S_\kappa^i(p), S_\kappa^i(p)) \approx 0$ the expression $IVS(S^i, S^j) \approx 0$. On the other hand if the times spent are very different then $\min(\frac{S_\tau^i(p)}{S_\tau^j(p)}, \frac{S_\tau^j(p)}{S_\tau^i(p)}) \approx 0$.

One important observation is that, even though the expression for *ICVS* is almost the same with the *IVS*, the results are totally different because, they use totally different processing vectors.

4.2 Analyzing the Visitor Behavior in a Web Site

We used a *Self Organizing Feature Map* (SOFM) as clustering method to discover significant patterns from the combination of the web pages' concepts and the visitors spent time per page. We select a toroidal topology because it maintain the continuity of the space [11,14,15].

The SOFM is randomly initialized. This means that the neurons feature vectors, which are normalized, are created with random values between [0, 1] in the epoch $t = 0$.

We use a Gaussian function that depends on the distance from the centroid to propagate the learning to the neighbor neurons. This function allows the centroid neuron to learn the pattern shown. Afterwards, the effect of the learning is passed to the neighborhood to a lesser degree, inversely proportional to the centroid distance.

We applied the *ICVS* shown in Eq.(2) to compare the sessions examples with the documents concepts.

At the end of the process we can take the SOFM and we applied a technique that we called Reverse Clustering Analysis (RCA). The RCA technique for WUM is explained in detail in [11]. This technique allows discover which are real pages that are in the clusters of the SOFM and for this way, to create content recommendations for the web site.

5 Experiments in a Real Web Site

The whole process explained before was applied to the web site of the School of Engineering and Sciences of the University of Chile.¹ This web site has 182 web

¹ <http://escuela.ing.uchile.cl>

pages and it is almost static throw the year (only the news page change continuously), thus we used the version of December 2005 of the web site. Besides, we took approximately 2 months of web logs November and December 2005.

The length of the ι -most important concept-based pages vector was set in three pages. Therefore, we needed the sessions which contain at least three pages visited to create those vectors in order to apply the Definition 1. To do so, we sorted the sessions by time spent on each page and then we only kept the three pages where the visitor spent more time.

This experiment was performed using a combination of web pages content and the visitors sessions. First, we clean the Web pages text with a stop words list, then we applied a stemming process. The concepts used in this case were the concepts form the titles of the web pages and links. We discover that several times we had titles that have different words however, this pages use synonyms of the words. Afterwards, we compute the TF*IDF for all the concepts.

The next step was to apply a sessionization process on the web log registers, i.e., to reconstruct the original visitor session.

As final step, a visitor behavior pattern extracting process was carry out by using a SOFM of 8×8 . The results are shown in Figure 1. The results in this case were just four main clusters for the 64 documents used before.

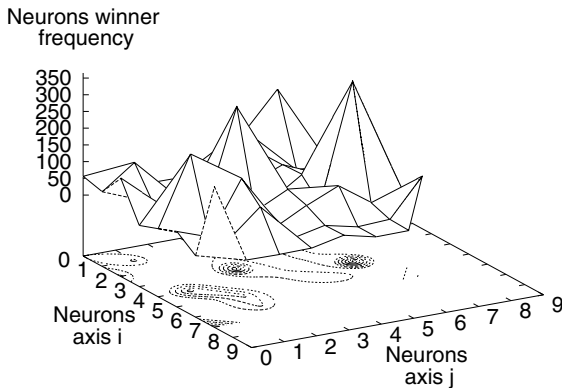


Fig. 1. Results using the concept-based approach

One example, we have many pages with title: “Faculty of Science Physics and Mathematics”, “School of Engineering”, “Faculty of Engineering”, etc. The vector that represent the words are very different one to each other. This is why in the traditional approach we obtained several clusters although in the concept based approach this clusters are all one, because all of them are referring to the School of Engineering.

The concepts reduce the amount of words to process in the creation of a vectorial representation for a web page. An example of this is what happen in the particular case of the analyzed web site. The compound words “Faculty of Science Physics and Mathematics”, “School of Engineering” and “Faculty of

Engineering” have the same semantic meaning, so they are similar when the Eq. (3) is applied.

From the clusters extracted, it is possible to extrapolate that the visitors are interested in:

- Test calendar, which is expressed for the concepts “prueba” (test), “control” (a monthly examination), and “examen” (the semester final examination). All of these words appear in different pages, then a recommendation is to create a unique page with the whole information.
- Educational material, which is expressed for the concepts “cátedra” (main lecture), “clase auxiliar” (lecture for resolving problems and exercises), “tareas” (homework) and “laboratorios” (laboratories). These concepts appear in different pages, which is no bad, but alternatively, could be necessary a unique page that concentrate the whole information

6 Conclusions

We show our first attempt to obtain concept-based mining technique which allows to obtain patterns that have more relation with the visitors goals. The process then can be used for the off-line personalization of a web site, in order provide text content and structure recommendations for modifying the web site.

This web mining approach allows to analyze a web site from the concept point of view, i.e., now the question about what the visitor is looking for change from “which words?” to “which idea?”.

Because before to apply our concept-base web mining algorithm it is necessary to reduce the page text content to concepts, there is a manual previous stage, i.e., a human being must to read the page text and to specify which concepts are inside.

As future work, we want to develop a semi-automatic preprocessing algorithm for extracting concepts from a web page text content.

Acknowledgment

This work has been funded partially by the Millennium Scientific Nucleus on Complexes Engineering Systems, Chile.

References

1. B. Berendt and M. Spiliopoulou. Analysis of navigation behavior in web sites integrating multiple information systems. *The VLDB journal*, 9:27–75, 2001.
2. S. Chakrabarti. Data Mining for Hypertext: A Tutorial Survey. *SIGKDD Explorations*, 1, 2000.
3. M. Eirinaki, C. Lampos, S. Paulakis, and M. Vazirgiannis. Web personalization integrating content semantics and navigational patterns. In *WIDM '04: Proceedings of the 6th annual ACM international workshop on Web information and data management*, pages 72–79, New York, NY, USA, 2004. ACM Press.

4. M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM Trans. Inter. Tech.*, 3(1):1–27, 2003.
5. S. Loh, L. K. Wives, and J. P. M. de Oliveira. Concept-based knowledge discovery in texts extracted from the web. *SIGKDD Explor. Newsl.*, 2(1):29–39, 2000.
6. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, 2000.
7. M. D. Mulvenna, S. S. Anand, and A. G. Buchner. Personalization on the net using web mining: introduction. *Commun. ACM*, 43(8):122–125, 2000.
8. J. Nielsen. User Interface directions for the web. *Communications of ACM*, 42(1):65–72, 1999.
9. M. Perkowit. *Adaptative Web Sites: Cluster Mining and Conceptual Clustering for Index Page Synthesis*. PhD thesis, Univerity of Washington, 2001.
10. M. Perkowit and O. Etzioni. Adaptive web sites. *Commun. ACM*, 43(8):152–158, 2000.
11. S. A. Ríos, J. D. Velásquez, H. Yasuda, and T. Aoki. Web Site Improvements Based on Representative Pages Identification. In S. Zhang and R. Jarvis, editors, *AI 2005: Advances in Artificial Intelligence: 18th Australian Joint Conference on Artificial Intelligence*, volume 3809, pages 1162–1166, Sydney, Australia, November 2005. Lecture Notes in Computer Science.
12. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM archive*, 18(11):613–620, November 1975.
13. M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS J. on Computing*, 15(2):171–190, 2003.
14. J. D. Velásquez, S. A. Ríos, A. Bassi, H. Yasuda, and T. Aoki. Towards the identification of keywords in the web site text content: A methodological approach. *International Journal of Web Information Systems*, 1(1):11–15, March 2005.
15. J. D. Velásquez, H. Yasuda, T. Aoki, and R. Weber. A new similarity measure to understand visitor behavior in a web site. *IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization*, E87-D(2):389–396”, February 2004.

New Network Management Scheme with Client's Communication Control

Kazuya Odagiri¹, Rihito Yaegashi², Masaharu Tadauchi², and Naohiro Ishii³

¹ Aichi Institute of Technology, 1-38-1 Higashiyamadouri Chikusa-ku
Nagoya-city Aichi, Japan
odagiri@toyota-ti.ac.jp

² Toyota Technological Institute, 2-12-1 Hisakata Tenpaku-ku
Nagoya-city Aichi, Japan
{rihito, tadauchi}@toyota-ti.ac.jp

³ Aichi Institute of Technology, 1-38-1 Higashiyamadouri Chikusa-ku
Nagoya-city Aichi, Japan
ishii@in.aitech.ac.jp

Abstract. Where customers with different membership and position, use computers as in the university network systems, it often takes much time and efforts for them to cope with the change of the system management. This is because the requirements for the respective computer usage are different in the network and security policies. In this paper, a new destination addressing control system (DACS) scheme for the university network services is proposed. The DACS Scheme performs the network services efficiently through the communication management of a client. As the characteristic of DACS Scheme, only the setup modification is required by a system administrator, when the configuration change is needed in the network server. Then, the setup modification is unnecessary by a customer, which shows a merit for both a system administrator and a customer. This paper describes the instruction and the prototype for DACS Protocol as the implementation of DACS Scheme. Then, the simplicity of the system management in DACS Scheme, is examined from the customer and the system administrator viewpoints.

1 Introduction

As the characteristic of the operation and management in the university network systems, it is pointed out that people with different membership and position as students, faculties, external persons, et al., use the network services comparatively freely. In the business corporations, it is comparatively easy to spread the information of the network usage based on a network policy or a security policy. However, in the university, it is often difficult to spread the information of the network usage, since the computer management section does not perform all operation and management for the respective needs. Although the system administrator of the computer section, carries out management and operation of the most network infrastructure and servers, the customer mainly performs the management of their clients [1]. Operation and management of the network system, are conventionally focused on the control in the infrastructure or server side [2] [3]. For example, DNS round robin [4], the control

using the load balancer and the load distribution of the server [5] [6] [7], are performed at the infrastructure or server side. When the configuration change of a server is carried out, it is necessary to make a setup change at the client side. For example, the environment where student uses a notebook-sized personal computer, is assumed. When comfortable internet environment is needed to be secured by establishing an outside telecommunication line for exclusive use of a classroom, it is necessary to reconnect to the PROXY Server, which is connected to the outside telecommunication line by setting change of the Web browser. In such a case, if the system administrator is able to control the communication freely between the server and client, it is not necessary to make setup change at the client side.

In this paper, a new DACS (destination addressing control system) scheme for the university network services, is proposed. The DACS Scheme performs the network services efficiently through the communication management. As the characteristic of DACS Scheme, only the setup modification is required by the system administrator, when the configuration change is needed in the network server. Then, the setup modification is unnecessary for the customer, which shows a merit for both a system administrator and a customer. This paper proposes the design of the DACS Scheme. The experimental evaluation is performed in the DACS Protocol.

2 Synopsis of DACS Scheme

2.1 Basic Principal of DACS Scheme

Fig.1 shows the basic principle of the network services by DACS Scheme.

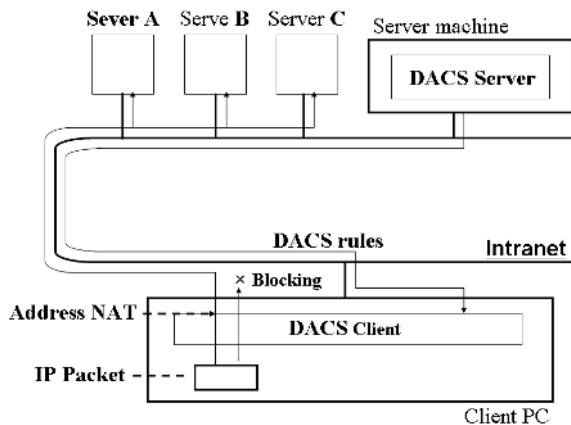


Fig. 1. DACS Scheme

At the timing of the (a) or (b) as shown in the following, DACS rules (rules defined by the user unit) are distributed from DACS Server to DACS Client.

- (a) At the time of a user logging in the client
- (b) At the time of a delivery indication from the system administrator

According to distributed DACS rules, DACS Client performs (1) or (2) operation as shown in the following. Then, communication control of the client is performed for every login user.

- (1) Destination information on IP Packet, which is sent from application program, is changed.
- (2) IP Packet from the client, which is sent from the application program to the outside of the client, is blocked.

An example of the case (1) is shown in Fig1. The system administrator can distribute a communication of the login user to the specified server among servers A, B or C. Moreover, the case (2) is described. For example, when the system administrator wants to forbid an user to use MUA (Mail User Agent), it will be performed by blocking IP Packet with the specific destination information. In order to realize DACS Scheme, the operation is done by DACS Protocol as shown in Fig.2. As shown by (a) in Fig.2, the distribution of DACS rules are performed on communication between DACS Server and DACS Client, which is arranged at the application layer. The application of DACS rules to DACS Control is shown by (b) in Fig.2. The steady communication control, such as a modification of the destination information or the communication blocking is performed at the network layer as shown by (c) in Fig.2.

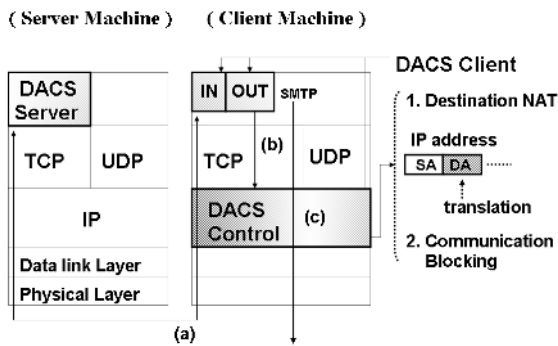


Fig. 2. Layer setting of DACS scheme

2.2 Comparison with the Existing Technology

Here, the difference between DACS Scheme and the existing technology is explained. Specifically, the difference from the technology of name resolution service (ex, WINS,DNS) and server load balancing is discussed. First, the difference from the name resolution service is explained. Although the mapping of a host name and an IP address is performed in the existing name resolution service, the mapping of the group of a host name, a user name and an IP address can be performed altogether by DACS Scheme. As the result, the IP address to be different for every user can be determined for the same host name. Next, the difference from server load balancing technology is explained. To realize server load balancing, there are methods by DNS round robin, and by the load balancer. Then, the difference from how to use the load balancer using Destination NAT is explained. The large difference from DACS

Scheme is the place which arranges Destination NAT. Although the load balancer arranges Destination NAT on the network course, it is arranged on the client in DACS Scheme. When Destination NAT is arranged on the network course, it cannot be specified whether IP Packet was sent by which user. For the reason, it is difficult to control communication per user. However, it can be guaranteed in DACS Scheme by arranging on the client that all IP Packet at the time of Destination Nat conversion is sent by the login user. But, when the client is multi-user system, the mechanism in the no login from remoteness is required. It is confirmed that the communication is sent by the user who sits down before a client and logs in directly, by the method of intercepting the unnecessary communication from the client outside.

3 Experimental Results by Prototype Construction

In order to prove the possibility of realization of the network services by DACS Scheme, the prototype was built. Then, the functional test was actually carried out under the operation. As the result of prototype construction, the function of changing a communicating PROXY server by a system administrator is realized as shown in Fig.3. When a PROXY Server A is set as reference PROXY server of the Web browser on a client, communication is done via PROXY Server B by the control of DACS Control. The confirmation by the way of PROXY Server B is identified in the access log of squid. The confirmation of no communication via PROXY Server A was also identified in the access log of squid.

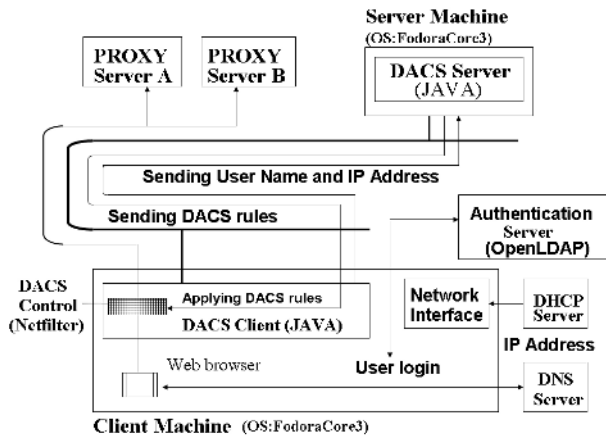


Fig. 3. Summary of prototype

3.1 Displayed Result of the Rule Table

At first, the result of a rule table on DACS Server is shown in Fig.4. The rule table is the thing for registering DACS rules. DACS rules which are registered to the user name transmitted from DACS Client are extracted. Then, they are transmitted to DACS client.

```
dacs_db=# select * from rule_table;
 user1 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user2 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user3 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user4 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user5 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user6 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user7 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user8 | tcp:192.168.1.1:3128-192.168.1.2:3128
 user9 | tcp:192.168.1.1:3128-192.168.1.2:3128
```

Fig. 4. Window results of the rule table

3.2 Displayed Result After the Application of DACS Rules

The result after the application of DACS rules from DACS Server to DACS Client (DACS Control) is shown in Fig.5. In this prototype, the functionality of Netfilter is used for DACS Control, and the iptables command is used for the application of DACS rules. The list of the rules is presented.

```
Chain POSTROUTING (policy ACCEPT)
target    prot opt source                destination

Chain OUTPUT (policy ACCEPT)
target    prot opt source                destination
DNAT     tcp  --  anywhere              192.168.1.1           tcp dpt:squid to:192.168.1.2:3128
DNAT     tcp  --  anywhere              192.168.2.1           tcp dpt:http to:192.168.2.2:80
DNAT     tcp  --  anywhere              192.168.3.1           tcp dpt:smtp to:192.168.3.2:25
```

Fig. 5. Window results after the application of DACS rules

4 Discussion of Effectiveness

In this chapter, a discussion is performed from the viewpoint of a customer and a system administrator about the effectiveness by DACS Scheme, which works well in the operation and management of network services. In DACS Scheme, the centralized management of communication by a system administrator is possible. Then, the effectiveness is evaluated from both sides of a customer and a system administrator as shown in the following.

At First, Effectiveness from the customer's viewpoint is described. By the communication control of a system administrator, the subsequent modification of setups is not needed. As the result, the user can use network services continuously without being conscious of a configuration change of the network server. In these days, university gives student a notebook-sized personal computer. At such a university, the time taken for a student to maintain a notebook-sized personal computer, becomes longer. If the network services by DACS Scheme is performed, the time and effort for students is saved, and the burden on student is reduced.

Next, Effectiveness from the system manager's viewpoint [8][9] is described. A modification of the existing system is unnecessary except building DACS Server on the server, and building DACS Client on each client. Furthermore, since a communication of the client can be performed by applying DACS rules and the existing system is continued without an outage of a server or a client, which shows affinity with the existing system. Therefore, affinity with the existing system is high.

Then, because the customer does not need to change the setups of network in DACS Scheme, the system administrator can realize system change easily. For example, the case where the network server software is changed, is assumed. When introducing new network server software, both verification of its function and verification to the load are required. By the conventional method, it is difficult to do a load examination besides using special verification software etc, after performing functional verification. By the DACS Scheme, it becomes possible to divide all the users into five equally, and to shift a user gradually every $1/5$ for example. Since the number of users is increased one by one gradually as a load examination in an actual environment, it can be checked whether it can bear to the number of users. Network environment can be changed safely.

In addition, there is a reduction effect of customers support. The support about the setups of network services is frequently taken place. When performing a update and configuration change of the server by conventional method, the change notice of network application may be needed to the customer, but it is not needed in DACS Scheme. Moreover, except for the initial installation and setups in the client, it is not needed that the customer changes the setups of the communication server of network. For this reason, the mistakes of the setups by the customer is not made after the initial introduction. As long as there are no mistakes by the system administrator, the inquiry from the customer about the setups of the communication server will be few. Then, the burden for the system administrator is reduced.

5 Communication Control on Client

From chapter 1 to chapter 4, the communication control on every user was given. However, it may be better to perform communication control on every client instead of every user. For example, it is the case where many and unspecified users use a computer room, which is controlled.

At First, the method of communication control on every client is described. When a user logs in to a client, the IP address of the client is transmitted to DACS Server from DACS Client. Then, if DACS rules corresponding to IP address, is registered into the DACS Server side, it is transmitted to DACS Client. Then, communication control for every client can be realized by applying to DACS Control. In this case, it is a premise that a client uses a fixed IP address. However, when using DHCP service, it is possible to carry out the same control to all the clients linked to the whole network or its subnetwork for example.

Then, the coexistence method with the communication control on every user is considered. When using communication control on every user and every client, communication control may conflict. In that case, a priority needs to be given. The judgment is performed in the DACS Server side as shown in Fig.6. Although not necessarily stipulated, the network policy or security policy exists in the organization such as a university (1). The priority is decided according to the policy (2). In (a), priority is given for the user's rule to control communication by the user unit. In (b), priority is given for the client's rule to control communication by the client unit. In (c), the user's rule is the same as the client's rule. As the result of comparing the conflict rules, one rule is determined respectively. Those rules and other rules not overlapping are

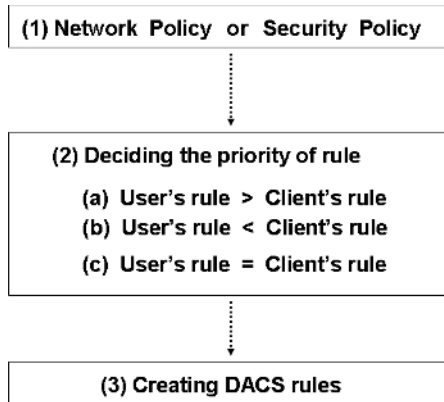


Fig. 6. Creating DACS rules in the DACS Server side

gathered, and DACS rules are created (3). DACS rules are transmitted to DACS Client. In the DACS Client side, DACS rules are applied to DACS Control. The difference between the user's rule and the client's rule is not distinguished.

6 Conclusion

As a way for increasing the efficiency of an operation and management for network services, DACS scheme is proposed here. The characteristic feature of the operation and management by DACS scheme is that the centralized administration by the system manager is possible after once the customer performs the initial setups. For the reason, it is not necessary to change the setups on client. Moreover, communication server is determined and selected services are given for every user by performing the management of a user and DACS rules. DACS protocol required to realize DACS scheme was described, and the prototype was actually built. Then, experimental result was shown. The study was discussed from the viewpoint of the customer and the system manager about the effectiveness of the operation and the management of the network services. For the customer, the load intensity of a management is reduced, such as changing the setups of the client, which shows as an advantage of the proposed DACS scheme. On the other hand, since affinity with the existing system is high, for a system manager, utility value is high at the following points.

- The operation and management after an initial introduction of DACS scheme or an introduction are very easy.
- After starting the operation and management by DACS scheme, a change of servers can be made freely and safely.
- There is an effect which reduces customer supports.

A construction of the whole system for the real operation, and implementation, will be done as a future project.

References

1. S.Heilbronner, and R.Wies,:"Managing PC networks", IEEE Commun.Mag.,vol.35,No.10,pp.112-117,Oct.,1997.
2. J.Chauki,M.andM.Shahsavari,:"Component-based distributed network management", Proc.Southeast con 2000, pp.460-466,IEEE Pub.,2000.
3. L.Raman,:"OSI systems and network management",IEEE Commun.Mag.,vol36,No.3,pp46-53, Mar.,1998.
4. T.Shimokawa,Y.Koba,I.Nakagawa,B.Yamamoto, and N.Yoshida,:"Server Selection Mechanism using DNS and Routing Information in Widely Distributed Environment", IEICE Tran. on Communications,vol.J86-B,No.8,pp.1454-1462,Aug.2003.
5. S.K.Das,D.J.Harvey, and R.Biswas,:" Parallel processing of adaptive meshes with load balancing", IEEE Tran.on Parallel and Distributed Systems, vol.12,No.12,pp.1269-1280,Dec.,2002.
6. M.E.Soklic,:"Simulation of load balancing algorithms: a comparative study", ACM SIGCSE Bulletin,vol.34,No.4,pp.138-141,Dec.,2002.
7. J.Aweya, M.Ouellette,D.Y.Montuno,B.Doray, and K.Felske,:"An adaptive load balancing scheme for web servers," Int.,J.of Network Management.,vol.12,No.1,pp.3-39,Jan/Feb.2002.
8. A.Konno,T.Yoshimura,H.Hashima,Y.Iwatani,T.Abe, and T.Kinoshita,:"Network Management Support System Based on the Activated Status Information",IPSJ Journal,vol.46,No.2,pp.493-505,Feb.,2005.
9. M.Kawashima, and M.Matsushita,:"Application of HTTP Protocol for Enterprise Network Management",IEICE Tran. on Communications, vol.J82-B,No.3,pp.339-346,Mar.,1999.

Prediction of Electric Power Generation of Solar Cell Using the Neural Network

Masashi Kawaguchi¹, Sachiyoichi Ichikawa¹, Masaaki Okuno¹, Takashi Jimbo²,
and Naohiro Ishii³

¹ Department of Electrical & Electronic Engineering, Suzuka National College of Technology,
Shiroko, Suzuka Mie 510-0294, Japan

{masashi, masaaki}@elec.suzuka-ct.ac.jp
<http://www.suzuka-ct.ac.jp/>

² Department of Environmental Technology and Urban Planning Graduate School of
Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan

jimbo.takashi@nitech.ac.jp

³ Department of Information Network Engineering, Aichi Institute of Technology,
Yachigusa, Yagusa-cho, Toyota, 470-0356 Japan

ishii@aitech.ac.jp

Abstract. We proposed the prediction system of electric power generation of solar cell using neural network. Recently, the solar cell system is developing in many fields. However this system is easily to influence by the weather condition. In the practical application, it has been required the prediction of electric power generation. By this system, it is possible to make the planning of supply and the security of alternative power source. This prediction system is used neural network system and it can predict the integral power consumption, largest electric power and time-serial prediction.

1 Introduction

In this study, we proposed the prediction system of electric power generation of solar cell using neural network. Recently, the solar cell system is developing in many fields. However this system is easy to influence by the weather condition. In the practical application, it has been required the prediction of electric power generation.

By this, it is possible to make the planning of supply and the security of alternative power source. This prediction system is using neural network and it can predict the integral power consumption, largest electric power and time-serial prediction[1].

2 Overview

The solar cell which we have used is shown in Fig. 1. We showed the rating of the solar cell in Table.1.



Fig. 1. The solar cell used for the measurement

Table 1. The rating of solar cell used in the experiment

	Value	Remarks
Max Power	70(W)	
Isc	4.6(A)	Short circuit current
Voc	21.0(V)	Open circuit voltage
Iop	4.1(A)	Operating current
Vop	17.0(V)	Operating Voltage
Area	5,400cm ²	

The fill factor(*F.F.*) means the optimal function point. Maximum output can be taken out from the solar cell. When *Voc* is equal to *Isc*, *F.F.* is almost equal to 1 on that time, a high output was obtained.

$$F.F. = \frac{Vop \times Iop}{Voc \times Isc} = \frac{Pmax}{Voc \times Isc} \tag{1}$$

2.1 Measurement

We measured current, voltage, short circuit current, open circuit voltage, fill factor, illuminance, temperature and weather in different conditions. We collected these data at 10:40A.M., 11:40A.M., 12:40P.M. and 1:40P.M.. The measurement days were March 29, May 24, June 19, August 6, December 11. We used these data as to leaning data. Another measurement days were April 26, May 6, July 8, July 22, September 4, November 22, January 13 and January 20. We used these data as to forecasting of power generation.

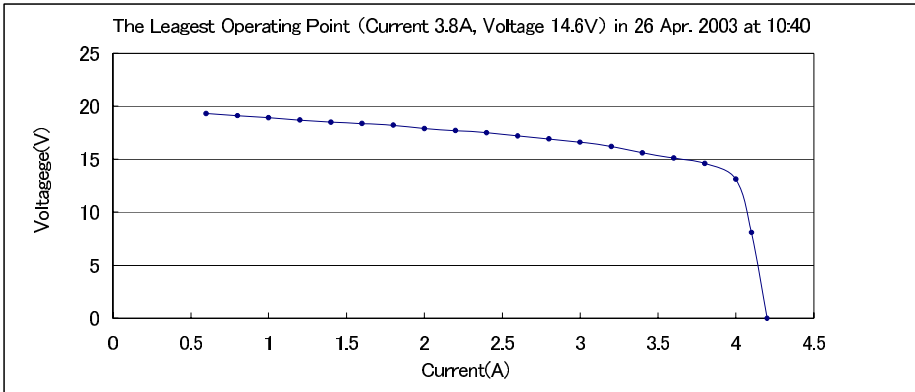


Fig. 2. The example of fill factor

2.2 Neural Network

We constructed neural network system as shown in Fig. 3. This neural network consists of three layers. There are 8 units in the first layer. 8 units mean the 8 measurement factors. There are another three units in the second and third layer. Three units in the third layer mean the generated electronic power at 11:40A.M., 12:40P.M. and 1:40P.M.. We will predict the power generation at these different time.

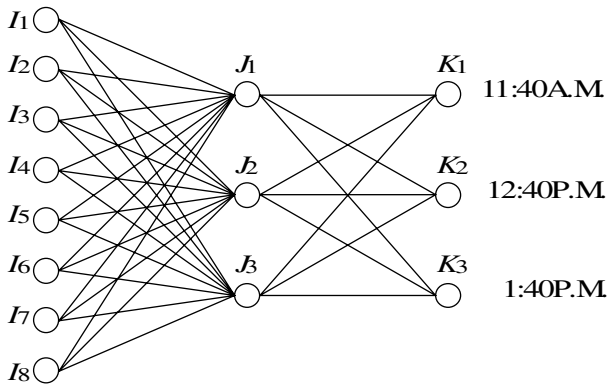


Fig. 3. Structure of Neural Network of this System

2.3 Convert the Measurement Data

In general, neural network can accept the input value in between 0 and 1. So, we converted the measurement data to the input data, which is between 0 and 1. These values are reported in Table. 2.

Table 2. The conversion style for changing the input signal from 0 within 1's

	Conversion style
Current	0.2*I (A)
Voltage	0.05*V(V)
Short circuit current	0.2*Isc(A)
Open circuit voltage	0.04*Voc(V)
Fill factor	F.F.
Illuminance	0.000005*S(lux)
Temperature	0.02*t(C)
Weather	1(fine), 0(rain)

Table 3. The example of learning data

	Mar.29	May.24	Jun.19	Aug.6	Dec.11
Current(I)	3.6	3.4	3	3.4	3
Voltage(V)	14.9	14.4	14.4	13.7	16.7
Short circuit current(A)	4.05	3.8	3.5	4	3.4
Open circuit voltage(V)	19.2	19.2	18.6	18.3	21
Fill factor	0.69	0.671	0.664	0.636	0.702
Illuminance(S)	101,600	101,700	93,600	105,100	79,000
Temperature(C)	21.5	29.6	33	38	7.2

Table 4. Power generation data as the teaching signal of neural network

	Mar.29	May.24	Jun.19	Aug.6	Dec.11
11:40	55.2	50.16	47.88	46.58	54.74
12:40	56	54	50.76	46.92	55.76
13:40	56	55.1	55.1	47.88	49.5

2.4 Learning and Prediction Using Neural Network

We constructed neural network system. We used here back-propagation model for data learning[2]. Input signal is measured each factor. At first, we used spring season data as teaching signal of this network learning. Then, we corrected data in all season as teaching signal. We predicted the power generation at 11:40A.M., 12:40P.M. and 1:40P.M. using 10:40A.M. data in the same day. We show the example of learning data in Table 3. In Table 4 means the teaching signal as for the power generation data of each time and date. We used equation (2) and equation (3) as the learning of this network.

$$w_{ij} = w_{ij} - 0.3 \sum (e_i j_j (1 - j_j) i_i) \tag{2}$$

w is connecting weight between input layer and middle layer. j is output value of middle layer. i is input value value of input layer. e is learning error.

$$v_{ij} = v_{ij} - 0.3 \sum (e_i k_j (1 - k_j) j_i) \tag{3}$$

v is connecting weight between middle layer and output layer. k is output value. j is output value of middle layer. e is learning error.

3 Experiment

We have shown the prediction error in Fig. 4. The average error rate is 8.54%. However, using all season data as teaching signal, the average error is only 4.15% shown in Fig. 5. Error means between predictive value and measured value[3].

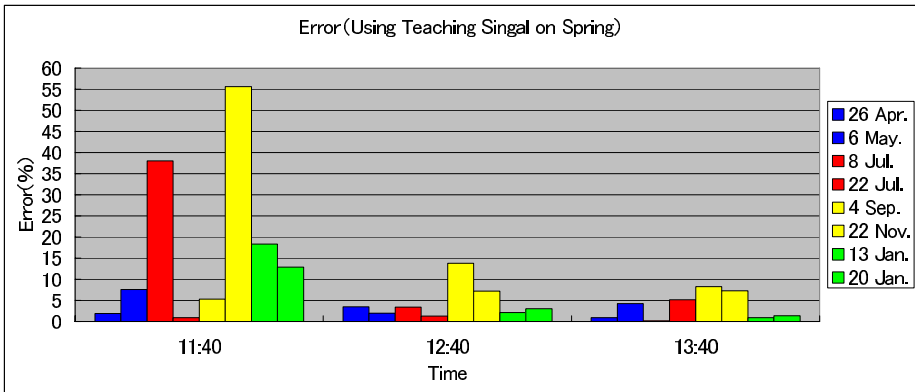


Fig. 4. Error for using teaching signal in spring

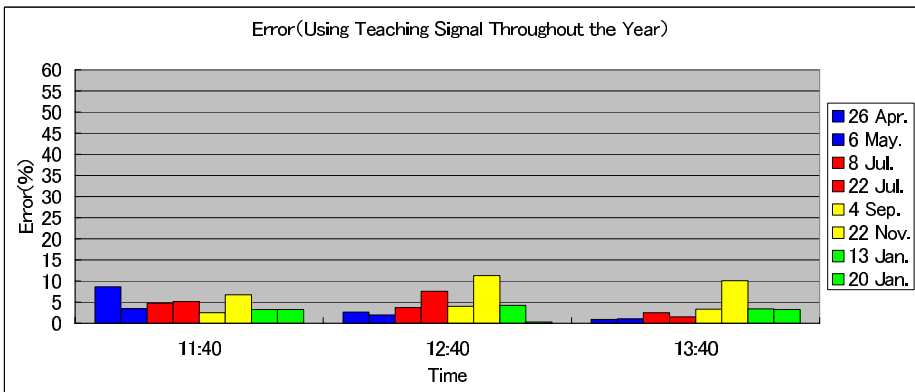


Fig. 5. Error for using teaching data throughout the year

4 Conclusion

The best average error rate for this system is 4.15%, using all season data as teaching signal. On the other hand, using teaching data in spring, average error rate is 8.54% however, error rate in the spring season is only 3.34%.

In the future study is as follows. We will collect more measurement data, training of this network in the every season. The error rate will be reduced.

Acknowledgements

This research was financially supported by Grant-in-Aid from the Japanese Ministry of Education, Science, Sports and Culture.

References

1. Yamamoto, S., Park, J.S., Takata, M., Sasaki, K., Hashimoto, T.: Basic study on the prediction of solar irradiation and its application to photovoltaic-diesel hybrid generation system, *Journal of Solar Energy Materials & Solar Cells*, Elsevier Science, Vol.75, No.3–4(2003) 577–584
2. Rumelhart, D. E., Hinton, G. E., Williams, R. J.: Learning internal representations by error propagation, *Parallel distributed processing*, MIT Press (1986) 318–362
3. Ichikawa, S., Kawaguchi, M., Okuno, M.: Measurement and Prediction of Electric Power Generation of Solar Cell Using the Neural Network, *Record of Tokai-Section Joint Conference of the Eight Institutes of Electrical and Relater Engineers* (2003) 34

Text Classification: Combining Grouping, LSA and kNN vs Support Vector Machine

Naohiro Ishii, Takeshi Murai, Takahiro Yamada, Yongguang Bao,
and Susumu Suzuki

Aichi Institute of Technology
Yachigusa, Yakusacho, Toyota, Japan 470-0392
ishii@aitech.ac.jp

Abstract. Text classification is a key technique for handling and organizing text data. The support vector machine(SVM) is shown to be better for the classification among well-known methods. In this paper, the grouping method of the similar words, is proposed for the classification of documents, which is applied to Reuters news and it is shown that the grouping of words has equivalent ability to the Latent Semantic Analysis(LSA) in the classification accuracy. Further, a new combining method is proposed for the classification, which consists of Grouping, LSA followed by the k-Nearest Neighbor classification (k-NN). The combining method proposed here, shows the higher accuracy in the classification than the conventional methods of the kNN, and the LSA followed by the kNN. Then, the combining method shows almost same accuracies as SVM.

1 Introduction

Text classification is a supervised learning task as assigning pre-defined category to new documents, which has become one of the key techniques for handling and organizing text data[1,2,5,6]. The text classification is used in not only automatic processing but also a wide field of internet search etc[7,8]. The treated document might be considered to be a vector that consists of the word and the appearance frequency of the word in the field of information retrieval and the information filtering. A typical classification technique includes the classification method with the rule base which decides for the document into the category to be classified. The support vector machine(SVM), which is a new learning method, is shown to be applicable to the text classification[10,11,12]. Among the developed classification methods, the k Nearest Neighbor (kNN) is a simple and useful technique. But, the kNN is expected to improve the classification accuracy, comparing with the accuracy of the classification by SVM.

The vector length of the document might have thousands of dimensions with words. The reduction of dimension has been paid to attention by analyzing the statistical appearance patterns by using the technique of the Latent Semantic Analysis (LSA)[3,4]. In this paper, the grouping method of the similar words, is proposed for the classification of documents, which is applied to Reuters international news and it is shown that the grouping of words has equivalent ability to the LSA in the

classification accuracy. Further, a new combining method is proposed for the documents classification, which consists of Grouping, LSA followed by the k-Nearest Neighbor classification (k-NN). The combining method shows the higher accuracy than the conventional method of the kNN, and the LSA followed by the kNN, which shows almost the same accuracy as SVM.

2 Vector Space Model

Documents are represented by vector space model. Document is characterized in a point of the vector space, since it consists of multiple words. The most commonly used document representation is so called vector space model. In the vector space model, a document is represented by a vector of words. Usually, one has a collection of documents which is represented by a word-by-document matrix A , where each entry represents the occurrences of a word in a document, i.e., $A = (a_{ik})$, where a_{ik} is the weight of the word i in the document k .

The number M of rows in matrix A , corresponds to the number of words in the dictionary. The k -th document is represented by the characteristic vector

$$x_k = (a_{1k}, a_{2k}, \dots, a_{Mk}),$$

where the suffix M of the vector component, a_{Mk} shows the number of the words in the dictionary, $dic = \{word_i | 1 \leq i \leq M\}$. Since the dimension of the vector space of documents, becomes very large in the vector space, it is necessary to reduce the dimension. To determine a_{ik} , the following approach is taken:

- (1) Term Frequency, which means by the occurrence of terms(words) in the document.
- (2) Document Frequency, which means by the occurrence of words in all the collected documents.

Let f_{ik} be the number of occurrence of the word i in the document k , N be the number of all the collected documents, M be the number of words after the removal of the unnecessarily words and n_i be the number of the occurrence of the word i in the all the collected documents. Then, the weight a_{ik} is determined by the following methods[1,2,7].

Boolean weighting. When there is the targeted word in the document, the weight a_{ik} becomes 1, otherwise 0 as follows,

$$a_{ik} = 1 \text{ if } f_{ik} > 0, \quad a_{ik} = 0 \text{ otherwise}$$

tf×idf – weighting. tf means by the term frequency, f_{ik} , while df means by the document frequency, n_i . The smaller n_i , will mean the high ability for

characterizing the word i . Thus, the inverse value of n_i , is computed. Then, $idf_i = \log(N/n_i)$ is defined. The weight of the word i in the document k , becomes

$$a_{ik} = f_{ik} \times \log(N/n_i)$$

3 k – Nearest Neighbor Classification

The k – nearest neighbor classification is carried out under the assumption that the similar documents will belong to the same category. The training of documents is shown by the 2-tuple $d = (\vec{x}, y)$, where \vec{x} shows the characteristic vector of the document and y shows the given category of the document for the trained documents set D . The algorithm for the k – nearest neighbor is as follows,

1) Let the trained documents be $d_{n1}, d_{n2} \dots d_{nk} \in D$, which are nearest neighbors of the unclassified document d_q . Then, the similarity of documents d_i and d_j , is computed as follows,

$$sim(d_i, d_j) = \cos(\text{vec } x_i, \text{vec } x_j),$$

where $\cos(\text{vec } x_i \cdot \text{vec } x_j)$ shows a inner product of vectors.

2) For the unclassified document d_q , the class category is determined from the normalized similarity computation denoted by $rank_{c_j}(d_q)$.

3) The unclassified document is assigned to all the class c_j , which satisfies the relation $rank_{c_j}(d_q) \geq \theta$, where θ is the threshold value.

4 Latent Semantic Indexing and Grouping of Words

Assuming that we have a $m \times n$ word-by-document matrix A , where m is the number of words, and n is the number of documents. The singular value decomposition of A is given by :

$$A = U \Sigma V^T,$$

where $U (m \times r)$ and $V (r \times n)$ have orthogonal columns and $(r \times r)$ is the diagonal matrix of singular values. $r \leq \min(m, n)$ is rank of A . If the singular values of A are ordered by size, the k largest may be kept and the remaining

smaller ones set to zero. The product of the resulting matrices is matrix A_k that is an approximation to A with rank k as shown in the following.

$$A_k = U_k \sum_k V_k^T,$$

where $\sum_k (k \times k)$ is obtained by deleting the zero rows and columns of \sum and $U_k (m \times n)$ and $V_k (n \times k)$, are obtained by deleting the corresponding rows and columns of U and V , respectively.

The similarity between words, is carried out by the cosine of two rows in the approximated matrix A_k or $U_k \sum_k$, which is derived from the word-by-document matrix A . The set of similar words, will be useful for the classification of documents. The set is made from the aggregation of similar words by the following procedure. Let the word in the document, be $k_i (i=0,1,\dots,n)$ and each set be $K_j (j=0,1,\dots,m (<n))$, where K_j is represented as follows,

$$K_j = \bigcup_{s \leq r(k_i, k_l)} k_i$$

The relation $s \leq r(k_i, k_l)$ shows the similarity relation $r(k_i, k_l)$ between words, k_i and k_l , which is computed from cosine of words described above. The s is the threshold of the similarity relation. By the similarity relation of the given threshold, the words are grouped into the same class, which is assumed as a new word K_j . Thus the aggregated set of the new words, is made as $\{K_j\}$. The schematic diagram of grouping process of words, is shown in Fig. 1. The d_1 shows a document consisting of words.

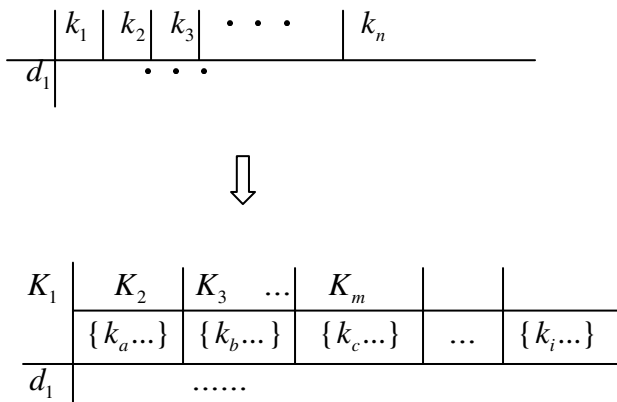


Fig. 1. Schematic diagram of grouping words

5 Support Vector Machine

SVM is a relatively new learning approach by Vapnik for solving two class pattern recognition problems[10,11,12]. In SVM, the problem is to find a decision surface, which separates data points in two classes.

6 Experimental Method and Results

The computer experiments were carried out by the well known news data called Reuters21578[9], which consists of international politics and economical news documents. Documents in Reuters21578, are represented in SGML and they are assigned to the class category. In this study, the documents are classified to 10 categories (cocoa, copper, cpi, gnp, rubber, fuel, gold, jobs, alum, coffee) as shown in Table 1[9]. Classification of Reuters news data, was carried out by the conventional kNN method and the proposed combining method of grouping, LSA and kNN.

To measure the classification accuracy in the class c_i , three indexes, $Recall_{c_i}$, $Precision_{c_i}$, and $Accuracy_{c_i}$ are defined as follows[1,5,7],

$Recall_{c_i}$. The ratio of documents classified to the class c_i within the total documents in the class c_i . $Precision_{c_i}$ The ratio of documents classified correctly in the class c_i within the documents assigned to the class c_i .

$Accuracy_{c_i}$. The ratio of documents classified correctly to the class c_i and other than c_i among all the documents. To measure and improve the inter-class accuracy, the indexes, micro-average, μ and macro-average, M are introduced. To compute the best results in those weighting, the classification in kNN with tfidf weighting, showed the highest values in the data weighting as shown in Table 1. Thus, we applied tfidf weighting method.

Table 1. Classification indexes in kNN (tfidf weighting)

Class	kNN (tfidf weighting)		
	Recall	Precision	Accuracy
alum	0.6316	0.9230	0.9523
cocoa	0.8461	1.0000	0.9881
coffee	1.0000	0.8276	0.9702
copper	0.7500	0.9231	0.9702
cpi	0.7500	0.9231	0.9702
fuel	0.4000	1.0000	0.9643
gnp	1.0000	0.5526	0.8988
gold	0.9259	0.8929	0.9702
jobs	0.7273	0.8000	0.9702
rubber	0.7273	0.8889	0.9762

Table 2. Classification indexes in kNN (tfxidf weighting)

micro-average	0.8155	0.8155	0.9631
macro-average	0.7758	0.8731	0.9631

The parameter k in the kNN, was chosen to be k=14, in the experiments. The bold numerals in Table 1 , shows the almost or same higher values than those in the combining Grouping, LSA and kNN in Table 3 . The dimension used in this experiment, becomes 2453 in the vector space. The data dimension is reduced to 210 in the Grouping ,LSA and kNN method.

Table 3. Classification indexes in combining Grouping, LSA and kNN

Class	kNN (Grouping + LSA + kNN)		
	Recall	Precision	Accuracy
alum	0.6842	1.0000	0.9643
cocoa	0.8461	1.0000	0.9881
coffee	1.0000	0.8571	0.9762
copper	0.8125	1.0000	0.9821
cpi	0.8125	0.9286	0.9762
fuel	0.5000	1.0000	0.9702
gnp	1.0000	0.5676	0.9048
gold	0.9630	0.8966	0.9762
jobs	0.7273	1.0000	0.9821
rubber	0.8182	0.9000	0.9821

Table 4. Classification indexes in combining grouping, LSA and kNN

micro-average	0.8512	0.8512	0.9702
macro-average	0.8164	0.9150	0.9702

Table 3 shows the classification indexes in the proposed method here, which consists of Grouping of data, followed by LSA and finally kNN processing. The bold numerals in Table 3, show the higher values of classification indexes in the proposed method than those indexes of the kNN method in Table 1. Similarly, the micro-average and the macro-average, in Table 4, shows the higher values than those in kNN in Table 2.

Classification results by SVM are shown in Fig.5. The bold character shows the better or same values than those by the proposed combining kNN method in Table 3 in the numerical values within four digits below the decimal point. Also, similar better results are shown in Table 6. However, the results by SVB are superior to the combining kNN within 0.5% in average of accuracy. This shows almost the same results between SVM and the combining kNN, which is improved in the accuracy than the kNN method only.

Table 5. Classification indexes in SVM

Class	Recall	Precision	Accuracy
alum	0.8421	1.0000	0.9821
cocoa	0.8462	1.0000	0.9881
coffee	1.0000	0.7058	0.9405
copper	0.9375	0.9375	0.9881
cpi	0.9375	0.9375	0.9881
fuel	0.5000	1.0000	0.9702
gnp	1.0000	0.7777	0.9643
gold	0.9629	1.0000	0.9940
jobs	0.8182	0.9000	0.9821
rubber	0.5454	0.8571	0.9643

Table 6. Classification indexes in SVM

micro-average	0.8810	0.8810	0.9762
macro-average	0.8390	0.9116	0.9762

7 Conclusion

Text and document classification still have received a lot of attention by the unsupervised manner. Classification accuracy is an important index to compare developed methods. It is desired to improve the classification accuracy as much as possible. This paper proposes a combining classification method, which consists of data grouping ,reduction of data dimension(LSI), followed by the kNN method. Text classification was carried out by the conventional kNN method, which is well-known method. But, the kNN method is weak in noises and is intolerant of the irrelevant attributes. In this paper, we developed a classification method, which consists of Grouping, followed by LSA and finally kNN processing. The combining method shows almost the same classification accuracy as SVM.

References

1. Grossman,D.A. and Frieder,O., Information Retrieval - Algorithms and Heuristics- , Springer-Verlag, pp.332, 2004
2. Sebastiani,F., “ A tutorial on automated text categorization “, Proc. of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, pp.7-35, Buenos Aires, 1999
3. Derrwester,S., Dumais,S.T., Furnas,G.W., Landauer,T.K. and Harshman,R., “ Indexing by latent semantic analysis “, Journal of the American Society for Information Science, No.41, pp.391-407, 1990
4. Landauer,P.W., Folz,T.K., and Laham,D., “ Introduction to latent semantic analysis”, Discourse Processes, No.25, pp.259-284, 1998

5. Sebastiani,F., “ Machine learning in automated text categorization”, ACM Computing Surveys, Vol.34, No.1, pp.1-47, 2002
6. Bao,B., and Ishii,N., “ Combining multiple k-nearest neighbor classifiers for text classification by reducts”, Proc. 5th Int. Conference on Discovery Science (Lecture Notes in Artificial Intelligence, Vol.2534, Springer- Verlag), pp.361-368, 2002
7. Sirmakessis,S., Text Mining and its Application, Springer-Verlag, pp.204, 2003
8. Baldi,P., Frasconi,P., and Smyth,P., Modeling the Internet and the Web, Wiley, pp.285, 2003
9. <http://www.research.att.com/~lewis/reuters21578.html>
10. Cortes,C., and Vapnik,V., “ Support vector networks”, Machine Learning, Vol.20, pp.273-297, 1995
11. Yang,Y., and Liu,X., “ A re-examination of text categorization methods”, Proc. of ACM SIGIR Cof. On Res. And Development in Information Retrieval, SIGIR’ 99,pp.42-49,1999
12. Joachims,T., “A statistical learning model of text classification for support vector machines”, Proc. of ACM SIGIR Cof. On Res. And Development in Information Retrieval, SIGIR’ 01,pp.128-136,2001

Particle Filter Based Tracking of Moving Object from Image Sequence

Yuji Iwahori¹, Toshihiro Takai², Haruki Kawanaka³,
Hidenori Itoh², and Yoshinori Adachi¹

¹ Chubu University, Matsumoto-cho 1200, Kasugai 487-8501, Japan
iwahori@cs.chubu.ac.jp, adachiy@isc.chubu.ac.jp
<http://www.cvl.cs.chubu.ac.jp>

² Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
toshi@center.nitech.ac.jp, itoh@ics.nitech.ac.jp

³ Aichi Prefectural University, Nagakute-cho, Aichi-gun 480-1198, Japan
kawanaka@ist.aichi-pu.ac.jp

Abstract. Object tracking is an important topic in computer vision and image recognition. The probabilistic approach using the particle filter has been recently used for the tracking of moving objects. Based on our trajectory recording system of the soccer scene with multiple video cameras at one view point, we propose the extended approach to increase the tracking robustness and accuracy using the particle filter. The proposed approach makes it possible to pass the necessary particle information using the color histogram and other key factors from one image to the next image, which are taken through the different camera scene with one PC. The performance of the proposed approach is evaluated in the experiments with real video sequence. It is shown that one PC can handle two video images in real-time.

1 Introduction

For motion tracking, particle filter [1]-[4] is one of the techniques for robust tracking in the presence of occlusion and noise. Particle filter is called Bayesian filter or Sequential Monte Carlo method (SMC), which are sophisticated model estimation technique based on simulation, and it is used to estimate Bayesian models.

Particle filter is a maximum posteriori estimation method based on the past and the present observations. It also achieves robust tracking for the case when the observation distribution is non-Gaussian. It approximates the discrete probability density where the random variables are represented by many particles. In this sense, the particle filter is used not only in the field of motion tracking but also in the field of speech recognition or other applications. Some researches treat the human motion tracking [5][6] or human head tracking. Combination of the particle filter with other algorithms makes the performance stronger.

In the application of motion tracking, the trajectory recording system of soccer playing game is one example of the motion tracking researches [7]. Our recent research proposes the motion extraction from three video cameras. This system

uses three video images with a single view point to cover the wide area of soccer playing field, and the fourth PC generates the trajectory based on the outputs obtained from three PCs via three video cameras.

The purpose of this system is to decrease the computation cost of image processing using multiple PCs. While, as the recent PC increases the performance of computation processing, one PC with the high performance has the ability to handle and process the multiple video sequences. Particle filter processing can decrease the processing cost rather than the background subtraction of image.

In this paper, we propose a particle filter based new approach for the moving object tracking. The proposed system supports the passing of information from one camera to another camera and the tracking results can be obtained using one PC. That is, the proposed approach can be also applied for the multiple cameras. The approach can perform the robust tracking with high speed. Experiments on real data are demonstrated.

2 Probability Based Tracking

Time Sequential Filtering: Time sequential filtering is a method to estimate the most suitable value from the past and present observation values. Let the state of tracking target at time t be \mathbf{x}_t , and let the observation result from image be \mathbf{z}_t . Let the observation results by time t be $\mathbf{Z}_t = (\mathbf{z}_1, \dots, \mathbf{z}_t)$. The probability density is discretely approximated by many particles with the state and the likelihood. The robust tracking to both the noise and the variation of environment is performed.

Weighting Sampling: Particle filter approximates the posterior $p(\mathbf{x}_t|\mathbf{Z}_t)$ at time t with N particles which consist of the state \mathbf{x} and its weight. Weight $\pi_t^{(i)}$ for the state $\mathbf{x}_t^{(i)}$ at time t for i -th hypothesis is evaluated by the likelihood function $p(\mathbf{z}_t|\mathbf{x}_t = \mathbf{x}_t^{(i)})$.

Particle Filter Based Tracking: Tracking with hypothesis is realized by repeating the following processes.

1. Sampling of hypotheses $\{\mathbf{s}_{t-1}^{(1)}, \dots, \mathbf{s}_{t-1}^{(N)}\}$ using the weight $\pi_{t-1}^{(i)}$ based on the state $\mathbf{x}_{t-1}^{(i)}$ of particles $\left\{ \left(\mathbf{x}_{t-1}^{(i)}, \pi_{t-1}^{(i)} \right), i = 1, \dots, N \right\}$ which approximates the posterior distribution $p(\mathbf{x}_{t-1}|\mathbf{Z}_{t-1})$ at time $t-1$.
2. Generate N hypotheses $\mathbf{x}_t^{(i)}$ at time t from the sampled hypotheses $\mathbf{s}_{t-1}^{(i)}$.
3. Likelihood function from $\mathbf{x}_t^{(i)}$ and the weight $\pi_t^{(i)}$ of $\mathbf{x}_t^{(i)}$ are calculated. Here, the weight is normalized so that $\sum_{i=1}^N \pi_t^{(i)} = 1$ holds.

Particles $\left\{ \left(\mathbf{x}_t^{(i)}, \pi_t^{(i)} \right), i = 1, \dots, N \right\}$ are obtained as a discrete approximation of posterior distribution $p(\mathbf{x}_t|\mathbf{Z}_t)$ at time t . Mean value of the hypotheses is used as the estimated state for the tracked target at time t .

3 Motion Tracking with Probabilistic Approach

3.1 Procedure

Four probability variables (*width, height, x, y*) of a rectangle are used. That is, a rectangle connotes the tracking target, whose center is located at the image coordinate (x, y) with *width* and *height*.

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ width \\ height \end{bmatrix} \tag{1}$$

where *width* means the length of a rectangle along *x*-axis direction, while *height* means the length of a rectangle along *y*-axis direction.

1. Particles are sampled for the hypothesis $\{\mathbf{s}'_{t-1}(1), \dots, \mathbf{s}'_{t-1}(N)\}$ based on the weights $\pi_{t-1}^{(i)}$ for the state $\mathbf{x}_{t-1}^{(i)}$ of particles $\left\{ \left(\mathbf{x}_{t-1}^{(i)}, \pi_{t-1}^{(i)} \right), i = 1, \dots, N \right\}$, which is the approximation of posterior distribution $p(\mathbf{x}_{t-1} | \mathbf{Z}_{t-1})$ at time $t - 1$. Sampling is done according to the following.

(a) Accumulative weight $c_{t-1}^{(i)}$ is calculated by

$$\begin{aligned} c_{t-1}^{(0)} &= 0 \\ c_{t-1}^{(i)} &= c_{t-1}^{(i-1)} + \pi_{t-1}^{(i)} \quad (n = 1, \dots, N) \end{aligned}$$

(b) Uniform random variable r whose $r \in [0, 1]$ is generated. Select the minimum j which holds $c_{t-1}^{(j)} \geq r$. Take the sampled hypothesis as $\mathbf{s}'_{t-1}(i) = \mathbf{x}_{t-1}^{(j)}$.

(c) Repeat (b) from $i = 1$ to $i = N$.

2. N hypotheses $\mathbf{x}_t^{(i)}$ at time t are generated based on the probabilistic dynamics $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ from the sampled hypothesis $\mathbf{s}'_{t-1}(i)$. $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ is defined as

$$p\left(\mathbf{x}_t = \mathbf{x}_t^{(i)} | \mathbf{x}_{t-1} = \mathbf{s}'_{t-1}(i)\right) = \begin{bmatrix} x'_{t-1}(i) + \omega_x \\ y'_{t-1}(i) + \omega_y \\ width_{t-1:t}^{(i)} \\ height_{t-1:t}^{(i)} \end{bmatrix} \tag{2}$$

where ω_x and ω_y are the Gaussian noises. $width_{t-1:t}^{(i)}$, $height_{t-1:t}^{(i)}$ are the *width* and *height* estimated from the result at time $t - 1$.

3. Calculate the likelihood function $p(\mathbf{z}_t | \mathbf{x}_t = \mathbf{x}_t^{(i)})$ from $\mathbf{x}_t^{(i)}$, then the normalized weight of $\mathbf{x}_t^{(i)}$ is calculated as $\pi_t^{(i)}$.

Likelihood is calculated using the color information included in the rectangle which covers the tracking target and it is given by.

$$p(\mathbf{z}_t | \mathbf{x}_t) = e^{kS^2} \tag{3}$$

where k is a constant and S is the similarity of the histogram using Swain's histogram intersection [8]. S is given by

$$S = \sum_{i=1}^M |h_{image}^i - h_{ref}^i| \tag{4}$$

where h_{image}^i is the color information within the rectangle \mathbf{x}_t , while h_{ref}^i is the reference color histogram obtained *a priori*. M is the size of the histogram.

Normalized weight for each hypothesis Eq.(3) is obtained as follows.

$$\pi_t^{(i)} = \frac{p(\mathbf{z}_t | \mathbf{x}_t = \mathbf{x}_t^{(i)})}{\sum_{i=1}^N p(\mathbf{z}_t | \mathbf{x}_t = \mathbf{x}_t^{(i)})} \tag{5}$$

Estimated result \mathbf{x}_t^e at time t is given by

$$\mathbf{x}_t^e = \begin{bmatrix} x_t^e \\ y_t^e \\ width_t^e \\ height_t^e \end{bmatrix} = \sum_{i=1}^N \pi_t^{(i)} \mathbf{x}_t^{(i)} \tag{6}$$

3.2 Estimation of Target Size

The proposed approach further introduces the estimation method of the target size to track for higher precision. Perspective projection assumes that the camera center coordinates (X, Y, Z) is projected to the image coordinates (x, y) . The perspective pinhole camera gives

$$x = f \frac{X}{Z'} \quad y = f \frac{Y}{Z'} \tag{7}$$

Z' is the distance of object from the lens along Z -axis. Let S be the area of the object projected onto the parallel plane to the image plane, then the corresponding area S' on the image plane is estimated by

$$S' = f^2 \frac{S}{Z'^2} \tag{8}$$

As far as S is obtained according to Z' and S' *a priori*, S' can be used as the estimated value of target size on the image plane.

3.3 Tracking over Multiple Camera

Tracking over multiple camera is proposed. [5] treats the multiple camera but it does not use the camera calibration. As a result, when the target passes to the next camera, the tracking precision becomes worse based on the data acquisition to calculate the visual region of each camera. This approach uses two fixed cameras, instead the necessary information for the tracking is passed through the camera calibration based on the assumption that the target moves on the plane. Fig.1 shows examples. When a target 1 enters into the overlapped region C, the tracking information of a target 1 is passed to the processing in camera 2. While, when a target 2 enters into the overlapped region B, the information is passed to the processing in camera 1.

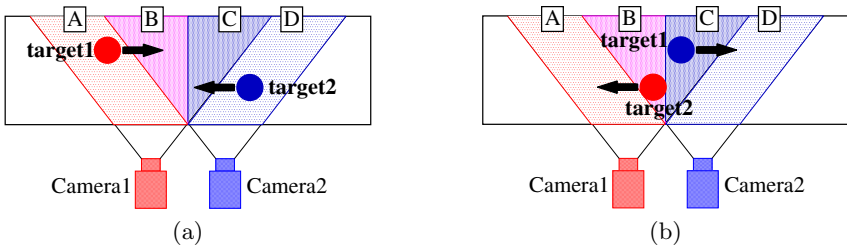


Fig. 1. Examples of Tracking for Overlapped Regions

Overlapped region is obtained from four corner points of the image of each camera by the projection and calibration. As far as the overlapped region for each camera is calibrated, this idea can be expanded to the system with multiple more cameras. The corresponding 3D world coordinate distance between each camera and each point of the overlapped region is calculated, then the camera for the nearest point processes the tracking for the overlapped region.

4 Experiments

The proposed approach is evaluated through the experiments. Video camera (SONY DCR-TRV22K) captures the scene of the running radio control cars with MPEG1 files of 720×480 pixels with 24 bits color. Specification of PC used is Athlon XP 2200+ CPU with MM 512MB.

Target size is estimated and evaluated for the tracking precision. Scene used in the evaluation is that a red radio control car runs inside the room with large movement. Number of particles for a tracking target was set to be 20, i.e., $k=20$ in Eq.(3) in the experiments. HS (Hue and Saturation) information was used as the reference color histogram.

Fig.2 shows the example of tracking. Rectangle of each particle and the estimated result \mathbf{x}_t^e at time t are shown. For the frames which a car passes through the feature point on the floor, the error was evaluated 10 times in comparison

Table 1. Mean Error for Probabilistic variables

	x	y	$width$	$height$
Estimation of Target Size	3.233333	2.566667	7.233333	4.100000
No estimation of Target Size	9.300000	5.933333	15.900000	11.666667

with the human eye check. Mean error of estimating the probabilistic variables in the various distance points is shown in Table 1.

Estimation of coordinate of tracking object was acceptable for both cases, but the rectangle (width and height) became more stable with estimating the center coordinate with the estimation of target size. In this case, target color information becomes more stable by capturing the precise rectangle. The performance of the proposed approach is 7 msec/frame. This means the tracking processing is done in real-time.



Fig. 2. Left: Efficiency of Knowledge for Target Size, Right: No Estimation for Frame No.741

Table 2. Mean Error for Probabilistic Variables

	x	y	$width$	$height$
Tracking with Multiple Cameras	4.975000	3.760000	4.250000	6.750000

Next, the experiment was done with one PC for two camera images. Fig. 3 shows the resulting images. Two (red and blue) radio control cars are running over the overlapped region through two camera images. Black rectangle represents the tracking result \mathbf{x}_t^e at time t . The evaluated result suggests that it took 8ms for one car, and 16ms for two cars. The real-time processing can be realized. Fig.3 shows the correct passing of target through the overlapped region.

The variables $width$ and $height$ are transmitted directly to another camera image. Table 2 shows the mean error of this case. Small error for $height$ is observed in Table 2. This error occurs because of the difference of camera angles. If camera angles are taken into account, further improvement is expected.

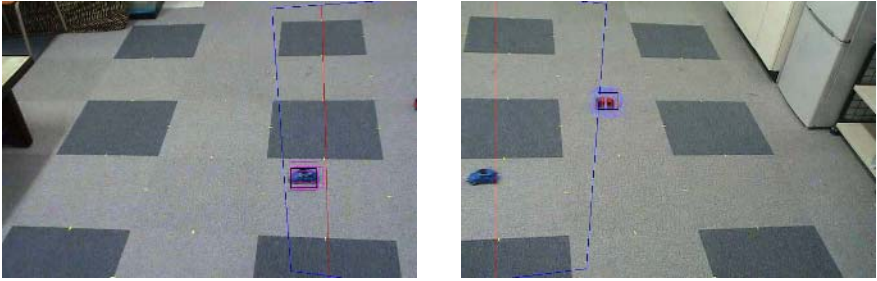


Fig. 3. Tracking with Multiple Camera (Left and Right Camera) for Frame No.766

5 Conclusion

A new approach is proposed for real-time tracking of moving objects to increase the accuracy under the fixed video cameras. The similarity of color histogram effectively worked in the particle filter based tracking approach. The approach proposed the idea to pass the variables of particle filter through the overlapped region between cameras. The accuracy was increased with the estimation of target size with 3D distance and the approach shows that one PC can handle two video images in real-time.

The assumption used in this paper is the color of the moving object is different from the background color. In this sense, tracking of multiple objects with the similar color information is still difficult and remains as the future subjects.

Acknowledgment

Iwahori's research is supported in part by Chubu University Grant, and the JSPS Grant No.16500108. The authors also would like to thank the related member of the Research Institute for Information Science, Chubu Univ. for their support.

References

1. N. J. Gordon, D. J. Salmond, A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation" *IEE PROCEEDINGS-F*, Vol.140, No.2, 1993.
2. M. Isard, A. Blake, "CONDENSATION - conditional density propagation for visual tracking" *International Journal of Computer Vision*, Vol.29, No.1, pp.5–28, 1998.
3. A. Doucet, S. Godsill, C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering" *Statistics and Computing*, vol.10, no.3, pp. 197–208, 2000.
4. J. S. Liu R. Chen, "Sequential Monte Carlo Methods for Dynamic Systems" *Journal of the American Statistical Association*, vol.93, no.443, pp.1032–1044, 1998.
5. S. Khan, M. Shah, "Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View" *IEEE Trans. on PAMI*, Vol.25, No.10, pp.1355–1360, 2003.

6. M. J. Hossain, K. Ahn, J. H. Lee, O. Chae, "Moving Object Detection in Dynamic Environment" KES2005, LNAI 3684, pp. 359–365, 2005.
7. A. Yamada, Y. Shirai, J. Miura, "Tracking Players and a Ball in Video Image Sequence and Estimating Camera Parameters for 3D Interpretation of Soccer Games" *Proceedings of ICPR2002*, 2002.
8. M. J. Swain, D. H. Ballard, "Color Indexing" *International Journal of Computer Vision*, Vol. 7, pp.11–32, 1991.
9. P. Perez, C. Hue, J. Vermaak, M. Gangnet, "Color-Based Probabilistic Tracking" *Proceedings of ECCV2002*, vol.1, pp.661–675, 2002.

Discrete and Continuous Aspects of Nature Inspired Methods

Martin Macaš, Miroslav Burša, and Lenka Lhotská

Gerstner Laboratory, Czech Technical University in Prague
Technická 2, 166 27 Prague 6, Czech Republic

mmacas@seznam.cz, bursam@fel.cvut.cz, lhotska@fel.cvut.cz

Abstract. In nature, industry, medicine, social environment, simply everywhere we find a lot of data that bear certain information. A dictionary defines data as facts or figures from which conclusions may be drawn. Data can be classified as either numeric or nonnumeric. The structure and nature of data greatly affects the choice of analysis method. Under the term structure we understand the facts that the data might be not a single number but n-tuples of measurements. Structure is also very closely linked to the reason of data collection and method of measurement. The paper describes the similarities and differences of nature inspired methods and their natural counterparts in light of continuous and discrete properties. Different examples of nature inspired methods are inspected in terms of data, problem domains and inner structure and principles.

1 Introduction

In nature, industry, medicine, social environment, simply everywhere we find a lot of data that bear certain information. A dictionary defines data as facts or figures from which conclusions may be drawn. Data can be classified as either numeric or nonnumeric.

In most real-world applications, we face the problem of heterogeneous data. Even if we leave structural or textual information out of consideration and assume presence of numerical data only, we find several types of data, namely binary, categorical, integer, continuous, fuzzy, and temporal data. For example, in medical domain the binary variable may represent presence or absence of a certain symptom, integer values may represent blood pressure, heart rate, continuous data represent measured signals.

The structure and nature of data greatly affects the choice of analysis method. Very frequently, development of any nature inspired method is based on attempts to model natural objects or processes. The resulting method has very often many properties similar to but also different from its natural counterpart. The paper describes these similarities and differences in light of continuous and discrete properties. Different examples of nature inspired methods are inspected in terms of data, problem domains and inner structure and principles.

2 Nature of Data

Data can be classified as either numeric or nonnumeric. With respect to this classification we can describe data in the following way. Qualitative data are nonnumeric and usually represent some descriptive features (e.g. colours, type of material, subjective view – poor, fair, good, better, best). Qualitative data are often termed categorical data. Some literature sources use the terms individual and variable to reference the objects and characteristics described by a set of data. They also stress the importance of exact definitions of these variables, including what units they are recorded in. The reason the data were collected is also important. Quantitative data are numeric. They are further classified as either discrete or continuous. Discrete data are numeric data that have a finite number of possible values. They are represented by integer or whole numbers. Continuous data have infinite possibilities and are represented by real numbers.

Data are collected by mapping entities in the domain of interest to symbolic representation by means of some measurement procedure, which associates the value of a variable with a given property of an entity. The relationships between objects are represented by numerical relationships between variables. Obviously the measurement process is crucial. It underlies all subsequent data analytic and data mining activities.

Let us briefly touch the issue of measurement. The experimental method depends on physically measuring things. The concept of measurement is related to the concepts of numbers and units of measurement. Statisticians categorize measurements according to levels. Each level corresponds to how this measurement can be treated mathematically. Measurements may be categorized in many ways. Some of the distinctions arise from the nature of the properties the measurements represent, while others arise from the use to which the measurements are put.

Nominal level of measurement is characterized by data that have no order and only give names or labels to various categories (for example marital status – single, married, widowed, divorced); they can be expressed in numerical representation. However, data cannot be arranged in an ordering scheme (such as low to high), or order is not meaningful. Ordinal level of measurement involves data that may be arranged in some order, but the interval between measurements is not meaningful (for example education scale) or cannot be determined. Interval level of measurement works with data that have meaningful intervals between measurements, but there is no natural zero starting point (where none of the quantity is present). Ratio level of measurement involves data that have the highest level of measurement. Ratios between measurements as well as intervals are meaningful because there is a starting point (zero).

There exist many taxonomies for measurement scales. Sometimes they are based not on the abstract mathematical properties of the scales but rather on the sorts of data analytic techniques used to manipulate them. Examples of such alternatives include counts versus measurements; nominal, ordinal, and numerical scales; qualitative versus quantitative measurements; metrical versus categorical measurements; and grades, ranks, counted fractions, counts, amounts, and balances. In most cases it is clear what is intended by these terms. Ranks, for example, correspond to an operational assignment of integers to the particular entities in a given collection on the

basis of the relative “size“ of the property in question: the ranks are integers which preserve the order property.

Sometimes raw data are not the most convenient form and it can be advantageous to modify them prior to analysis. There is a duality between the form of the model and the nature of the data. Certain transformations of the data may lead to the discovery of structures that were not at all obvious in the original scale. Discretization means transformation of continuous values into intervals. There are numerous discretization methods available in the literature. The choice is mostly dependent on the nature of the solved problem. In both transformation and discretization processes, it is necessary to satisfy the condition of minimum useful information loss.

3 Neural Networks

The real neuron differs from other cells in capability of altering its resting membrane potential and conducting signal. Generally, the neuron is a continuous processing element with inputs (dendrites) and an output (axon). The input synaptic potential perturbs the resting membrane potential and causes the neuron to be fired. An action potential is generated if the electric signal provided by the synaptic (or receptor) potential is bigger than certain threshold. In that case, the action potential is generated and further conducted along the axon towards a synaptic terminal. The axon potential has form of a spike of depolarization. Its amplitude and duration does not depend on amplitude and duration of the input signal. The amplitude of the input signal is transformed into frequency of an action potential sequence. At the end of axon, the number and frequency of action potentials stimulate the release of neurotransmitters – the frequency-modulated information is decoded into an amount of neurotransmitters which diffuses to the postsynaptic cell. The neurotransmitter finally causes an output synaptic potential to be generated. This mechanism invokes some discussion about discrete properties of real neuron. The threshold inside the neuron operates like a trigger and the neuron could be considered like having two discrete states – active and non-active. Although the inputs and outputs are continuous, the inner structure has certain discrete properties.

The artificial neuron is very simplified, but also very similar model of one type of real neuron. There are various neural models, however, when we consider the McCulloch and Pitts neuron based on logical calculus or Rosenblatt’s perceptron for real domain, the majority of the models includes a nonlinear activation function corresponding to the threshold element of real neuron described above [1]. It can be concluded that both the real neuron and its model have continuous inner structure with some discrete elements. Concerning the inputs and outputs, while the pre and postsynaptic potentials have continuous character, the artificial neurons and the majority of their networks can have any type of input and output data. However, the situation is more difficult and complex. Therefore we cannot analyse the whole spectrum of real neural structures and also a large number of artificial neural networks.

4 Particle Swarm Optimization

The PSO approach was originally introduced for multidimensional parameter optimization in [2]. The algorithm is inspired by social behaviour of bird flocking or fish schooling. The method uses group (swarm) of problem solutions (particles). Each solution consists of real valued vector of parameters and represents a point in multidimensional space. The particles fly through the space in an organized manner, measure their fitness and search for the global optima. Two kinds of information are available to the particles. The first is their own experience – they have tried the choices and know which state has been better so far and how good it was. The second information is social knowledge – the particles know how the other individuals in their neighbourhood have performed. Such information is used by particles for updating their velocities and position in order to effectively search the parameter space and reach the global optima.

It is clear, that the real bird moves in the real space continuously and in the view of PSO algorithm, it is searching for the best position which could be represented by 3-dimensional real valued vector. Maybe, this inspiration implicated the continuous character of problems mostly solved by PSO. However, there are also many discrete optimization problems and thus many attempts for solving these problems in terms of PSO. Two sorts of approaches exist. The first one is to modify the classical continuous PSO algorithm to work in binary or discrete domain. This approach comprises binary PSO or discrete PSO with crisp representation described below. The second approach is to propose special continuous representation of the discrete problem and use it with the classical (or modified) continuous PSO algorithm.

Binary PSO. As it was mentioned above, the search process in continuous PSO is an analogy of movement of individuals in real world space. On the other hand, the binary form of PSO has quite different real-world analogy. The work of binary PSO evokes rather a model of binary decision [3], where individuals are represented by their decision vector and the goal is to find optimal binary pattern of some choices. So, the binary PSO could be considered as inspired by a discrete natural decision process wherein the probability of making a binary decision is a function of personal and social factors. Although the binary PSO is designed for searching a binary pattern, the inner structure of the algorithm has some continuous properties. Each particle has its velocity, which represents continuous predisposition to make one or the other choice.

Discrete PSO. Let us have a discrete combinatorial problem: Let n_1, \dots, n_c be c finite lists of integer numbers. Let n_{ij} be the j^{th} value in the list n_i . The main goal is to find an optimal combination of c such values which are optimal in every sense.

Clerk [4] proposed two representations, the *crisp* representation and the *fuzzy* representation. The first one introduced special form of velocity and special operators for PSO equations. The position x was a list of c possible values (c -dimensional vector). The PSO using such representation and modifications worked (found the optimum), however the results were very sensitive to PSO parameter settings and the algorithm was not robust – even a small change in coefficients in the velocity update formula could give a far better or far worse results.

The problem with robustness in crisp representation was partly solved by the use of fuzzy representation. In this representation, each component in the position vector

was not only a single number, but also a fuzzy set on the possible values. For example, if the first component of searched vector could take k discrete values, the first component of fuzzy position was composed of k confidence coefficients on $(0, 1)$. This enabled the particles to move more continuously. The velocity was represented similarly, but the confidence coefficients could have any real value. During each iteration, two additional operators were needed – normalization, which ensured all coefficients in a fuzzy position to be in $\langle 0, 1 \rangle$ and defuzzification which transformed the fuzzy position to crisp position onto which a fitness function could be applied. The resulting algorithm was much more robust – the results were less dependent on settings of PSO parameters.

The utilization of continuous PSO for solving a discrete combinatorial problem was also demonstrated in other studies [5] and experimental results have shown that it was worthwhile to transform the discrete problem to continuous fuzzy space. However, for the fuzzy representation, the particle does not consist of c -dimensional vector, but of $\sum_{i=1 \dots c} k_i$ coefficients, where k_i is a number of possible values on the i -th position in the searched combination. It significantly increases the dimensionality of the searched space.

The PSO involves another difference between real and artificial swarms. This difference is comprised in temporal properties of fitness evaluation. While the birds and other real creatures evaluate their current situation continuously, in PSO, we can only evaluate the fitness of the current position at discrete time intervals [6]. This difference causes something like aliasing effect and in connection with introduction of “short range forces” can be used for some improvements of behaviour of PSO.

5 Genetic Algorithms

Genetic algorithms are inspired by evolution process in nature. The natural evolution is based on natural selection, mutation and crossover. The individuals are represented by their genomes. Each genome carries the complete set of instruction for making an organism. The information is coded using four bases included in DNA. The bases are guanine, cytosine, thymine and adenine and form four-letter alphabet. Thus, an individual is represented using discrete (ternary) coding. The similar approach is used in genetic algorithm, where the binary coding is mostly used. However, the selection of type of chromosome coding is quite dependent on concrete task to be solved. For example TSP problem is often solved using permutation encoding or GA based neural network training uses real-valued encoding. However, representation of an individual is very often discrete.

On the other hand, inner structure of natural and artificial evolution processes seem continuous from all points of view. The natural selection is consequence of complex mixture of many processes with mostly continuous character. The selection in GA is based on fitness evaluation of individuals. The fitness function has in most cases continuous values and the selection process itself (roulette wheel, rank selection, steady-state selection, etc.) has not any discrete property too. The mutation is random process without any discrete properties, therefore the evolution, whether natural or artificial, can be considered as continuous process.

6 Ant Systems

This nature inspired method is based on observation of *foraging* (food gathering) of ants [7]. As the ants move, they leave (secret) a substance called *pheromone* on their trail. The amount of the pheromone laid is proportional to the number of ants that used the route. Thus, the pheromone might be considered as being updated in discrete steps. However, the pheromone not only evaporates in time but often the ants deposit different amounts of pheromone depending on the quality of solution (food source) found. Therefore this value might better be considered continuous. Pheromone evaporation helps to avoid local minima and allows dynamic adaptation when the problem (route, food source) changes.

Combinatorial problems. The continuous space search can be reduced to a graph search. When an ant makes decision on its next move, it considers the amount of the pheromone on the route and performs a stochastic decision (natural decision process). It chooses with higher probability the route where the amount of pheromone is higher. In real nature, the ant performs a continuous space search. The ant colony optimization approach [9] makes a simplification of the continuous space to a space of interconnected locations – a graph structure. This model can be effectively used to solve the well-known Travelling Salesman Problem [9] – a well-known static NP-hard problem or (for example) an adaptive routing problem in telephone networks [11] – a dynamic problem.

Continuous optimization. The continuous space search (continuous optimization) using the model of an ant colony has also been performed [10]. However, this approach uses certain form of discretization. The authors [10] suggest using a finite set of regions in each iteration of the algorithm: agents are sent to these regions from which they randomly explore selected directions within the range of exploration. Thus this search performs a coarse-grained search with local optimization, which is much more informed than the real nature method – the real ants have to traverse the search space continuously.

Data clustering. Several species of ant workers have been reported to form piles of corpses (cemeteries) to clean up their nests. This aggregation phenomenon is caused by attraction between dead items mediated by the ant workers. The workers deposit (with higher probability) the items in the region with higher similarity (when more similar items are present within the range of perception). This approach has been modelled [8] to perform a clustering of data. The agents move in two-dimensional space and with probability based on the local similarity they randomly pick up or deposit items (data vectors). The space might be considered continuous; the data vectors can contain real values of practically any dimension. However, this approach (as all the clustering methods) is very sensitive to the similarity measure used (e. g. Euclidean distance, etc.) and the range of agent perception. No pheromone is used in this method.

The ants exist (search and optimize) in continuous space. However, they have to make some decisions, which can be considered as a binary or discrete component of their solution. On the other hand, the models use certain simplification of the search space (discretization, graph or network paradigm), which significantly improves the model efficiency and leads to elegant solution of the problem.

7 Conclusions

The aim of the paper was to present a survey of the most important nature inspired methods used in classification and optimization tasks in relation to their continuous and discrete properties and discrete and continuous data. Different examples of nature inspired methods have been inspected in terms of data, problem domains and inner structure and principles. Basically, there are certain similarities between PSO, GA and Ant Systems. However, there are also differences. The most important one concerns the problem of finding global optimum in optimization tasks.

The main task of natural neural network is to process and transfer information included in continuous signal. The information is imparted between neurons by the help of continuous synaptic potentials. However, receptors for example can represent a source of binary information. The inner structure of both the natural and artificial neuron is continuous with some discrete features (thresholds and activation functions). On the other hand, there are both the discrete and continuous types of tasks solved by artificial neural networks. For example, while the function approximation is continuous, the classification task could be considered as discrete. The artificial neural networks can handle arbitrary data types.

The natural evolution is continuous process from all points of view except the data representation by a ternary encoding. While the goal of Darwinian evolution is to evolve a well-adapted organism, the genetic algorithms can perform on combinatorial problems. The PSO method working in real values is more suitable for problems from continuous domain; however using a special representation can enable the PSO to solve discrete problems. The process of making a decision is discrete. However the mechanism of choice of a decision is a complex problem, which is far from being purely discrete. The inner structure of binary PSO inspired by these processes has continuous probabilistic character with certain thresholds.

In ant system modelling, the goal is to find a robust and efficient method to solve a problem and not to model the natural process as a whole. In the real nature world, the main goal of an ant system is to survive. This is accomplished by positive feedback and stigmergic communication. The main tasks performed (and modelled) are: food search, food transport and item clustering (brood sorting, item clustering). The food search is nothing more than a space search dealing with an infinite search space. The model [10] performs some discretization, because it is memory intensive (if even possible) to store pheromone information for the whole continuous space. The second task, food transport, is solved using pheromone routes, which are located in the continuous space and effectively change with time (to involve new food sources, avoid obstacles, etc.). The modelled optimization [9, 11] reduces the two-dimensional route space to a graph structure. The third main task modelled, clustering [8], is based on the similarity measure. The two dimensional continuous space is reduced to discrete space in the model.

In conclusion we can say that unlike other machine learning algorithms all mentioned above are able to work with both discrete and continuous data. When we want to perform discretization for some reason we have to keep in mind that potential information loss caused by application of unsuitable method can be crucial for the success or failure of the solving process.

Acknowledgement

The research was supported by the research program No. MSM 6840770012 "Transdisciplinary Research in the Area of Biomedical Engineering II" of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of the Czech Republic and by the FP6-IST project No. 13569 NiSIS (Nature-inspired Smart Information Systems).

References

- [1] McCulloch, W. S., Pitts, W. H.: A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, (1943) 5:115-133.
- [2] Kennedy, J. and Eberhart, R., "Particle Swarm Optimization", Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia, 1995, pp. 1942-1945.
- [3] Boyd, R, Richardson, PJ, Culture and Evolutionary Process. Chicago: University of Chicago Press. 1985
- [4] Clerc, M.: Discrete Particle Swarm Optimization: A Fuzzy Combinatorial Black Box, 2000, http://clerc.maurice.free.fr/ps0/Fuzzy_Discrete_PSO/Fuzzy_DPPO.htm
- [5] Fuzzy Discrete Particle Swarm Optimization for Solving Travelling Salesman Problem, Wei Pang, Kang-ping Wang, Chun-guang Zhou, Long-jiang Dong, Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04), (2004).
- [6] Hendtlass, T. and Rodgers, T., Discrete Evaluation and The Particle Swarm Algorithm, The 7th Asia-Pacific Conference on Complex Systems, 2004.
- [7] Deneubourg, J.-L., S. Aron, S. Goss, and J.-M. Pasteels. "The Self-Organizing Exploratory Pattern of the Argentine Ant." *J. Insect Behavior* 3 (1990)
- [8] Deneubourg, J.-L., S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chretien. "The Dynamics of Collective Sorting: Robot-Like Ant and Ant-Like Robot." In Proceedings First Conference on Simulation of Adaptive Behavior: From Animals to Animats, edited by J. A. Meyer and S. W. Wilson, 356-365. Cambridge, MA: MIT Press, 1991
- [9] Dorigo, M., V. Maniezzo, and A. Colomi. "The Ant System: optimization by a Colony of Cooperating Agents." *IEEE Trans. Syst. Man Cybern. B* 26 (1996): 29-41
- [10] Bilchev, G., and I. C. Parmee. "The Ant Colony Metaphor for Searchin Continuous Design Spaces." In Proc. of AISB Workshop on Evolutionary Computing Lecture Notes in Computer Science 993, edited by T. C. Fogarty, 25-39. Berlin: Springer-Verlag, 1995
- [11] Schoonderwoerd, R. O. Holland, J. Bruten, and L. Rothkrantz. "Ant-Based Load Balancing in Telecommunications Networks." *Adapt. Behav.* 5 (1996): 169-207

Social Capital in Online Social Networks

Przemysław Kazienko¹ and Katarzyna Musiał^{1,2}

¹ Wrocław University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

² Blekinge Institute of Technology, S-372 25 Ronneby, Sweden

kazienko@pwr.wroc.pl, weaving@wp.pl

Abstract. The problem of social capital in context of the online social networks is presented in the paper. Not only the specific elements, which characterize the single person and influence the individual's social capital like static social capital, activity component, and social position, but also the ways of stimulation of the social capital are described.

1 Introduction

Social networks are the information systems available online that have recently been rapidly developed. A social network consists of the set of even millions of people worldwide, who are in mutual relationships, e.g. *Friendster* or *LinkedIn*. Each its participant can be characterized by the value called *social capital* that can be stimulated in order to either make the network more persistent or spread it. In most cases, both, the community as a whole and each individual that belongs to the social network will benefit if the social capital of the members grows [2]. However, definitions of an online social network and social capital are not well established. Some research in this area should be done not only to specify what these concepts are, but also to enable the appropriate evolution of the social network. Stanley Milgram conducted the small-world experiment, which conclusion was that people in USA form the social network and they are connected with “six degrees of separation” [15]. A social network is the set of actors, i.e. group of people or organizations, which are nodes of the network, and ties, called also relationships that link the nodes [1, 9]. Social networks indicate the ways in which actors are related. The nodes and ties are usually presented by graphs (sociograms) or matrices [10]. The evolution of the social network depends on the mutual experience, knowledge, relative interpersonal interests, and trust of human beings [8, 12]. Some measures can be defined to investigate the number and quality of the relationships within the network. The crucial methods, which are currently used to identify the structure of a social network, are: full network method, snowball method, and ego-centric methods [10]. Since there is no established classification of social networks, we propose our own taxonomy of social networks: dedicated (e.g. dating or business networks, networks of friends, graduates, fun clubs), indirect (online communicators, address books, e-mails), common activities (e.g. co-authors of scientific papers, co-organizers of events), local networks (e.g. people living in the neighbourhood), families, employees networks, hyperlink networks (links between home pages), etc. Additionally, the classification of social networks can be based not only on the type of the relations that occur in the network, but also on the type of the

communication between members i.e. they can be either online (virtual, via the computer network) or offline (tangible, with personal contact). Only the first ones will be considered in this paper. Online social networks (called also virtual communities) enable and support the communication between people who are in different places and on different schedules [18]. This makes the relationships not as tangible as those from the real world. The ways of communication within the online networks vary depending on the functionality of the network. The following ones can be distinguished: email, chat, forum, blog, comments, testimonials, photo/movie album, etc. (Fig.1). All social networks are defined by the static attributes of actors like their interests or demographic data as well as the description of the relationships between actors. All this data create the user profile [14], which can be analyzed in order to define and measure the social capital – one of the most important factors of social networks.

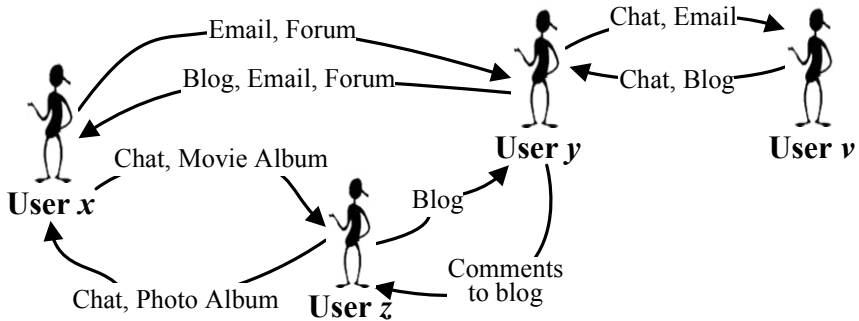


Fig. 1. An example of the online social network

2 Social Capital

There are many approaches to the concept of social capital and each of them presents the social capital in a bit different way [2, 7, 11]. However, all the definitions have something in common – they are implicitly [16, 17] or explicitly [6] functional i.e. social capital is described by its function rather than by its nature. Putnam defined social capital as “features of social organization such as networks, norms, and social trust that facilitate coordination and cooperation for mutual benefits” [17]. The scientist who defined the social capital as explicitly functional concept was Coleman. According to his theory “social capital is defined by its function. It is not a single entity, but a variety of different entities having two characteristics in common: they all consist of some aspect of social structure, and they facilitate certain actions of individuals who are within the structure“ [6].

The social capital can be seen as “metaphor about advantage” [5]. In other words, people have to make an effort in order to receive higher returns [5, 11]. In general, social capital usually grows with the use of all its component resources. Although social capital generally brings benefits, it happens sometimes the investment in its

building is bigger than the outcomes on the efforts [2]. The social capital of the group that form a social network is aggregation of social capital of all individuals.

The definition that will serve as the basis for the further consideration was formulated by Bourdieu and Wacquant: the social capital is “the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition” [3]. Although this definition does not concern explicitly the online social network, we applied its general idea in our further research.

3 Components of Social Capital in Online Social Networks

3.1 User Social Capital

The user social capital can be defined as the set of features that describe the ability of cooperation between the people. The social capital consists of two main parts: static and dynamic (Fig.2). The users themselves deliver the information about the former while the latter is monitored by the system.

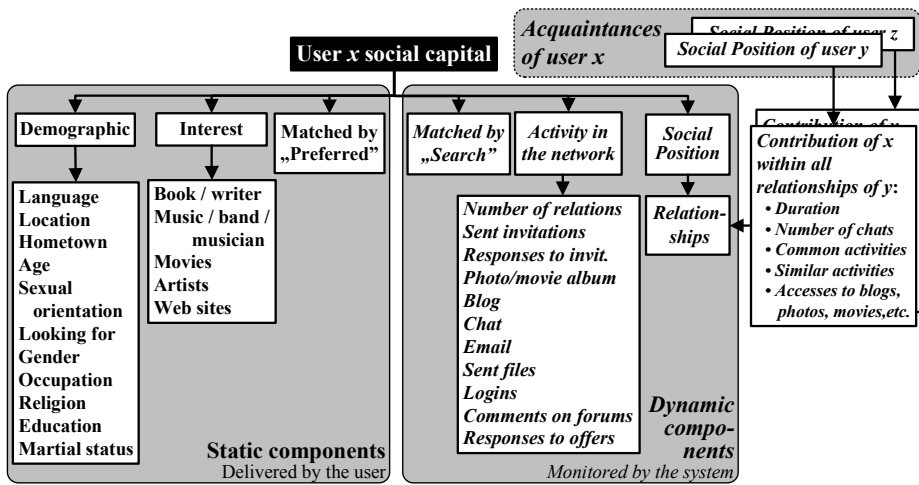


Fig. 2. Social capital of user x grouped in components. *Location* is the place of residence, *Hometown* is the place of origin, *Looking for* denotes the purpose of inviting new friends.

The static social capital $S(x)$ provides information about *Interest* and *Demographic* features of user x . Additionally, each user can deliver the data about people sought by them and the system counts the number of participants who match these constraints – *Matched by “Preferred”* component. The more overall these conditions are, the more open-minded and flexible the user is, and that is why their social capital should be greater. The users themselves maintain all this data and neither their behaviour within the social network nor the system itself can change these static components.

The dynamic part of social capital depends on the activity and position of the user within social network. Three components of the dynamic social capital for user x can

be distinguished: *Matched by "Search"* – $MS(x)$, *Activity* – $A(x)$, and *Social Position* – $SP(x)$. The social capital $SK(x)$ for user x is defined as follows:

$$SK(x) = \alpha \cdot S(x) + \beta \cdot MS(x) + \gamma \cdot A(x) + \delta \cdot SP(x) \quad (1)$$

where α , β , γ , δ – coefficients of importance for particular components, which are constant at the beginning. Latter on, their values can be recalculated based on the user feedback [13]. The analysis of user choices within the network will be done to establish their most appropriate values. They can be either the same for all the network members or separate for each individual.

The *Matched by "Search"* $MS(x)$ component is derived from all searches made by user x within the network. The set of participants that match search criteria is used to estimate the range of the network that covers user's interest. The more network members fit to performed searches, the more sociable and open-minded the user for new acquaintances is, and in consequence $MS(x)$ should also be higher.

The *Activity* part characterizes the relative activity of the user compared to all other community members. It consists of many components that describe user activity within certain domain like frequency of login, sent emails, invitations, files, chats, as well as total number of relationships, or published multimedia. Some of these elements are calculated for fixed periods (1 month, half a year) – frequencies, e.g. sent emails, while some others are absolute numbers e.g. sent invitations, number of relationships. Nevertheless, all of them are normalized according to the highest values among all network members.

The *Social Position* $SP(x)$ component describes the general position of user x in the network and it can be derived from their relationships. $SP(x)$ increases with the number of acquaintances who are in relationship with x . Moreover, if acquaintance y of user x possesses the high *Social Position* $SP(y)$, then also $SP(x)$ is high. In such case we can say that member x has the "significant" friend y , i.e. x gets on with real authorities. If user y allocates most contribution of their activity directly to user x , then user x is the best friend of y . As a result, the greater part of $SP(y)$ is inherited by user x . This contribution i.e. the significance of x for y consists of many partial activities of y like duration of their acquaintance, number of direct communications e.g. chats, common and similar activities as well as number of accesses of y to the content published by x (blogs, photos, movies). Common activities are for example co-authorships in projects or studies, co-organization of events, etc. Similar activities are derived from usually unintentional meetings of x by y in the network, e.g. comments on the same forums or to the same news, auctions or purchases similar products, trips to the same places, etc.

The values of all the components are summed in (1) and social capital is additive because each of its elements have positive influence on its final value. We also considered multiplication but in this case each of the components would have crucial influence on social capital.

3.2 Static Components

The static component $S(x)$ of the social capital define the user x characteristic, which does not change over time. Note that the conventional definition does not contain such elements since they do not "accrue by virtue of possessing a durable network" [3]. Nevertheless, all of them are the resources which enable to improve individuals'

social networks. Based on the *interests* and *demographic* features of the network member x , the system can find and recommend another person y who will fit their expectations. Obviously, this is mutually beneficial for social capital of both participants x and y . Moreover, some static features can be utilized at direct calculation of social capital, e.g. location. A person from a bigger city is a member of larger local community so they are able to invite more new network members or co-operate with larger number of people also out of the network. People supposed to be more mobile, for whom the country of origin and residence are different, have greater opportunity to benefit from this fact. In online communities the accessibility and quality of the access to the Internet play an important role in assessing the social capital and people from highly developed regions have usually better access. Furthermore, a user who speaks more languages and especially speaks world-wide English has more possibilities to create new relationships. Well-educated people usually do better in the sense of receiving higher return on investment and additionally they are well-connected [5].

The search preferences characterize network members as well. If the participant provides very precise description of the people they want to be in contact, then there may not be enough members who match these expectations.

3.3 Activity of the User

The *Activity* component $A(x)$ of social capital for user x respects the relative frequency of all possible activities performed by user x within the social network, separately for all activity types, like updates of blogs or multimedia albums, number of comments on forums, sent invitations, etc (Fig. 2). $A(x)$ describes how active is user x compared to the most active members in the particular activity type, e.g. in updating their blogs:

$$A(x) = \frac{1}{N} \sum_{i=1}^N \frac{A_i(x)}{A_i^{\max}} \quad (2)$$

where $A_i(x)$ – measure of activity type i for user x ; A_i^{\max} – maximum value of activity i ; N – number of all activity types.

In addition, the user can be more active in one period and less in another one and the older periods should have less influence on the final measure of activity $A_i(x)$. The precise formula for $A_i(x)$ was presented in [14]. Since all elements $A_i(x)$ and A_i^{\max} are dynamic and change over time, the calculation of $A(x)$ should be periodically repeated for all network members.

3.4 Social Position

The position and value of the user in the network tightly depends on the strength of the relationships that this user maintains. In order to assess the general strength of the relationships of user x *Social Position* function $SP(x)$ has been introduced. Its inspiration was the PageRank algorithm, which is the basic method used by Google to determine the page's relevance or importance [4]. *Social Position* $SP(x)$ of user x respects both $SP(y)$ value of user x acquaintances as well as the activity of y with relation to x :

$$SP(x) = (1 - \varepsilon) + \varepsilon \cdot (SP(y_1) \cdot C(y_1 \rightarrow x) + \dots + SP(y_m) \cdot C(y_m \rightarrow x)) \quad (3)$$

where: ε – constant coefficient from the range $[0,1]$; y_1, \dots, y_m – acquaintances of x , i.e. members who are in the direct relation to x ; m – the number of acquaintances of user x ; $C(y_1 \rightarrow x), \dots, C(y_m \rightarrow x)$ – the function that describes the contribution which has user x in the set of the relationships of user y_1, \dots, y_m , respectively.

The general concept of $SP(x)$ is the inheritance of *Social Position* from all acquaintances of user x , especially from those for whom x is the really good friend. This goodness of the friendship is derived from the contribution of acquaintances' activities directed to x . User x possesses the high $SP(x)$ if x would have the real authorities as friends and these authorities would really be in contact with x . Moreover, $SP(x) \in [1-\varepsilon, 1]$ and it equals $1-\varepsilon$ when user x does not have any relationships within the network. $SP(x)$ for all users is calculated iteratively and this process should be repeated periodically. The number of iterations can be fixed to constant value l or the calculation stops when the differences in values of $SP(x)$ between the following iterations are below the given threshold.

Note that the same *Social Position* can be achieved by x if user x would have many relationships with people who have medium $SP(y)$ or if x would have only few relationships but with participants with high $SP(y)$ (Fig.3).

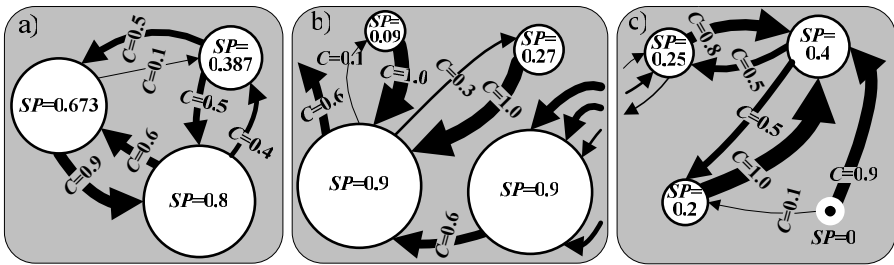


Fig. 3. Examples of social networks with calculated $SP(x)$ and $C(y \rightarrow x)$ for each user, $\varepsilon=1$

The contribution $C(y \rightarrow x)$ of user x within activity of their acquaintance y is the sum of all contacts, cooperation, and communications from y to x in relation to all activities of y :

$$C(y \rightarrow x) = \frac{\mu_1 \cdot \frac{a_1(y \rightarrow x)}{a_1^{sum}(y)} + \mu_2 \cdot \frac{a_2(y \rightarrow x)}{a_2^{sum}(y)} + \dots + \mu_k \cdot \frac{a_k(y \rightarrow x)}{a_k^{sum}(y)}}{\mu_1 + \mu_2 + \dots + \mu_k} \tag{4}$$

where: μ_1, \dots, μ_k – fixed coefficients of importance for the particular activity type; k – the number of criteria (activity types) that describe strength of relationship; $a_1(y \rightarrow x), \dots, a_k(y \rightarrow x)$ – the number of activities common for both user x and y of the first, k -th type, respectively, e.g. the number of common projects; $a_1^{sum}(y), \dots, a_k^{sum}(y)$ – the total number of the activities of type 1, k that performed user y .

There is only one requirement for activity functions: $\sum_{j=1}^m a_i(y \rightarrow x_j) = a_i^{sum}(y)$.

One of the activity types is communication via chat. In this case, $a_i(y \rightarrow x)$ is the number of chats that are common for x and y ; and $a_i^{sum}(y)$ is the number of all chats in which y took part in. If user x has many common chats with y in comparison to the number of all chats with y , then x has greater contribution within activities of y i.e.

$C(y \rightarrow x)$ will have greater value and in consequence *Social Position* of user x will grow. Note that $C(y \rightarrow x)$ will have value 1 when user x is the only friend of user y .

However, not all of the elements can be calculated in such simple way. Much more complex are similar activities, e.g. comments on forums. Each forum consists of many threads where people can submit their comments. In this case, $a_i(y \rightarrow x)$ is the number of user's x comments in the threads in which y has also commented, whereas the function $a_i^{sum}(y)$ is the number of comments that have been made by all friends of y on these threads.

4 Social Capital in Data Mining and Recommendation Systems

People behave in the network community similarly to their attitude in the real life. Thus, we can say that the analysis of online communities will deliver the interesting knowledge about human beings. Social capital is the combination of single features that characterize participants within the online society. The research on social capital and its dynamic over the course of time for the entire online social as well as for its users will gain in better understanding of human behaviours and limits.

Based on the historical data related to social capital of individuals in the given network we can build a model for its prediction. It can be also used for prediction of new desirable relationships between network members. The social capital is depleted by lack of communication rather than by activities within relationships. Building new or strengthening the existing relationships causes that both individuals and group will benefit and the social capital will increase. The number of relationships can be increased by tailoring and utilizing the concepts known as bonding and bridging social capital [17]. The goal of the former is the interconnection of two or more homogeneous and similar but separate groups whereas in the latter the different heterogeneous groups are linked. Usually, the joined groups are internally very close. Both bonding and bridging enable to achieve the larger community in which the associations between human beings are permanent. Another approach is to stimulate new relationships within groups. In this way the social network contains many disconnected but very close groups of members that know one another very well. Practically, the creation of new relationships and in consequence the growth of the social capital can be stimulated by various types of recommendation systems that process social capital values using diverse data mining techniques [12, 14].

Association rules and clustering, typical data mining methods can be utilized to identify interesting groups of people in the network, especially those that share the same level of social capital or some of its components: static components, *Activity* or *Social Position*. On the other hand, by means of sequential patterns and time sequences we can analyse general or unusual human behaviours with respect of time. Next, this knowledge could be harnessed at developing more "human sensitive" information systems.

5 Conclusions and Future Work

Social capital is an important indicator of significance and position of particular members within the online social network. The above presented method for estimation of social capital consists of several components that cover wide range of static

and dynamic features of network users. It also includes *Social Position* of the participant in the context of the number and strength of direct relationships with other participants as well as *Social Position* of these acquaintances.

The future work will focus on the application of the social capital concept to artificial environments like multi-agent systems and on the development of specialized recommendation systems that would stimulate the expansion of the network [14].

References

1. Adamic L.A., Adar E.: Friends and Neighbors on the Web. *Social Networks* 25(3) (2003) 211-230.
2. Adler P., Know S.: Social Capital: Prospects for a new concept. *Academy of Management Review* 27(1) (2002) 17-40.
3. Bourdieu P., Wacquant L. J. D.: *An invitation to reflexive sociology*. University of Chicago Press (1992).
4. Brin B., Page L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. *WWW7, Computer Networks* 30 (1-7) (1998) 107-117.
5. Burt R.S.: *The Network Structure of Social Capital*. Research in Organizational Behavior 22, JAI Press (2000).
6. Coleman J.S.: *Foundations of Social Theory*. Harvard University Press, Cambridge, MA, (1990).
7. Fukuyama F.: *Social Capital and Development: The Coming Agenda*. *SAIS Review* 22 (1) (2002) 23-37.
8. Golbeck J.: *Computing and Applying Trust in Web-Based Social Networks*. Ph.D. Thesis. University of Maryland (2005) <http://www.cafepress.com/trustnet.20473616>.
9. Golbeck J., Hendler J.A.: Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks. *EKAU 2004, Springer Verlag, LNCS 3257* (2004) 116-131.
10. Hanneman R., Riddle M.: *Introduction to social network methods*. Online textbook. (2005), <http://faculty.ucr.edu/~hanneman/nettext/>.
11. Kadushin C.: Too Much Investment in Social Capital? *Social Networks* 26(1)(2004)75-90.
12. Kazienko P., Kiewra M.: Personalized Recommendation of Web Pages. Chapter 10 in: Nguyen T. (ed.): *Intelligent Technologies for Inconsistent Knowledge Processing*. Advanced Knowledge International, Adelaide, South Australia (2004) 163-183.
13. Kazienko P., Kołodziejcki P.: Personalized Integration of Recommendation Methods for E-commerce. *Int. Journal of Computer Science and Applications* 3(3) (2006) to appear.
14. Kazienko P., Musiał K.: *Recommendation Framework for Online Social Networks*. AWIC 2006, Springer Verlag, (2006).
15. Kumar R., Raghavan P., Rajagopalan S., Tomkins A.: *Social Networks: From the Web to Knowledge Management*. Chapter 17 in Zhong N. Liu J., Yao, Y. (eds): *Web Intelligence*, Springer-Verlag (2003) 367-379.
16. Lin N.: *Social Capital: A Theory of Social Structure and Action*. Cambridge University Press, New York (2001).
17. Putnam, R. D.: *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York (2000).
18. Wellman B., Salaff J., Dimitrova D., Garton L., Gulia M., Haythorntwaite C.: *Computer Networks As Social Networks: Collaborative Work, Telework, and Virtual Community*. *Annual Review of Sociology* 22 (1996) 213-238.

Nature-Inspiration on Kernel Machines: Data Mining for Continuous and Discrete Variables

Francisco J. Ruiz¹, Cecilio Angulo², and Núria Agell³

¹ GREC - UPC. Technical University of Catalonia,
Pau Gargallo 5, 08028 Barcelona, Spain
`francisco.javier.ruiz@upc.edu`

² ERIC - UPC. Technical University of Catalonia,
Rambla de l'Exposició s/n, 08800 Vilanova i la Geltrú, Spain
`cecilio.angulo@upc.edu`

³ GREC - ESADE. Ramon Llull University,
Avda. Pedralbes 60-62, 08034 Barcelona, Spain
`nuria.agell@esade.edu`

Abstract. Kernel Machines, like Support Vector Machines, have been frequently used, with considerable success, in situations in which the input variables are given by real values. Furthermore, the nature of this machine learning algorithm allows extending its applications to deal with other kinds of systems with no vectorial information such as facial images, hand written texts, micro-array gene expressions, or protein chains. The behavior of a number of systems could be better explained if artificial infinite-precision variables were replaced by qualitative variables. Hence, the use of ordinal or interval scales on input variables would allow kernels to be defined for nature-inspired systems directly. In this contribution, two new kernels are designed for applying kernel machines to such systems described by qualitative variables (orders of magnitude or intervals). In addition, the structure of the feature space induced by this kernel is also analyzed.

1 Introduction

Qualitative reasoning is the area of Artificial Intelligence (AI) which creates representations for continuous aspects of the world closer to human reasoning [1]. It's well known that no more than seven divisions of data may be retained in human short-term memory. However with this apparent restriction, many complex tasks can be carried out by human beings better and faster than computers and robots designed with exact mathematics.

Machine learning algorithms based on kernel functions (Kernel Machines) have been frequently used on real-valued input variables. Implicitly, when a kernel function is employed, input variables are projected onto a different space, called feature space, usually omitted during the training process, because it is not explicitly required [2,3]. Kernels are very appealing because the associated projection mapping avoids the 'curse of the dimensionality' problem when learning. Furthermore, this projection enables learning machines for input variables

living in non-euclidian spaces, providing that a certain metrics exists in the feature space. For these occasions, the procedure is to define some designed feature space and a projection mapping, driving to an appropriated kernel.

Following the later procedure, two different feature spaces associated to continuous variables, a Reproducing Kernel Hilbert Space (RKHS) and another one associated to the Hilbert space $L^2(\mathbb{R})$ with the usual inner product, are analyzed in Section 2. The novel approach considering the $L^2(\mathbb{R})$ space allows a generalization to kernels in the space of interval. In Section 3, the Qualitative Space of the Absolute Orders of Magnitude is briefly introduced. In a similar form to *string kernels* [4], a kernel is proposed to be used when available information is about variables' order of magnitude. This kind of information naturally arises when information is treated by people according to the principle of simplicity or parsimony. Next, it is demonstrated in Section 4 that the proposed kernel converges to the continuous exponential kernel when the granularity of the discrete kernel tends to infinity, where it is possible to apply this function on both, real values and intervals. Finally, some conclusions are extracted and a list of further research topics to be developed is enumerated.

2 Kernel Functions

Some learning machines, among them Support Vector Machines (SVMs), deal with a dual formulation of the learning problem so that the learned decision function can be expressed by the inner product of the input data and the training patterns. Furthermore, the possible non-linearly separable original problem can be transformed into a separable problem by projecting original space to the so-called *feature space*, $\mathcal{F} = \phi(\mathcal{X})$. Dual formulation allows to replace this double step procedure by the *kernel function* $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. Hence, the discriminant function will stand,

$$f(\mathbf{x}) = \sum_{i=1}^l \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \quad (1)$$

Function $k(\cdot, \cdot)$ is a suitable kernel, i.e. it represents an inner product in a feature space, when Gram's matrix $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j}$ is symmetric semi-definite positive, that is, their eigenvalues are non-negative. For the general case, a symmetric function is a kernel when the Mercer condition is met,

$$\int_{\mathcal{X} \times \mathcal{X}} k(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) f(\mathbf{z}) d\mathbf{x} d\mathbf{z} \geq 0 \quad \forall f \in L^2(\mathcal{X}) \quad (2)$$

where $L^2(\mathcal{X})$ is the Hilbert space of the squared integrable functions in \mathcal{X} . This requirement is equivalent to showing that for any finite subset of \mathcal{X} , the associated matrix \mathbf{K} is a symmetric semi-definite positive matrix.

2.1 The RKHS Feature Space

A fixed kernel does not univocally determines the representation of the map ϕ , nor the feature space. The most frequent feature space associated to a fixed

kernel is that named *Reproducing Kernel Hilbert Space* (RKHS). This space is a subset of the whole set of real functions defined on \mathcal{X} , noted $\mathbb{R}^{\mathcal{X}}$, built on the map,

$$\begin{aligned} \phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ \mathbf{x} &\rightarrow k(\cdot, \mathbf{x}) \end{aligned} \tag{3}$$

Definition 1. *The RKHS is the complete set of all the maps into $\mathbb{R}^{\mathcal{X}}$ being a linear combination of $\phi(\mathbf{x}_i)$, with $\mathbf{x}_i \in \mathcal{X}$, $f(\cdot) = \sum_{i=1}^m \alpha_i \phi(\mathbf{x}_i) = \sum_{i=1}^m \alpha_i k(\cdot, \mathbf{x}_i)$, with $m \in \mathbb{N}$, $\mathbf{x}_1, \dots, \mathbf{x}_m \in \mathcal{X}$. Let $f, g \in \text{RKHS}$ be two functions in the form,*

$$f = \sum_{i=1}^{m_1} \alpha_i k(\cdot, \mathbf{x}_i), \quad g = \sum_{i=1}^{m_2} \beta_i k(\cdot, \mathbf{x}_i) \tag{4}$$

then, the inner product in the RKHS \mathcal{F} is defined as,

$$\langle f, g \rangle = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{j=1}^{m_2} \beta_j f(\mathbf{x}_j) = \sum_{i=1}^{m_1} \alpha_i g(\mathbf{x}_i) \tag{5}$$

Each element in this space is determined by a not unique finite set of real numbers, α_i or β_i in Eqn. 4, but the inner product is well defined. It is direct to demonstrate that $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$ is satisfied.

2.2 A Feature Space Associated to $L^2(\mathbb{R})$

The RKHS defined above is the more usual feature space designed from a fixed kernel. In the following, a different methodology is described. Inversely to the RKHS design methodology, feature space \mathcal{F} will be built from no kernel as starting point, but methodology will drive to obtain a kernel.

Definition 2. *Let $\phi : \mathbb{R} \rightarrow L^2(\mathbb{R})$ be a map such that $\phi(x_0) = f_{x_0, \sigma}(x)$, with $f_{x_0, \sigma}(x) = F_{\sigma}(|x - x_0|) = F_{\sigma}(z)$, being F a decremental function with respect to $z = |x - x_0|$ in \mathbb{R}^+ , with $F_{\sigma}(x_0) = 1$ and σ a parameter or set of parameters. Then, function $f_{x_0, \sigma}(x)$ is called influence function.*

The proposed feature space, designed in $L^2(\mathbb{R})$, contains a set of influence functions $f_{x_0, \sigma}(x)$ symmetric with respect to $x = x_0$, having decreasing output along the distance between x_0 and x increases. The associated kernel is now defined by using the usual inner product in the Hilbert space $L^2(\mathbb{R})$,

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \int_{-\infty}^{\infty} f_{x_i, \sigma}(x) f_{x_j, \sigma}(x) dx \tag{6}$$

Let's illustrate the use of this kernel-building methodology based on influence functions by defining four possibilities: *hard*; *triangular*; *Gaussian*; and *exponential*. The top row of graphs in Figure 1 shows the shape of these functions, the shaded areas representing $k(x_i, x_j)$ for two particular values. The graphs below show $k(x_i, x_j)$ with respect to $|x_i - x_j|$ for each one of the four cases.

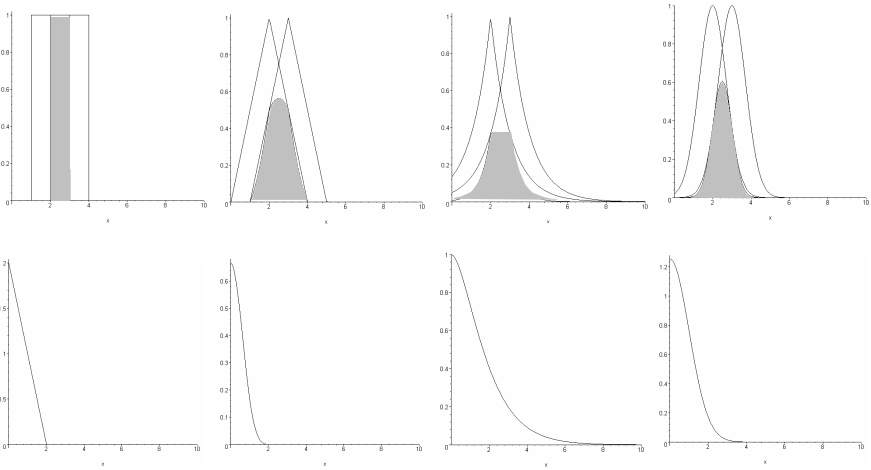


Fig. 1. Top, *hard, triangular, exponential and Gaussian* influence functions are represented. Also illustrated is the interpretation of the associated kernel for two fixed values. Bottom, the $k(x_0, x_0 + x)$ value is represented according to the x -axis.

Definition 3. *The hard (top-left), the triangular (bottom-left), the Gaussian and the exponential (top,bottom-right) influence functions are defined as,*

$$\begin{aligned}
 f_{x_0,\sigma}(x) &= \begin{cases} 1 & \text{if } |x - x_0| \leq \sigma \\ 0 & \text{otherwise} \end{cases} & f_{x_0,\sigma}(x) &= e^{-\frac{(x-x_0)^2}{\sigma^2}} \\
 f_{x_0,\sigma}(x) &= \begin{cases} \frac{|x - x_0| - \sigma}{\sigma} & \text{if } |x - x_0| \leq \sigma \\ 0 & \text{otherwise} \end{cases} & f_{x_0,\sigma}(x) &= e^{-\frac{|x-x_0|}{\sigma}}
 \end{aligned} \tag{7}$$

Proposition 1. *The associated kernels in $L^2(\mathbb{R})$ for the aforementioned functions are, respectively,*

$$k(x_i, x_j) = \begin{cases} 2\sigma - |x_i - x_j| & \text{if } |x_i - x_j| \leq 2\sigma \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$= \begin{cases} \frac{-6|x_i - x_j|^2 \sigma + 3|x_i - x_j|^3 + 4\sigma^3}{6\sigma^2} & \text{if } |x_i - x_j| \leq \sigma \\ \frac{(2\sigma - |x_i - x_j|)^3}{6\sigma^2} & \text{if } \sigma < |x_i - x_j| \leq 2\sigma \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$= \sigma \sqrt{\frac{\pi}{2}} e^{-\frac{(x_i - x_j)^2}{2\sigma^2}} \tag{10}$$

$$= (|x_i - x_j| + \sigma) \cdot e^{-\frac{|x_i - x_j|}{\sigma}} \tag{11}$$

In the third case, the well-known Gaussian kernel is obtained in a different form than the standard RKHS use. In both methods, the feature space generated from

the Gaussian kernel is also composed by Gaussian functions. However a lower width is obtained in the introduced novel approach,. On the other hand, this is a new demonstration that the Gaussian function is effectively a kernel.

2.3 A Feature Space for Intervals

The use of influence functions defining kernels can be extended to the set of intervals on the real line.

Definition 4. Let $I(\mathbb{R}) = \{[a, b] \mid a \in \mathbb{R}, b \in \mathbb{R}, a \leq b\}$ be the set of all the intervals on \mathbb{R} . Let $\phi : I(\mathbb{R}) \rightarrow L^2(\mathbb{R})$ be a map defined in such a form that $\phi([a, b]) = f_{[a,b],\sigma}(x)$, with $f_{[a,b],\sigma}(x) = f_{a,\sigma}(x)$ for $x < a$, $f_{[a,b],\sigma}(x) = 1$ for $a \leq x \leq b$, and $f_{[a,b],\sigma}(x) = f_{b,\sigma}(x)$ for $x > b$, being $f_{x_0,\sigma}$ some influence function on \mathbb{R} . Function $f_{[a,b],\sigma}(x)$ is named the interval influence function.

For instance, if the exponential influence function is used, then

$$f_{[a,b],\sigma}(x) = \begin{cases} e^{-\frac{|x-a|}{\sigma}} & \text{if } x < a \\ 1 & \text{if } a \leq x \leq b \\ e^{-\frac{|x-b|}{\sigma}} & \text{if } x > b \end{cases} \tag{12}$$

By using the exponential function it is possible to express $f_{[a,b],\sigma}(x)$ as

$$f_{[a,b],\sigma}(x) = \frac{f_{a,2\sigma}(x)f_{b,2\sigma}(x)}{f_{a,2\sigma}(a)f_{b,2\sigma}(a)} \tag{13}$$

So the calculation of the Gram’s matrix \mathbf{K} is easy because

$$\begin{aligned} k([a, b], [c, d]) &= \int_{-\infty}^{\infty} \frac{f_{a,2\sigma}(x)f_{b,2\sigma}(x)f_{c,2\sigma}(x)f_{d,2\sigma}(x)}{f_{a,2\sigma}(a)f_{b,2\sigma}(a)f_{c,2\sigma}(c)f_{d,2\sigma}(c)} dx \\ &= A \cdot \int_{-\infty}^{\infty} e^{-\frac{|x-a|+|x-b|+|x-c|+|x-d|}{2\sigma}} dx \end{aligned} \tag{14}$$

with $A = (f_{a,2\sigma}(a)f_{b,2\sigma}(a)f_{c,2\sigma}(c)f_{d,2\sigma}(c))^{-1}$

The image shape for an interval according to the map ϕ in Definition 4 can be observed on the left in Figure 2. The shaded area corresponds to the value of the kernel for two fixed intervals. In the same figure, on the right, the value for $k([0, 1], [x, 1+x])$ is represented with respect to the x -axis. It can be appreciated how $k(\cdot, \cdot)$ diminishes when the distance between intervals increases.

3 Absolute Orders of Magnitude Model

A main goal of Qualitative Reasoning, similar to the principle of parsimony in nature-inspired applications, is just to tackle problems in such a way that the principle of relevance is preserved; that is to say, each variable involved in a real problem is valued with the required level of precision [5].

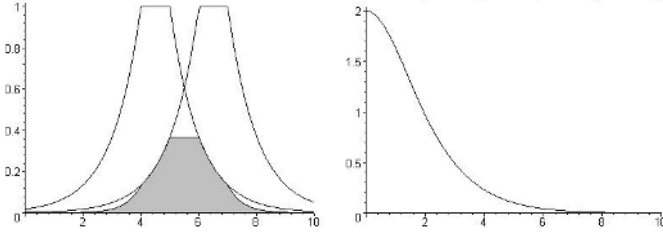


Fig. 2. On the left, representation in the feature space of two intervals by using the interval influence function as mapping. The shaded area represents the kernel value. On the right, a representation of the $k([0, 1], [x, x + 1])$ w.r.t. x .

For this reason, the absolute orders of magnitude models deal with a finite set of symbols or qualitative labels obtained via a partition of the real line, where any element of the partition is a basic label. These models provide a mathematical structure, which unifies signed algebra, and interval algebra through a continuum of qualitative structures built from the rougher to the finest partition of the real line.

In particular, the absolute orders of magnitude model of granularity n , $OM(n)$, is defined by a symmetric partition of the real line in $2n + 1$ classes, from the real numbers $\{-a_{n-1}, \dots, -a_1, 0, a_1, \dots, a_{n-1}\}$. Each class is named basic description or basic element, and it is represented by a label of the set S_1 ,

$$S_1 = \{N_n, N_{n-1}, \dots, N_1, 0, P_1, \dots, P_{n-1}, P_n\} \tag{15}$$

where

$$\begin{aligned} N_n =]-\infty, -a_{n-1}[, N_i =]-a_i, -a_{i-1}[, N_1 = [a_1, 0[, \\ 0 = \{0\}, \\ P_1 =]0, a_1[, P_i =]a_{i-1}, a_i[, P_n =]a_{n-1}, +\infty[\end{aligned} \tag{16}$$

Usually a linguistic label is associated to each one of these classes, for instance small positive, medium positive, etc. Finally, this set is extended with all the possible convex subsets of the real line defined from the basic elements. So, the Quantity Space, S , is obtained by considering all the labels of the form,

$$I = [X, Y]; \forall X, Y \in S_1 \text{ with } X < Y, \tag{17}$$

where $X < Y$ represents $x < y \forall x \in X, \forall y \in Y$.

An order relation, \leq_P , is defined in S , to be more precise than, for any pair $X, Y \in S$, $X \leq_P Y$ if $X \subseteq Y$. From this relation, the concept of *base of a qualitative label* can be defined $\forall X \in S - \{0\}$, as the set $B_X = \{B \in S_1 - \{0\} \mid B \leq_P X\}$. On the other hand, the qualitative equality or *q-equality* is defined on pairs $X, Y \in S$, as $X \approx Y$ when $X \cap Y \neq \emptyset$. This qualitative equality reflects the possibility that labels X and Y represent the same value. Kernel construction will be induced from the concept of remoteness.

Definition 5. Given a fixed $U \in S$, the remoteness with respect to U , $a_U : S \rightarrow \mathbb{N}$, is defined as follows: for all $X \in S$, $a_U(X) = \text{Card}(B_{X_U}) - \text{Card}(B_X)$.

Remoteness represents the number of non-null basic labels added to a label X for obtaining a qualitatively equal element to basic label U . The remoteness concept will allow us to define an application indicating global position of a qualitative label with respect to all the basic labels.

Definition 6. *Given $X \in S - \{0\}$ and decay factor $\lambda \in [0, 1]$, it is defined the map $\phi : S \rightarrow \mathcal{F} \subseteq [0, 1]^{2^n}$ so that $\phi(X) = (\lambda^{a_{N_n}(X)}, \dots, \lambda^{a_{P_n}(X)})$*

Defined mapping ϕ allows us to build a kernel in $OM(n)$ in a similar way to string kernels [4] as, $k(X, Y) = \langle \phi(X), \phi(Y) \rangle$ where $\langle \cdot, \cdot \rangle$ is the usual inner product in \mathbb{R}^{2^n} . This construction is directly extensible to a k -dimensional qualitative space $[OM(n)]^k$ to be used in the case that patterns are given by k descriptions. In such a case, map ϕ will be a function $\phi : S^k \rightarrow \mathcal{F} \subseteq [0, 1]^{2^{nk}}$. For $X = (X_1, \dots, X_k) \in S^k$, $\phi(X) = (\phi(X_1), \dots, \phi(X_k))$.

4 Continuous Limit of the Kernel in $OM(n)$

Feature space \mathcal{F} associated to the qualitative labels space S , defined in the previous section, can be considered a function space, with functions in the form $f_{\lambda, X_0} : S_1 - \{0\} \rightarrow [0, 1]$. Hence, $\phi(X_0) = f_{\lambda, X_0}(X) = \lambda^{a_x(X_0)}$ with $X \in S_1 - \{0\}$ and $X_0 \in S - \{0\}$.

Function $f_{\lambda, X_0}(X)$ can be considered an influence function, such as those introduced in Subsection 2.3. It can be represented by associating each qualitative label X to a rectangle of height $\lambda^{a_{X_0}(X)}$. Consequently, applying the kernel in $OM(n)$ to a pair of qualitative labels, X_1 and X_2 , is the same as adding the areas of the rectangles obtained from all the possible products of the heights of the rectangles associated to the basic labels in $\phi(X_1)$ and $\phi(X_2)$.

For the case of equal width labels, $a_{i+1} - a_i = \Delta, \forall i$, and x_i denoting the midpoint of any basic label X_i , it can be written that,

$$\phi(X_0) = f_{\lambda, X_0}(X) = \lambda^{\frac{|x-x_0|}{\Delta}} \tag{18}$$

with $X \in S_1 - \{0\}$. This influence function can be related to the exponential function example in Subsection 2.3, by using the equivalence $\Delta = -\sigma \cdot \ln \lambda$.

5 Example

An example using the ‘Iris’ database illustrates the application of the developed kernels. Data set contains 3 classes, 50 instances each, described by 4 continuous attributes, discretized using a supervised algorithm [6].

For the exponential kernel, parameters $\sigma = 0.5$ and $C = 10$, the obtained mean misclassification error was lower than 7.5% after 30 repetitions in a 2-fold cross validation procedure. In a second experiment, using the second presented kernel with parameter $\lambda = 0.5$ and same regularization and cross-validation parameters, reached mean misclassification error was 4.88%, better than in the first experiment and similar or better than other classification techniques using continuous variables.

6 Conclusions and Further Work

A new methodology to obtain kernel functions has been proposed. Similarly to human data treatment, it allows to deal with whether continuous variables, orders of magnitude, or those defined on real intervals. These kernels are obtained via a mapping from the original data set to the Hilbert space $L_2(\mathbb{R})$ with its usual inner product. In this form, by using a Gaussian function, it is possible to reproduce the Gaussian kernel. It has been pointed out a special case, the exponential function, because it has been demonstrated that it corresponds to the continuous limit of a kernel defined on the qualitative space of the absolute orders of magnitude. The obtained relationship opens a new challenge to develop learning methods based on kernel to be applied when data is simultaneously expressed in not vectorial form, such as intervals, orders of magnitude, and real values.

Acknowledgments

This work has been partially supported by the coordinated research project AURA (Qualitative Reasoning-based Machine Learning, TIN 2005-08873-C02-01,02) from the Spanish Ministry of Education and Science.

References

1. Forbus, K.: Commonsense physics: A survey. *Annual Review of Computer Science* **3** (1988) 197–232
2. Cristianini, N., Shawe-Taylor, J.: *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press (2000)
3. Schölkopf, B., Smola, A.J.: *Learning with Kernels*. The MIT Press, Cambridge, MA (2002)
4. Lodhi, H., Shawe-Taylor, J., Cristianini, N., Watkins, C.J.C.H.: Text classification using string kernels. In: *NIPS*. (2000) 563–569
5. Travé-Massuyès, L., Dague, P., Guerrin, F.: *Le raisonnement qualitatif pour les sciences de l'ingénieur*. Ed. Hermes, Paris (1997)
6. Ruiz, F., Angulo, C., Agell, N.: A supervised discretization method for quantitative and qualitative ordered variables. *Computación y Sistemas* (2006) Accepted for publication.

Testing CAB-IDS Through Mutations: On the Identification of Network Scans

Emilio Corchado, Álvaro Herrero, and José Manuel Sáiz

Department of Civil Engineering, University of Burgos, Spain
{escorchado, ahcosio, jmsaiz}@ubu.es

Abstract. This study demonstrates the ability of powerful visualization tools (based on the use of connectionist models) to identify network intrusion attempts in an effective and reliable manner. It presents a novel technique to test and evaluate a previously developed network-based intrusion detection system (IDS). This technique applies mutant operators and is intended to test IDSs using numerical data sets. It should be made clear that some mutations were discarded as they did not all provide real life situations. As an application example of the proposed testing model, it has been specially applied to the identification of network scans and mutations of these. The tested Connectionist Agent-Based IDS (CAB-IDS) is used as a method to investigate the traffic which travels along the analysed network, detecting anomalous traffic patterns. The specific tests performed in this study were based on the mutation of one or several variables analysed by CAB-IDS.

1 Introduction

Intrusion Detection Systems (IDSs) are tools designed to monitor and analyse computer system or network events in order to detect suspect patterns that may relate to a network or system attack. An IDS that analyses packets travelling over an entire network is referred to as a network-based IDS.

Visualization techniques are starting to be applied in the field of IDSs [1], [2], [3], [4], [5], [6] and they are generally applied to numeric data. However, in the field of Computer Security, traffic data sets normally have a categorical and/or textual nature and their conversion into a data type to which visualization techniques (such as scatter plot or projectionist models) may be applied is not always obvious. Previous attempts are presented in [1], [4], [5], [6].

IDS evaluation is not a clear cut task [7]. Previous works have presented several techniques to test and evaluate misuse detection models for network-based IDSs. Some of these techniques were based [8] on a mechanism that generates a large number of variations on a known exploit by applying mutant operators to its template. In this study, a method is proposed to apply such a mutation technique for visualization techniques using numerical data sets.

In this case, the method is used to analyse the response of CAB-IDS (Connectionist Agent-Based IDS) [4], [5], [6] in the detection of a network scan. The ability to detect such scans can help to identify wider and potentially more dangerous threats to a

network. The main advantage of this testing model is that it allows analysis of IDSs based on numerical data sets.

A port scan may be defined as series of messages sent to different port numbers to gain information on its activity status. These messages can be sent by an external agent attempting to access a host to find out more about the network services this host is providing. A port scan provides information on where to probe for weaknesses, for which reason scanning generally precedes any further intrusive activity. This work focuses on the identification of network scans, in which the same port is the target for a number of computers. A network scan is one of the most common techniques used to identify services that might then be accessed without permission [3].

The principal research interest and novelty of this work lies in the development of a testing method. The main goal of this method is to prove the effectiveness and capability of any IDS based on numerical data to confront unknown attacks. In this particular study it has been used to test CAB-IDS.

2 CAB-IDS

CAB-IDS (Connectionist Agent-Based Intrusion Detection System) is a tool that has previously been described [4], [5] and can be defined as an IDS formed of different software agents [9] that work in unison [6] in order to detect anomalous situations by taking full advantage of an unsupervised connectionist model.

To detect anomalous situations, CAB-IDS consists of different kinds of agents:

- Sniffer Agent (S-A): this type of agent "controls" each segment (in which the network is divided).
- IDS Agent (IDS-A): there is only one agent of this kind, which is in charge of processing the information sent by S-As and alerting the network administrator.

The different functions performed by these agents are:

- 1st step.- Network Traffic Capture: captures packets travelling over the network segments where S-As are located.
- 2nd step.- Data Pre-processing: the captured data is selected, pre-processed and sent to the IDS-A. A set of packets and features contained in the headers of the captured data is selected from the raw network traffic.
- 3rd step.- Data Analysis: once the IDS-A receives the pre-processed data, a connectionist model (see Sect. 2.1) is applied to analyse the data and identify anomalous patterns.
- 4th step.- Visualization: the projections are presented to the network administrator.

2.1 The Unsupervised Connectionist Model

The data analysis task performed by the IDS-A is based on the use of a neural Exploratory Projection Pursuit (EPP) [10], [11] model called Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [12], [13], [14]. It was initially applied in the field of Artificial Vision [12], [13] to identify local filters in space and time. In CAB-IDS it is applied in the field of Computer Network Security. CMLHL is based on

Maximum Likelihood Hebbian Learning (MLHL) [15], [16] adding lateral connections [12], [13] which have been derived from the Rectified Gaussian Distribution [17]. The resultant net can find the independent factors of a data set but does so in a way that captures some type of global ordering in the data set.

Considering an N-dimensional input vector (x), an M-dimensional output vector (y) and with W_{ij} being the weight (linking input j to output i), CMLHL can be expressed [12], [13], [14] as:

1. Feed-forward step:

$$y_i = \sum_{j=1}^N W_{ij} x_j, \forall i . \tag{1}$$

2. Lateral activation passing:

$$y_i(t+1) = [y_i(t) + \tau(b - Ay)]^+ . \tag{2}$$

3. Feedback step:

$$e_j = x_j - \sum_{i=1}^M W_{ij} y_i, \forall j . \tag{3}$$

4. Weight change:

$$\Delta W_{ij} = \eta \cdot y_i \cdot \text{sign}(e_j) |e_j|^{p-1} . \tag{4}$$

Where: η is the learning rate, τ is the "strength" of the lateral connections, b the bias parameter, p a parameter related to the energy function [13], [15], [16] and A a symmetric matrix used to modify the response to the data. The effect of this matrix is based on the relation between the distances among the output neurons.

3 A Mutation Testing Model for Numerical Data Sets

Testing an IDS tool is the only way to establish its effectiveness. In order to test CAB-IDS, it was decided to measure its results confronting unknown anomalous situations. Furthermore, it was decided to compare it alongside other models such as Principal Component Analysis (PCA) [18] or MLHL [10], [11] as no other IDS, as far as the authors are aware, shares similar characteristics. It is noticeable that few unsupervised methods have been applied to the field of IDSs. Examples include PCA [1], EPP [4], [5] and Self-Organizing Maps (SOM) [19], [20]. Projectionist models such as PCA, EPP, MLHL or CMLHL have one important advantage over SOM in the field of computer network security in that they use time as a key variable when analysing the evolution of the packets in the traffic data set.

Misuse IDSs based on signatures rely on models of known attacks. The effectiveness of these IDSs depends on the "goodness" of their models. This is to say, if a model of an attack does not cover all the possible modifications, the performance of the IDS will be greatly impaired.

Our mutation testing model is inspired by previous testing models [8], [21], but this is the first one for IDSs based on numerical data sets. In general, a mutation can

be defined as a random change. In keeping with this idea, the testing model modifies different features of the numerical information extracted from the packet headers.

The modifications created by this model may involve changes in aspects such as: attack length (amount of time that each attack lasts), packet density (number of packets per time unit), attack density (number of attacks per time unit) and time intervals between attacks. The mutations can also concern both source and destination ports, varying between the different three ranges of TCP/UDP port numbers: well known (from 0 to 1023), registered (from 1024 to 49151) and dynamic and/or private (from 49152 to 65535).

Time is another fascinating issue of great importance when considering intrusions since the chance of detecting an attack increases in relation to the duration of it. There are therefore two main strategies:

- Drastically reduce the time used to perform a scan.
- Spread the packets out over time, which is to say, reduce the number of packets sent per time unit that are likely to slip by unnoticed.

It should be taken into account and will be explained further on that any of the possible mutations may be meaningless such as a sweep of less than 5 hosts in the case of a network scan.

Several tests have been designed to verify the performance of CAB-IDS. Each test is related to a data set obtained by mutating the original one (see Sect. 4). Changes were made to the traffic related to the sweeps to take the following points into account:

- Number of sweeps in the scan (that is, number of scanned ports).
- Destination port numbers at which sweeps are aimed.
- Time intervals when sweeps are performed.
- Number of packets (density) forming the sweeps (number of scanned hosts).

Taking these issues into account, the collection of data sets designed for the research (see Sect. 4) covers the majority of the different scan-related situations with which a network might be confronted. Despite the fact that this technique is unable to provide a formal evaluation, it represents in our opinion a good approximation.

4 Data Sets and Tests

It was previously indicated that the proposed CAB-IDS [4], [5], [6] is able to identify a network scan contained in a data set with the following attributes:

- Three different sweeps to several hosts.
- Each sweep aimed at port numbers 161, 162 and 3750.
- A time difference between the first and the last packet included in each sweep of 17 866 ms for port number 161, 22 773 ms for port number 162 and 17 755 ms for port number 3750.
- An MIB (Management Information Base) information transfer event. This anomalous situation and its potential risks are fully described in [4], [5].

As previously explained, several testing data sets containing the following key features were presented to CAB-IDS following their mutation in order to measure the performance of CAB-IDS:

- Case 1 (modifying both the amount of sweeps and the destination ports):
 - Data set 1.- only one sweep: port 3750.
 - Data set 2.- two sweeps: ports 161 and 162.
 - Data set 3.- only one sweep: port 1734.
 - Data set 4.- two sweeps: ports 4427 and 4439.
- Case 2 (modifying both time and the number of sweeps):
 - Data set 5.- three time-expanded sweeps: ports 161, 162 and 3750.
 - Data set 6.- three time-contracted sweeps: ports 161, 162 and 3750.
 - Data set 7.- one time-expanded sweep: port 3750.
- Case 3 (modifying both the amount of packets and the destination ports):
 - Data set 8.- two 5-packet sweeps: ports 4427 and 4439.
 - Data set 9.- two 30-packet sweeps: ports 1434 and 65788.

The first issue to consider is the amount of sweeps in the scan. Data sets containing 1 sweep (Data sets 1, 3 and 7), 2 sweeps (Data sets 2, 4, 8 and 9) or 3 sweeps (Data sets 5 and 6) have been used. Each sweep is aimed at a different port number. The implications are crystal clear; hackers can check the vulnerability of as many services/protocols as they want. The number of sweeps (ranging from 1 to 65 536) can be modified from one scan to another.

A scan attempting to check port protocol/service can be aimed at any port number (from 0 to 65535). The data sets contain sweeps aimed at port numbers such as 161 and 162 (well known ports assigned to Simple Network Management Protocol), 1434 (registered port assigned to Microsoft-SQL-Monitor, the target of the W32.SQLExp.Worm), 3750 (registered port assigned to CBOS/IP ncapsulation), 4427 and 4439 (registered ports, as yet unassigned) and 65788 (dynamic or private port).

In order to check our system in relation to the time-related strategies, data sets 5, 6 and 7 were used. Data set 5 was obtained by spreading the packets contained in the three different sweeps (161, 162 and 3750) over the captured session. In this data set, there is a time difference of 247 360 ms between the first (in the sweep aimed at port 161) and the last scan packet (in the sweep aimed at port 3750). The duration of the captured session (all the packets contained in the data set) is 262 198 ms, whereas in the original data set the scan lasts 164 907 ms. In the case of data set 7, the same mutation has been performed but only for packets relating to the sweep aimed at port 3750. On the other hand, the strategy of reducing the time was used to obtain data set 6. In this case, the time difference between the first and the last packet is about 109 938 ms.

Finally, the number of packets contained in each sweep was also considered. In the case of a network scan, each packet means a different host included in the scan. Data sets 8 and 9 were designed with this issue in mind. Data set 8 contains low-density sweeps given that they have been reduced to only 5 packets. It was decided that a sweep scanning less than 5 hosts should not constitute a network scan. This is a fuzzy

lower limit because it could also be set as 4 or 6 packets. On the other hand, data set 9 contains medium-density sweeps. In this case, each one of them has been extended to 30 packets. This is also a fuzzy upper limit.

Apart from identifying the mutated sweeps, the detection of the MIB information transfer contained in all the data sets also represented a serious test for the performance of CAB-IDS. The experimental results obtained for these data sets are shown in the following section.

5 Results and Comparison

All the results were obtained by training the connectionist model for each new data set. The application of our model to the different scenarios (see Sect. 4) led to the results that are shown in Figs.1 to 6. Only figures for the most representative cases are presented. Through these figures, it may be seen how CAB-IDS is able to identify the different mutated anomalous situations, even though some are identified with greater clarity than others. Apart from traffic related to the scan, these figures also show packets interrelated with the rest of the traffic.

Situations are labelled anomalous whenever they tend not to resemble parallel and smooth directions (normal situations). In Fig. 1 two anomalous situations are highlighted (MIB information transfer and the network scan), both identified by CMLHL. As previously explained in [4], [5], [6], those situations are identified by CMLHL as anomalous by taking account of such aspects as traffic density or "anomalous" traffic directions.

Considerable experience is required to identify the sweep in the case of the projection for data set 7 (Fig. 2). Conversely, the other anomalous situation (the MIB information transfer) is identified with far greater clarity than in any of the other cases.

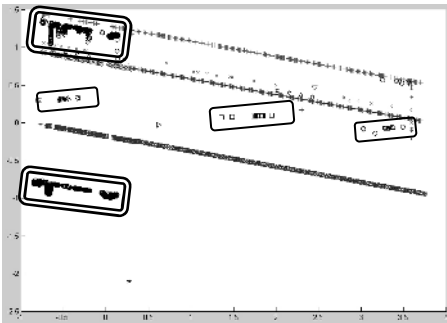


Fig. 1. CMLHL projection for data set 5

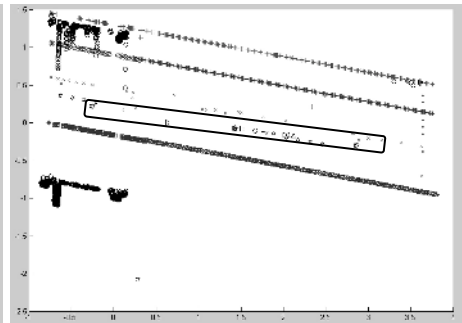


Fig. 2. CMLHL projection for data set 7

When sweeps contain only 5 packets (Data set 8 – Fig. 3), an expert is once again required to identify the anomalous scan situations. On the other hand, CAB-IDS very clearly detects high-density sweeps (Fig. 4).

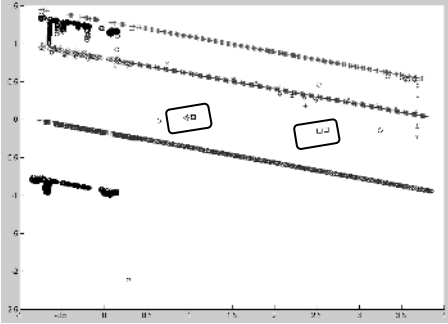


Fig. 3. CMLHL projection for data set 8

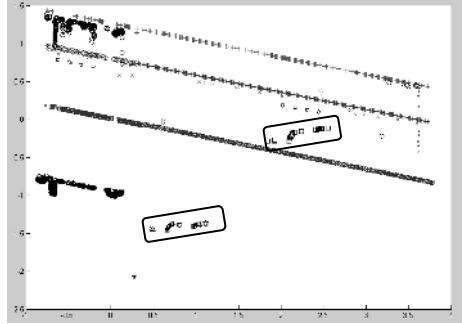


Fig. 4. CMLHL projection for data set 9

For comparison purposes, we have also applied PCA to the previous mutated data. As it can be seen in Fig. 5, the best PCA projection (Factor pair 1-3) is capable of identifying the 3-sweep scan but it is not capable of identifying the MIB information transfer. The projection of the two first principal components (Factor pair 1-2) obtained by applying PCA is unable to detect these anomalous situations. On the other hand, Fig. 6 shows how CMLHL is capable of identifying both situations.

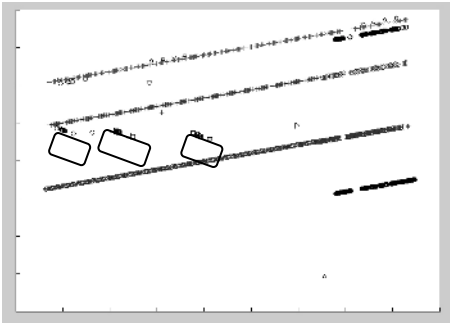


Fig. 5. PCA projection for data set 6

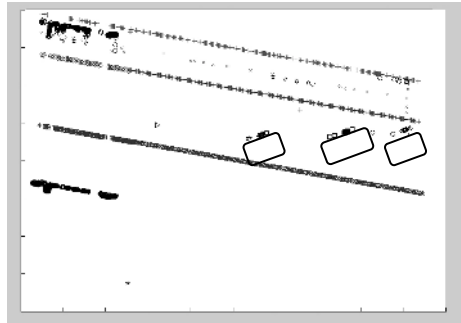


Fig. 6. CMLHL projection for data set 6

6 Conclusions and Future Work

This paper has introduced a novel mutation testing model for IDSs oriented to analyse numerical traffic data sets. It was used to test CAB-IDS and demonstrate its ability to identify most of the anomalous situations it confronted. The identification of these mutated scans can, in broad terms, be explained by the generalization capability of the connectionist model applied in this work. That is to say, through the use of one of these models, the IDS is capable of identifying not only the real anomalous situations contained in the data sets (known) but also the mutated (unknown) ones which may be real. This generalization capability of CAB-IDS represents its main advantage over

the majority of signature-based IDSs. Future work will be based on the application of new learning rules to improve CMLHL.

Acknowledgments

This research has been supported by the MCyT project TIN2004-07033 and the project BU008B05 of the JCyL.

References

1. Goldring, T.: Scatter (and Other) Plots for Visualizing User Profiling Data and Network Traffic. ACM Workshop on Visualization and Data Mining for Computer Security (2004) 119–123
2. Muelder, Ch., Ma, K-L., Bartoletti: Interactive Visualization for Network and Port Scan Detection. 8th International Symposium on Recent Advances in Intrusion Detection (RAID). Lecture Notes in Computer Science, Vol. 3858. Springer-Verlag, Berlin Heidelberg New York (2005) 265–283
3. Abdullah, K., Lee, Ch., Conti, G., Copeland, J.A.: Visualizing Network Data for Intrusion Detection. IEEE Workshop on Information Assurance and Security (2002) 100–108
4. Herrero, A., Corchado, E., Sáiz, J.M.: Identification of Anomalous SNMP Situations Using a Cooperative Connectionist Exploratory Projection Pursuit Model. International Conference on Intelligent Data Engineering and Automated Learning (IDEAL). Lecture Notes in Computer Science, Vol. 3578. Springer-Verlag, Berlin Heidelberg New York (2005) 187–194
5. Corchado, E., Herrero, A., Sáiz J.M.: Detecting Compounded Anomalous SNMP Situations Using Unsupervised Pattern Recognition. International Conference on Artificial Neural Networks (ICANN). Lecture Notes in Computer Science, Vol. 3697. Springer-Verlag, Berlin Heidelberg New York (2005) 905–910
6. Corchado, E., Herrero, A., Sáiz, J.M.: A Feature Selection Agent-Based IDS. First European Symposium on Nature-Inspired Smart Information Systems (2005)
7. Ranum, M.J.: Experiences Benchmarking Intrusion Detection Systems. NFR Security (2001)
8. Vigna, G., Robertson, W., Balzarotti, D.: Testing Network-Based Intrusion Detection Signatures Using Mutant Exploits. ACM Conference on Computer and Communication Security (ACM CCS) (2004) 21–30
9. Wooldridge, M.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, Gerhard Weiss (1999)
10. Friedman J., Tukey, J.: A Projection Pursuit Algorithm for Exploratory Data Analysis. IEEE Transaction on Computers, Vol. 23 (1974) 881-890
11. Hyvärinen A.: Complexity Pursuit: Separating Interesting Components from Time Series. Neural Computation, Vol. 13(4) (2001) 883-898
12. Corchado, E., Han, Y., Fyfe, C.: Structuring Global Responses of Local Filters Using Lateral Connections. Journal of Experimental and Theoretical Artificial Intelligence, Vol. 15(4) (2003) 473–487
13. Corchado, E., Fyfe, C.: Connectionist Techniques for the Identification and Suppression of Interfering Underlying Factors. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 17(8) (2003) 1447–1466

14. Corchado, E., Corchado, J.M., Sáiz, L., Lara, A.: Constructing a Global and Integral Model of Business Management Using a CBR System. First International Conference on Cooperative Design, Visualization and Engineering (CDVE). Lecture Notes in Computer Science, Vol. 3190. Springer-Verlag, Berlin Heidelberg New York (2004) 141–147
15. Corchado, E., MacDonald, D., Fyfe, C.: Maximum and Minimum Likelihood Hebbian Learning for Exploratory Projection Pursuit. Data Mining and Knowledge Discovery, Vol. 8(3), Kluwer Academic Publishing (2004) 203–225
16. Fyfe, C., Corchado, E.: Maximum Likelihood Hebbian Rules. European Symposium on Artificial Neural Networks (2002) 143–148
17. Seung, H.S., Socoli, N.D., Lee, D.: The Rectified Gaussian Distribution. Advances in Neural Information Processing Systems, Vol. 10 (1998) 350–356
18. Oja, E.: Neural Networks, Principal Components and Subspaces. International Journal of Neural Systems, Vol. 1 (1989) 61–68
19. Hättönen, K., Höglund, A., Sorvari, A.: A Computer Host-Based User Anomaly Detection System Using the Self-Organizing Map. International Joint Conference of Neural Networks (2000) 411–416
20. Zanero, S., Savaresi, S.M.: Unsupervised Learning Techniques for an Intrusion Detection System. ACM Symposium on Applied Computing (2004) 412–419
21. Marty, R.: Thor: A Tool to Test Intrusion Detection Systems by Variations of Attacks. ETH Zurich. Diploma Thesis (2002)

Nature Inspiration for Support Vector Machines

Davide Anguita and Dario Sterpi

Dept. of Biophysical and Electronic Engineering,
University of Genoa, 16145 Genoa, Italy
{anguita, sterpi}@dibe.unige.it

Abstract. We propose in this paper a new kernel, suited for Support Vector Machines learning, which is inspired from the biological world. The kernel is based on Gabor filters that are a good model for the response of the cells in the primary visual cortex and have been shown to be very effective in processing natural images. Furthermore, we build a link between energy-efficiency, which is a driving force in biological processing systems, and good generalization ability of learning machines. This connection can be the starting point for developing new kernel-based learning algorithms.

1 Introduction

In the last decades two very inter-related, but in some way separate research areas have been developed and have generated a large amount of successful methods and algorithms for data mining: Neurocomputing [19,10] and Machine Learning [23,13]. While both frameworks have been very successful in many real-world applications the former relies more on biological inspiration for deriving connectionist architectures and learning algorithms, while the latter builds on solid mathematical foundations like, for example, Statistical Learning Theory (SLT) [24].

One of the state-of-the-art techniques in Machine Learning is the Support Vector Machine (SVM) [6], which has been applied to a vast amount of data mining problems. This algorithm has been developed in the framework of SLT but it can be easily related to the well-known Radial Basis Function (RBF) network [15] as pointed out, for example, in [21]. While RBF networks cannot be considered a biologically plausible computational framework, nevertheless their connection to the biological world is quite clear, as they can be linked to several biological mechanisms [16].

We believe that the cross-fertilization between Neurocomputing and Machine Learning can and must be pursued: this paper shows some preliminary result in this direction. In particular, we show in the following section that Gabor filters can be used as an admissible kernel for SVMs. Section 4 shows the relation between energy efficient computation, which is a characteristic of biological processing systems [26], and Vapnik's Structural Risk Minimization (SRM) framework, which is the basis of the SVM.

2 The Support Vector Machine

We briefly revise here the SVM algorithm for classification, which will be the target of our proposals in the following sections.

Let us consider a dataset composed by l patterns $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\}$ where $\mathbf{x}_i \in \mathfrak{R}^n$ and $y_i = \pm 1$. The SVM learning phase consists in solving the following Constrained Quadratic Programming (CQP) problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2} \boldsymbol{\alpha}^T Q \boldsymbol{\alpha} + \mathbf{r}^T \boldsymbol{\alpha} & (1) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C \quad \forall i \in [1, \dots, l] \\ & \mathbf{y}^T \boldsymbol{\alpha} = 0 \end{aligned}$$

where $r_i = -1 \ \forall i$, and Q is a symmetric positive semidefinite $l \times l$ matrix $q_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ defined through a Mercer kernel $K(\cdot, \cdot)$ [6].

Once the parameters $\boldsymbol{\alpha}$ and the bias b are found, by solving the above CQP problem, the SVM feed-forward phase can be computed as:

$$y(\mathbf{x}) = \sum_{i=1}^l y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{2}$$

where \mathbf{x} is a new sample. Note that, in classification tasks, we are interested in $\text{sign}(y)$ and not in the actual value of y .

3 A Nature-Inspired Kernel

Two-dimensional Gabor filters are a simple model for the responses of simple cells in the primary visual cortex [12]. From an image processing point of view, they possess optimal localization properties in both spatial and frequency domain and have been used in a large variety of applications (e.g. texture segmentation, target detection, document analysis, edge detection, retina identification, image coding, image representation, etc.). The conventional 2D-Gabor filter is a complex sinusoidal wave modulated by a Gaussian envelope:

$$g(x, y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2})} e^{-j2\pi(u_0x + v_0y)}. \tag{3}$$

Note that Eq. (3) is an anisotropic 2D function, which can be easily extended to the multidimensional case, but at the expense of a linear increase of the number of hyperparameters. We propose here a simplified version of this extension, which allows us to derive the SVM kernel. Let us consider the real part of an isotropic Gabor filter (which is a standard form of SVM kernels) and let us introduce the rotation invariance property by considering the spherical version of (3), which makes it robust against changes caused by rotation angles [27]. Then, we obtain a Gabor kernel:

$$G(\mathbf{x}, \mathbf{y}) = e^{-\gamma t^2} \cos(\omega_0 t) \tag{4}$$

where $t = \|\mathbf{x} - \mathbf{y}\|$. It can be shown that Eq. (4) defines a Mercer's kernel by making use of the following lemma:

Lemma 1. *A stationary kernel is positive definite if and only if it is the Fourier transform of a positive finite measure.*

which is a special case of the Bochner's theorem (see [7] and the references therein). Then, we can state the following

Proposition 1. *the kernel described by Eq. (4) is positive definite and, therefore, is an admissible Mercer's kernel.*

Proof. We have to show that a positive function $f(\omega) > 0$ exists, such that its Fourier transform is the Gabor kernel. We can find $f(\omega)$ by computing the inverse Fourier transform of the kernel

$$f(\omega) = \frac{1}{2\pi} \int e^{j\omega t} K(t) dt = \sqrt{\frac{\pi}{\gamma}} e^{-\frac{\omega^2 + \omega_0^2}{4\gamma}} \cosh\left(\frac{\omega_0}{2\gamma}\omega\right) \quad (5)$$

which is positive for any value of the hyperparameters γ and ω_0 . \square

4 Nature-Inspired Kernel Machines

Recent studies on biological computation show that natural systems are energy efficient: retinal processing, for example, are efficient in terms of minimizing synaptic activity [26]. Obviously, energy efficiency is of paramount importance in biological systems because, as pointed out in [26]:

... it seems reasonable that minimizing energy consumption has shaped the evolution of neural processing of the brain. If equivalent computations could be carried out using less energy, savings from these energy efficient mechanisms could be used for other vital processes such as growth or reproduction.

Energy efficiency is also a major issue in modern (digital) computing systems. The design of low-power electronic devices has been developed in the last decade [5] and is forecasted to be a major issue in the future, where "the requirements for processing power will be 1000x in the next ten years, while the requirement for dynamic power consumption will not change noticeably" [11].

In a conventional digital system, given a fixed amount of computation time (e.g. a clock period), energy efficiency can be directly connected to the computational complexity of the implemented algorithm [14]. Therefore, when targeting a particular digital implementation, computational complexity is the focus of our proposal. In particular, it can be shown that the feed-forward phase of a kernel machine can be computed using only shift-and-add operations, so that the size (and therefore the power requirements) of the corresponding digital architecture grows linearly with the number of bits required by the computation [1] (see also [2]). Obviously, by minimizing the number of bits and maximizing the sparsity of the parameters of the SVM, low energy requirements can be met.

The question is if this complexity (or, equivalently, simplicity) measure is important not only for satisfying energy constraints, but is also beneficial to desired goal of good classification performance. Translating this question to the field of Machine Learning: can computational complexity be related in some way to the generalization ability ?

The short answer is yes: the connection between several complexity measures and the generalization ability has already being addressed in the past, following the well-known *Occam's razor* principle. Schmidhuber showed the connection between algorithmic simplicity (i.e. low Kolmogorov Complexity) and high generalization capability of artificial neural networks [21]; Vapnik's seminal works on SRM generated a large amount of research and interesting results, relating low complexity hypothesis spaces (e.g. learning machines with low Vapnik-Chernovenkis dimension) as a measure of high generalization [24]. These results were the basis for further refined theories relating the simplicity of the learning machine to good generalization, like, for example, the size of the weights in Multi Layer Perceptrons [4] or their sparsity [8].

The SVM is a direct consequence of these approaches: the rationale behind them is the construction of a learning machine which is sparse in the number of parameters and has low VC dimension.

We propose to build a kernel machine, where the connection between simplicity and high generalization capability is made explicit. In the following section we sketch the *Bit Sparing Machine* (BSM) and prove that an upper bound of the probability of error of the BSM is related to two quantities, which are a measure of its computational complexity: the size and the sparsity of its parameters. Our proposal goes in the direction indicated by Vapnik (see [25], p.295):

The challenge is to find refined concepts containing more than one parameter (say two parameters) that describe some properties of capacity, ... , by means of which one can obtain better bounds.

Similar refined concepts have been presented, for example, in [9], where the generalization ability of a kernel machine has been related to both the attained margin and the sparsity of the parameters. Here, however, we focus on a discrete setting that is the natural way of dealing with digital systems.

4.1 The BSM

We consider, for simplicity, Eq. (2) with $b = 0$, as this restriction does not affect most of the properties of a kernel machine if a strictly positive definite kernel is chosen [17].

Let us describe each parameter as an integer value of m bits:

$$\alpha_i = \sum_{j=0}^{m-1} a_i^j 2^j \quad (6)$$

where $a_i^j = \{0, 1\}$ and, therefore, $0 \leq \alpha_i \leq 2^m - 1$.

Given that each parameter can assume only a limited number of values, then, for a fixed training set of l samples and a fixed kernel, the number of classifiers N_f^l , that can be built through Eq. (4), is finite. This property allows us to use a family of well-known generalization bounds for finite hypothesis sets [3], which uses N_f^l as a measure of the classifier complexity. Classifiers with continuous parameters, instead, define an infinite hypothesis set, for which very refined concepts, like the VC dimension, are needed.

Let d be the number of non-zero parameters ($\alpha_i > 0$), then

$$N_f^l(m, d) \leq \sum_{i=1}^d \binom{l}{d} [(2^m - 1)^d - (2^{m-1} - 1)^d] \tag{7}$$

where the second term inside the square brackets take in account the fact that if all the parameters are even numbers, then they can be divided by 2 without changing the class estimate.

According to the SRM framework [24], a nested structure of hypothesis sets ($\mathcal{H} = h_1 \subseteq h_2 \dots \subseteq h_i \subseteq h_{i+1} \subseteq \dots$), with increasing complexity (e.g. increasing VC dimension), must be constructed before seeing the data. Then, the generalization capability of the classifiers can be controlled by choosing the appropriate set and, therefore, by finding the correct compromise between the learning error and the complexity so that a good generalization ability on unseen data can be guaranteed.

In our case, the set structure \mathcal{H} can be defined through two quantities: the sparsity d and the size of the parameters m . Starting from the set h_1 , which contains $N_f^l(1, 1)$ classifiers, the subsequent set h_2 is built by adding the classifiers described by an increased number of bits ($m \rightarrow m + 1$) or decreased sparsity ($d \rightarrow d + 1$) so that its cardinality is minimum. In other words, the cardinality of the sets is forced to grow as parsimoniously as possible by choosing

$$\min_{i,j} N_f^l(1 + i, 1 + j) \tag{8}$$

classifiers, with $i, j \in \{0, 1\}$ ($i \neq j$) and adding them to h_1 for building h_2 . Then, this procedure is iterated for building subsequent sets.

Obviously, the complexity of the classifiers grows proportionally to d and m , so a classifier belonging to h_i is more energy-efficient than a classifier taken from h_j , where $i < j$, because $h_i \subseteq h_j$.

The parsimony of the set growth has direct consequences on the generalization ability of the classifiers, according to the following bound [3], which holds with probability $1 - \delta$:

$$\pi \leq \nu + \sqrt{\frac{\ln N_f^l(d, m) - \ln \delta}{2l}} \tag{9}$$

where π is the generalization error and ν is the error obtained by the learning machine on the training set.

Fig. 1 shows the growth of the term $\frac{\ln N_f^l}{l}$, according to the set structure described above, for $l = 1000$ and $m \leq 8$. As a reference, a straight (dotted) line connecting 0 and $\frac{\ln 2^{md}}{l}$ is also plotted, showing that the proposed structure grows at a sub-exponential pace. Note that, as it clearly appears from Fig. 1 and

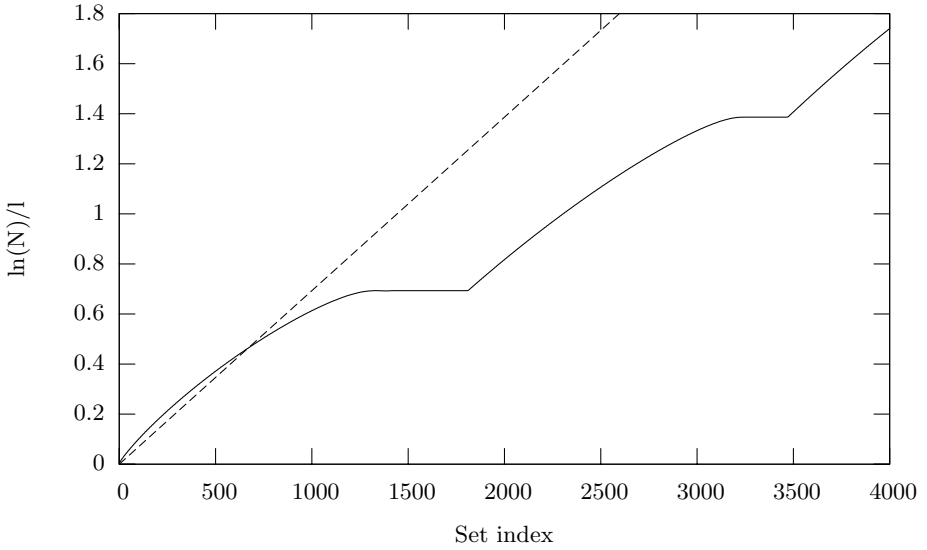


Fig. 1. Growth of $\ln N_f^l/l$ with $l = 1000$ and $m \leq 8$

as pointed out by several authors, the bound of Eq. (9) is of no practical use in predicting the generalization ability of a classifier, due to its looseness. However, like other similar or even more refined bounds obtained in the SRM framework [22], it gives a good explanation of the generalization phenomenon in learning machines. In our case, it connects the complexity of the machine, in terms of number of bits needed for the computation and, therefore, its energy-efficiency, to its generalization ability.

4.2 Some Preliminary Experiments

A direct application of the theory detailed in the previous section is being developed. The learning must be performed in a discrete setting, so the main challenge is to find an efficient learning algorithm, which exploits the complexity structure defined on both the sparsity and the size of the parameters. Some preliminary results are encouraging and will be described in a forthcoming work. Here, instead, we show the connection mentioned above through the use of the conventional SVM algorithm.

The experiments are performed using the datasets collected and prepared by G.Rätsch for the purpose of benchmarking machine learning algorithms [18]. The parameters of the SVM are found by solving the quadratic programming

Table 1. The classification error rate for the floating-point (FP) and decreasing (m bit) precision of the parameters

Name	FP	24	20	16	12	10	8
Banana	10.6	10.6	10.6	10.6	10.6	10.6	10.9
Breast-Cancer	33.8	33.8	33.8	32.5	26.0	28.6	28.6
Diabetis	24.0	24.0	24.0	23.7	23.7	23.3	24.3
Flare-Solar	33.0	33.0	33.0	33.0	33.0	33.0	33.0
German	22.0	22.0	22.0	21.7	24.3	25.3	50.3
Heart	18.0	18.0	18.0	18.0	18.0	18.0	18.0
Image	1.8	1.8	1.8	1.9	2.2	3.6	17.0
Ringnorm	2.5	2.5	2.5	2.5	2.5	2.7	3.6
Splice	5.2	5.2	5.2	5.2	5.4	5.8	6.2
Thyroid	5.3	5.3	5.3	5.3	5.3	5.3	5.3
Titanic	22.7	22.7	22.7	22.7	23.0	32.0	32.0
Twonorm	2.8	2.8	2.8	4.4	50.0	50.0	50.0
Waveform	10.1	10.1	10.1	10.1	10.1	10.1	10.3

described by Eq. (2), then each parameter is coded using only m bits, eventually truncating it to zero.

Table 1 shows the effect of using decreasing values of m . As expected, the error rate increases with the decrease in precision, however the results clearly show that, in many cases, only few bits are needed for computing the feed-forward phase of the SVM, without any loss in generalization performance and, at least in one case, the precision decrease is even beneficial. These results indicate that a low-complexity kernel machine can provide good results (at least as good as a conventional SVM).

5 Conclusion

We have sketched some research lines, which exploit some nature-inspired characteristics for suggesting new approaches in the Machine Learning field. A lot of work is still to be performed on these topics, but we believe that cross-fertilization of the two fields can be of interest for building new and effective approaches to data classification.

References

1. Anguita, D., Pischiutta, S., Ridella, S., Sterpi, D.: Feed-forward Support Vector Machines without Multipliers. *IEEE Trans. on Neural Networks* (2006) in press
2. Anguita, D., Boni, A., Ridella, S.: A Digital Architecture for Support Vector Machines: Theory, Algorithm and FPGA Implementation. *IEEE Trans. on Neural Networks* **14** (2003) 993–1009
3. Anthony, M., Bartlett, P.L.: *Neural Networks Learning: Theoretical Foundations*. Cambridge University Press (1999)

4. Bartlett, P.L.: The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network. *IEEE Transactions on Information Theory* **44** (1998) 525–536
5. Chandrakasan, A., Brodersen, R.: Minimizing power consumption in digital CMOS circuits. *Proc. of the IEEE* **83** (1995) 498–523
6. Cortes, C., Vapnik, V.: Support–vector networks. *Machine Learning* **27** (1991) 273–297
7. Genton, M.G.: Classes of kernel for machine learning. *Journal of Machine Learning Research* **2** (2001) 299–312
8. Herbrich, R.: *Learning Kernel Classifiers: Theory and Algorithms*. The MIT Press (2002)
9. Herbrich, R., Graepel, T., Shawe-Taylor, J.: Sparsity vs. Margins for Linear Classifiers. *Proc. of the 13th Conf. on Computational Learning Theory* (2000) 304–308
10. Hertz, J., Krogh, A., Palmer, R.G.: *Introduction to the Theory of Neural Computation*. Addison–Wesley (1997)
11. The International Technology Roadmap for Semiconductors. ITRS (2005) <http://public.itrs.net>
12. Jones, J.P., Palmer, L.A.: An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58** (1987) 1233–1258
13. Mitchell, T.: *Machine learning*. McGraw Hill (1997)
14. Parhami, B.: *Computer arithmetic: algorithms and hardware design*. Oxford University Press (2000)
15. Poggio, T., Girosi, F.: Networks for approximation and learning. *Proc. of the IEEE* **78** (1987) 1481–1497
16. Poggio, T., Girosi, F.: *A Theory of Networks for Approximation and Learning*. Technical Report 1140, MIT AI Lab (1989)
17. Poggio, T., Mukherjee, S., Rifkin, R., Rahklin, A., Verri, A.: b. Technical Report 198, MIT CBCL (2001)
18. Rätsch, G., Onoda, T., Müller, K.R.: Soft margins for AdaBoost Machine Learning, **42** (2001) 287–320
19. Rumelhart, D.E., McClelland, J.L.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. The MIT Press (1986)
20. Schmidhuber, J.: Discovering neural nets with low Kolmogorov complexity and high generalization capability. *Neural Networks* **10** (1997) 857–873.
21. Schölkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V.: Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. on Signal Processing* **45** (1997) 2758–2765
22. Shawe-Taylor, J., Bartlett, P.L., Williamson, R.C., Anthony, M.: Structural Risk Minimization over Data-dependent Hierarchies. *IEEE Trans. on Information Theory* **44** (1998) 1926–1940
23. Valiant, L.G.: A theory of the learnable. *Comm. of the ACM* **27** (1984) 1134–1142
24. Vapnik, V.: *Statistical Learning Theory*. John Wiley & Sons (1998)
25. Vapnik, V.: *The Elements of Statistical Learning Theory* (2nd Ed.). Springer (2000)
26. Vincent, B.T., Baddeley, R.J.: Synaptic energy efficiency in retinal processing. *Vision Research* **43** (2003) 1283–1290
27. Wang, Y., Chua, C.-S.: Face recognition using 2D and 3D images using 3D Gabor filters. *Image and Vision Computing* **23** (2005) 1018–1028

The Equilibrium of Agent Mind: The Balance Between Agent Theories and Practice

Nikhil Ichalkaranje¹, Christos Sioutis², Jeff Tweedale², Pierre Urlings², and Lakhmi Jain¹

¹ School of Electrical and Information Engineering
University of South Australia

Nikhil.Ichalkaranje, Lakhmi.Jain@unisa.edu.au

² Airborne Mission Systems, Defence Science and Technology Organisation
Jeffrey.Tweedale, Christos.Sioutis,
Pierre.Urlings@dsto.defence.gov.au

Abstract. This paper outlines the abridged history of agent reasoning theories as ‘agent mind’ from the perspective of its implementation inspired by new trends such as ‘teaming’ and ‘learning’. This paper covers how the need for such new notions in agent technology introduced a change in fundamental agent theories and how it can be balanced by inducing some original cognitive notions from the field of ‘artificial mind’. This paper concentrates on the popular agent reasoning notion of Belief Desire Intention (BDI) and outlines the importance of the human-centric agent reasoning model as a step towards the next generation of agents to bridge the gap between human and agent. The current trend including the human-centric nature of agent mind and human-agent teaming is explained, and its needs and characteristics are also explained. This paper reports add-on implementation on BDI in order to facilitate human-centric nature of agent mind. This human-centric nature and concepts such as *teaming agreements* are utilised to aid human-agent teaming in a simulated environment. The issues in order to make agents more human-like or receptive are outlined.

1 Introduction

One of the modern Artificial Intelligence (AI) approach leads towards the most talked-about trend called ‘Intelligent agents’. Many researchers believe that agent technology is the result of convergence of many notions and trends within computer science namely AI, cognitive science, object oriented programming and distributed computing [1, 2]. The result of such convergence led to the birth of a modern AI field known as Distributed Artificial Intelligence (DAI), which focuses on agents and their ‘interactions’ with environments and peers (Multi-Agent Systems-MAS). In order to simplify the development and study of agent technology, popular categorisation based on agent theories, agent system architectures and agent languages by leading researchers such as Wooldridge [3] will help significantly in forming a basic understanding of the agent area. Agent theories define and address reasoning within agents. Agent system architectures facilitates the implementation of agent theories

within specified environment, and agent languages are similar to programming language which facilitates the theories and system architecture to construct and compile [4].

Agent theories form the first building block for agent technology development to achieve autonomy, proactiveness and intelligence. Agent theories provide a means to exhibit these attributes to develop intelligent agent systems with the help of various reasoning and decision making models. The introduction of distributed systems introduced ‘interaction’ as an additional must-have agent characteristic. This forced *agent theories* to review its models to include interaction in agents. DAI encapsulates such revision by the introduction of notions such as Multi-agent systems (MAS) by combining traditional AI notions and paradigms. MAS mainly focus on achieving intelligence by means of interaction. MAS are based on the principle that an interaction between simple agents can achieve a highly complex intelligence. The ideal role of MAS is to complement the intelligence by interaction. Agent theories have been popular in providing intelligence but have failed to include interaction in their core models. Current trends in agent theories are facilitating the inclusion of interaction by hybridising different existing models from traditional AI. MAS’ scope of interaction is limited to its agent-only interaction, meaning inter-agent and agent-environment interaction. However, when it comes to systems where human involvement is crucial such interaction fails to include human due to technical and technological constraints.

Technological constraints such as the agent’s ability to include and interact with humans by exhibiting human social norms such as learning, trust and respecting human privacy etc. also need to be addressed in its core theories. In order to address technological constraints one needs to focus on agent theories and ways to include social norms. Learning or adaptation is an important step forward to complement intelligence. Traditional AI notions such as Machine Learning and cognitive science theories could be of great help to facilitate such social attributes in current agent theories.

This paper briefly focuses on agent theories and attempts to unfold their origins from cognitive science theories (such as ‘artificial mind’). By doing so, it establishes how this strong relationship can be utilised to aid new trends in agent technology such as ‘teaming and learning’ and an agent’s human-centric nature.

2 Agent Trends: Teaming, Learning, and Human-Centric Agent

Recent popular trends in agent technology include teaming, and adaptation or learning. In the DAI community, the most common term used to describe multi-agent interaction is ‘teaming’. Teaming broadly covers MAS interaction and its resultant child attributes such as *communication* and *coordination*, along with sub-notions such as *cooperation* and *collaboration* with peers, which we like to describe as *teaming agreements* utilising communication and coordination. We will discuss these notions briefly in the following paragraphs.

Communication: Agents have to communicate in order to convey their intentions. Communication is an integral part of interaction but does not have to be direct. It can be indirect by means of a resulting action. Communication in MAS can be implemented either as message passing or using shared variables [5]. A variety of protocols exist for agent communication based on agent knowledge manipulation, naturalistic human-like communication. Amongst these, those of significance are Knowledge Query and Manipulation Language (KQML) and Knowledge Interchange Format (KIF) [6], and FIPA's (Foundation of Intelligent Physical Agents) Agent Communications Language (ACL) [7]. Such research contributions in agent communication are close to reaching a standardised state.

Coordination: Coordination is crucial as a means of organising agents, their resources and tasks and thus improving agent performance and resolving conflicts. Ehlert [5] discusses a simple way of managing coordination via task allocation methods. Ehlert classifies task allocations as centralised, distributed, and emergent. In *centralised task allocation*, one central 'leader' conducts task distribution either by imposing tasks upon agents (hierarchical) or by trading/brokering tasks. In *distributed task allocation*, each agent attempts to obtain the services it requires from other agents either by sending requests to agents whom it knows have the required services or by sending requests to all agents and accepting the best offer. Distributed task allocation can be separated further in two ways, allocation by *acquaintances* or by *contract net*. Lastly in *emergent* task allocation, which is characteristic of reactive systems, each agent is designed to perform a specific task, therefore no negotiation is necessary. From Ehlert's categorisation, it is evident that two other important attributes arise, namely '*negotiation*' and '*competition*'. These attributes may be utilised to coordinate the agent's activities and resources.

Teaming agreements: Sub-notions such as coordination and collaboration are often confused in terms of definitions and implementation. We like to simplify such notions by stating that communication and coordination are parent class attributes and important to any agent who decides to *interact* with other agents. Thus, no matter what teaming agreements one follows, every entity has to communicate and coordinate their resources, goals, and skills to act as a 'team'. Teaming agreements such as cooperation and collaboration become child class attributes by utilising communication and coordination, either directly or indirectly. In simple terms, one can take the meaning of cooperation as coexisting with other entities with the obligation to share one or more resources as a part of a coexistence agreement. On the other hand, collaboration is something that encapsulates parts of coexistence with self-induced motivation to share resources and/or skills to achieve a common goal.

Human-centric agents, an answer to early automation pitfalls: Early machine automation and its techniques largely failed to address the human and the human cognitive process [8-10]. This was due to the aggressive introduction of automation based on perceived needs and the tools available at that time. Agent technology may aid in such human-centric automation by means of its inherited attributes from cognitive science.

The human-like reasoning and decision making theories such as Belief Desire Intention (BDI) [11] are attractive candidates of agent technology for human-centric automation. These theories could make the agent a stand-alone substitute for the human by replacing him or her. Although these theories exhibit human-like intelligence, they fall short of human interaction abilities. When achieving such human replacement it is imperative that the human should have final ‘control’ along with being able to interact with the agent to regain control in critical situations. The answer to such a trade-off in *control* is human-centric agents. Recent research efforts define this area as *human-agent teaming*. The possibility of considering the human as an equal part of any system and interacting in co-existence would give agent technology the leading edge, which traditional AI systems have failed to give. Human interaction inherits the same issues in MAS interaction with an added focus on naturalistic and proactive communication with the human [12] and adaptability. Along with these issues, involving the human brings to the fore new issues such as respecting the social obligations of human society. These social obligations and norms include control, trust, loyalty, and privacy [13].

Learning: The next stage in human-agent teaming would be to demonstrate the adaptive nature of agents. This adaptive nature (learning) will portray the agent as a smart team member especially when dealing with human counterparts. ‘Learning’ is one of the important attributes for allowing the human to ‘feel’ comfortable to communicate, cooperate and adapt to the environment. Modern reasoning models may utilise ‘learning’ to make agents more human-like. Current learning methods have a strong lineage with Machine Learning from traditional AI. Hybridising such traditional AI techniques with new reasoning models with the help of cognitive science theories and reward-based learning methods can result in making the hybrid model specifically catered to be more human-like. Such a notion is reported in [14, 15], combining existing evolving methods such as reinforcement learning [16] and cognitive science theories such as Observe, Orient, Decide and Act Loop (OODA) [17] and Rasmussen’s decision ladder [18] into the BDI reasoning model.

3 Agent Mind: Reasoning and Decision Making Model

What is *agent mind*? Is it another new notion following the trend of relabelling the traditional AI field into agent technology? The answer is yes and no. We like to refer to agents reasoning and decision making ability or model as ‘agent mind’. Most current agent models are inherited from work in ‘artificial mind’ based on human thinking models, with the difference being an evolutionary paradigm change to suit its implementation - evolutionary changes excluding social abilities such as learning, emotions and so on. Due to these exclusions, current reasoning models do not qualify to be addressed as ‘mind’. We believe that introducing some social ability into such reasoning models will aid the human-centric nature of agents, thus, partially qualifying the title ‘mind’. Agent mind aims to utilise earlier efforts of human-like mind in cognitive science and psychology, eventually becoming AI fundamentals for cognition-based problem solving [19]. Davis [19] illustrates artificial mind in a simplistic way by saying that it can be seen as a control system based on belief- and desire-like attributes to achieve a control state like goal. Current efforts such as

COJACK [20], to support psychologically plausible models of human variability (e.g. fatigue and time constraints) in agent paradigms proves that agent technology is taking its first steps towards ‘artificial mind’.

3.1 Agent Mind: Current Reasoning Theories

The key notions in agent technology based on Intentions (BDI) and Production theories and knowledge-based paradigms have a traditional AI lineage and have proven to be stable and suitable in the development of commercial strength applications. One such popular technology, namely BDI, is the result of this strong lineage with an extendable nature. BDI theory caters well for the definition of an intelligent agent with a few exceptions in implementing notions such as *learning* and *interaction* in multi-agent systems (MAS). Early developments in BDI implementation did not consider MAS interactions or learning as crucial until the recent evolution of the distributed nature of systems. The development of BDI progressed (Plan, Capabilities [21], Obligations [22]) as the software development community started to take an interest in BDI as autonomous programs having control over their own execution of action. This BDI development is slowly maturing with the support of new tools, platforms and notions (obligation [22], learning). These implementations achieve autonomy and intelligence solely based on reasoning models developed by early BDI theorists.

Another step up in BDI is to address MAS interaction (teaming) issues. Many recent advancements have begun to address these issues but their focus is on inter-agent interactions while ignoring human involvement. Teaming in MAS has attracted a lot of interest in the agent research community but there are no fixed theories standardised as yet. There are three dominant research efforts which are becoming standard teaming theories for implementation. These are Shared Plans [23], Joint Intentions [24] (in BDI), and economic theoretic based COM MTDP (Communicative Multiagent Team Decision Problems) [25]. Communication, coordination and cooperation are seen as important parts of MAS interactions. Ongoing standardisation efforts, as made by FIPA in MAS interaction issues (communication and coordination), are truly in place to address and guide further development along with the above mentioned theories. Recent efforts in BDI, such as JACK teams [26], partially address MAS interaction issues (inter-agent teaming only) and demonstrate BDI’s extendable nature.

3.2 BDI: Paradigm of Choice for Complex Reactive Systems

The early development of DAI and cognitive science research streams such as artificial mind [19], and practical reasoning theories gave birth to a rigorously formalised [11] theory named Beliefs, desire and intentions (BDI). We believe that BDI was a result of achieving the means of deliberative (symbolic AI) and reactive reasoning. BDI is closely related to the practical human reasoning process (sometimes referred to as folk psychology) as it possesses three mental attitudes – i.e. belief, desire and intention, which represent information, motivation and deliberative states of the agent, respectively [11]. These three mental attitudes are responsible for an

agent's behaviour and achieving optimal performance over resource hungry deliberative processes. It is believed that early work in philosophical theory on human practical reasoning such as that by Bratman [9] led to the foundation of BDI, which was again formalised by Rao and Georgeff [11]. A new breed of BDI implementation, such as JACK [27], Beliefs, Obligations, Intentions and Desires (BOID) [22] and others, continued bringing agent oriented methodology to object oriented software development by enhancing BDI logics with new notions such as capabilities [21], teaming, obligations [22] etc., according to the needs of the application area.

4 Towards Human Centric Agent Mind: CHRIS BDI Extension

A collaborative case study by University of South Australia and Defence Science and Technology Organisation [29] presented in this section describes our proposed first steps in understanding how to implement human-agent teaming in an intelligent environment [29] through *learning* and *teaming agreements* [28]. It focuses on extending the framework realised by the *Cognitive Hybrid Reasoning Intelligent Agent System* (CHRIS) as reported in [15]. CHRIS indicates how to achieve agent learning by combining three popular human reasoning models: Rasmussen's Decision Ladder (RDL) [18], the Observe Orient Decide and Act (OODA) [17] loop and the BDI [11] model.

It is proposed that in order to successfully implement human-agent teaming, agents must be able to enter into team-like agreements (*partnerships*) during operation, without requiring a pre-defined role obligation structure [28]. *Partnerships* allow agents to work towards achieving their own goals (dictated by their own team leader) by negotiating how to receive help from agents that belong in other teams. Two types of partnerships have been initially identified: *cooperation* and *collaboration* [28]. It is proposed that partnerships are implemented as a new *Partnership* module in the CHRIS Reasoning framework. The partnership module has been identified to be linked within the Orientation stage, specifically between the State and the Identification operation of CHRIS model.

5 Conclusions

From a theoretical point of view, agent technology has matured by learning from the shortcomings of traditional AI systems. However, when it comes to practical implementation of agents, the popularity graph of such implementation attempts is slowly increasing. Although agent technology is believed to be mature enough for real world applications, it is not yet ready for widespread implementation in everyday life. The important factor in this major step is to consider human interaction. Agent technology needs to consider human interaction factors such as social abilities and norms (e.g. learning, trust, and privacy) to become 'user friendly'. User friendliness is not the only factor that agent technology needs to live up to - intelligence is also important. The term intelligence often has high expectations when it is compared with human counterparts. It is believed that agent technology is the next generation of traditional AI paradigms. The technology has already begun to step forward in

considering human interaction. More and more applications with agents as intelligent assistants are already in existence, with some of these showing limited respect for human social norms. Day by day, agents such as interface agents and decision support agents are becoming a bridge between core AI paradigms and humans.

Key notions and theory in agent technology, such as MAS and BDI, have good inheritance from human cognition theories as well as a nice blend of improvement over traditional AI paradigms. BDI is proving to be a popular choice of solution for complex systems and is gaining commercial support in terms of tools and a maturing development environment. However, there is still a long way to go in terms of including human interactions. The balance between agent theories and its implementation will bring agent technology to the software developer's use for the creation of more 'user friendly' human-centric applications. We believe that to achieve such equilibrium, a paradigm shift is needed to include key social abilities such as adaptation and collaboration with humans. The new cognition concepts will significantly assist in achieving such social ability for agent technology to be mature enough to include the human and his or her interaction needs at its focus.

References

- [1] Wooldridge, M. J. (Ed): *An introduction to multiagent systems*. Chichester: J. Wiley, (2002).
- [2] Decker, K.: A Vision for Multi-agent Systems Programming, presented at Programming Multi-agent Systems, First International Workshop, PROMAS 2003, Dastani, M., Dix, J., and El Fallah Seghrouchni, A., Eds., Lecture Notes in Artificial Intelligence, vol. 3067, Springer. (2004), 1-17.
- [3] Wooldridge, M., Jennings, N. R.: Theories, Architectures, and Languages : A Survey, Intelligent Agents, presented at ECAI-94 Workshop on Agent Theories, Architectures, and Languages, Wooldridge, M., Jennings, N. R, Ed., Lecture Notes in Artificial Intelligence, vol. 890, Springer-Verlag ,Heidelberg. (1995), 1-39.
- [4] Wooldridge, M., Jennings, N. R.: Intelligent agents: Theory and practice. The Knowledge Engineering Review, vol. 10(2).(1995), 115-152.
- [5] Ehlert, P., Rothkrantz, L.: Intelligent Agents in an Adaptive Cockpit Environment, Delft university of technology, Netherlands, Research Report Ref. DKE01-01,Version 0.2, (October 2001).
- [6] The Arpa-Sponsored Knowledge Sharing Effort ,*KQML Specification Document*, at web-address.:<http://www.cs.umbc.edu/kqml/>, Last accessed. Jan, (2006).
- [7] Foundation for Intelligent Physical Agents (Fipa): *FIPA Agent Communication specifications*, at web-address.:<http://www.fipa.org/repository/aclspecs.html>, Last accessed. Jan, (2006).
- [8] Urlings, P.: *Teaming Human and Machine*, PhD Thesis, University of South Australia, Adelaide (2003).
- [9] Bratman, M. E. (Ed): *Intention, Plans, and Practical Reason*. Cambridge, MA,: Harvard University Press, (1987)
- [10] Russel, S. J., and Norvig, P (Ed): *Artificial Intelligence: A Modern Approach*, 2nd ed. New Jersey: Prentice-Hall, (2006)
- [11] Rao, A. S. ,Georgeff, M. P.: BDI Agents: From Theory to Practice, *Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS-95)*, San Francisco, USA.(June 1995), 312-319.

- [12] Yen, J., Yin, J., Ioerger, T. R., Miller, M., Xu, D., Volz, R. A.: CAST: Collaborative Agents for Simulating Teamwork, presented at Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), Seattle, WA. (August 2001), 1135-1142.
- [13] Tambe, M., Bowring, E., Pearce, J., Varakantham, P., Scerri, P., Pynadath, D.: Electric Elves: What Went Wrong and Why presented at AAAI Spring Symposium on "What Went Wrong and Why" (2006).
- [14] Sioutis, C., Ichalkaranje, N.: Cognitive Hybrid Reasoning Intelligent system Design, presented at 9th International Conference, on Knowledge-Based Intelligent Information and Engineering Systems, Melbourne, Australia, Springer-Verlag, Germany. (September 2005), 838-843.
- [15] Sioutis, C.: *Reasoning and Learning for Intelligent Agents*, PhD Thesis, University of South Australia, Adelaide (2006).
- [16] Sutton, R. S., Barto, A. G. (Ed): *Reinforcement Learning*: The MIT Press, (March 1998).
- [17] Boyd, J. R.: *The essence of winning and losing*, at web-address: <http://www.mindsim.com/MindSim/Corporate/ODA.html>, Last accessed. 17 Dec, (2005).
- [18] Rasmussen, J. (Ed): *Cognitive Systems Engineering*. Brisbane: Wiley, (1994).
- [19] Davis, D. N. (Ed): *Visions of Mind: Architectures for Cognition and Affect, Collected work*: IDEA Group Publishing, (March 2005).
- [20] Norling, E., Ritter, F. E.: Towards Supporting Psychologically Plausible Variability in Agent-Based Human Modelling, presented at AAMAS (2004), 758-765.
- [21] Padgham, L., Lambrix, P.: Agent Capabilities: Extending BDI Theory, presented at Seventeenth National Conference on Artificial Intelligence - AAAI 2000. (2000), 68-73.
- [22] Broersen, J., Dastani, M., Huang, Z., Hulstijn, J., Torre, L. V. D.: The BOID architecture, presented at fifth international conference on Autonomous Agents (Agents2001), Montreal. (2001).
- [23] Grosz, B., Kraus, S.: The Evolution of SharedPlans, in *Foundations and Theories of Rational Agencies*, Wooldridge, A. R. a. M., Ed., (1999), 227-262.
- [24] Cohen, P. R., Levesque, H. J.: Confirmation and joint action, presented at International Joint Conference on Artificial Intelligence. (1991).
- [25] Pynadath, D., Tambe, M.: The communicative multiagent team decision problem: Analyzing teamwork theories and models, *Journal of AI Research (JAIR)*, vol. 16. (2002), 389-423.
- [26] *JACK Intelligent Agents™: JACK Teams Manual*, Agent Oriented Software Pty. Ltd, at web-address: <http://www.agent-software.com/shared/resources/index.html>.
- [27] *JACK Intelligent Agents™ User Guide*, Agent Oriented Software Pty. Ltd, at web-address: <http://www.agent-software.com/shared/resources/index.html>, Last accessed. Dec, (2005).
- [28] Ichalkaranje, N., Sioutis, C., Jain, L. C., Tweedale, J.: Innovations in Human-Agent Teaming, University of South Australia-DSTO Collaborative Project, Internal Research Report Ref. DSTO, EIE-KES-IAAMS-LJNICS, (November 2005).
- [29] Sioutis, C., Tweedale, J., Urlings, P., Ichalkaranje, N., Jain, L. C.: Teaming Humans and Agents in a Simulated World, presented at 8th International Conference, on Knowledge-Based Intelligent Information and Engineering Systems, Wellington, New Zealand, Springer-Verlag, Germany, (September 2004), 80-86.

Trust in LORA: Towards a Formal Definition of Trust in BDI Agents

Bevan Jarvis and Lakhmi Jain

School of Electrical and Information Engineering, University of South Australia
bevan.jarvis@postgrads.unisa.edu.au, lakhmi.jain@unisa.edu.au

Abstract. Trust plays a fundamental role in multi-agent systems in which tasks are delegated or agents must rely on others to perform actions that they themselves cannot do. Dating from the mid-1980s, the Belief-Desire-Intention architecture (BDI) is the longest-standing model of intelligent agency used in multi-agent systems. Part of the attraction of BDI is that it is amenable to logical formalisms such as Wooldridge's Logic Of Rational Agents (LORA). In a previous paper the present authors introduced a model of trust, here named the Ability-Belief-Commitment-Desire (ABCD) model, that could be implemented within the BDI framework. This paper explores the definition of the ABCD model within the LORA formalism.

1 Introduction

In this paper we expand on our Ability-Belief-Commitment-Desire (ABCD) model of trust, which we introduced (but did not so name) in a previous paper [13]. In particular, we begin to explore how the ABCD model might be integrated into the most recent and comprehensive of the formal logical explications of the Belief-Desire-Intention (BDI) model of intelligent agency, namely the Logic of Rational Agents (LORA) [16].

LORA provides a rigorous logical basis for BDI agent theory. In so doing it maintains a link with research topics such as the logical specification and automatic verification of programs, as discussed in [16]. It is a desirable feature of any extension or application of BDI that this link should be maintained. Two approaches are possible to achieve this: to extend LORA (or another logical definition of BDI), or to define the extension in terms of the chosen definition. We select the latter approach. It is in this context that we wish to establish a definition of the ABCD model in LORA.

The ABCD model is our theory that trust is a belief about another agent's abilities, beliefs, commitments and desires. It thus represents a kind of theory of mind of the other agent. This theory is set in the context of BDI, the longest-standing of intelligent agent architectures [16], which combines a sound basis in philosophy [3] with rigorous formal logical explications (such as LORA) and numerous implementations [1,7,9,12].

Trust is fundamental to delegation. In delegating a task to another person (when so doing is a free choice made by us), we are, among other things, expressing our confidence in a set of beliefs about that person's ability, knowledge and motivations.

We refer to this set of beliefs as a 'model of trust' of the other person, and say that our confidence in that model is a measure of how trustworthy we consider them to be. Trustworthiness is thus a measure of how predictable a person or agent is, while the model of trust is our means of predicting the behaviour of an agent. O'Hara and Shadbolt [15] make a similar distinction between trustworthiness and trust. Trustworthiness is often modelled as a probability that an act of delegation will result in success, and in this context may, confusingly, be referred to as trust.

In our previous paper explicating the ABCD model, we distinguished between intention and commitment, and chose to consider the latter as a distinct mental attitude. In Bratman's theory [3], the role of commitment is to convert desire into intention. Commitment can also affect other mental attitudes, in particular norms and obligations, to turn these also into intentions [6]. In order to align with the terminology of LORA, we here return to considering intentions.

A final consideration before we discuss a definition of trust in LORA is the definition of ability, which is not defined in the BDI model. Wooldridge provides a definition in LORA terminology [16], as we outline below.

In related work, Dignum et al. [6], among others, extend the BDI model to include new modalities for norms and for obligations. Broersen et al. [4] define a dynamic logic that incorporates trust and commitment.

In the following sections, we first look at norms and obligations, which together with desires form a (nearly) complete explanation of the sources of intention. Next we consider LORA and the definition of ABCD in terms of LORA. Finally, we conclude and consider future research.

2 Norms and Obligations

The BDI model enables the definition of intelligent agents that have beliefs about the world, desires that they would like to have achieved, intentions – desires to which the agent has made some commitment – and plans that if successful will achieve its intentions [16]. The BDI model describes individual agents, and so desires are by default private to the agent. To describe group and team behaviour it is necessary to consider sources of intention that are extrinsic to the agent. This means introducing concepts such as obligations and norms.

Obligations arise from communications between individual agents [6]. For the purpose of this paper, an obligation is a communicated desire that another agent wants the potentially obliged agent to achieve. Often, it will be a part of a plan that the requesting agent wishes to delegate. While accepting an obligation may often benefit the doer, this benefit is not the primary reason for the action (although it may contribute to the decision to accept). Obligations, by our definition, are explicit, but often may be private to the agents involved.

Norms (normative behaviours or social conventions) arise from social relationships among groups of agents [6]. In human societies, norms are often implicit, and may remain unstated until somebody (perhaps a naïve foreigner) unwittingly breaks a rule. Other norms – for instance the various prescriptions and proscriptions that make up the law – are explicit. In this paper we regard norms as desires that are shared by a group, with no single agent being the instigator.

Both norms and obligations (as we have described them) initially have a status similar to desires. As with desires, they become intentions when the agent applies some level of commitment to their achievement. The nature of this commitment may differ, however. For instance, where an ordinary intention may be maintained until it is achieved, or dropped when it is believed no longer possible, an obligation intention may also be dropped when the requesting agent no longer requires it. A norm intention, on the other hand, might never be dropped.

3 LORA and ABCD

In order to relate the ABCD model with LORA, we must define each component of ABCD in LORA terms. Belief and desire are common to both, which leaves ability and commitment. First, however, we describe LORA.

3.1 LORA

LORA comprises four components: first-order logic; a temporal logic; a multiply-modal (BDI) logic; and a dynamic logic:

1. The first component is classical first-order logic with the quantifiers \forall and \exists .
2. The temporal component is the branching-time logic CTL* [8].
3. The BDI component introduces modal operators, representing beliefs, desires and intentions, that connect possible worlds, which are each represented by a CTL* tree. (For an accessible introduction to modal logics, see [10]. A thorough, modern approach is provided by Blackburn et al. [2].)
4. The action (or dynamic logic) component consists in labelling the state transitions in the CTL* tree with actions. The labelled tree is essentially a program specification.

As well providing a rigorous basis for reasoning about individual agents, LORA allows for reasoning about interaction between agents. To do this requires further operators, for example to indicate that a plan achieves a particular condition (Achvs) and that an agent is able to perform a particular plan (Agt).

Wooldridge goes on to explore one scenario of how a group might come together to perform a task to mutual benefit, and defines a number of stages in LORA syntax.

3.2 LORA and Ability

We now consider the definition of ability in the LORA context. Wooldridge provides such a definition, adapted from Moore [14], which we here summarise.

First we define the zero-order ability that agent λ has to achieve state φ as follows. There is some action α which agent λ believes it can perform and believes achieves φ , which agent λ actually can perform and which does achieve φ . In LORA notation this is:

$$(\text{Can}^0 \lambda \varphi) \equiv \exists \alpha : (\text{Bel } \lambda (\text{Agt } \lambda \alpha) \wedge (\text{Achvs } \alpha \varphi)) \wedge (\text{Agt } \lambda \alpha) \wedge (\text{Achvs } \alpha \varphi) \quad (1)$$

Wooldridge notes that the action α is quantified *de re* with respect to the belief modality. This in fact requires that agent λ knows about action α – in other words, it has detailed knowledge of the actions within its own plans.

Next we note that it is easily conceivable that while no single action can bring about φ , there may be an action which brings about a situation where another action brings about φ . If I want to buy a round of drinks but have no ready cash, I may first extract money from a nearby automatic teller machine (or "hole-in-the-wall"). The possibility of such a conjunction of actions should also be regarded as the ability to bring about φ . Thus we define 1-order ability:

$$(\text{Can}^1 \lambda \varphi) \equiv (\text{Can}^0 \lambda \varphi (\text{Can}^0 \lambda \varphi)) \quad (2)$$

By extension, we define the k -order ability ($k > 0$) when $k+1$ actions are required:

$$(\text{Can}^k \lambda \varphi) \equiv (\text{Can}^{k-1} \lambda \varphi (\text{Can}^0 \lambda \varphi)) \text{ for } k > 0 \quad (3)$$

Finally we can define ability as meaning that there is some finite sequence of actions that leads to state φ :

$$(\text{Can} \lambda \varphi) \equiv \bigvee_{k \geq 0} (\text{Can}^k \lambda \varphi) \quad (4)$$

Wooldridge does not extend this idea to encompass plans. In the BDI model, plans are hierarchical: depending on circumstances, certain actions may be inappropriate or may fail, and other actions will take their place. It may be assumed that the successful completion of a plan means that the objective has been achieved. That is, each possible path in the plan represents a series of actions, as above, that will achieve the objective. However, if an action fails and the plan has no alternative action to take, then the plan fails. Depending on its level of commitment to the objective, the agent may choose another plan.

We simply note that the existence of a plan to achieve something is equivalent to a number of sequences of actions as defined above, each of which will (in the circumstances in which it is executed) achieve the objective. This is again equivalent to having multiple solutions to equation (4). Thus, the existence of a plan implies ability.

3.3 LORA and Commitment

In our previous paper explicating the ABCD model [13], we distinguished between intention and commitment, choosing to consider the latter as a distinct mental attitude that converts desires, norms and obligations into intentions [3,6]. In order to align with the terminology of LORA, we here return to considering intentions.

3.4 LORA and Trust

We can now propose a definition of trust in LORA. Let us say that agent κ trusts agent λ with respect to sets A, B, I and D corresponding to abilities, beliefs, intentions and desires. (Recall that we have defined obligations and norms in such a way that they may be regarded as desires.) Each set actually comprises logical propositions representing states to be achieved, believed, intended or desired. This means that agent κ believes that:

1. Agent λ can achieve any state in A;
2. For any state in B, agent λ either already believes it or can achieve this belief;
3. Agent λ has some level of intention to achieving every state in I;
4. Agent λ desires every state in D.

For convenience, we call $M = A \times B \times I \times D$ the model of trust that agent κ has for agent λ – this being the set of tuples $(\alpha, \beta, \iota, \delta)$, where α , β , ι , and δ are respectively abilities, beliefs, intentions and desires that agent κ believes of agent λ .

The above definition can then be written as:

$$\begin{aligned} (\text{Trust } \kappa \lambda M) &\equiv (\text{Bel } \kappa \forall (\alpha, \beta, \iota, \delta) \in M: \\ (\text{Can } \lambda \alpha) \wedge ((\text{Bel } \lambda \beta) \vee (\text{Can } \lambda (\text{Bel } \lambda \beta)) \wedge (\text{Int } \lambda \iota) \wedge (\text{Des } \lambda \delta)) \end{aligned} \quad (5)$$

Trustworthiness, from the point of view of the trusting agent, can be regarded as the level of confidence it has in its model of trust of the trusted agent. In multi-agent systems, trustworthiness is generally implemented as a probability that delegating to the trusted agent will achieve a successful outcome. Level of confidence is not a concept that exists in LORA, but perhaps it could be implemented as a new modality, by adding commitment to belief. A high level of commitment to belief in the model of trust would mean a high level of trustworthiness, and similarly a low level or no commitment would indicate low trustworthiness. This, however, is beyond the scope of the present paper.

4 Conclusion and Future Research

In this paper we have looked at trust as being based around actions that an agent delegates to others. The natural way to delegate an action is to ask or oblige the second agent to perform it. Time and effort may be saved by considering the second agent's own desires (including norms and obligations) and intentions (desires that it has committed to), which may give an indication of whether the agent would be inclined to perform the action required. Additionally, it is rational also to consider the agent's beliefs (i.e., knowledge) and abilities (including the ability to gain knowledge) in order to predict whether it is capable of performing the required action.

The definition of trust that we develop from this is a kind of theory of mind – a set of beliefs about another agent's beliefs, abilities and motivations. We have specified this in terms of LORA, a logic for describing the behaviour and interaction of BDI agents. Importantly, we have not needed to adapt or extend LORA in order to do this. We have, however, had to compromise on our conception of separating commitment from intention.

Trustworthiness we regard as a level of confidence in the model of trust, and we have not attempted to define it in LORA terminology. We suggest that specifying 'level of confidence' might be possible by defining a new modality to LORA, which we provisionally identify as belief with commitment.

While we have established trust on the basis of agent-to-agent interaction, this does not immediately relate to group or team trust – for example, the trust an agent might have in a group, which may or may not include itself. One way forward is to consider the SimpleTeam approach [11], in which team behaviour is determined and bound by individual obligations of team members to an agent that coordinates the team.

References

1. AOS: JACK Intelligent Agents: JACK Manual, Release 5.0, Agent Oriented Software, Pty Ltd (2005)
2. Blackburn, P., de Rijke, M., and Venema, Y.: *Modal Logic*, Cambridge University Press (2001)
3. Bratman, M. E.: *Intention, Plans, and Practical Reason*, Harvard University Press (1987)
4. Broersen, J., Dastani, M., Huang, Z., and van der Torre, L.: 'Trust and Commitment in Dynamic Logic', in *Eurasia-ICT 2002: Information and Communication Technology*, LNCS 2510, Springer-Verlag (2002)
5. Cohen, P. R. and Levesque, H. J.: 'Intention is choice with commitment', in *Artificial Intelligence*, 42 (1990), 213-256
6. Dignum, F., Kinny, D., and Sonenburg, L.: 'From Desires, Obligations and Norms to Goals', in *Cognitive Science Quarterly*, 2(3-4) (2002) 407-430
7. d'Inverno, M., Kinny, D., Luck, M. and Wooldridge, M.: 'A Formal Specification of dMARS', in *Intelligent Agents IV: Proceedings of the International Workshop, ATAL 1997*, LNAI 1365, Springer-Verlag (1998)
8. Emerson, E. A. and Halpern, Y.: '"Sometimes" and "not ever" revisited: on branching time versus linear time temporal logic', in *Journal of the ACM*, 33(1) (1986) 151-178
9. Georgeff, M. P., and Lansky, A. L.: 'Procedural Knowledge', in *Proceedings of the IEEE*, 74 (1986) 1383-1398
10. Girle, R. A.: *Modal Logics and Philosophy*, Acumen (2000)
11. Hodgson, A., Rönnquist, R., and Busetta, P.: 'Specification of Coordinated Agent Behavior (The SimpleTeam Approach)', in *Proceedings of the Workshop on Team Behaviour and Plan Recognition at IJCAI-99* (1999)
12. Huber, M.: 'JAM: A BDI-theoretic mobile agent architecture', in *Proceedings of the Third International Conference on Autonomous Agents (Agents'99)*, New York, ACM Press (1999) 236-243
13. Jarvis, B., Corbett, D., and Jain, L.: 'Beyond Trust: A BDI Model for Building Confidence in an Agent's Intentions', in *Proceedings of the 9th International Conference on Knowledge-based and Intelligent Information and Engineering Systems, KES 2005*, Melbourne, LNAI 3682, Springer-Verlag, Berlin (2005).
14. Moore, R. C.: 'A formal theory of knowledge and action', in Allen, J. F., Hendler, J. and Tate, A., (eds): *Readings in Planning*, Morgan Kaufmann Publishers (1990) 480-519
15. O'Hara, K., and Shadbolt, N.: 'Knowledge technologies and the semantic web', in *Trust and Crime in Information Societies*, Edward Elgar (2005)
16. Wooldridge, M.: *Reasoning About Rational Agents*, The MIT Press, Cambridge, MA (2000)

Agent Cooperation and Collaboration

Christos Sioutis and Jeffrey Tweedale

Air Operations Division,
Defence Science and Technology Organisation,
Edinburgh SA 5111, Australia
{Christos.Sioutis, Jeffrey.Tweedale}@dsto.defence.gov.au

Abstract. This paper describes preliminary work performed to gain an understanding of how to implement collaboration between intelligent agents in a multi-agent system and/or humans. The paper builds on previous research where an agent-development software framework was implemented based on a cognitive hybrid reasoning and learning model. Agent relationships are formed using a three-layer process involving communication, negotiation and trust. Cooperation is a type of relationship that is evident within structured teams when an agent is required to cooperate with and explicitly trust instructions and information received from controlling agents. Collaboration involves the creation of temporary relationships between different agents and/or humans that allows each member to achieve their own goals. Due to the inherent physical separation between humans and agents, the concept of collaboration has been identified as the means of realizing human-agent teams. A preliminary demonstration used to confirm this research is also presented.

1 Introduction

Agent oriented development can be considered the successor of object oriented development when applied in the Artificial Intelligence (AI) problem domains. Agents embody a software development paradigm that attempts to merge some of the theories developed in AI research within computer science. Bratman's Beliefs-Desires-Intentions (BDI) reasoning model [1] has demonstrated the potential of becoming the method of choice for realizing truly autonomous agents. *Beliefs* represent the agent's understanding of the external world, *desires* represent the goals that it needs to achieve, and *intentions* are the courses of action that the agent has committed to follow in order to satisfy its desires [2].

When defining the intelligence of agents, researchers mention the properties that a system should exhibit. Firstly, *autonomy* means operating without the direct intervention of humans. Secondly, *social ability* means interacting with other agents. Thirdly, *reactivity* means perceiving their environment and responding to any changes that occur in it. Finally, *pro-activeness* means exhibiting goal-directed behavior.

This paper is concerned with social ability because it provides agents with the potential to function stand-alone or cooperate with other agents as required. Different techniques have been developed allowing agents to form teams. Agents

are dynamically assigned a particular role depending on the situation and their suitability. Recent advances in this field have focused on the formation of rather unique teams with human and machine members based on cognitive principles. The major advantage of such teams is an improved situation awareness capability for the human when dealing with unknown or military hostile environments [3].

The structure of teams is traditionally defined during the system design and is required to remain constant during operation. Within teams, agents are required to cooperate and explicitly trust other team members. This paper explores the idea of introducing dynamic, temporary team-like links that are established and destroyed at runtime with no input from the original designer. This approach allows the achievement of greater autonomy since different systems, each executing different agent teams are able to collaborate in order to achieve their goals.

Section 2 of the paper describes some background research on agent teams. Section 3 then describes how agents within teams are required to cooperate with each other. Section 4 introduces the idea of having agents collaborate. Section 5 briefly describes how collaboration can be used to facilitate human-agent teaming, and Section 6 provides some concluding remarks and future work.

2 Forming Agent Teams

There are three underpinning problems to consider in order to effectively form agent teams, these are *Communication*, *Negotiation* and *Trust*. Communication is concerned with the means of communication between agents such that they can understand each other. In-fact, early agent development relied on the idea that intelligence is an emergent property of complex interactions between many simple agents. For example: The *Open Agent Architecture (OAA)* [4] defines agents as any software process that meets the conventions of the OAA society, communication between agents is managed by facilitators, which are responsible for matching requests with the capabilities of different agents. *Aglets* [5] are Java objects that can move from one host on the network to another, hence also called mobile agents. Finally, *Swarm* [6] provides a hierarchical structure that defines a top level *observer swarm* and a number of *model swarm* are then created and managed in the level below it.

The second problem is Negotiation, this problem is concerned with how to form teams. There are many aspects to consider in regards to negotiation. Generally, development of teams involves separating the requirements of a team from the requirements of individual agents. This includes assigning goals to a team as a whole, and then letting the team figure out how to achieve it autonomously. A team is constructed by the definition of a number of roles that are required in order to achieve the goals of the team. Additionally, agents can be specifically developed to perform one or more roles. An important feature of this approach is that agents are assigned with roles at runtime and can also change roles dynamically when required. Hence, one agent may need to perform one or more roles during its operation. Examples of agent development platforms that follow

this approach are: *MadKit* [7] is a multi-agent platform written built upon the an organizational model called *Agent/Group/Role* and agents may be developed in many third party languages. JACK Teams [8] is an extension to the JACK Intelligent Agents platform that provides a team-oriented modeling framework, specifically this allows the designer to specify features such as team functionality, roles, activities, shared knowledge and possible scenarios.

The third problem is Trust, it involves deciding how an agent should handle trust in regards to other agents. For example, should an agent trust information given by another agent, or trust another agent to perform a particular task. The level of trust is not easily measured, although loyalty can be used to weight information and consequently the strength of bond that is created. The fragility of that bond reflects on the frequency and level of monitoring required for the team to complete the related portion of a task. For further details on trust, the reader may refer to [9].

3 Agent Cooperation

Current implementations of multi-agent systems are very good at defining highly structured teams. A team is created by firstly defining a number of roles and secondly by assigning agents to perform these roles. Although it is possible to assign the agents to roles at runtime, it is not possible to also change the *structure* of the team dynamically.

In a structured team agents have no choice but to cooperate with other team members. Trust plays a very small role as it is implicitly defined by the designer who structures the team. In other words, when a designer defines one agent to be the subordinate of another agent, it has no choice but to trust all information provided by the controlling agent. Although, recent research [10] shows that trust becomes an important factor when there is more than one agent available to perform a particular role. The question then becomes, which agent should be trusted to be tasked with the role.

Having agents being able to change links with a structured team is not desired due to the inherent uncontrollable nature of that approach. On the other hand, there is no reason why agents should not be able to negotiate limited, temporary (and supervised) links outside of their own teams. Such relationships are based on collaboration and are described in detail in the following section.

4 Agent Collaboration

Collaboration involves the creation of dynamic links between agents without requiring a pre-defined role structure. Such links are weaker than cooperation links because they are established and destroyed at runtime with no input from the original designer. In general terms, they allow agents to work toward achieving their own goals by negotiating contracts with agents that belong in other teams. The contracts oblige the agent to provide some resource(s) to the other team

while obtaining a similar return. Two types of collaboration have currently been identified.

The first one involves agents negotiating a *mutual goal* that when achieved will benefit both agents. This allows two agents to help each other by committing a certain amount of *shared effort* to achieving this goal. The second one involves agents negotiating how to *share resources* available to them for the benefit of each other. The key difference with the first one is that, in this case, there is no need to negotiate a common goal. Each agent has its own goals while having access to some resources available to another agent that is not normally available.

The usefulness of this approach becomes clear when considering highly autonomous systems where agents are given particular goals and are then required to operate completely autonomously in order to achieve them. The problem with such cases is that the designer of the system is not able to predict all of the resources required by the system to complete its goals. Consequently, the system will need to collaborate with other autonomous systems in order for both to achieve their objectives. Collaboration can occur between any combinations of single autonomous agents, multi-agent teams, and humans.

Collaboration is useful when realizing human-agent teams because humans cannot be assumed to be part of a structured team indefinitely. When a human breaks out of the loop, an appropriate collaboration contract would normally be destroyed in order to notify the system to assign another internal agent to perform the required goals and vice versa.

4.1 Prototype Implementation

A prototype implementation framework has been developed that allows an agent to establish collaboration with another agent or human. The framework is based on *CHRIS* [11], an agent reasoning and learning framework developed as an extension of JACK at the University of South Australia through PhD research [12]. *CHRIS* equips a JACK agent with the ability to learn from actions that it takes within its environment, it segments the agent reasoning process into five stages based on Boyd's Observation, Orientation, Decision and Action (OODA) loop [13], Rasmussen's Decision Ladder [14] and the BDI model. The collaboration module itself has been implemented within the Orientation stage, specifically between the State and the Identification operation as shown in figure 1.

The path on the left shows the process involved for establishing a collaboration contract between two or more agents. The path on the right indicates that agents need to continuously perform assessment in order to ensure that collaboration is progressing as previously agreed and also whether the collaboration is yielding the required effect toward achieving their own goals. Both of these operations are highly dependent on trust, which is updated accordingly.

This implementation is based on using JACK team agents. Negotiation is performed using an authoritative Collaboration Manager Agent (CMA). Subordinate agents simply need to be able to perform the **Cooperation** role. The current implementation only supports goal-based collaboration relationships, where an

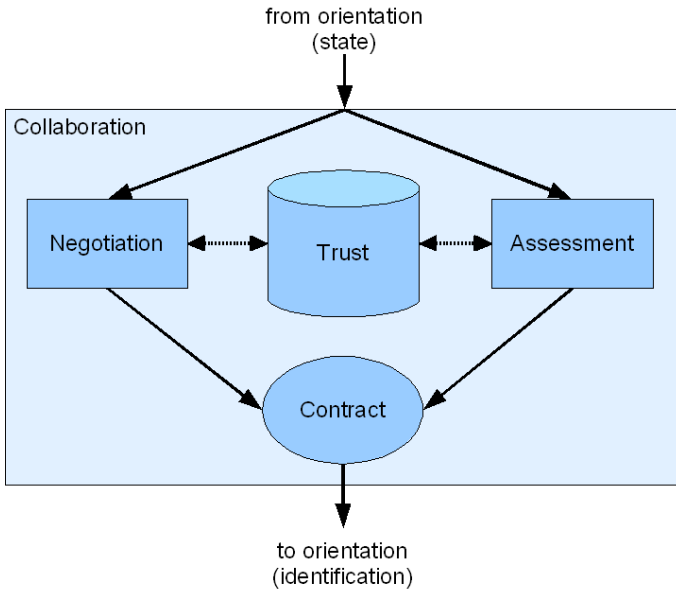


Fig. 1. Reasoning and Collaboration

agent negotiates for another agent to achieve a particular goal. Finally, an event called `RequestCollaboration` is used to ask the CMA for collaboration.

5 Human-Agent Teaming

A demonstration program was written that provides limited human-agent collaboration. It uses two agents, the first agent called `Troop` which connects to a computer game called Unreal Tournament (UT) using `UtJackInterface` [15] and controls a player within the game. The second agent is called `HumanManager` and is used to facilitate the communication with humans encountered within the game. The demonstration shows that the `Troop` agent is given the goal hierarchy shown in figure 2. The `Defend` and `Attack` goals are executed in parallel, also, considering that it is not possible for the `Troop` agent to perform both the `Defend` and `Attack` goals, it decides to handle the `Attack` goal and then asks the CMA to organise for any friendly human player seen in the game to take responsibility for the `Defend` goal. The sequence of operations for the demonstrations is:

1. The agents `Troop`, `CMA` and `HumanManager` are created and a `scenario.def` file is used to form a team with the `Cooperation` role between the CMA and the `HumanManager`.
2. The `Win` goal is activated and the `Defend` and `Attack` sub-goals are subsequently activated automatically in parallel. `Attack` is handled by the `Troop` agent which subsequently attacks any enemy that comes within the field of

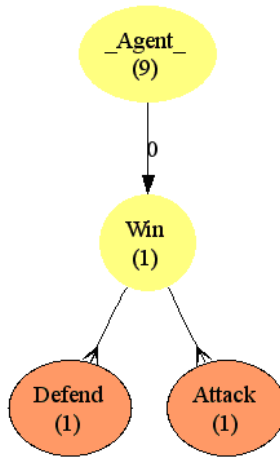


Fig. 2. Goal hierarchy used for demonstration

view. For demonstration purposes, the **Attack** goal succeeds after the agent attacks five enemy players.

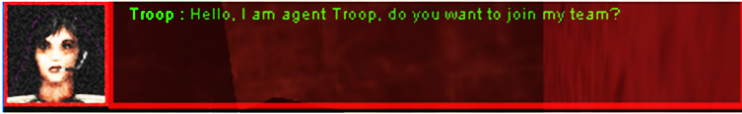
3. A **RequestCollaboration** message is sent to the CMA for the **Defend** goal. The CMA then executes an **@team_achieve** for any sub-teams that perform the **Cooperation** role. The **HumanManager** agent then negotiates and performs assessment with the human in order to satisfy the **Defend** goal.

5.1 The Human's Point of View

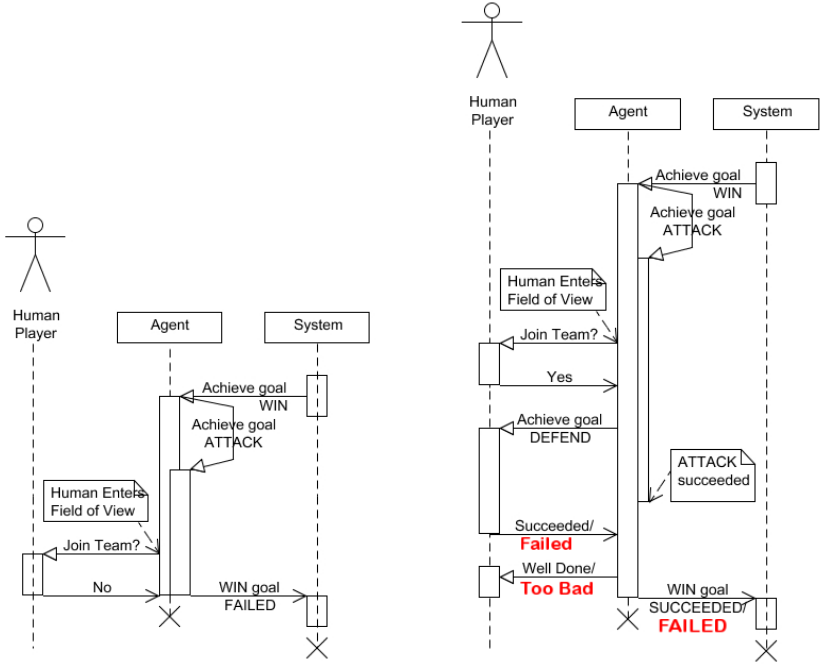
The human is able to communicate with agents via text messages through UT. Figure 3a illustrates what appears on the human's monitor when a message is received from the agent. The sequence diagram shown in figure 3b illustrates that if the human refuses to join the team, the collaboration fails and hence both the **Defend** and **Win** goals both fail. On the other hand, the sequence diagram shown in figure 3c illustrates that when the human accepts to join the team the collaboration is formed and the human is assigned with the **Defend** goal. The result of the **Win** goal then depends on whether the human reports that he/she was successful in achieving the **Defend** goal.

6 Conclusion

Agent collaboration provides the opportunity for agents to share resources during their execution. Such resources are not normally available within current MAS designs because resources are allocated for the use of specific teams. Team structures and team members are defined explicitly when the system is being designed. Using collaboration, agents are able to recognize when additional resources are needed and negotiate with other teams to obtain them. Collaboration



(a) Asking the Human



(b) The Human Refuses

(c) The Human Accepts and Succeeds/Fails

Fig. 3. The Human's point of view

is also a natural way to implement human-agent teaming due the temporary and unpredictable nature of human team members. Future work for the Collaboration implementation includes: The implementation of a collaboration role where agents can negotiate some resources without the involvement of goals. The implementation of generic beliefs and associated monitoring behaviors to manage contracts, resources and the progress of sub-teams in achieving their given roles.

References

1. Bratman, M.E.: Intention, Plans, and Practical Reason. Center for the Study of Language and Information (1999)
2. Rao, A., Georgeff, M.: Bdi agents: from theory to practice. In: Proceedings for the 1st International Conference on Multi-Agent Systems (ICMAS-95), AAAI Press, California (1995) 312-319

3. Urlings, P.: Teaming Human and Machine. PhD thesis, School of Electrical and Information Engineering, University of South Australia (2003)
4. Cheyer, A., Martin, D.: The open agent architecture. *Journal of Autonomous Agents and Multi-Agent Systems* 4 (2001) 143–148
5. Lange, D.B.: Java aglet application programming interface (j-aapi) white paper - draft 2. Technical report, IBM Tokyo Research Laboratory (1997)
6. SwarmDevelopmentGroup: Documentation set for swarm 2.2. [Online Accessed: 19 April 2006] <http://www.swarm.org/swarmdocs-2.2/set/set.html> (2006)
7. Ferber, J., Gutkecht, O., Michel, F.: Madkit development guide. [Online accessed: 19 April 2006] <http://www.madkit.org/madkit/doc/devguide/devguide.html> (2006)
8. AgentOrientedSoftware: Jack intelligent agents teams manual. [Online, accessed 19 April 2006] <http://www.agent-software.com.au/shared/resources/index.html> (2006)
9. Tweedale, J., Cutler, F.: Human-computer trust in multi-agent systems. In: Proceedings of the 10th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES 2006), Oct 9-11, Bournemouth, England. (2006) accepted
10. Jarvis, B., Corbett, D., Jain, L.C.: Beyond trust: A belief-desire-intention model of confidence. In Khosla, R., Howlett, R.J., Jain, L.C., eds.: Proceedings of the 9th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES 2005), Sep 14-16, Melbourne, Australia. Volume 2 of Lecture Notes in Artificial Intelligence., Springer Verlag, Heidelberg (2005) 844–850
11. Sioutis, C., Ichalkaranje, N.: Cognitive hybrid reasoning intelligent agent system. In Khosla, R., Howlett, R.J., Jain, L.C., eds.: Proceedings of the 9th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES 2005), Sep 14-16, Melbourne, Australia. Volume 2 of Lecture Notes in Artificial Intelligence., Springer Verlag, Heidelberg (2005) 838–843
12. Sioutis, C.: Reasoning and Learning for Intelligent Agents. PhD thesis, School of Electrical and Information Engineering, University of South Australia (2006)
13. Hammond, G.T.: *The Mind of War: John Boyd and American Security*. Smithsonian Institution Press: Washington, USA. (2004)
14. Rasmussen, J., Pejtersen, A., Goodstein, L.: *Cognitive Systems Engineering*. Wiley and Sons, New York, NY (1994)
15. Sioutis, C., Ichalkaranje, N., Jain, L.: A framework for interfacing bdi agents to a real-time simulated environment. In Abraham, A., Koppen, M., Franke, K., eds.: *Design and Application of Hybrid Intelligent Systems*. Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, The Netherlands (2003) 743–748

Teamwork and Simulation in Hybrid Cognitive Architecture

Jinsong Leng¹, Colin Fyfe², and Lakhmi Jain¹

¹ School of Electrical and Information Engineering,
Knowledge Based Intelligent Engineering Systems Centre,
University of South Australia, Mawson Lakes SA 5095, Australia
`Lenjy002@students.unisa.edu.au`,
`Lakhmi.Jain@unisa.edu.au`

² Applied Computational Intelligence Research Unit,
The University of Paisley, Scotland
`colin.fyfe@paisley.ac.uk`

Abstract. Agents teamwork is a sub-area of multi-agent systems that is mainly composed of artificial intelligence and distributed computing techniques. Due to its inherent complexity, many theoretical and applied techniques have been applied to the investigation of agent team architecture with respect to coordination, cooperation, and learning skills. In this paper, we discuss agent team architecture and analyse how to adapt the simulation system for investigating agent team architecture, learning abilities, and other specific behaviors.

1 Introduction

Over the last few decades, Artificial Intelligence (AI) [8] and agent-based systems have found wide applications in many areas, for example, Unmanned Aerial Vehicles (UAVs) [24], electronic commerce [6], entertainment [11], and some other agent-based fields. AI and agent systems are tightly combined in some application areas. AI is concerned with studying the components of intelligence, while agent studies focus on integrating components with some properties of AI.

An agent is defined as a hardware and/or software-based computer system displaying the properties of autonomy, social adaptiveness, reactivity, and proactivity [25]. The ability of an agent can be described in terms of agency, intelligence, and mobility. Agency is the degree of autonomy and authority residing in the agent. Intelligence is the degree of reasoning and learned behaviour. Mobility is the degree to which agents themselves travel through a network [10]. Every agent has its own knowledge, capability, and goals. The ability to act without human, or other intervention (Autonomy) is a key feature of an agent.

Multi-agent systems (MASs) differ from single-agent systems in which they are composed of several agents that can interact both among themselves and with the environment. The environment of multi-agent systems can be changed dynamically by other agents. MASs inherit many distributed AI motivations, goals and potential benefits, and also inherit those of AI technologies that may have the ability to deal with incomplete information, or the capability for each

agent to exhibit distributed control with decentralised data and asynchronous computation [14].

Agent teamwork is a sub-research area of MASs that deals with the joint intentions of agents. Agent researchers concentrate mainly on deliberative-type agents with symbolic internal models, dealing with interaction and communication between agents, the decomposition and distribution of tasks, coordination and cooperation, conflict resolution via negotiation, etc [17]. Hence, to make agents work as a team brings out some challenges related to complexity, communication, and dynamic environment. Reinforcement learning and Q-learning [15] are well-known techniques for improving the agent learning skills. Bayesian networks (BN), also called probabilistic networks or belief networks, are proving to be powerful tools for reasoning under uncertainty [4,16]. In fact, a BN is a directed graph with a set of nodes and links. Each node represents a variable that links to a conditional probability table. Bayesian networks are one of the most popular formalisms for reasoning under uncertainty. For example, Dearden et al. [9] adopts a Bayesian approach for maintaining the uncertain information that requires to assess the the agent's uncertainty about its current value estimates for states. A Bayesian approach is used to represent the uncertainty the agent has about its estimate of Q-value of each state.

In this paper, we discuss the teamwork and the cooperative learning based on the BDI (Beliefs, Desires, Intentions) architecture [18]. The behaviours of agents team are investigated using a simulation system. The learning abilities are analysed using Bayesian networks [22]. The rest of paper is organised as follows: section 2 discusses the BDI architecture and agent teamwork. A simulation system is used as the testbed for agent teamwork in section 3. Finally, we present conclusion and future work.

2 Cognitive Architecture and Agent Teamwork

2.1 BDI Architecture

Shoham proposed an Agent-Oriented Programming (AOP) [19] architecture, in which he presents an agent computing paradigm. The agent states (agenthood) consists of mental components such as knowledge, capabilities and goals. One popular approach derived from philosophy and cognitive science is that the agents' mental states are represented as beliefs, desires, and intentions (BDI).

BDI architecture has found large-scale applications in agent-based systems. The BDI architectures are recognised as one of the more successful development architectures for developing complex systems in dynamic environments. The BDI agent model is goal-oriented providing both reactive and proactive behaviour. Each agent has its mental state including beliefs, desires, and intentions. Beliefs are the knowledge of the world or the state of the world, which can be represented as a set of variables, or a relational database, or symbolic logic expressions, or some other data structures. Desires are the goals or objectives that the agent wants to accomplish. Intentions are the deliberative state of the agent, and lead to actions to pursue the goal. As the agent is goal-oriented and it has at least

one goal to achieve. An agent perceives the environment and updates its beliefs, and then executes an action selected from a plan based on its beliefs, resulting in the further changes to the environment.

2.2 Agents Teamwork

A team can be defined as a set of agents having a shared objective and a shared mental state [7]. The aims of agent teamwork research are to improve the concept understanding, to develop some reusable algorithms, and to build high-performance teams in dynamic and possibly hostile environments. Coordination manages dependencies between activities. Coordination and cooperation are necessary for agents in a team to achieve a common goal that may not achieve by individual agent. Bradshaw et al proposed three requirements for effective coordination [5]:

- Common ground: Common ground refers to the pertinent mutual knowledge, beliefs, and assumptions that support interdependent actions in the context of a given joint activity. This includes initial common ground prior to engaging in the joint activity as well as mutual knowledge of shared history and current state that is obtained while the activity is underway.
- Interpredictability: In highly interdependent activities, it becomes possible to plan ones own actions (including coordination actions) only when what others will do can be accurately predicted. Skilled teams become interpredictable through shared knowledge and idiosyncratic coordination devices developed through extended experience in working together; bureaucracies with high turnover compensate for experience by substituting explicit pre-designed structured procedures and expectations.
- Directability: Directability refers to the capacity for deliberately assessing and modifying the actions of the other parties in a joint activity as conditions and priorities change.

Mutual understanding, mutual prediction, and conflict-resolution are important to construct an effective team. In a dynamic and non-deterministic environment, uncertainty is a common property and makes it infeasible to predict all possible situation a priori. Automation and learning are two significant challenges, especially for human-agent interactions or intra-agent team.

3 Simulation Environment

MASs are often extremely complex and it can be difficult to formally verify their properties [13]. In order to test complex behaviors for a given task, it is impossible to use algorithms or other formal methods to predict the whole state space in advance in the dynamic environment. Thus, simulation in MASs plays an important role in the MASs' research area, which can be used for developing the AI techniques and investigating the behaviors of agent systems. The required scenarios can be simulated using testbeds to envision the real

environment. The aspects of the teamwork such as communication, knowledge-sharing, and effective coordination and cooperation will be tested.

Normally, computer games are dynamic and use real-time environments that include complex 2D or 3D space, multiple interacting characters, and an uncertain environment. Unreal Tournament (UT) [2] and RoboCup soccer [3] are considered to be the best-known simulation systems for a variety of research in MASs. UT has been used as the testbed for teaming human-agents in our group [21,23].

3.1 Simulation System

In order to evaluate the most realistic performance of the agents teamwork, we adopt the small size soccer league SoccerBots, which is one of a collection of application of TeamBots [1]. Each soccer team can have 2-5 players (see Fig 1).

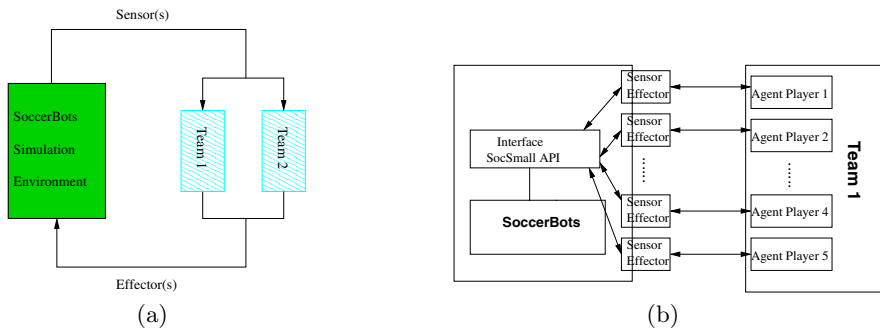


Fig. 1. The SoccerBots and Agents team. (a) SoccerBots. (b) Agent Team Architecture

Each player can observe the behaviors of other objects such as the ball, a teammate, an opponent, and their locations and motions via a sensor. The soccer game players have to interact, coordinate, and cooperate with each other. The action that is sent back to each player is the deliberative outcome of the whole team.

The SoccerBots promotes the following issues to be solved:

- Each player has a local and limited view of the world. Moreover, an agent team may have incomplete and imperfect knowledge of the world. The research question to be addressed is: How do we create techniques to filter the raw data from the individual teammate or also to respond to deceptive agents in the hostile environment?
- In the real-time situation environment, performance is a high priority to be considered to follow the changes of motions of the ball and the players. Due to dynamic changes of location and motion between the ball, the goal, and the players, we cannot completely predict or control events in order to take an action to deal with them. The question arises as to how the locations and motions can be mutually predicted by teammates let alone how they may be

predicted by the opponent. Also how the optimal combination of behaviors can be achieved by adopting some approximate algorithms is a question to be addressed.

- Due to the uncertainty in the dynamic environment, we're investigating Bayesian learning within the cooperative context i.e. learning of optimal joint behaviors when the agents are on the same team. Moreover, we're investigating two sets of competing Bayesian learning teams when these teams are in opposition but not taking into account deceptive strategies.

3.2 JACK Intelligent Agent

JACK [12] is an agent development environment that extends the Java programming language with agent-oriented concepts, including: Agents, Capabilities, Events, Plans, Knowledge Bases (Databases), Resource and Concurrency Management. JACK intelligent Agents has been applied to some application domains such that agents have beliefs about the world and desires to satisfy, driving it to form intentions to act. An intention is a commitment to perform a plan. A plan is an explanation of how to achieve a particular goal, and it is possible to have several plans for same goal. This is one of the key characteristics of the BDI architecture so that the system is able to select alternative approaches in case a particular strategy (plan) fails.

JACK also provides support for team oriented programming, which provides a unique coordination model that is particularly powerful for large scale applications with hierarchical structures. Unique concepts only available in such a teamwork model are teams, roles, and team plans.

We would then be looking at the team programming model provided by JACK teams as a core technology. The JACK Teams model provides a natural way to specify team structure and coordinated behavior related to team structure in terms of BDI model entities.

JACK TeamsTM (Teams) is an extension to JACK agents that provides a team-oriented framework. The JACK agent player is used to control the actions of a soccer player performed in the SoccerBots simulation environment. A team reasoning entity (team agent) is regarded as another JACK agent with its own beliefs, desires, and intentions. Every agent player can be regarded as a sub-team and performs a certain role that defines the relationship between team and sub-teams.

The beliefs of a team agent has the ability to combine the propagated sub-team beliefs. The team plan is used to specify the coordination activities. The @team-achieve statement is used to send an event to the sub-team a sub-team for coordinating the activities among the agent players. Meanwhile, teamplan has the ability to specify the parallel actions using @parallel.

Furthermore, JACK is particularly suitable to extend new agent architectures. The feasibility has been shown in [20], in which the extension to the JACK agent development environment has been developed to provide an abstract way to create learning agents without the need to worry about how the internal

learning algorithms operate. We will concentrate on the new agent teamwork architecture and learning algorithms based on JACK development environment.

4 Conclusion and Future Work

Specific aspects such as team architecture, team performance can be tested in the SoccerBots simulation environment. The ultimate aim of our research is to develop efficient learning algorithms for agent teamwork. The test-suite is to help us analyse the learning abilities of the team. Stochastic techniques such as Bayesian learning possibly combining with reinforcement learning and Q-learning algorithms are adopted for knowledge reconstruction and decision making. We will compare the centralised and distributed approach to agent teamwork and create an effective agent team architecture.

References

1. *TeamBotsTM Domain: SoccerBots*.
<http://www-2.cs.cmu.edu/~trb/TeamBots/Domains/SoccerBots/>.
2. InfoGrames Epic Games and Digital Entertainment. Technical report, Unreal tournament manual, 2000.
3. Humanoid Kid and Medium Size League, Rules and Setup for Osaka 2005. Technical report, Technical report, Robocup, 2005.
4. Michael Berthold and David J. Hand, editors. *Intelligent Data Analysis*. Springer, 1999.
5. J.M. Bradshaw, P. Feltovich, Hyuckchul Jung, Shri Kulkarni, J. Allen, L. Bunch, N. Chambers, L. Galescu, R. Jeffers, M. Johnson, M. Sierhuis, W. Taysom, A. Uszok, and R. Van Hoof. Policy-based coordination in joint human-agent activity. 2:2029–2036, 2004.
6. A. Chavez and P. Maes. Kasbah: An agent marketplace for buying and selling goods. In *Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM-96)*, pages 75–90, 1996.
7. P. Cohen, H. Levesque, and I. Smith. On Team Formation. *Contemporary Action Theory*, 1998.
8. Randall Davis. What Are Intelligence? And Why? 1996 AAAI Presidential Address. *AI Magazine*, 19(1):91–111, 1998.
9. Richard Dearden, Nir Friedman, and Stuart J. Russell. Bayesian Q-Learning. In *AAAI/IAAI*, pages 761–768, 1998.
10. D. Gilbert, M. Aparicio, B. Atkinson, S. Brady, J. Ciccarino, B. Grosz, P. O'Connor, D. Osisek, S. Pritko, R. Spagna, and L. Wilson. IBM Intelligent Agent Strategy. Technical report, IBM Corporation, 1995.
11. S. Grand and D. Cliff. Creatures: Entertainment Software Agents with Artificial Life. *Autonomous Agents and Multi-Agent Systems*, 1(2), 1998.
12. Nick Howden, Ralph Rönquist, Andrew Hodgson, and Andrew Lucas. JACK Intelligent AgentsTM—Summary of an Agent Infrastructure. In *the 5th International Conference on Autonomous Agents*, Montreal, Canada, 2001.
13. N. R. Jennings and M. Wooldridge. Applications of Intelligent Agents. pages 3–28, 1998.

14. Nicholas R. Jennings, Katia Sycara, and Michael Wooldridge. A Roadmap of Agent Research and Development. *Autonomous Agents and Multi-Agent Systems*, 1(1):7–38, 1998.
15. Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, pages 4:237–285, 1996.
16. David J. C. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
17. Hyacinth S. Nwana. Software Agents: An Overview. *Knowledge Engineering Review*, 11(3):205–214, 1996.
18. E. C. Olivéira, K. Fischer, and O. Stepankova. Multi-Agent Systems: Which Research for which Application? *Journal of Robotics and Autonomous Systems*, 27:1-2:3–13, 1999.
19. Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.
20. C. Sioutis. *Reasoning and Learning for Intelligent Agents*. PhD thesis, School of Electrical and Information Engineering, University of South Australia, 2005.
21. C. Sioutis, J. Tweedale, P. Urlings, N. Ichalkaranje, and LC. Jain. Teaming Humans and Agents in a Simulated World. In *Proceedings of the 8th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES 2004)*, pages 80–86. Springer-Verlag, 2004.
22. Michael E. Tipping. Bayesian Inference: An Introduction to Principles and Practice in Machine Learning. In *O. Bousquet et al. (eds.): Machine Learning 2003. LNCS 3176*, pages 41–62. Springer-Verlag, 2004.
23. P. Urlings, J. Tweedale, C. Sioutis, and N. Ichalkaranje. Intelligent Agents and Situation Awareness. In *Proceedings of the 7th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES 2003), Part II*, pages 723–733. Springer-Verlag, 2003.
24. Peter Wallis, Ralph Rönnquist, Dennis Jarvis, and Andrew Lucas. The Automated Wingman - Using JACK for Unmanned Autonomous Vehicles. In *IEEE Aerospace Conference, Big Sky MT*, 2002.
25. Michael Wooldridge and Nick Jennings. Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10, 1995.

Trust in Multi-Agent Systems

Jeffrey Tweedale and Philip Cutler

Air Operations Division,
Defence Science and Technology Organisation,
Edinburgh SA 5111, Australia
{Jeffrey.Tweedale, Philip.Cutler}@dsto.defence.gov.au

Abstract. Research in Multi-Agent Systems has revealed that Agents must enter into a relationship voluntarily in order to collaborate, otherwise that collaborative efforts may fail [1,2]. When examining this problem, trust becomes the focus in promoting the ability to collaborate, however trust itself is defined from several perspectives. Trust between agents within Multi-Agent System may be analogous to the trust that is required between humans. A Trust, Negotiation, Communication model currently being developed, is based around trust and may be used as a basis for future research and the ongoing development of Multi-Agent System (MAS).

This paper is focused on discussing how the architecture of an agent could be designed to provide it the ability to foster trust between agents and therefore to dynamically organise within a team environment or across distributed systems to enhance individual abilities. The Trust, Negotiation, Communication (TNC) model is a proposed building block that provides an agent with the mechanisms to develop a formal trust network both through cooperation¹ or confederated or collaborative associations². The model is conceptual, therefore discussion is limited to the basic framework.

1 Introduction

Many challenges are presented to researchers in Artificial Intelligence as they attempt to increase the level of personification in intelligent systems. These challenges are both technical and psychological in nature. Agent technologies, and in particular agent teaming, are increasingly being used to aid in the design of “intelligent” systems [2,3]. In the majority of the agent-based software currently being produced, the structure of agent teams have been defined by the programmer or software designer. In the current paper, however, a description of work being undertaken to develop a model that extends the architecture of an agent to provide the adaptable functionality required to maintain agent teams in an Multi-Agent System. Work on this adaptive concept is being undertaken by the Defence Science and Technology Organisation (DSTO) and University of

¹ With parent or descendant agents.

² With sibling or external agents at the same level.

South Australia. The mechanism used to manage each partnership is based on trust.

The current paper proposes why the chosen mechanism for trust is appropriate, describes the form and methods of how trust may be used to form successful partnerships³. Section 2 describes the basic architecture of agents, while Sections 3 and 4 discuss the significance of trust to this model. Section 5 of the paper provides a brief description of the model. Future work is presented in Section 6.

2 Agent Architectures

Forming agent teams generally requires prior knowledge of the resources and skills required by teams to complete a goal or set of decomposed tasks. This will be determined by the maturity and composition of the team, its members and any associated teams. The maturity could be measured across each phase of “Team Development” as described by Tuckman [4]⁴. Such personified characteristics/functions are possible in Beliefs, Desires, Intentions (BDI) agent architectures, although each consumes resources that may only be used by relatively few agents, many sporadically during its life. To reduce these overheads, only the resources/functionality required for a specified period would be instantiated and then released some time after that function is no longer required. Based on this premise, interactions between agents within teams and between teams can generally be catalogued. In this paper, a generic structure embodying TNC functionality is discussed. This structure must be capable of being expanded by agents during run-time to enable agents to form teams and to interact with other agents, teams or even distributed across agent systems. Trust is a key characteristic used to mediate this interaction, therefore a discussion on building and measuring trust⁵ is required.

3 The Nature of Trust

Trust is encountered across a broad domain of applications and perspectives with levels of inconsistent influence. A unified definition of trust is elusive as it is context specific, however a number of notions appear commonly in the literature on trust including: complexity [5], reliability [6], flexibility, predictability

³ The agent architecture is based on the proposed Trust, Negotiation, Communication model and implements trust as a bond that can be strengthened via the exchange of certified tokens gained through previous encounters (similar to Microsoft’s cookies). The problems of measuring trust are two fold; the form trust should take, and the measures used need to be quantitative.

⁴ These where initially; Forming, Storming, Norming and Performing, but later (1975) Tuckman added Adjourning.

⁵ Possibly by measuring the strength of the bond or the fragility of the loyalty established.

[7], credibility [8], complacency [9], consistency, situational, experiential, security, accuracy [8], dependence, responsibility and uncertainty. The notions cover a spectrum from the system (local) level, e.g. the usability or reliability of a specific piece of software, through to a grand social level, e.g. responsibility, dependence [10]. In the following sections some of these issues are discussed and their relevance to the current work considered.

In Kelly *et al.* [9] trust is defined simply as the confidence placed in a person or thing, or more precisely, the degree of belief in the strength, ability, truth or reliability of a person or thing. They also list a number of elements of trust identified from the research literature, including faith, robustness, familiarity, understandability, usefulness, self-confidence, reputation and explication of intention. As is evident, the dimensions of trust are many, and it will be important to narrow them down for the current context.

Perhaps the most important factor in trust is risk. The ability to reason about trust and risk allows humans to interact even in situations where they may only have partial information [11]. If the risk is believed to be too great the interaction may not take place. Three distinct components or levels of trust are presented in which at each level there is an increasing tendency to believe that a person is trustworthy. The levels are: predictability (Used in the early stages as a basis for trust), dependability (Corresponds to the trust placed in the qualities attributed to the other entity), and faith (Reflects an emotional security). The dynamic nature of trust is self-preserving and self-amplifying [11]. It is increased through successful interactions and degraded through unsuccessful outcomes. Some level of trust will be gained if a system meets certain expectations which may be different between individuals, and the situation.

4 Measure of Trust

One approach to building trust is to provide a confirmation/feedback mechanism. A description of a socio-cognitive, agent based, model of trust, using Fuzzy Cognitive Maps is provided by Castelfranchi *et al.* [12]. Trust, in any model, is based on different beliefs which are categorised as internal and external attributes. The TNC model is based on the premise that the origin and the justification of the strength of beliefs come from the sources of the beliefs. Four possible types of belief sources are considered: direct experience, categorisation, reasoning and reputation. The credibility value given to an important belief is shown to influence the resulting decision, and the trust worthiness of the system. In addition to these beliefs is loyalty and fragility will be a measure of the priorities used to form the bond⁶. By examining Figure 1 the collaborative bond between siblings within the team (MAS) would have a higher priority to that of another team or

⁶ Kelly [9] discusses this process based on objective and subjective measures. The latter can be measured if its origins can be rated. It is proposed that loyalty will be used to measure the strength of the partnership and determine the frequency and level of monitoring required to maintain that partnership. A measure of the level of loyalty can also be used to weight the bond and merit of information exchanged.

the human-computer interface. The co-operative bond, however, is hierarchical and implicit within that structure i.e. This is analogous to load sharing verses task management/processing.

5 The Conceptual TNC Model

The concept of the TNC model is shown in Figure 1, which details the type of interaction possible between single agents [13,14,15]. The model works for a hierarchy (team) of agents and by loose association for another agent team or even, in the future, human beings within a complex system. The goal of the model being to provided a flexible structure that enables agents to team together without prior configuration (adaptation). Trust is the centre attribute used to form and maintain he partnership(s) and is required for agents joining or already within a team.

The model is effectively a series of wrappers for agents, that extend the communication ability inherent in agents to incorporate an ability to ‘independently’ negotiate partnerships. The model provides a communications interface, negotiation mechanism and a trust monitoring capacity.

The application of agents implemented in the form of the model would be as such. An environment would be built with predefined resources and skills

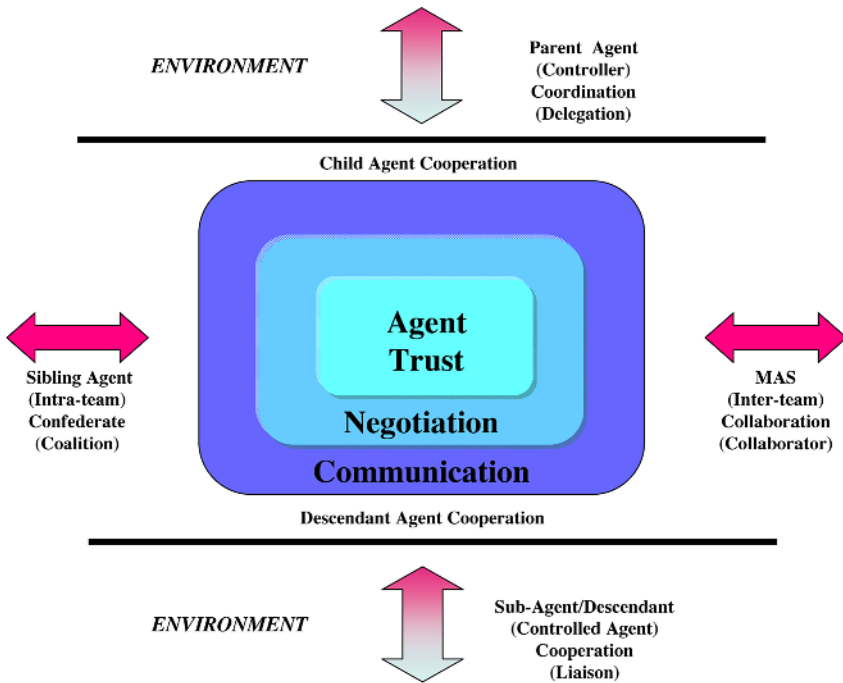


Fig. 1. MAS Trust via Cooperation and Collaborations

which may be context dependent. When a resource is called upon, by another agent or environment, a team (comprising of one or many agents in the form of the proposed model) would be instantiated by the resource manager. Each team would have a hierarchy, where each agent would fulfil a specific skill. The partnerships within the team would be coordinated by a controller agent and the level of control based on loyalty through the bond established⁷.

The approach is analogous to that used by humans. When Human decision makers are presented a problem they are not able to solve themselves, they may choose to form a team of people they trust to assist in making that decision. The team is likely to have a hierarchical structure of some form in order to effectively manage the project in terms of scheduling and tasking. The structure may include multiple levels. The TNC model is based on a similar concept. When decisions are required to be made within an environment, a team would be assembled with the current resources and skills considered necessary to achieve the goal. If either resources or skills are lacking within the environment, the necessary skills or resources may be obtained from another environment or source.

Once a team has been instantiated, agents within the team at each level are optionally able to collaborate with other agents in the team at the same level or with agents at a similar level in another team. Ordinarily, agents are controlled using team hierarchy, with commands being issued from controller agents, at a level above, or delegated to agents at a level below.

5.1 The Trust Layer

As mentioned, the trust layer within the TNC Model is the basis by which teams are structured and by which the partnerships between agents are formed and maintained. Using trust is again analogous to human teams. Bonds or partnerships will be formed based on the trust that members have sufficient capabilities to solve the problem, will provide a solution within a sufficient amount of time, and so on.

Trust is required to initiate a partnership, to remain within a partnership and to resume/re-initiated a partnership. Further detail of this trust layer will be explored as the model matures.

5.2 The Negotiation and Communication Layers

Due to space and existing research on these topics only a brief summary of each will be presented.

The communication layer within the TNC model extends the existing communication abilities of agents by incorporating an ability for an agent to autonomously handle communications based on a variety of different protocols. Analogous to a member of team dealing with another member by voice or with

⁷ The agents within the team would also have the ability to seek partnerships outside of the direct team when they require or are able to provide additional resources. The partnerships established would be based on trust, which would strengthen or wain over time.

another by email. The establishment of an ontology in order to pass information within a partnership is also enabled within the communications layer. This is analogous to establishing the language in which the information will be exchanged within a partnership.

The negotiation layer is acknowledged as a key element of team formation through task allocation and accomplishment. Details on this topic will also be published as the model matures.

6 Future Research

The TNC model is based on *Trust* being *Negotiated* through a common framework of *Communication*. Presently the model is only at a conceptual stage and is expected to be refined over the next few years. However, work will commence on the implementation of the model to provide the end goal of agents having the capability to autonomously implement the administrative aspects of team formation (including human operators) and task/resource management of any agent in an MAS.

For the formation of a bond to occur between two agents the communications layers of both agents must jointly facilitate establishment of the protocol to be used. Establishing and implementing this process across a finite set of protocols will constitute the first phase of the implementation of the model.

The second phase will be to establish and implement methods of measuring the loyalty to establish the strength of the trust within a bond. Loyalty can be used as a weight that can be applied to determine the fragility of the collaborative bond between siblings or inter-team entities. Cohen [16] describes a situation specific trust (SST) model that comprises of a qualitative (informal core) that describes the mental attributes and includes a quantitative (prescriptive extension). This could be used to refine trust estimates formed during the negotiation phase. These estimates must be presented or exchanged using a flexible taxonomy that is easy to interpret and be used in the formative stages prior to the agents establishing bonds between themselves.

The third phase will see the implementation of agents will the abilities to establish bonds and begin the process of providing the ability for agents to form teams in order to solve specific tasks within defined environments.

References

1. d'Inverno, M., Luck, M.: Understanding agent systems. Springer (2001)
2. Wooldridge, M., Jennings, N.R.: The cooperative problem-solving process. *Journal of Logic and Computation* **9** (1999) 563–592
3. Urlings, P.: Teaming Human and Machine. PhD thesis, School of Electrical and Information Engineering, University of South Australia (2003)
4. Mann, R.: Interpersonal styles and group development. *American Journal of Psychology* **81** (1970) 137–140

5. Corritore, C.L., Wiedenbeck, S., Kracher, B.: The elements of online trust. In: Extended Abstracts on Human Factors in Computing Systems (CHI '01), ACM Press (2001) 504–505
6. Prinzel, L.J.: The relationship of self-efficacy and complacency in pilot-automation interaction. Technical Report TM-2002-211925, NASA, Langley Research Center, Hampton, Virginia (2002)
7. Gefen, D.: Reflections on the dimensions of trust and trustworthiness among online consumers. *SIGMIS Database* **33** (2002) 38–53
8. Frankel, C.B., Bedworth, M.D.: Control, estimation and abstraction in fusion architectures: Lessons from human information processing. In: Proceedings of the Third International Conference on Information Fusion (FUSION 2000). Volume 1. (2000) MOC5–3 – MOC5–10
9. Kelly, C., Boardman, M., Goillau, P., Jeannot, E.: Guidelines for trust in future ATM systems: A literature review. Technical Report 030317-01, European Organisation for the Safety of Air Navigation (May 2003)
10. Hoffman, R.R.: Whom (or what) do you (mis)trust?: Historical reflections on the psychology and sociology of information technology. In: Proceedings of the Fourth Annual Symposium on Human Interaction with Complex Systems. (1998) 28–36
11. Cahill, V., Gray, E., Seigneur, J.M., Jensen, C.D., Chen, Y., Shand, B., Dimmock, N., Twigg, A., Bacon, J., English, C., Wagealla, W., Terzis, S., Nixon, P., Serugendo, G.D.M., Bryce, C., Carbone, M., Krukow, K., Nielson, M.: Using trust for secure collaboration in uncertain environments. *Pervasive Computing, IEEE* **2** (2003) 52–61
12. Castelfranchi, C., Falcone, R., Pezzulo, G.: Trust in information sources as a source for trust: A fuzzy approach. In: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems, ACM Press (2003) 89–96
13. Nwana, H.S.: Software agents: An overview. *Knowledge Engineering Review* **11** (1996) 1–40
14. Wooldridge, M., Jennings, N.: Agent theories, architectures, and languages: A survey. In Wooldridge, M., Jennings, N.R., eds.: *Intelligent Agents - Theories, Architectures, and Languages*. Volume 890 of Proceedings of ECAI94 Workshop on Agent Theories, Architectures. Springer-Verlag (1995) 403
15. Consoli, A., Tweedale, J.W., Jain, L.: The link between agent coordination and cooperation. In: 10th International Conference on Knowledge Based Intelligent Information and Engineering Systems. Lecture Notes in Computer Science, Bournemouth, England, Springer Verlag (2006)
16. Cohen, M.S.: A situation specific model of trust to decision aids. *Cognitive Technologies* (2000)

Intelligent Agents and Their Applications

Jeffrey Tweedale¹ and Nihkil Ichalkaranje²

¹ Air Operations Division,
Defence Science and Technology Organisation,
Edinburgh SA 5111, Australia

Jeffrey.Tweedale@dsto.defence.gov.au

² School of Electrical and Information Engineering,
Knowledge Based Intelligent Engineering Systems Centre,
University of South Australia, Mawson Lakes, SA 5095, Australia
Nihkil.Ichalkaranje@unisa.edu.au

Abstract. This paper introduces the invited session of Intelligent Agents and their Applications being presented at the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. This session concentrates on discussing Intelligent Agents and uses some applications to demonstrate the theories reported. An update on last years session [1] is included to report on several key advances in technology that further enable the ongoing development of multi-agent systems. A brief summary of the impediments presented to researchers is also provided to highlight why innovation must be used to sustain the evolution of Artificial Intelligence. This year a variety of challenges and concepts are presented, together with a collection of thoughts about the direction and vision of Intelligent Agents and their Applications.

1 Introduction

During the 20th Century management has pursued “scientific management principles” as a means of providing productivity and improving efficiency within the workplace [2]. Early management focus within this period relied on making machines that concentrated on productivity (making man part of the machine). Mayo challenged this theory by attempting to find out if fatigue and monotony effected productivity and what could be controlled to improve the situation¹, such as; optimising rest breaks, work hours, temperature and humidity. Maslow further investigated the concept that these experiments were more about mechanisation[3] and less about the workers hierarchy of needs². Observers like Adam Smith the economist (1723-1790) and Charles Babbage the mathematician (1792-1871) (Babbage (1835)) have equally displayed powers

¹ He achieved this using new management techniques captured during his Hawthorn Experiments.

² Man’s basic needs are physiological, for example, hunger, thirst and sleep. When these are satisfied it is easier to impart the need for safety, reflecting the personal desire to protection against danger and/or deprivation.

of analysis and observation on which the future machine-based developments were to be based. For these reasons and many more, the average worker has created a distrust [4] of machines and automation. An accepted definition of *automation* is: “a device or system that accomplishes (partially or fully) a function that was previously carried out (partially or fully) by a human operator.” A number of important points are emphasised about this definition. First, *automation* is a continuum ranging from the level of manual control to full automation. Second, is the level of the balance of control assumed when implementing these systems i.e. Automation must cater for intensive communication, management and co-ordination between the different members of the human-machine team [5].

Not only has management alienated its workers against machines, cinema has portrayed machines as evil and capable of replacing people. This concept is brilliantly portrayed in Charlie Chaplin’s satire about the mechanised world [6] and again in George Orwell’s [7] prophetic nightmarish vision of a “Negative Utopia” in his book titled *1984*. Beginning in the late nineteen eighties, management were also subject to the limitations experienced during failed attempts to deliver Expert Systems (ES). More recently pilots and operators of complex equipment, like the Chernoyvl Nuclear Reactor, have experience cataclysmic failure due to mode confusion or missed information. Future developments in Artificial Intelligence (AI) need to concentrate on building trust and scalable agent architectures capable of true autonomous behaviour [8,9,10] and team behaviour in integrated environments.

Since last year a number of changes have enabled advances in agent design to continue; i.e. computer architectures are beginning to change, with bus speeds, memory and storage capacities all increasing. Research has already indicated that manufacturers are becoming more creative to maintain the growth in microprocessor speeds, although significant architectural changes are still required. For instance the Von Neuman architecture is not suited to pattern matching or massively parallel processing³. Last year Intel conceded that increasing power consumption and problems with heat dissipation proved to be a significant impediment in providing further speed enhancements to its current Pentium 4tm based on its “NetBurst” architecture [11]. Intel has introduced a new dual core AI-32 Pentium Mtm microprocessor, with a “wider instruction pipeline”, enabling it to run at lower speeds and dissipate less heat⁴. Intel predicts this new design can achieve microprocessor densities of over 100 cores on a single chip using lithography fabrication techniques based on 32 nm in a plant due to open

³ It is not clear how many threads the new series of microprocessors can spawn and maintain simultaneously i.e. Agents require multiple thread machines with multiple Arithmetic Logic Unit (ALU)/Floating Point Unit (FPU) capacity. Time will also reveal how new compiler will address large monolithic programs.

⁴ This chip is also being migrated from 90 nm to 65 nm lithography fabrications techniques. This achievement is based on improved lithographic fabrication techniques that have produced an order of magnitude reduction in track sizes from 1 μm to under 130 nm. Techniques, such as “state of the art” Extreme Ultra Violet (EUV) techniques are now being investigated that surpass 32 nm barrier [12].

by the end of 2008⁵. Existing research is being realised to maintain *Moore's Law* [13]. Technologies, such as; Atomic Layer Deposition (ALD), will allow the self-assembly of molecules one mono-atomic layer at a time⁶. To gain some perspective of these gains, 100 of these new transistor gates would fit inside the diameter of a single human red blood cell [13]. Eventually technology will surpass the needs for AI and fully scalable multi-agents systems capable of solving real world problems. At present we need to ensure our designs are tested and capable of this function and scale. Hence the range of papers being presented in this session.

2 Session Papers

During the current KES conference, two representatives from the “Knowledge And Intelligent-Agent Centre of Excellence” (KAICE) group, one from the Defence Science and Technology Organisation (DSTO) and the other from University of South Australia (Uni-SA), hosted an invited session titled “Intelligent Agents and their Applications”. The session includes five papers which present discussion on research being conducted by the group, which covers new or emergent areas of agent technology such as teaming, learning and inherent agent components such as trust, negotiation and communication.

The first paper by Tweedale and Cutler [14], focuses on trust in multi-agent systems (MAS). This paper is focused on discussing how the architecture of an agent could be designed to provide it the ability to foster trust between agents and therefore to dynamically organise within a team environment or across distributed systems to enhance individual abilities. A conceptual model called the Trust, Negotiation and Communication (TNC) model is a proposed building block to facilitate an agent with the mechanisms to develop a formal trust network both through cooperation or confederated or collaborative associations.

The second paper by Leng, Fyfe and Jain [15], focuses on multi-agent teaming. In this paper the authors address different aspects of teaming and agent teaming architecture from simulation point of view. Further more authors discuss agent team architecture and analyse how to adapt the simulation system for investigating agent team architecture, learning abilities, and other specific behaviours. The authors utilise a simulation platform namely Soccer bots (Team bots) to exercise feasibility of simulating teaming notions.

The third paper by Sioutis and Tweedale [16] describes preliminary work performed to gain an understanding of how to implement collaboration between intelligent agents in a multi-agent system and/or with humans. The paper builds on previous research where an agent-development software framework was implemented based on a cognitive hybrid reasoning and learning model. Authors demonstrate how agent relationships are formed by utilising a three-layer

⁵ Transitioning from 45 nm using new three dimensional High-k Gate Dielectric transistors by 2007 [12].

⁶ Enabling manufacturers to produce silicon circuits with in excess of 50 Quadrillion Nano-transistors.

process involving communication, negotiation and trust. Cooperation is a type of relationship that is evident within structured teams when an agent is required to cooperate with and explicitly trust instructions and information received from controlling agents. Collaboration involves the creation of temporary relationships between different agents and/or humans that allows each member to achieve their own goals. Due to the inherent physical separation between humans and agents, the concept of collaboration has been identified as the means of realising human-agent teams. The authors support their claims using a preliminary demonstration.

The fourth paper by Jarvis and Jain [17], focuses on Trust in a BDI architecture. The authors choose a well defined BDI model, namely Logic of Rational Agents (LORA), to demonstrate and extend an existing model called the Ability-Belief-Commitment-Desire (ABCD) [18]. The ABCD model forms the basis of the author's theory that *trust* is a *belief* about another agent's abilities, beliefs, commitments and desires. Furthermore the authors introduce trust as being an integral step in delegation.

The fifth paper by Ichalkaranje et.al [19], outlines the abridged history of agent reasoning theories as 'agent mind' from the perspective of how the implementation can be inspired using new trends, such as; 'teaming' and 'learning'. This paper covers how the need for such new notions in agent technology introduced a change in fundamental agent theories and how it can be balanced by inducing some original cognitive notions from the field of 'artificial mind'. This paper concentrates on the popular agent reasoning notion of BDI. Furthermore this paper outlines the importance of the human-centric agent reasoning model as a step towards the next generation of agents to bridge the gap between human and agent. The authors explain the current trend including the human-centric nature of agent mind and human-agent teaming, along with its needs and characteristics. This paper reports add-on implementation on BDI in order to facilitate human-centric nature of agent mind. This human-centric nature and concepts such as teaming agreements are utilised to aid human-agent teaming in a simulated environment. Finally the authors discuss the issues in order to make agent more human-like or receptive and the widespread implementation in software community.

3 Conclusion

The main challenges faced by researchers in the field of multi-agent systems include: the ability to solve problems with the architectures currently available, being able to scale those solutions and finding an agent architecture suitable of providing the variety of functionality required to achieve those outcomes. Given that agents are now commonly found in areas, such as; gaming, education, medicine, transportation, commerce, banking, on the web and in many automated recreational devices, we are also faced with the growing need to mentor sufficient advocates to cover the diversity of applications being conceived.

These people will need to become increasingly adaptive in order to develop hybrid solutions to complex problems. There will be a growing movement towards applications capable of fusing data sources to provide intelligent outputs.

Systems already exist that can source a variety of providers for the best price on a specified item, or determine the most cost effective method of attending a conference on given dates. Allied systems can even assist with hotel, taxis or vehicle bookings and provide a weather forecast. Future extensions for these concepts may be developed for the providers, who could ascertain those people, or those most likely to attend a conference this year, generate a tailored package, complete with a host of options, including the form of travel, bands of investment, tours and entertainment all based on information found on-line.

We may even be able to re-target areas such as automation or autonomous processes using new technologies and architectures. In a hostile military environment, this concept extends to create a team of agents capable of providing a seamless flow of information, together with integrated connectivity and automated processes. The TNC model is embryonic to this style of functionality and contributes to the vision for future research in this field.

The topics covered throughout the invited session present the audience with a host of concepts and sufficient detail to pursue further study in the fields discussed. The papers presented are listed in the bibliography and published in the proceedings of the 10th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES06) by Springer-Verlag in 2006.

Acknowledgements

We would like to thank the reviewers for contributing their valuable time and effort. Their work contributed to the quality of the papers and is gratefully appreciated.

References

1. Tweedale, J., Ichalkaranje, N.: Innovations in intelligent agents. In Khosla, R., Howlett, J., Jain, L., eds.: 9th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES05), Melbourne, Australia, Springer-Verlag, Berlin, Germany (2005) 821–824
2. Druker, P.F.: The key decisions, in managing for results, Harper Business, New York (1993) 195–202
3. Amabile, T.M.: Motivating creativity in organisations: On doing what you love and loving what you do. *California Management Review* 1 (1997) 39–58
4. Hardin, R.: *Distrust*. Russel Sage Foundation, New York (2004)
5. Kelly, C., Boardman, M., Goillau, P., Jeannot, E.: Guidelines for trust in future atm systems: A literature review. Technical Report 030317-01, European Organisation for the Safety of Air Navigation, Naval Weapons Center, China Lake, CA (May 2003)
6. Chaplin, C.: *Modern times*. Movie (1936)

7. Orwell, G.: *Nineteen Eighty Four*. Clarendon Press, Oxford (1949)
8. Urlings, P.: *Teaming Human and Machine*. PhD thesis, School of Electrical and Information Engineering, University of South Australia (2003)
9. Klein, G.A.: *Decisionmaking in complex military environments*. Technical report for the naval command, control and ocean surveillance center, san diego, ca, Klein Associates Inc., Fairborn, OH (1992)
10. Klein, G.A., Calderwood, R., MacGregor, D.: *Critical decision method for eliciting knowledge*. *IEEE Transactions on Systems, Man and Cybernetics* **19** (1989)
11. P.Gelsinger, P.: *Technology and Research at Intel: Architectural Innovation for the Future* for the Future Technology and Research at Intel. Intel Press Release, Remond, America (2004)
12. Holt, B.: *Intel 45nm Technology will take future platforms to new performance-per-Watt Levels*. Intel Press Release, Remond, America (2006)
13. Moore, G.E.: *Intel silicon innovation: Fueling new solutions for the digital planet*. Eelctronic, Remond, America (2005)
14. Tweedale, J., Cutler, P.: *Trust in Multi-Agent Systems*. In Gabrys, B., Howlett, J., Jain, L., eds.: *10th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES06)*, Bournemouth, England, Springer-Verlag, Berlin, Germany (2006) in Press.
15. Leng, P., Fyfe, C., Jain, L.: *Teamwork and simulation in hybrid cognitive architecture*. In Palade, V., Howlett, J., Jain, L., eds.: *10th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES06)*, Bournemouth, England, Springer-Verlag, Berlin, Germany (2006) in Press.
16. Sioutis, C., Tweedale, J.: *Agent cooperation and collaboration*. In Gabrys, B., Howlett, J., Jain, L., eds.: *10th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES06)*, Bournemouth, England, Springer-Verlag, Berlin, Germany (2006) in Press.
17. Jarvis, B., Jain, L.: *Trust in LORA: Towards a formal definition of trust in bdi agents*. In Gabrys, B., Howlett, J., Jain, L., eds.: *10th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES06)*, Bournemouth, England, Springer-Verlag, Berlin, Germany (2006) in Press.
18. van der Hoek, W., Wooldrige, W.: *Towards a logic of rational agency* (2003)
19. Ichalkaranje, N., Jain, L., Sioutis, C., Tweedale, J., Urlings, P.: *The equilibrium of agent mind: The balance between agent theories and practice*. In Gabrys, B., Howlett, J., Jain, L., eds.: *10th International Conference on Knowledge Based Intelligent Information and Engineering Systems (KES06)*, Bournemouth, England, Springer-Verlag, Berlin, Germany (2006) in Press.

From Community Models to System Requirements: A Cooperative Multi-agents Approach

Jung-Jin Yang and KyengWhan Jee

School of Computer Science and Information Engineering
The Catholic University of Korea
San 43-1 YuckGok 2-Dong WonMi-Gu BuCheon-Si KyungGi-Do, South Korea
Tel: +82-02-2164-4678, Fax: +82-02-2164-4777
{jungjin, sshine106}@catholic.ac.kr

Abstract. The ubiquitous computing technology utilizing multi-agent system provides services suitable to the given situation without the restraint of time and space. Agent-oriented software engineering (AOSE) emerges to build a multi agent system in a large scale and capable of accommodating communications between diverse agents. Building on a meta-model of the community computing for the abstraction of agent system design in a previous work, our work focuses on a structure of community computing middleware leading to the extension of the system. The middleware proposed enables the framework of community computing, that is, a cooperative multi-agent approach, to be constituted in a dynamic fashion.

1 Introduction

As different informational devices and applications sink into our life, the ubiquitous computing paradigm postulated by Mark Weiser [1] is being realized one by one. For autonomous entities such as agents to interact with one another, they need to know, beforehand, what kinds of interfaces they need to know, what kinds of interfaces they support and what protocols or commands they understand. In a truly distributed scenario, such as the ubiquitous computing environment, it may not be reasonable to assume that such agreement exists [2]. Although researchers of the ubiquitous computing share the presupposition that smart entities' autonomous and adaptive handling of the change of their surroundings improves the user's life [3], an agent, which is the primary entity that constitutes the Ubiquitous Computing, is not adaptively correspond to the distributed and dynamic environment as much as its cognitive ability of the environment to which it belongs. Agents that play a main role in the new computing environment are various agents utilizing the knowledge base in the Ubiquitous Computing environment. In order for the agents to solve problems cooperatively in a disparate and distributed environment, it is necessary they pro-actively change for appropriate processes on the basis of communication and data and knowledge expression etc. that can be integrated. The middleware for the Ubiquitous Computing should be a base structure on which the relationship between an agent and agents can be adaptively reconstructed and be capable of flexibly accommodating new sources of

knowledge and entities. Building on a meta-model of the community computing for the abstraction of agent system design in a previous work [4], our work focuses on a structure of community computing middleware leading to the extension of the system. The middleware proposed enables the framework of community computing, that is, a cooperative multi-agent approach, to be constituted in a dynamic fashion.

2 Related Work

Work of Multi-agent system in ubiquitous computing environment often needs to deal with contextual information to provide appropriate services while transcending physical location and computing platform. The situation should be recognized and inferred to provide the context-aware service. And the model is required to re-represent the situation, and these situations (circumstances) need to be controlled which are modified and generated from diverse individuals. Román et al [5] suggested GAIA, a middleware to recognize the situation in ubiquitous environment and perform the appropriate task. GAIA is the model to re-represent the recognized situation, and re-expresses a variety of circumstances by using the technical terms, classifying their conformation and employing the restricting factors. GAIA provides the Context Provider Agent to recognize and create the situation, Context Consumer Agent to consume the created situation, and Context Provider Lookup Service. GAIA also offers Context Synthesizer Agent to provide the inferred (deduced) situation based on the recognized circumstance, and Context History to perform the data-mining by storing the circumstantial information. u-Health Care needs the intelligent services to provide the medical treatment suitable for the diverse situations. In this regard, Kirn et al [6] suggested SESAM (ShEll for Simulated Agent systeMs) to obtain the analytical results according to the scenario which was modeled in the absence of embodiment of Hospital Information System. SESAM defines the characteristics of Agent, and simulates the application of Agent basis without programming by describing the Behavior of Agent and utilizing UML behavioral diagram. SESAM can be an example which demonstrates the formulated language and manufacturing devices for production and analysis of Agent. However, the Agent which acquires and embodies the modeling restricted to the designing capability of simulator showed the properties dependent on a certain platform.

It is also necessary to improve the efficacy of agent design, embodiment and analysis by introducing the standardized language, manufacturing device and monitoring tool for building multi-agent system. There are many methodologies and models based on the organizational and social abstractions. A list of methodologies and their evaluations can be found in [7,8]. Gaia is originally proposed by Wooldridge et al [9] and extended by Zambonelli et al [10]. Gaia is a complete methodology based on the organization metaphor for the development of MAS and views the process of analyzing and designing MAS as one of constructing computational organizations. The extensions of Gaia enriches the modeling capability in the all agents level with the three high-level organization abstractions and concepts in order to deal with the complex open MAS. However, it is still weak in the level of two or more agents acting in a coordinated way.

Our work take perspectives of both agent oriented software engineering and MAS in ubiquitous computing environment into account massively to come up with a meta-model and middleware architecture of community computing in MAS.

3 A Meta-model for Community Computing

The meta-model in [4] is to actively organize circumstantial communities and provides the base structure for the performance of assigned tasks. Also, the model presented is supposed to lead to the adaptive and circumstantial reconstitution of communities.

- **Agent and Agent Role Classifier:** Agent is the smallest unit constituting the community computing and cannot be divided further. The Agent Role Classifier in Fig. 1 is a classifier assorting agents, and agents are classified by at least one agent role classifier. The agent classifier can form the hierarchy of agents according to their features.
- **Community Manager:** The smallest unit needed to complete a task in the community computing is 'community.' The community manager administers resources consumed in the community computing and constitutes communities. The community manager casts communities or agents classified by the role classifier as the community members. Communities performing a task inform the community manager of their demise upon the accomplishment of their aims so as to return their resources.
- **Community and Community Role Classifier:** A community has a group communication of certain individuals, whose original purpose can be achieved by subdividing a certain task and constituting a community fitted to the subdivided purposes. A community can be cast with a *community member to communicate with other members*.
- **Structured Community and Emergence Community:** The Structure Community carries out the communication templetized according to the designed purpose, and the Emergence Community communicates according to the dynamically assigned purpose of the community members.

4 Requirements for Community Computing Infrastructure

The requirements of constructing community computing infrastructure based on our meta model are following. Role Classifier that classifies Agent and Community is formalized and guaranteed to get validity of being stored and retrieved in a knowledge base. Role Classifier also describes relevant community features to support community members to take different roles in different communities. For purpose of casting community members, Role Classifier needs to admit Match-making requirements [11]. Community Description Model needs to be described in a standardized language for describing an organized community. It needs to represent roles of carrying a task, knowledge required, and information for communication. A well-defined model provides modeling interfaces for designer and improves re-usability of an organized community. Communications occurring in Community Computing can be also

formalized and described in a knowledge base for semantic interoperability of information. When information generated and consumed by an entity is represented in a formalized model, information required itself can be a goal of community call. It could improve casting for community members and the productivity of community design. For this to be done, information model needs to be described to allow Discovery [12] requirements. Ultimately, all entities of Community Computing, except Community Manager that takes a role of kernel in Community Computing, need to be a body of knowledge base and Community Computing Infrastructure should be a framework of inducing knowledge evolution.

4.1 Framework for Community Computing

Community Computing Infrastructure as in Fig. 1, constitutes middleware library activating based on FIPA-compliant Multi-agent system. Community Computing Middleware Library provides core modules of Community Observer, observing a community and collects information, and Community Manager, managing a community. It also constitutes basic agents for an effective communication among entities. Ontology Agent Factory for agentifying an entity and Utility Agent Factory for reusing of a community, for instance.

Ontology Repository is an entity of providing knowledge participating in Community computing within a dynamic and open distributed environment. Diverse ontology repositories and multi-agent system needs to interact with each other in an internal structure-independent fashion.

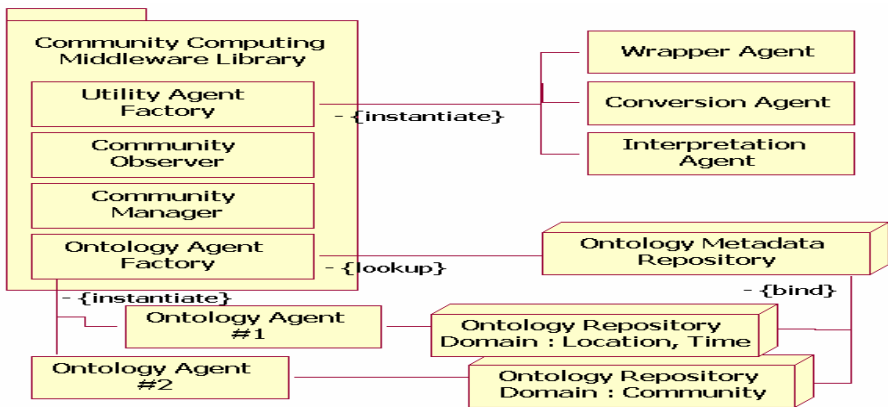


Fig. 1. Community Computing Middleware Library

4.2 Ontologies for Community Computing

When entities of community computing are turned into ontologies, the sharing and reuse of knowledge is ensured and the extension of knowledge is facilitated through logic-based inference and validity checking of knowledge model. In order to meet the requirements of community computing mentioned above, we tried to agentify ontology repositories, (that is, an ontology agent,) generalize the way of using ontology

through agent ACL messages, and allow agents request a query and get a response using ontology repositories and standardized semantic query language. If semantic queries and responses are carried through unified protocol to access ontology repositories within a community computing infrastructure, diverse entities such as ontology repositories and agent system are ensured of interoperability.

The general methodology of the utilization of ontology can be obtained if the community manager circumscribes the type of necessary information to realize the community computing. To utilize the ontology repositories in diverse fields shown in Fig. 1, the ontology agent factory obtains the query and connection privilege optimized to the ontology structure by interacting with the Ontology Repository Naming Service. For instance, the ontology agent factory can constitute an ontology agent by collecting the query, which enables the community manager to retrieve the community description and role classifier to be used, from the Community Computing Ontology Repository. Or it can do so by acquiring the privilege from the Location Ontology Repository or the Healthcare Ontology Repository in order to retrieve the context to be used in the community. The ontology agent produced subsequently is cast as a community member with the role of utilizing the ontology repository and communicates with other members.

It is to ensure the access to a variety of ontology repositories that the ontology agent factory is used since the adaptability of the services offered by the community computing proposed in this paper is governed by the ontology structure.

4.3 Utility Agents for Communication

Utility Agent Factory takes a role of generating agents that are required to carry harmonious communication among entities participating in community computing.

- **Wrapper Agent:** The Wrapper Agent plays the role of controlling diverse existing systems like those agents producing/spending agreed contexts and makes the agent-based community computing possible. The Wrapper Agent acquires information about the required context, the related service shown in Yellow Page, and service description by the communication with the ontology agent. On the ground of the acquired information, passive entities are activated by the Object Broker expressed in Fig.2.

The ontology repository has to ensure the role of services provided by passive entities and the semantic interoperability of the produced/consumed data, as well as furthermore to function as a knowledge store for the performance of the Service Composition through the community computing. A number of agents organizing a community produce and consume various types of context. However, an agent may not be provided with a context consumed to perform a certain function in diverse and constantly changing situations. On the other hand, it has to produce a context to be used to design an agent that shall play a new role.

- **Conversion Agent:** Likewise, the Conversion Agent, as shown in Fig.1, functions to produce the necessary context out of the context produced and consumed by an agent performing a similar function, and it produces the intended context through such changing processes as filtration, merger, composition, translation,

and calculation. As in Fig.1, a context is included in a message to be transferred and produces the object message out of at least more than one message. The Conversion Agent retrieves XSLT-oriented conversion rules, which are employed in the converting task, from the ontology repository and carries out conversion using the XSLT parser on the ground of the rules. The conversion agent is cast as a source-context-producing agent or as a community member together with the community.

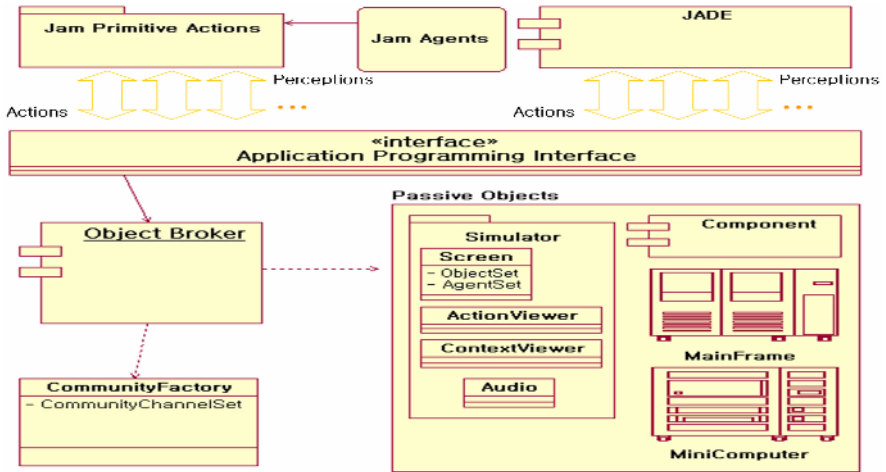


Fig. 2. Mechanism of Agentifying Passive Entity

- **Interpretation Agent:** The Interpretation Agent, as depicted in Fig.1, intermediates translatable protocols between agents. The proxy agent retrieves translation rules described by the ontology agent and is cast into a community to enable the community members of different protocols to communicate with one another.

5 Discussion

We experimented on our meta-model of the community computing and the requirements of the architecture by conducting a child safety scenario in IDIS simulator [4] and gained the support of realization of our system.

The generation of community suitable for a situation and the communication and activities of community members are implemented in JAM [13] that is a representative BDI agent architecture. The simulator is designed and implemented for monitoring the virtual environment. The simulator is embodied in Java and consists of passive objects as described in Fig. 1. that are Wrapper agents.

Class diagrams of the simulator are depicted in Fig. 3 and the simulator controls the simulator screen, sounds, and global storage of messages. Particular entities such as agents and objects are generated through simulator screen control. Their movements can be controlled through the screen and message transferring and both generating and destroying of community can be monitored.

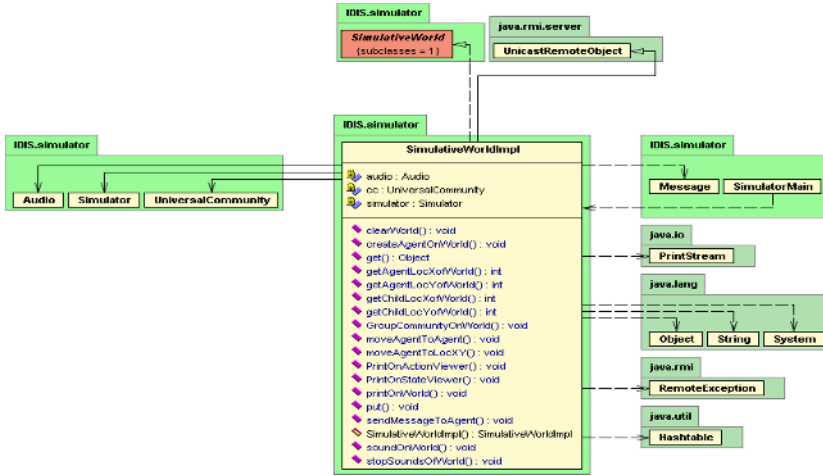


Fig. 3. Class Diagrams of Simulative World

The simulator provides translational message channels, and the method of message passing is conceptually the same as that of the Emergence Community broadcasting.

The embodied agents performed autonomous communications as agents of independent mind, and there have been a lot of trials and errors in constituting their BDI. Although the number of agents is relatively small in the simulator, the value of Beliefs, which changes simultaneously in concept, and uncontrolled communications have sometimes shown entirely unintended results. Consequently, it has been learned that behaviors should be constituted in small units to encompass diverse uncertainties since the Emergence Community possesses the uncertainty each community member possesses at Behavior level.

Additionally, it has also been found out that the Emergence Community can be re-constituted into the Structured Community by monitoring and tracking, and that the utilization of a tool that induces the Structured Community from the Emergence Community enables community computing more feasible.

6 Conclusion

Communities with uncertainties should be capable of being transformed into communities free of those uncertainties, and can be formalized into Community Description Model. It will be the subject of our future research. Also, the Emergent Community can be used in research in a variety of fields and poses the necessity of monitoring tools and translators that converts them into the Organized Community.

In conclusion, our work, building on the preceding research in this area, induced a meta-model of the community computing and the requirements of the middleware architecture. While previous researches constituted a meta-model for the agent group reflecting the viewpoint of the designer, this research provides a meta-model susceptible to result in an uncertain outcome by guaranteeing the autonomy of the agent. The ultimate aim of the meta-model is to induce the change of knowledge through

uncertain communications and support agent systems that can induce the evolution of knowledge.

Acknowledgement. This research is supported by the KOCCA 2005 CRC under grant number 1-05-4005-002-2401-00-002 to by the ubiquitous Autonomic Computing and Network Project, the Ministry of Information and Communication(MIC) 21st Century Frontier R&D Program in Korea.

References

- [1] Weiser M., The computer for the 21st Century. *Scientific American*, 265(3), 66-75 1991
- [2] Ranganathan A. , McGrath R.E., Campbell R.H., and Mickunas M.D., Ontologies in a Pervasive Computing Environment, In Workshop on Ontologies and Distributed Systems (part of the 18th International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco, Mexico, 2003
- [3] Dey A.K., Salber D., and Abowd G.D., A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction*, vol.16. 2001
- [4] KyengWhan Jee, Jaeho Lee, and Jung-Jin Yang, From Agents to Communities: A Meta-model for Community Computing in Multi-Agent System, to appear in Proceedings of 2nd Int. Workshop on Massively Multi-Agent System, Hakodate, Japan, 2006
- [5] Gaia: A Middleware Infrastructure to Enable Active Spaces. Manuel Román, Christopher K. Hess, Renato Cerqueira, Anand Ranganathan, Roy H. Campbell, and Klara Nahrstedt, In *IEEE Pervasive Computing*, pp. 74-83, Oct-Dec 2002.
- [6] Stefan Kirn, Christian Heine, Rainer Herrler, Karl-Heinz Krempels. Agent.Hospital - agent-based open framework for clinical applications, *wetice*, p. 36, Twelfth International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2003.
- [7] G. Weib, Agent Orientation in Software Engineering, *The Knowledge Engineering Review*, 16(4):349-373, 2003
- [8] O.Arazy and C.Woo, Analysis and Design of Agent-Oriented Information Systems, *The Knowledge Engineering Review*, 17(3):215-260, 2002
- [9] M. Wooldridge, N. Jennings, and D. Kinny. The Gaia Methodology for Agent-Oriented Analysis and Design, *International Journal of Autonomous Agents and Multi-agent System*, 3(3):285-312, 2000
- [10] F. Zambonelli, N. Jennings, and M. Wooldridge, Developing Multiagent Systems: The Gaia Methodology, *ACM Transactions on Software Engineering Methodology*, 12(3):317-370, 2003
- [11] David Trastour, Claudio Bartolini, and Javier Gonzalez-Castillo, A Semantic Web Approach to Service Description for Matchmaking of Services, HP Laboratories Bristol, Bristol HPL-2001-183, <http://www.hpl.hp.com/techreports>, July 30 2001.
- [12] Robert E. McGrath, Discovery and Its Discontents: Discovery Protocols for Ubiquitous Computing, Department of Computer Science University of Illinois Urbana-Champaign, Urbana UIUCDCS-R-99-2132, March 25 2000.
- [13] Marcus J. Huber, Ph.D., Usage Manual for the Jam! Agent Architecture, November 2001, <http://www.marcush.net/IRS/Jam/Jam-man-01Nov01.doc>

Scheduling Jobs on Computational Grids Using Fuzzy Particle Swarm Algorithm

Ajith Abraham^{1,3}, Hongbo Liu², Weishi Zhang³, and Tae-Gyu Chang¹

¹ IITA Professorship Program, School of Computer Science and Engineering,
Chung-Ang University, Seoul 156-756, Korea

ajith.abraham@ieee.org

² Department of Computer, Dalian University of Technology, 116023, Dalian, China
lhb@dlut.edu.cn

³ School of Computer Science, Dalian Maritime University, 116024, Dalian, China
teesiv@dlmu.edu.cn

Abstract. Grid computing is a computing framework to meet the growing computational demands. This paper introduces a novel approach based on Particle Swarm Optimization (PSO) for scheduling jobs on computational grids. The representations of the position and velocity of the particles in the conventional PSO is extended from the real vectors to fuzzy matrices. The proposed approach is to dynamically generate an optimal schedule so as to complete the tasks within a minimum period of time as well as utilizing the resources in an efficient way. We evaluate the performance of the proposed PSO algorithm with Genetic Algorithm (GA) and Simulated Annealing (SA) approaches.

1 Introduction

A computational grid is a large scale, heterogeneous collection of autonomous systems, geographically distributed and interconnected by low latency and high bandwidth networks [1]. Job sharing (computational burden) is one of the major difficult tasks in a computational grid environment [2]. Grid resource manager provides the functionality for discovery and publishing of resources as well as scheduling, submission and monitoring of jobs. However, computing resources are geographically distributed under different ownerships each having their own access policy, cost and various constraints. The job scheduling problem is known to be NP-complete. Recently genetic algorithms were introduced to minimize the average completion time of jobs through optimal job allocation on each grid node in application-level scheduling [3],[5]. Because of the intractable nature of the problem and its importance in grid computing, it is desirable to explore other avenues for developing good heuristic algorithms for large scale problems.

Particle Swarm Optimization (PSO) is a population-based optimization tool, which could be implemented and applied easily to solve various function optimization problems, or the problems that can be transformed to function optimization problems [4]. In this paper, fuzzy matrices are used to represent the position and velocity of the particles in the PSO algorithm for mapping the job

schedules and the particle. Our approach is to dynamically generate an optimal schedule so as to complete the tasks within a minimum period of time as well as utilizing all the resources.

Rest of the paper is organized as follows. In Section 2, issues related to grid resource management and scheduling is provided following by the proposed PSO based algorithm in Section 3. Experiment results are presented in Section 4 and some conclusions are provided towards the end.

2 Grid Resource Management and Scheduling Issues

The grid resource broker is responsible for resource discovery, deciding allocation of a job to a particular grid node, binding of user applications (files), hardware resources, initiate computations, adapt to the changes in grid resources and present the grid to the user as a single, unified resource [5]. To formulate the problem, we consider J_j ($j \in \{1, 2, \dots, n\}$) independent user jobs on G_i ($i \in \{1, 2, \dots, m\}$) heterogeneous grid nodes with an objective of minimizing the completion time and utilizing all the computing nodes effectively. The speed of each grid node is expressed in number of Cycles Per Unit Time (CPUT), and the length of each job in number of cycles. Each job J_j has its processing requirement (cycles) and the node G_i has its calculating speed (cycles/second). Any job J_j has to be processed in the one of grid nodes G_i , until completion. Since all nodes at each stage are identical and preemptions are not allowed, to define a schedule it suffices to specify the completion time for all tasks comprising each job.

To formulate our objective, define $C_{i,j}$ ($i \in \{1, 2, \dots, m\}$, $j \in \{1, 2, \dots, n\}$) as the completion time that the grid node G_i finishes the job J_j , $\sum C_i$ represents the time that the grid node G_i completes the processing of all the jobs. Define $C_{max} = \max\{\sum C_i\}$ as the makespan, and $\sum_{i=1}^m (\sum C_i)$ as the flowtime. An optimal schedule will be the one that optimizes the flowtime and makespan. The conceptually obvious rule to minimize $\sum_{i=1}^m (\sum C_i)$ is to schedule Shortest Job on the Fastest Node (SJFN). The simplest rule to minimize C_{max} is to schedule the Longest Job on the Fastest Node (LJFN). Minimizing $\sum_{i=1}^m (\sum C_i)$ asks the average job finishes quickly, at the expense of the largest job taking a long time, whereas minimizing C_{max} , asks that no job takes too long, at the expense of most jobs taking a long time. Minimization of C_{max} will result in maximization of $\sum_{i=1}^m (\sum C_i)$.

3 Dynamic Grid Job Scheduling Based on PSO

PSO is a population-based optimization algorithm, which could be implemented and applied easily to solve various function optimization problems, or the problems that can be transformed to function optimization problems [4]. In this Section, we design a fuzzy scheme based on discrete particle swarm optimization [6] to solve the job scheduling problem.

Suppose $G = \{G_1, G_2, \dots, G_m\}$, and $J = \{J_1, J_2, \dots, J_n\}$, then the fuzzy scheduling relation from G to J can be expressed as follows:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{bmatrix}$$

Here s_{ij} represents the degree of membership of the i -th element G_i in domain G and the j -th element J_j in domain J with reference to S . The fuzzy relation S between G and J has the following meaning: for each element in the matrix S , the element

$$s_{ij} = \mu_R(G_i, J_j), i \in \{1, 2, \dots, m\}, j \in \{1, 2, \dots, n\}. \tag{1}$$

μ_R is the membership function, the value of s_{ij} means the degree of membership that the grid node G_j would process the job J_i in the feasible schedule solution. In the grid job scheduling problem, the elements of the solution must satisfy the following conditions:

$$s_{ij} \in [0, 1], i \in \{1, 2, \dots, m\}, j \in \{1, 2, \dots, n\}. \tag{2}$$

$$\sum_{i=1}^m s_{ij} = 1, i \in \{1, 2, \dots, m\}, j \in \{1, 2, \dots, n\}. \tag{3}$$

According to fuzzy matrix representation of the job scheduling problem, the position X and velocity V are re-defined as follows:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}; \quad V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{bmatrix}$$

The elements in the matrix X above have the same meaning as (1). Accordingly, the elements of the matrix X must satisfy the constraint conditions given in (2), (3). We get the equations (4) and (5) for updating the positions and velocities of the particles based on the matrix operations.

$$V(t+1) = w \otimes V(t) \oplus (c_1 * r_1) \otimes (X^\#(t) \ominus X(t)) \oplus (c_2 * r_2) \otimes (X^*(t) \ominus X(t)). \tag{4}$$

$$X(t+1) = X(t) \oplus V(t+1). \tag{5}$$

The position matrix may violate the constraints given in (2) and (3) after some iterations, so it is necessary to normalize the position matrix. First we make all the negative elements in the matrix to become zero. If all elements in a column of the matrix are zero, they need be re-evaluated using a series of random numbers within the interval [0,1] and then the matrix undergoes the following transformation without violating the constraints:

$$X_{normal} = \begin{bmatrix} x_{11}/\sum_{i=1}^m x_{i1} & x_{12}/\sum_{i=1}^m x_{i2} & \cdots & x_{1n}/\sum_{i=1}^m x_{in} \\ x_{21}/\sum_{i=1}^m x_{i1} & x_{22}/\sum_{i=1}^m x_{i2} & \cdots & x_{2n}/\sum_{i=1}^m x_{in} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}/\sum_{i=1}^m x_{i1} & x_{m2}/\sum_{i=1}^m x_{i2} & \cdots & x_{mn}/\sum_{i=1}^m x_{in} \end{bmatrix}$$

Since the position matrix indicates the potential scheduling solution, we choose the element which has the max value, then tag it as “1”, and other numbers in the column are set as “0” in the scheduling array. After all the columns have been processed, we get the scheduling solution from the scheduling array and the makespan (solution). The scheme based on fuzzy discrete PSO for the job scheduling problem is summarized as Algorithm 1, in which the job lists and grid node lists are follows:

- $JList_1$ = Job list maintaining the list of all the jobs to be processed.
- $JList_2$ = Job list maintaining only the list of jobs being scheduled.
- $JList_3$ = Job list maintaining only the list of jobs already allocated ($JList_3 = JList_1 - JList_2$).
- $GList_1$ = List of available grid nodes (including time frame).
- $GList_2$ = List of grid nodes already allocated to jobs.
- $GList_3$ = List of free grid nodes ($GList_3 = GList_1 - GList_2$).

4 Experiment Settings, Results and Discussions

In our experiments, Genetic Algorithm (GA) and Simulated Annealing (SA) were used to compare the performance with PSO. Specific parameter settings of all the considered algorithms are described in Table 1. Each experiment (for each algorithm) was repeated 10 times with different random seeds. Each trial had a fixed number of $50 * m * n$ iterations (m is the number of the grid nodes, n is the number of the jobs). The makespan values of the best solutions throughout the optimization run were recorded. And the averages and the standard deviations were calculated from the 10 different trials. In a grid environment, the main emphasis was to generate the schedules as fast as possible. So the completion time for 10 trials were used as one of the criteria to improve their performance.

First we tested a small scale job scheduling problem involving 3 nodes and 13 jobs represented as (3,13). The node speeds of the 3 nodes are 4, 3, 2 CPU, and the job length of 13 jobs are 6,12,16,20,24,28,30,36,40,42,48,52,60 cycles, respectively. Fig. 1(a) shows the performance of the three algorithms. The results (makespan) for 10 GA runs were {47, 46, 47, 47.3333, 46, 47, 47, 47, 47.3333, 49}, with an average value of 47.1167. The results of 10 SA runs were {46.5, 46.5, 46, 46,46, 46.6667, 47, 47.3333, 47, 47} with an average value of 46.6. The results of 10 PSO runs were {46, 46, 46, 46, 46.5, 46.5, 46.5, 46, 46.5, 46.6667}, with an average value of 46.2667. The optimal result is supposed to be 46. While GA provided the best results twice, SA and PSO provided the best result three and five times respectively. Table 2 shows one of the best job scheduling results for (3,13), in which “1” means the job is scheduled to the respective grid node. Further, we tested the three algorithms for other three (G, J) pairs, i.e. (5, 100), (8, 60) and (10, 50). All the jobs and the nodes were submitted at one time. Fig. 1(b) illustrate the performance curves for the three algorithms during the search process for (10, 50). The average makespan values, the standard deviations and the time for 10 trials are illustrated in Table 3. Although the average makespan value of SA was better than that of GA for (3,13), the case was reversed for

Algorithm 1. A scheduling scheme based on fuzzy discrete PSO

- 0 If the grid is active and ($JList_1 = 0$) and no new jobs have been submitted, wait for new jobs to be submitted. Otherwise, update $GList_1$ and $JList_1$.
 - 1 If ($GList_1 = 0$), wait until grid nodes are available. If $JList_1 > 0$, update $JList_2$. If $JList_2 < GList_1$ allocate the jobs on a first-come-first-serve basis and if possible allocate the longest job on the fastest grid node according to the LJFN heuristic. If $JList_1 > GList_1$, job allocation is to be made by following the fuzzy discrete PSO algorithm detailed below. Take jobs and available grid nodes from $JList_2$ and $GList_3$. If $m * n$ (m is the number of the grid nodes, n is the number of the jobs) is larger than the dimension threshold D_T , the jobs and the grid nodes are grouped into the fuzzy discrete PSO algorithm loop, and the single node flowtime is accumulated. The LJFN-SJFN heuristic is applied alternatively after a batch of jobs and nodes are allocated.
 - 2 At $t = 0$, represent the jobs and the nodes using fuzzy matrix.
 - 3 Begin fuzzy discrete PSO Loop
 - 3.0 Initialize the size of the particle swarm n and other parameters.
 - 3.1 Initialize a random position matrix and a random velocity matrix for each particle, and then normalize the matrices.
 - 3.2 While (the end criterion is not met) do
 - 3.2.0 $t = t + 1$;
 - 3.2.1 Defuzzify the position, and calculate the makespan and total flowtime for each particle (the feasible solution);
 - 3.2.2 $X^* = argmin_{i=1}^n (f(X^*(t-1)), f(X_1(t)), f(X_2(t)), \dots, f(X_i(t)), \dots, f(X_n(t)))$;
 - 3.2.3 For each particle, $X_i^\#(t) = argmin_{i=1}^n (f(X_i^\#(t-1)), f(X_i(t)))$
 - 3.2.4 For each particle, update each element in its position matrix and its velocity matrix according to equations (9, 10 and 6);
 - 3.2.5 Normalize the position matrix for each particle;
 - 3.3 End while.
 - 4 End of the fuzzy discrete PSO Loop.
 - 5 Check the feasibility of the generated schedule with respect to grid node availability and user specified requirements. Then allocate the jobs to the grid nodes and update $JList_2$, $JList_3$, $GList_2$ and $GList_3$. Un-allocated jobs (infeasible schedules or grid node non-availability) shall be transferred to $JList_1$ for re-scheduling or dealt with separately.
 - 6 Repeat steps 0-5 as long as the grid is active.
-

bigger problem sizes. PSO usually had better average makespan values than the other two algorithms. The makespan results of SA seemed to depend on the initial solutions extremely. Although the best values in the ten trials for SA were not worse than other algorithms, it had larger standard deviations. For SA, there were some “bad” results in the ten trials, so the averages were the largest. In general, for larger (G, J) pairs, the time was much longer. PSO usually spent the least time to allocate all the jobs on the grid node, GA was the second, and SA had to spent more time to complete the scheduling. It is to be noted that PSO usually spent the shortest time to accomplish the various job scheduling tasks and had the best results among all the considered three algorithms.

Table 1. Parameter settings for the algorithms

Algorithm	Parameter name	Parameter value
GA	Size of the population	20
	Probability of crossover	0.8
	Probability of mutation	0.02
	Scale for mutations	0.1
SA	Number operations before temperature adjustment	20
	Number of cycles	10
	Temperature reduction factor	0.85
	Vector for control step of length adjustment	2
	Initial temperature	50
PSO	Swarm size	20
	Self-recognition coefficient c_1	1.49
	Social coefficient c_2	1.49
	Inertia weight w	$0.9 \rightarrow 0.1$

Table 2. An optimal schedule for (3,13)

Grid Node	Job												
	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9	J_{10}	J_{11}	J_{12}	J_{13}
G_1	0	0	1	0	0	0	1	1	0	1	0	0	1
G_2	1	0	0	1	1	0	0	0	1	0	1	0	0
G_3	0	1	0	0	0	1	0	0	0	0	0	1	0

Table 3. Performance comparison of the three algorithms

Algorithm	Item	Instance			
		(3,13)	(5,100)	(8,60)	(10,50)
GA	Average makespan	47.1167	85.7431	42.9270	38.0428
	Standard Deviation	± 0.7700	± 0.6217	± 0.4150	± 0.6613
	Time	302.9210	2415.9	2263.0	2628.1
SA	Average makespan	46.6000	90.7338	55.4594	41.7889
	Standard Deviation	± 0.4856	± 6.3833	± 2.0605	± 8.0773
	Time	332.5000	6567.8	6094.9	6926.4
PSO	Average makespan	46.2667	84.0544	41.9489	37.6668
	Standard Deviation	± 0.2854	± 0.5030	± 0.6944	± 0.6068
	Time	106.2030	1485.6	1521.0	1585.7

Table 4. Run time performance comparison for large dimension problems

(G,J)	PSO	GA
(60,100)	1721.1	1880.6
(100,1000)	3970.80	5249.80

It is possible that (G, J) is larger than the dimension threshold D_T . We considered two large-dimensions of (G, J) , $(60,500)$ and $(100,1000)$ by submitting the jobs and the nodes in multi-stages consecutively. In each stage, 10 jobs were allocated to 5 nodes, and the single node flowtime was accumulated. The LJFN-SJFN heuristic was applied alternatively after a batch of jobs and nodes were allocated. Fig. 2 and Table 4 illustrates the performance of GA and PSO during the search process for the considered (G, J) pairs. As evident, even though the performance were close enough, PSO generated the schedules much faster than GA as illustrated in Table 4.

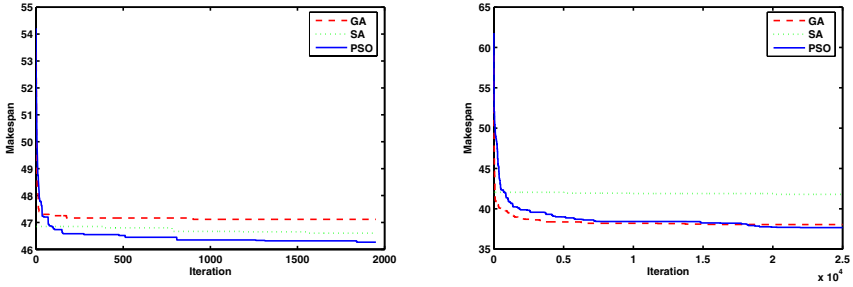


Fig. 1. Performance for job scheduling [a] (3,13) and [b] (5,100)

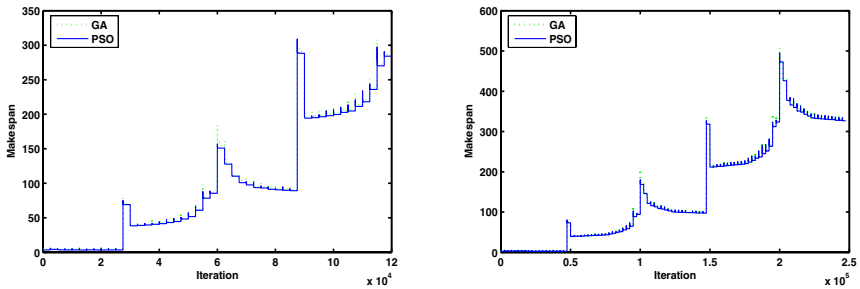


Fig. 2. Performance for job scheduling [a] (60,500) and [b] (100,1000) for GA and PSO

5 Conclusions

In this paper, we evaluated the performance of a fuzzy particle swarm algorithm for grid job scheduling and compared its performance with genetic algorithms and simulated annealing. Empirical results reveal that the proposed approach can be applied for job scheduling. When compared to GA and SA, an important advantage of the PSO algorithm is its speed of convergence and the ability to obtain faster and feasible schedules.

Acknowledgements

This research was supported by the International Joint Research Grant of the Institute of Information Technology Assessment foreign professor invitation program of the Ministry of Information and Communication, Korea.

References

1. Foster,I., Kesselman,C.: *The Grid: Blueprint For A New Computing Infrastructure*. Morgan Kaufmann, USA (2004).
2. Laforenza,D.: *Grid Programming: Some Indications Where We Are Headed* Author. *Parallel Computing*, 28(12) (2002) 1733–1752
3. Gao,Y., Rong,H.Q., Huang,J.Z.: *Adaptive Grid Job Scheduling With Genetic Algorithms*. *Future Generation Computer Systems*, 21 (2005) 151–161.
4. Kennedy,J., Eberhart,R.: *Swarm Intelligence*. Morgan Kaufmann (2001)
5. Abraham,A., Buyya,R., Nath,B.: *Nature's Heuristics For Scheduling Jobs on Computational Grids*. In: *Proceedings of the 8th International Conference on Advanced Computing and Communications*, Tata McGraw-Hill, India, (2000) pp. 45-52.
6. Pang,W., Wang,K., Zhou,C., Dong,L.: *Fuzzy Discrete Particle Swarm Optimization for Solving Traveling Salesman Problem*. In: *Proceedings of the Fourth International Conference on Computer and Information Technology*, IEEE CS Press (2004) 796–800.

Twinned Topographic Maps for Decision Making in the Cockpit

Steve Thatcher¹ and Colin Fyfe²

¹ Aviation Education, Research and Operations Laboratory (AERO Lab),
University of South Australia,
South Australia

² Applied Computational Intelligence Research Unit,
The University of Paisley,
Scotland

Abstract. There is consensus amongst aviation researchers and practitioner that some 70% of all aircraft accidents have human error as a root cause [1]. Thatcher, Fyfe and Jain [2] have suggested an intelligent landing support system, comprising of three agents, that will support the flight crew in the most critical phase of a flight, the approach and landing. The third agent is envisaged to act as a pattern matching agent or an ‘extra pilot’ in the cockpit to aid decision making. This paper will review a new form of self-organizing map which is based on a nonlinear projection of latent points into data space, identical to that performed in the Generative Topographic Mapping (GTM) [3]. But whereas the GTM is an extension of a mixture of experts, our new model is an extension of a product of experts [4]. We show visualisation results on some real and artificial data sets and compare with the GTM. We then introduce a second mapping based on harmonic averages and show that it too creates a topographic mapping of the data.

1 Introduction

There is universal agreement that air travel in modern heavy turbo-jet aircraft has become extremely safe. Improvements in on-board automated systems and the wide spread implementation of flight crew training in team management and group effectiveness has produced a significant improvement in safety [5]. Crew Resource Management (CRM) training is now used by airlines all over the world in an effort to increase the safety of their airline operations. However, an estimated 70% of all accidents are caused by human error [1]. The majority of these can be attributed to the flight crew. A Flight Safety Foundation (FSF) report [6] concluded that from 1979 through 1991 Controlled Flight Into Terrain (CFIT) and approach-and-landing accidents (ALAs) accounted for 80% of the fatalities in commercial transport aircraft accidents. The FSF Approach-and-landing Accident Reduction Task Force Report [7] concluded that the two primary causal factors for such accidents were ‘omission of action/inappropriate action’ and ‘loss of positional awareness in the air’. Thatcher, Fyfe and Jain have suggested using

a trio of intelligent software agents to assist the flight crew with the complex task of approaching and landing at an airport [2].

This paper discusses two techniques for one of the agents, the pattern matching agent, which acts as a pilot assistant or third crew member. This agent will analyse a set of crew behavioural markers and determine whether the flight situation is safe [8]. This technique uses a new form of self-organizing map which is based on a nonlinear projection of latent points into data space, identical to that performed in the Generative Topographic Mapping (GTM) [3]. But whereas the GTM is an extension of a mixture of experts, our new model is an extension of a product of experts [4]. A topographic mapping (or topology preserving mapping) is a transformation which captures some structure in the data so that points which are mapped close to one another share some common feature while points which are mapped far from one another do not share this feature. The most common topographic mappings are Kohonen's self-organizing map (SOM) [9] and varieties of multi-dimensional scaling [10]. The SOM was introduced as a data quantisation method but has found at least as much use as a visualisation tool. It does have the disadvantage that it retains the quantisation element so that while its centres may lie on a manifold, the user must interpolate between the centres to infer the shape of the manifold.

In this paper, we briefly review two new topology preserving mappings; the first of which we call the Topographic Products of Experts (ToPoE), and the second we call the Harmonic Topographic Map (HaToM). Based on a generative model of the experts we show how a topology preserving mapping can be created from a product of experts (in a manner very similar to that used by [3]) to convert a mixture of experts to the Generative Topographic Mapping (GTM). Contrary to the SOM, neither of these mappings quantises but spreads the points across the manifold. We then show how they may be twinned in order to extract information from both data sets simultaneously. One of us [11] previously twinned other self organising maps and used this for prediction.

2 Topographic Products of Experts

Hinton [4] investigated a product of K experts with

$$p(\mathbf{x}_n|\Theta) \propto \prod_{k=1}^K p(\mathbf{x}_n|k) \quad (1)$$

where Θ is the set of current parameters in the model. Hinton notes that using Gaussians alone does not allow us to model some distributions (e.g. multi-modal distributions), however the Gaussian is ideal for our purposes. Thus our base model is

$$p(\mathbf{x}_n|\Theta) \propto \prod_{k=1}^K \left(\frac{\beta}{2\pi} \right)^{\frac{D}{2}} \exp \left(-\frac{\beta}{2} \|\mathbf{m}_k - \mathbf{x}_n\|^2 \right) \quad (2)$$

Fyfe [12] allows latent points to have different responsibilities, r , depending on the data point presented:

$$p(\mathbf{x}_n|\Theta) \propto \prod_{k=1}^K \left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}} \exp\left(-\frac{\beta}{2}\|\mathbf{m}_k - \mathbf{x}_n\|^2 r_{kn}\right) \tag{3}$$

where r_{kn} is the responsibility of the k^{th} expert for the data point, \mathbf{x}_n . Thus all the experts are acting in concert to create the data points but some will take more responsibility than others. Note how crucial the responsibilities are in this model: if an expert has no responsibility for a particular data point, it is in essence saying that the data point could have a high probability as far as it is concerned. We do not allow a situation to develop where no expert accepts responsibility for a data point; if no expert accepts responsibility for a data point, they all are given equal responsibility for that data point (see below).

We now turn our attention to the nature of the K experts which are going to generate the K centres, \mathbf{m}_k . We envisage that the underlying structure of the experts can be represented by K latent points, t_1, t_2, \dots, t_K . To allow local and non-linear modeling, we map those latent points through a set of M basis functions, $f_1(), f_2(), \dots, f_M()$. This gives us a matrix Φ where $\phi_{kj} = f_j(t_k)$. Thus each row of Φ is the response of the basis functions to one latent point, or alternatively we may state that each column of Φ is the response of one of the basis functions to the set of latent points. One of the functions, $f_j()$, acts as a bias term and is set to one for every input. Typically the others are gaussians centered in the latent space. The output of these functions are then mapped by a set of weights, W , into data space. W is $M \times D$, where D is the dimensionality of the data space, and is the sole parameter which we change during training. We will use \mathbf{w}_i to represent the i^{th} column of W and Φ_j to represent the row vector of the mapping of the j^{th} latent point. Thus each basis point is mapped to a point in data space, $\mathbf{m}_j = (\Phi_j W)^T$.

We may update W either in batch mode or with online learning. To change W in online learning, we randomly select a data point, say \mathbf{x}_i . We calculate the current responsibility of the j^{th} latent point for this data point,

$$r_{ij} = \frac{\exp(-\gamma d_{ij}^2)}{\sum_{k=1}^K \exp(-\gamma d_{ik}^2)} \tag{4}$$

where $d_{pq} = \|\mathbf{x}_p - \mathbf{m}_q\|$, the euclidean distance between the p^{th} data point and the projection of the q^{th} latent point (through the basis functions and then multiplied by W). If no centres are close to the data point (the denominator of (4) is zero), we set $r_{ij} = \frac{1}{K}, \forall j$.

We wish to maximise the likelihood of the data set $X = \{\mathbf{x}_n : n = 1, \dots, N\}$ under this model. The ToPoE learning rule (5) is derived from the minimisation of $-\log(p(\mathbf{x}_n|\Theta))$ with respect to a set of parameters which generate the \mathbf{m}_k . Define $m_d^{(k)} = \sum_{m=1}^M w_{md} \phi_{km}$, i.e. $m_d^{(k)}$ is the projection of the k^{th} latent point on the d^{th} dimension in data space.

Similarly let $x_d^{(n)}$ be the d^{th} coordinate of \mathbf{x}_n . These are used in the update rule

$$\Delta_n w_{md} = \sum_{k=1}^K \eta \phi_{km}(x_d^{(n)} - m_d^{(k)}) r_{kn} \tag{5}$$

where we have used Δ_n to signify the change due to the presentation of the n^{th} data point, \mathbf{x}_n , so that we are summing the changes due to each latent point’s response to the data points. Note that, for the basic model, we do not change the Φ matrix during training at all.

3 Harmonic Averages

Harmonic Means or Harmonic Averages are defined for spaces of derivatives. For example, if you travel $\frac{1}{2}$ of a journey at 10 km/hour and the other $\frac{1}{2}$ at 20 km/hour, your total time taken is $\frac{d}{10} + \frac{d}{20}$ and so the average speed is $\frac{2d}{\frac{d}{10} + \frac{d}{20}} = \frac{2}{\frac{1}{10} + \frac{1}{20}}$. In general, the Harmonic Average is defined as

$$HA(\{a_i, i = 1, \dots, K\}) = \frac{K}{\sum_{k=1}^K \frac{1}{a_k}} \tag{6}$$

3.1 Harmonic K-Means

This has recently [13,14] been used to robustify the K -means algorithm. The K -Means algorithm [10] is a well-known clustering algorithm in which N data points are allocated to K means which are positioned in data space. The algorithm is known to be dependent on its initialisation: a poor set of initial positions for the means will cause convergence to a poor final clustering. Zhang and Zhang et al [13,14] have developed an algorithm based on the Harmonic Average which converges to a better solution than the standard algorithm.

The algorithm calculates the Euclidean distance between the i^{th} data point and the k^{th} centre as $d(\mathbf{x}_i, \mathbf{m}_k)$. Then the performance function using Harmonic averages seeks to minimise

$$Perf_{HA} = \sum_{i=1}^N \frac{K}{\sum_{k=1}^K \frac{1}{d(\mathbf{x}_i, \mathbf{m}_k)^2}} \tag{7}$$

Then we wish to move the centres using gradient descent on this performance function

$$\frac{\partial Perf_{HA}}{\partial \mathbf{m}_k} = -K \sum_{i=1}^N \frac{2(\mathbf{x}_i - \mathbf{m}_k)}{d(\mathbf{x}_i, \mathbf{m}_k)^4 \{ \sum_{l=1}^K \frac{1}{d(\mathbf{x}_i, \mathbf{m}_l)^2} \}^2} \tag{8}$$

Setting this equal to 0 and "solving" for the \mathbf{m}_k 's, we get a recursive formula

$$\mathbf{m}_k = \frac{\sum_{i=1}^N \frac{\mathbf{x}_i}{d_{i,k}^4 (\sum_{l=1}^K \frac{1}{d_{i,l}^2})^2}}{\sum_{i=1}^N \frac{1}{d_{i,k}^4 (\sum_{l=1}^K \frac{1}{d_{i,l}^2})^2}} \tag{9}$$

where we have used $d_{i,k}$ for $d(\mathbf{x}_i, \mathbf{m}_k)$ to simplify the notation. There are some practical issues to deal with in the implementation; details of which are given in [13,14].

Zhang et al [13] have extensive simulations showing that this algorithm converges to a better solution (less prone to finding a local minimum because of poor initialisation) than both standard K -means or a mixture of experts trained using the EM algorithm.

3.2 The Harmonic Topographic Map

The above can now be used with the latent variable model. Since

$$\frac{\partial Perf_{HA}}{\partial W} = \frac{\partial Perf_{HA}}{\partial \mathbf{m}_k} \frac{\partial \mathbf{m}_k}{\partial W} = \frac{\partial Perf_{HA}}{\partial \mathbf{m}_k} \Phi_k \tag{10}$$

we could use the algorithm directly in a learning rule as with the ToPoE. However an alternative method is suggested in this paper.

With this learning rule on the same model as above, we get a mapping which has elements of topology preservation but which often exhibits twists, such as are well-known in the SOM [9]. We therefore opt to begin with a small value of K (for one dimensional latent spaces, we tend to use $K=2$, for two dimensional latent spaces and a square grid, we use $K=2*2$) and grow the mapping. As noted earlier, we do not randomise W each time we augment K . The current value of W is approximately correct and so we need only continue training from this current value. Also for this paper we have implemented a pseudo-inverse method for the calculation of W from the positions of the centres, rather than (10).

3.3 Twinned Harmonic Topographic Maps

With the same data as previously, we investigate twinning HaToMs (Figure 1). We note that the mapping is rather good with much less of a pull towards the centre of the data set. Typical mean squared errors on the first data set are

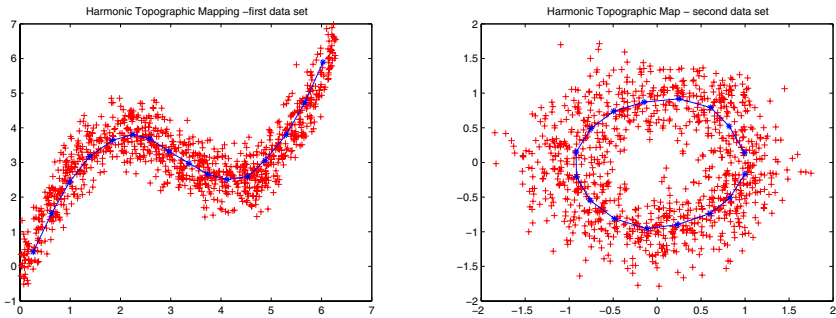


Fig. 1. The two diagrams show the two data sets ('+'s) and the projections of the latent points ('*'s)

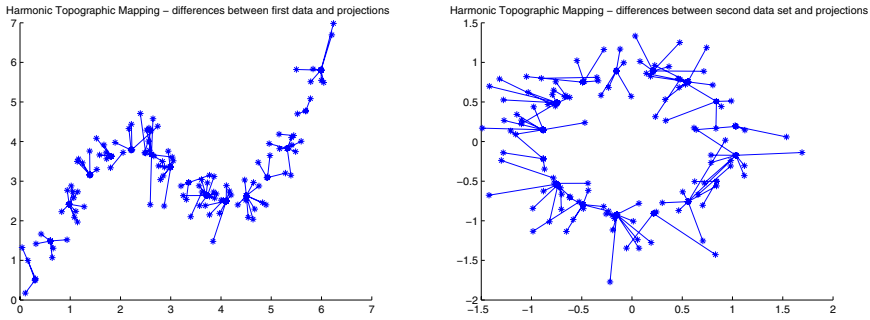


Fig. 2. The two diagrams show 100 samples of each of the two data sets and the HaToM's estimates of the projections of the data points onto the manifold

0.0141 and on the second data set are 0.0132. The differences between the data points and the HaToM's estimate of their positions is shown in Figure 2.

4 Conclusion

We have discussed how flight crew on heavy jet transport aircraft might benefit from an intelligent agent that acts as a pilot assistant in the cockpit. We have described a pattern matching technique using Twinned Harmonic Topographic Maps to pattern match crew behaviour to a safe or unsafe flight outcome. This technique has been trialled on artificial and real data sets. The results show that the mapping is good with much less deviation towards the centre of the data set than other techniques. The data set comprising flight crew behavioural markers is being collected during simulator sessions and will be used in future analysis.

References

1. R. L. Helmreich and H. C. Foushee. Why crew resource management? In E. L. Wiener, B. G. Kanki, and R. L. Helmreich, editors, *Cockpit Resource Management*. San Diego: Academic Press, 1993.
2. S. J. Thatcher, L. C. Jain, and C. Fyfe. An intelligent aircraft landing support system. In *Lecture Notes in Computer Science: The Proceedings of the 8th International Conference on Knowledge-based Intelligent Information and Engineering*. Berlin:Springer-Verlag, 2004.
3. C. M. Bishop, M. Svensen, and C. K. I. Williams. Gtm: The generative topographic mapping. *Neural Computation*, 1997.
4. G. E. Hinton. Training products of experts by minimizing contrastive divergence. Technical Report GCNU TR 2000-004, Gatsby Computational Neuroscience Unit, University College, London, <http://www.gatsby.ucl.ac.uk/>, 2000.
5. H. C. Foushee and R. L. Helmreich. Group interaction and flight crew performance. In E. L. Wiener and D. C. Nagel, editors, *Human Factors In Aviation*. San Diego: Academic Press, 1988.

6. Khatwa and R. L. Helmreich. Analysis of critical factors during approach and landing in accidents and normal flight., data acquisition and analysis working group, flight safety foundation approach-and-landing accident reduction task force. *Flight Safety Digest*, Nov 1998-Feb 1999.
7. Flight Safety Foundation, <http://fsf.com/>. *Controlled Flight into Terrain-Flight Safety Foundation Report*, 2001.
8. S. J. Thatcher, L. C. Jain, and C. Fyfe. An intelligent aircraft landing support paradigm. In R. S. Jensen, editor, *The Proceedings of the 13th International Symposium on Aviation Psychology*. Oklahoma City, OK:Ohio State University, 2005.
9. Tuevo Kohonen. *Self-Organising Maps*. Springer, 1995.
10. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
11. Y. Han, E. Corchado, and C. Fyfe. Forecasting using twinned principal curves and twinned self organising maps. *Neurocomputing*, (57):37–47, 2004.
12. C. Fyfe. Two topographic maps for data visualisation. Technical report, School of Computing, the University of Paisley, 2005.
13. B. Zhang, M. Hsu, and U. Dayal. K-harmonic means - a data clustering algorithm. Technical report, HP Laboratories, Palo Alto, October 1999.
14. B. Zhang. Generalized k-harmonic means – boosting in unsupervised learning. Technical report, HP Laboratories, Palo Alto, October 2000.

Agent-Enabled Decision Support for Information Retrieval in Technical Fields

Gloria Phillips-Wren

Loyola College in Maryland, 4501 N. Charles Street, Baltimore, MD 21210 USA
gwren@loyola.edu

Abstract. Information retrieval (IR) is challenging for a non-expert who operates in a technical area such as medical terminology. Since IR is essentially a decision making process, experience with the design of intelligent decision support systems is relevant. Intelligent agents can assist the user during decision making, and, by extension, in IR to locate the desired information. This paper presents an extension to an IR system for a medical application in which the user lacks the descriptive vocabulary needed to retrieve the resources. Agents continuously seek information on behalf of the user and are autonomous, proactive, communicative, mobile, goal-driven and persistent in order to retrieve information on behalf of the user.

1 Introduction

Information retrieval (IR) is increasingly difficult due to the plethora of information sources, particularly in Web-based environments. In cases in which the information retrieval task requires expertise with the subject matter, the user can benefit from artificial intelligence technologies such as intelligent agents. Agents are ideally suited for IR since they can perform many useful functions such as mediating the environment, brokering, and categorizing information. [1]

1.1 Information Retrieval and Decision Making

Information retrieval is essentially a decision making process. [2, 3, 4] The decision making process consists of the three steps identified by Nobel prize winner Simon [5] together with a fourth step identified by later researchers. [6] A comparison between the processes of decision making and information retrieval is shown in Table 1 using the Information Search Process suggested by Kuhlthau. [7, 4] During the intelligence phase the user gathers information, formulates the problem statement, and determines criteria on which to base decision. The design phase consists of evaluating the criteria and exploring alternatives. The user then makes a selection during the choice phase, and implements the selection. The IR tasks are similar to the decision making tasks. In IR the user explores information and narrows the search by formulating criteria, evaluates the various results from the information search, makes a selection, and finally retrieves the desired information. As in decision making, the process is iterative with feedback loops, although the steps roughly proceed sequentially.

Table 1. Steps in the decision making process compared to IR

Decision-Making Process	Description	Information Search Process [7]
Intelligence	Recognize problem; Gain problem understanding; Seek and acquire information	Task Initiation Topic Selection
Design	Develop criteria; Specify relationships; Explore alternatives	Prefocus Exploration Focus Formulation
Choice	Evaluate alternatives; Develop recommendations; Make decision	Information Collection
Implementation	Weigh consequences; Implement decision	Search Closure

A number of studies have indicated that support can be provided for the decision making process with intelligent agents (see, for example, [8, 9, 10]). Intelligent agents have wide applicability (see, for example, [11]) and have been utilized to support advanced tasks in IR. [12] This paper explores a new implementation of intelligent agents to assist decision making and IR for the non-expert user who requires information in a technical field in which the terminology is not part of his/her lexicon. The intelligent IR system is developed for a suicide prevention website developed for the U.S. National Institute of Mental Health to, in part, provide an interface to the U.S. National Library of Medicine.

2 Information Retrieval for the Non-expert

In the Internet environment in which many users searching for information reside, IR is accomplished with search engines. One of the best known search engines, Google, bases relevancy on a democratic voting algorithm. As stated on their site, "Google works because it relies on the millions of individuals posting websites to determine which other sites offer content of value. ... By analyzing the full structure of the web, Google is able to determine which sites have been 'voted' the best sources of information by those most interested in the information they offer." [13] As evidenced by the success of Google, this technique is highly successful in IR tasks that involve descriptive words or phrases that are commonly known. However, in highly technical fields such an approach is not useful for the non-expert who does not have the necessary vocabulary to describe the needed information.

2.1 Technical Database Search

IR is particularly difficult for the non-expert in a technical field. [2, 3] The user may be unfamiliar with terminology that permits access to the desired information or have difficulty describing the needed data. As a case in point, we utilize the U.S. National Library of Medicine that is accessed through an interface called the Gateway. [14] The Gateway, shown in Figure 1, is a search box that allows the user to enter terms to search a number of databases in the NLM. As shown in Figure 1, the databases

consist of: Professional Journals, Citations, Abstracts; Consumer Health Information such as medical encyclopedias; Books, Audiovisual Materials; Clinical Research Studies; Online Support Organizations; Full-Text Documents; Research on Drug Toxicological Effects.



Fig. 1. U.S. National Library of Medicine Gateway [14]

In the Medline database alone, NLM lists 13 million references from 4800 biomedical journals in both the U.S. and 70 foreign countries. [14] The medical information in the NLM is catalogued using a MeSH® system whose keywords are developed, and whose articles are categorized, by medical subject area specialists. [14] This specialized medical vocabulary is unfamiliar to the non-expert and inaccessible to the average user searching for information using the open search box in Figure 1. Intelligent agents provide an appropriate interface to the databases for the category of users in our research.

2.2 Preventing Suicide Network

The Preventing Suicide Network (PSN) was developed under contract to the NIMH to provide access to a specialized category of resources on suicide. The intended users are intermediaries, that is, people who are concerned about someone who may be suicidal. Users typically lack experience with medical terminology and yet have a compelling need for information that may crucial in preventing a suicide. The homepage is shown in Figure 2, and the text is directed to non-expert users. The suite of search tools and databases included Google-accessed information, information developed specifically for the PSN, and information from the NLM. [15]



Fig. 2. Welcome screen in the Preventing Suicide Network [16]

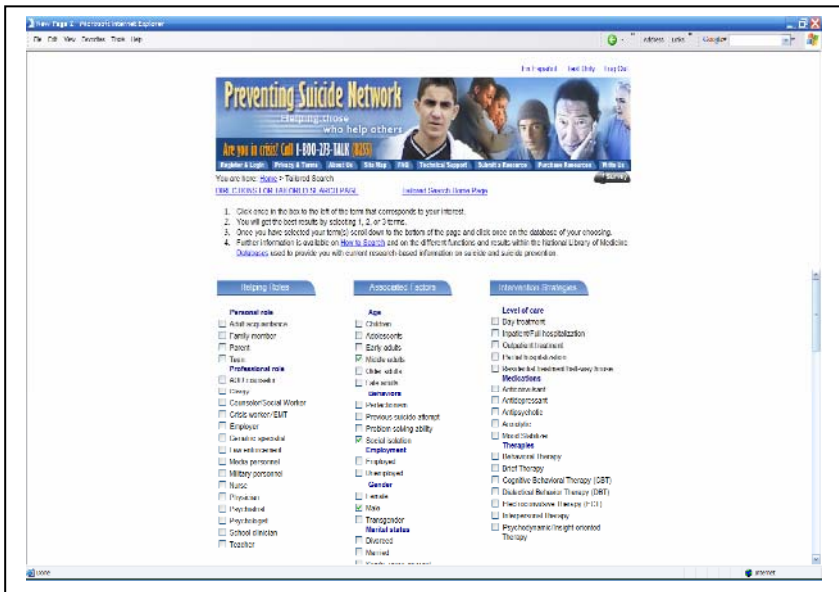


Fig. 3. Portion of the tailored search in the Preventing Suicide Network [16]

Recognizing that the intended users are non-experts in the medical field of suicide, one of the options allows the user to perform a tailored search of the NLM. The tailored search was developed by utilizing the MeSH® system to identify appropriate keywords and then presenting these keywords to the user through more commonly-known descriptors as shown in Figure 3. Previous research discussed the use of intelligent agents to mediate decision making in the search process. [2, 3] In that research it was shown that users need assistance in the design phase to narrow the search and produce a successful outcome, that is, a convergent search strategy enabled with intelligent agents. Intelligent agents can also provide support during the intelligence phase of the decision making process, and that aspect is discussed in this paper.

3 Intelligent Agents During the Intelligence Phase

The PSN allows a user to register and save personal data that describes the type of suicide information of interest. In terms of decision making, these data can be utilized as input to an on-going intelligence phase carried out by intelligent agents.

3.1 Registered Users

Users can register in the PSN by providing a profile and search preferences on a Personal Page in the PSN. Registration provides access to news, a discussion board and search capabilities. The user also selects items of interest from predetermined terms as shown in Figure 3. The terms are descriptive for the non-expert user since the underlying information is technical in nature. By registering these search preferences, the user initiates an intelligent agent to monitor, seek and report new information as it becomes available. Search preferences can be updated at any time.

3.2 Implementation of Agents

Intelligent agents act on behalf of the non-technical user in the IR process in the PSN. Characteristics of intelligent agents [17, 18, 19, 20] are:

- Autonomous: capable of working without human supervision.
- Adaptive: ability to learn and change behavior as their knowledge base increases.
- Proactive: ability to take an initiative on its own.
- Communicative: ability to communicate with other systems, agents and the user.
- Cooperative: as an advanced capability, ability to act in coordination with other agents.
- Mobile: ability to travel throughout computer systems to gain knowledge or perform tasks.
- Goal-Directed: ability to work toward achieving a specific goal.
- Persistent: ability to persist and maintain state over long periods of time.

In the PSN, intelligent agents are autonomous, proactive, communicative, mobile, goal-driven and persistent. Agents communicate with remote databases at the NLM,

specifically the PubMed database, moving between the user and the NLM. The agents undertake their goal proactively, and check periodically to see if any additional information is available from the NLM in the user's declared topics of interest. To do so, agents are goal-driven and translate the user's interests into technical terms in accordance with the MeSH®. In this way the agents in the PSN are unique in IR since they need to infer the user's technical medical interests based on a non-technical description.

From an implementation perspective, information on the current state of the user's IR needs and available information must be stored, the agent must proactively check for information, compare old with current information, determine which information is new to the user, and communicate the results to the user. In practice, the agents send a non-technical email to the user's registered address communicating that new information is available as shown in Figure 4, and reminding them to access the new information by visiting the PSN site. The new results are accessed by selecting the Get New Search Results button on the user's MyPage.

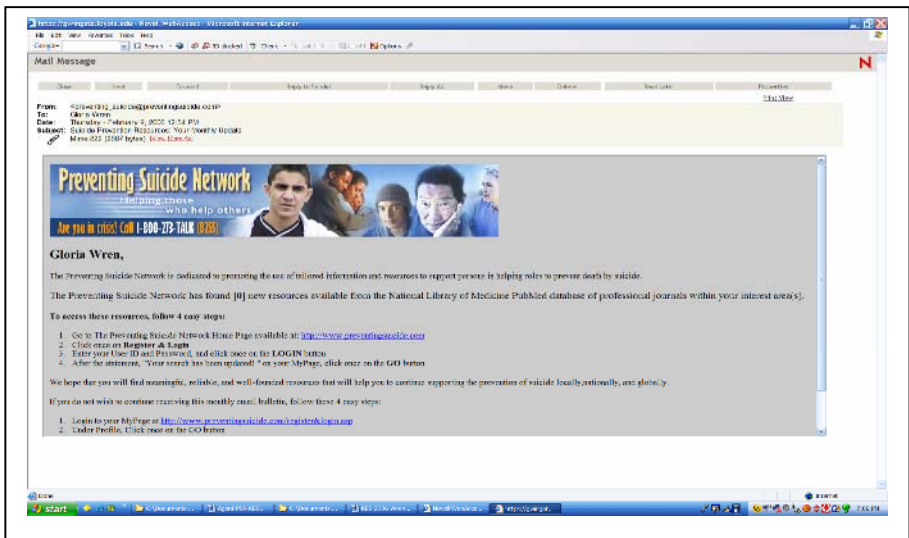


Fig. 4. Email proactively sent by agent to registered user indicating new results for search preferences

4 Summary

This paper presents a novel application of intelligent agents to support the intelligence phase of decision making during information retrieval for the non-expert who is searching in a technical field. Agents persist in autonomously and proactively locating remote information by moving between resources, retrieving information, and communicating with the user. As currently implemented, agents do not have the characteristic of self-learning, and this feature can be added in the future so that

agents learn the interests of the user and suggest potential items of interest to the user accordingly.

Acknowledgements

The authors would like to thank Florence Chang, chief of the Specialized Information Services Branch of the National Library of Medicine, for her invaluable assistance with the NLM databases. This work was supported in part by iTelehealth, Inc., and Consortium Research Management, Inc., under a Small Business Innovation Research contract # N44MH22044 from the National Institutes of Mental Health for an Intermediary-Based Suicide Prevention Website Development Project.

References

1. Sobroff, I.: Agent-based Information Retrieval. Accessed on February 6, 2006, from <http://www.cs.umbc.edu/abir/>
2. Phillips-Wren, G.E., Forgionne, G.: Aided search strategy enabled by decision support. *Information Processing and Management* **42** (2006) 503-518
3. Phillips-Wren, G.E., Forgionne, G.: Intelligent decision making in information retrieval. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.): *Knowledge-Based Intelligent Information and Engineering Systems 8th International Conference Proceedings, Lecture Notes in Artificial Intelligence* 3213 (2004) 103-109
4. Wang, Y.D.: *A Decision Theoretic Approach to the Evaluation of Information Retrieval Systems*, unpublished Ph.D. dissertation, University of Maryland Baltimore County, Baltimore, MD (2006)
5. Simon H.: *Administrative behavior*, fourth edition (Original publication date 1945), The Free Press, New York, NY (1977)
6. Forgionne, G. A.: Decision Technology Systems: A Vehicle to Consolidate Decision Making Support. *Information Processing and Management* **27** (1991) 679-797
7. Kuhlthau, C.: *Seeking Meaning: A Process Approach to Library and Information Services*. Ablex Publishing Corporation, Norwood, NJ (1993)
8. Hess, T., Rees, L., Rakes, T.: Using Autonomous Software Agents to Create the Next Generation of Decision Support Systems. *Decision Sciences* **31** (2000) 1-31
9. Phillips-Wren, G.E., Jain, L.C. (eds.): *Intelligent Decision Support Systems in Agent-Mediated Environments. Frontiers in Artificial Intelligence and Applications* 115, IOS Press, Amsterdam, The Netherlands (2005)
10. Forgionne, G., Mora, M., Gupta, J. (eds.): *Intelligent Decision-making Support Systems: Foundations, Applications and Challenges*. Springer-Verlag, Berlin, Germany (2006)
11. Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.): *Knowledge-Based Intelligent Information and Engineering Systems 8th International Conference Proceedings, Lecture Notes in Artificial Intelligence* 3213-3214 (2004)
12. Rhodes, B., Maes, P.: Just-in-time information retrieval agents. *IBM Systems Journal* **39** (2000) 685-704
13. Google. Accessed on February 2, 2006, from <http://www.google.com/corporate/tenthings.html>
14. NLM: National Library of Medicine Gateway. Accessed on February 1, 2006, from <http://www.nlm.nih.gov/>

15. Wang, Y.D., Phillips-Wren, G.E., Forgionne, G.: E-delivery of personalized healthcare information to intermediaries for suicide prevention. *Int. J. Electronic Healthcare* **1** (2006) 396-412
16. PSN. Accessed on February 2, 2006, from <http://www.preventingsuicide.com>
17. Bradshaw, J. (ed.): *Software Agents*. The MIT Press, Cambridge, MA (1997)
18. Huhns, M., Singh, M. (eds.): *Readings in Agents*. Morgan Kaufmann Publishers, Inc., San Francisco, CA (1998)
19. Design-Ireland. Accessed on February 1, 2006, from <http://www.design-ireland.net/index.php?http%3A//www.design-ireland.net/internet/browsing-13.php>
20. Jennings, N., Woolridge, M. (eds.): *Agent Technology: Foundations, Applications and Markets*. Springer-Verlag, Berlin, Germany (1998)

Is There a Role for Artificial Intelligence in Future Electronic Support Measures?

Phillip Fitch

Knowledge-Based Intelligent Engineering Systems Centre
School of Electrical and Information Eng., University of South Australia
fitps001 @students.unisa.edu.au

Abstract. This paper provides a description of pulse processing and signal identification techniques in Radar Warning Receiver and Electronic Support Measures systems with the objective of describing their similarity to certain Artificial Intelligence techniques. It also presents aspects for which future developments in artificial intelligence based techniques could support the objectives of such systems, both during operation and during the more detailed analysis of data after operations and in counteracting future trends in radar developments. These include parameter optimization, learning and predicting outcomes related to unseen data and so on.

1 Introduction

Electronic Warfare comprises a range of disciplines which can be considered as comprising Communications Interception, Detection and Identification of Non-communications signals and Self Protection against threats. Within the Detection and Identification of Non-communications signals domain there is the category of Electronic Support Measures (ESM) and within the Self Protection domain there is the category of Radar Warning Receivers. ESMs and RWRs have a basic similarity in that they are fundamentally concerned with the reception of electromagnetic signals and the identification of the source.

ESM systems are more interested in what the source of a signal might be, and the provision of tools to enable a detailed analysis of the signals, than RWRs, which are more concerned with answering the question “Could this source represent a threat?”. However both ESMs and RWRs must undertake similar activities in performing their respective functions. These comprise:

- a. Reception of electromagnetic transmissions;
- b. Separation of differing transmissions; and
- c. The identification of possible sources.

Artificial Intelligence (AI) is about the limited mimicking the behaviour of humans. A brief review of the successful applications of AI is contained in [1].

AI is a collection of many techniques including expert systems, neural networks, reasoning techniques, intelligent agents and evolutionary computing [2]. Expert systems can emulate the performance of a human expert by incorporating their acquired knowledge into a computer system, usually via software. Neural nets and

topologies contain elements that behave in a similar manner to the nerve cells in the brain. Evolutionary computing techniques are procedures for solving problems using the principles inspired by natural population genetics. The trend is to fuse these paradigms to combine the benefits of each technique.

To date, the techniques and knowledge bases of AI have not been utilised by Electronic Warfare to the extent which the author believes is possible.

2 Signal Reception

Receivers and digitisers are used to convert the RF energy received by antennas into streams of digital data which comprise a mix of all signals received within a controlled period of time. The data from this process is referred to Pulse Descriptor Words (PDWs) [3]. The typical contents of a PDW include:

- a. Frequency;
- b. Pulse amplitude;
- c. Pulse width;
- d. Time Of Arrival (TOA);
- e. Angle Of Arrival (AOA);
- f. Data Flags to indicate the presence of phase and/or frequency modulation.

3 Transmission Separation

There have been simple pulse sorters for many years but often the environment in which they must work is too intense for their reliable operation. Consequently pulse sorters, (deinterleavers) have become progressively more capable and also more complex. Traditionally the first stage in the process has been to sort pulse based on key known parameters. In low density environments this often resolves many pulse trains, particularly for uncomplicated sources.

Figure 1 shows an example of a Histogram deinterleaver, which is often used for the first stage of deinterleaving. The key aspects of this deinterleaver is that two basic pulse parameters are used, viz. frequency and Angle of Arrival, which enables the grouping of a range of pulses. To reject unintentional noise induced pulses a threshold level is applied as it is not common for the noise pulse count to not exceed the pulse count threshold in the time for which data is collected. The pattern which results from this type of deinterleaver provides both groupings of pulses and, in some cases, an indication of the signal type.

More complex signals exist and these require more computerised processing. The objective of all the various deinterleaving techniques [4] is to:

- a. allocate the maximum number of PDWs to a source;
- b. generate as few candidate sources as is reasonable; and
- c. minimise the variations in the differences between successive TOAs.

Wiley [4] describes a variety of techniques for this process and, with the growth in the complexity of some radar signals and the corresponding growth in computing power, combinations of these techniques are often applied.

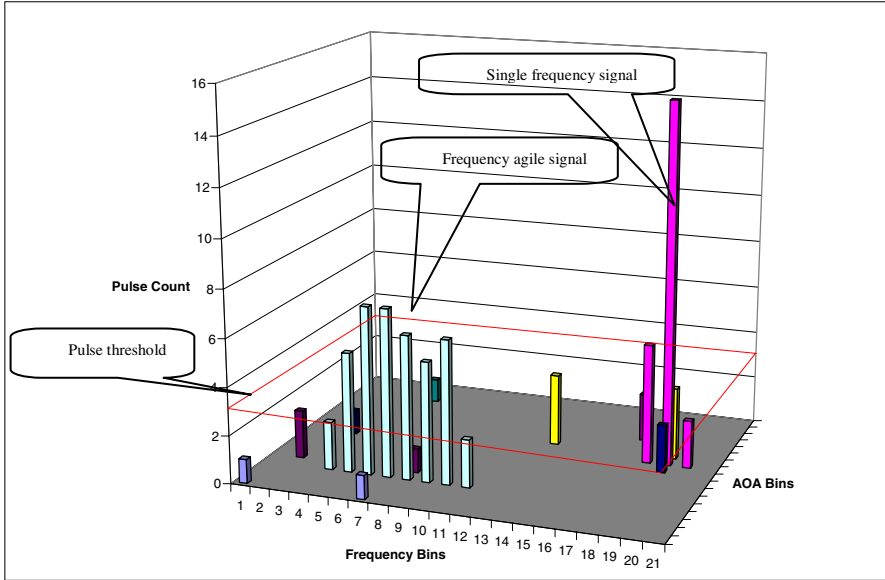


Fig. 1. Histogram Deinterleaver

After these techniques have been applied groups of correlated PDWs exist, and there is a remainder of PDWs which have not been correlated with others.

In dense environments, a data set of uncorrelated pulses after initial deinterleaving, will contain multiple emitter pulse trains. Figure 2 shows such a set, which comprises three pulse trains. One set comprises a constant Pulse Repetition Interval (PRI) train, the second is 2 stage Dwell and Switch PRI train and the third is a 4 pulse Pulse Group PRI.

Templates of known PRI types are applied and reference oscillator frequencies[5] are sought to identify pulses from a common source. Traditionally the techniques for this process has been considered as being purely digital signal processing. Partly this is a consequence of the rule based techniques used and the high speed processing without human intervention. However, the pattern recognition skills of an operator are often required to discriminate between sources in complex pulse vs time patterns.

The role of AI techniques, such as the application of rule-based and fuzzy expert systems, can be argued on the basis that the current techniques do not always provide absolute answers. For example, the TOA differences between successive pulses, the PRI, within a grouping are not constant and will have variations due to transmission effects such as reflections and noise. Hence tolerances are placed on these differences, which can affect the validity of the pulse groupings. Learning systems based on the repeatability of patterns, and incorporating the cognitive knowledge of experts may provide more accurate determinations, albeit in a greater time than the current approach. A combinational approach is worthy of research and evaluation.

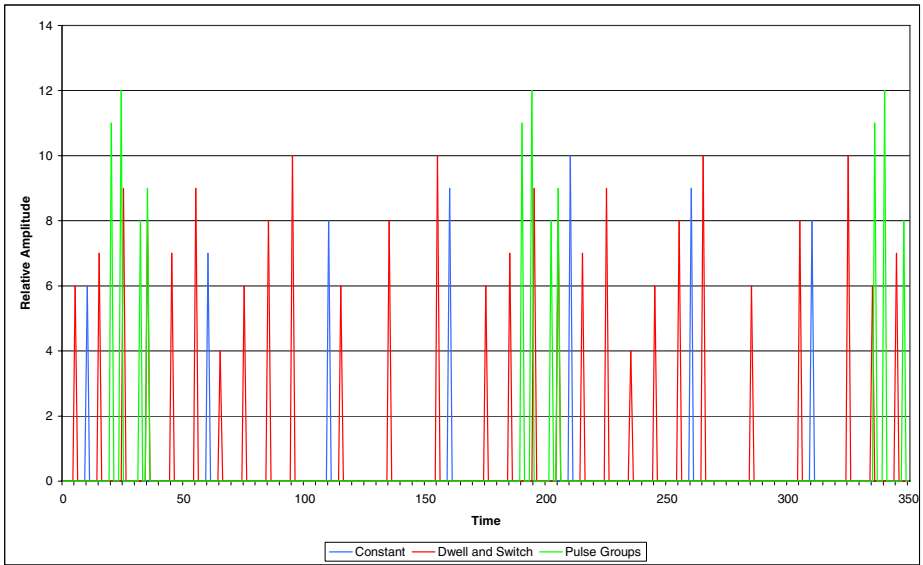


Fig. 2. Post Initial Deinterleaver Pulse Set

4 Source Identification

Traditionally the identification of a source is made from the PDWs groupings. This is often achieved by comparing the frequency and extracted PRI of a signal grouping with a template of known sources. If there is a match, within predetermined tolerances, then a match is declared. In practice there are many sources which have similar characteristics. Table 1 shows some of common sources which are very similar.

If a source were the RDR-1400C and a RDR-1500B were transmitting at the same time then it is probable that they would both be considered as candidate identifications. Likewise for the APS-504 and the ORB-32.

To reduce the list of ambiguous solutions to identification further activities can be implemented. These include the processing of the amplitude of the signal received to determine scan and beamwidth characteristics. This is because as the radiation pattern moves in space a fixed point will see a modulation effect, over time, of the pulse amplitude. By processing this long term variation an estimation of the scan period and the type can be made. Even within this approach there can be ambiguities due to different effects combining to generate a similar effect at the receiver.

For a RWR, further processing is often not a practical option. If a potential threat is identified within the possible candidate identifications, a RWR often declares the existence of the threat. This is a consequence of the need to take timely action. An ESM has an emphasis on knowing what is radiating and hence it may have more time to reduce the identification options. Even so, there remains the probability that multiple candidate identifications will exist.

This approach is akin to a frame-based expert system[9]. It has the same characteristics in that a data structure exists both for the data collected and for the library of potential sources.

The techniques of AI could provide more information about a source and possibly provide a weighting factor on the candidate identifications. Traditionally operators, who have been trained extensively in the subtle nuances of sources, have made decisions as to whether a candidate source is not practical. With the rapid development of neural network based and frame-based expert systems, AI techniques could provide valuable support to decisions made regarding source identification. This is an area in which further research could yield positive results.

Table 1. Source Characteristics

Characteristic	RDR-1400C[5]	RDR-1500B[6]	APS-504[7]	ORB -32[8]
	Search and Rescue and Weather Avoidance	Surface Search	Maritime Patrol	Maritime Patrol
Country of Origin	US	US	Canada	US
Frequency	9.375 GHz	9.375 GHz	8.6 – 9.4 GHz 16 channels	8.5 – 9.6 GHz
Power	10 kW	10 kW	100 kW	108 kW
Pulse Width	0.5, 2.35 μ sec	0.1, 0.5, 2.35 μ sec	0.5, 2.4 μ sec	0.25, 1.5 μ sec
PRF	740, 240 pps	200,800, 1600 pps	1600, 1200, 400, 200 pps	1200, 600, 300 pps)
Elevation Beamwidth	10 degs	10.5 degs	5 degs	7.5 degs
Azimuth Beamwidth	1.5 degs (estimated)	2.6 degs	2.3 degs	2.6 degs
Scan Type	Sector 60 or 120 degs	Sector 120 degs or Circular	Circular and Sector 30 – 120 degs selectable	Circular and Sector 60, 120, 180 or 240 degs
Scan Rate	28 degs/sec	28 degs/sec for Sector, 45 – 90 degs/sec for Circular	30, 12 RPM Sector rate 72 degs/sec	20, 40 RPM Sector rate not provided

5 Post Operation Analysis

Modern systems record the PDWs and other more detailed data when available. After operational use, this data can be further analysed in a less time constrained environment. The data can be reprocessed using a variety of techniques and possibly better identifications can be made. In this more benign environment multiple techniques can be used either in series or in parallel. The use of fuzzy logic expert

systems, neural network based expert systems, especially those utilising heteroassociative networks, and frame-based expert systems seem to have the greatest potential to introduce a new and innovative approach to source discrimination and identification.

A consequence of the post operation analysis is the identification of previously undetected or unknown sources and the expansion, and refinement, of the source library.

6 Future Trends

Modern radar signals [10] are different to those of the past. In the ongoing technological contest between the radar designers and the ESM designers, modern radars are becoming more difficult to detect, much less, identify. These radars are called Low Probability of Intercept (LPI) radars and they are often characterised by:

- a. the transmission of unexpected parameters;
- b. using parameters which are characteristics of different transmitters; and
- c. low power transmission using correlation techniques to enhance radar performance.

Whilst advances in digital signal processing will assist ESM systems in the handling of some aspects of LPI radars, the nature of these signals makes identification, once detection and data collection has occurred, even more difficult. This represents an opportunity for the application of AI techniques.

7 Conclusion

In summary, AI techniques can enhance future ESM systems. It is anticipated that expert systems could assist in significantly enhancing the performance of the next generation of ESM systems. Such enhancement could assist in overcoming the difficulties future radars present for such systems. AI support would assist in the timely determination, during operation of such systems and enable operators to make more informed decisions regarding the electromagnetic environment in which they are working. The separation of differing transmissions will be achieved using fuzzy clustering techniques [11 – 14]. We plan to optimize the parameters related to the transmission separation and source identification using evolutionary computing paradigms [15 – 17]. The reception of electromagnetic signals and identification of their sources play a very important role in early warning receivers. We will exploit the learning abilities of artificial neural networks by using their learning abilities for mimicking the human expertise [18 – 21].

To support the operators in the field, the analysis of data collected can be enhanced by the used of a range of expert systems [14]. This process would enhance the store of knowledge about sources, in particular, the range of tolerances which should apply to the observed characteristics.

Acknowledgements

I wish to express our appreciation to Professor Lakhmi Jain for his valuable contribution in this work and his encouragement to prepare this paper. I would also like to thank Professor Colin Fyfe for seriously questioning my belief that the line between digital signal processing and some types of experts systems significantly overlap in the area of pulse signal processing in ESM systems and thus causing me think more deeply about the opportunity to use expert systems in such tasks.

References

1. Jain, L.C. and Chen, Z., Industry, Artificial Intelligence in, Encyclopedia of Information Systems, Elsevier Science, Volume 2, (2003), pp. 583-597.
2. Jain, L.C. and Jain, R.K. (Editors), Hybrid Intelligent Systems, World Scientific, (1997).
3. Tsui, James B.: Digital Techniques for Wideband Receivers, Scitech Publishing Inc, (2004), pp. 18.
4. Wiley, Richard G.: Electronic Intelligence: The Analysis of Radar Signals, Artech House Inc, Second Edition (1993), pp. 237-249.
5. Data Sheet for RDR-1400c Search and Rescue (SAR) and Weather Avoidance Radar from http://www4.janes.com/subscribe/jrew/doc_view.jsp
6. Data Sheet for RDR-1500B Multimode Surveillance Radar from http://www4.janes.com/subscribe/jrew/doc_view.jsp
7. Friedman, Norman.: The Naval Institute Guide to World Naval Weapons Systems 1997 – 1998, Naval Institute Press, (1997), pp 193.
8. Friedman Norman.: The Naval Institute Guide to World Naval Weapons Systems 1997 – 1998, Naval Institute Press, (1997), pp 196.
9. Negnevitsky, Michael.: Artificial Intelligence, A Guide to Intelligent Systems, Addison-Wesley, Second edition (2005), pp 131.
10. Wiley, Richard G.: The Future of EW and Modern Radar Signals, pp 4, from <http://www.ewh.ieee.org/r5/dallas/aes/IEEE-AESS-Nov04-Wiley.pdf>
11. Jain, L. C. and Martin, N.M. (Editors, Fusion of Neural Networks, Fuzzy Sets, and Genetic Algorithms, CRC Press USA, (1999).
12. Sato, M., Sato, Y. and Jain, L.C., Fuzzy Clustering Models and Applications, Springer-Verlag, Germany, 1997.
13. Sato, M. and Jain, L.C., Fuzzy Clustering Models and Applications, Springer-Verlag, Germany, (2006), in press.
14. Jain, L.C. (Editor), Soft Computing Techniques in Knowledge-Based Intelligent Engineering Systems, Springer-Verlag, Germany, (1997).
15. Jain, L.C. and De Wilde, P. (Editors), Practical Applications of Computational Intelligence Techniques, Kluwer Academic Publishers, USA, (2001).
16. Seiffert, U. and Jain, L.C. (Editors), Self-Organising Neural Networks, Springer-Verlag, Germany, (2002).
17. Jain, L.C. and Kacprzyk, J.(Editors), New Learning Paradigms in Soft Computing, Springer-Verlag, Germany, (2002).
18. Abraham, A., Jain, L.C., and Kacprzyk, J.(Editors), Recent Advances in Intelligent Paradigms and Applications, Springer-Verlag, Germany, (2003).

19. Fulcher, J. and Jain, L.C.(Editors), Applied Intelligent Systems, Springer-Verlag, Germany, (2004).
20. Ghosh, A. and Jain, L.C.(Editors), Evolutionary Computation in Data Mining, Springer-Verlag, Germany, (2005).
21. Russo, M. and Jain, L.C.(Editors), Fuzzy Learning and applications, CRC Press USA, (2001).

Artificial Intelligence for Decision Making

Gloria Phillips-Wren¹ and Lakhmi Jain²

¹ Loyola College in Maryland, 4501 N. Charles Street, Baltimore, MD 21210 USA
gwren@loyola.edu

² University of South Australia, School of Electrical and Information Engineering
KES Centre, Adelaide
Mawson Lakes Campus, South Australia SA 5095
Lakhmi.Jain@unisa.edu.au

Abstract. Artificial Intelligence techniques are increasingly extending and enriching decision support through such means as coordinating data delivery, analyzing data trends, providing forecasts, developing data consistency, quantifying uncertainty, anticipating the user's data needs, providing information to the user in the most appropriate forms, and suggesting courses of action. This session of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems focuses on the use of Artificial Intelligence to enhance decision making.

1 Introduction

A refined class of Artificial Intelligence techniques is revolutionizing the support of decision making, especially under uncertain conditions by such means as coordinating data delivery, analyzing data trends, providing forecasts, developing data consistency, quantifying uncertainty, anticipating the user's data needs, providing information to the user in the most appropriate forms, and suggesting courses of action. This session focuses on Artificial Intelligence techniques and applications that can support decision making. Papers in the session explore advances in methods such as intelligent agents and fuzzy logic that can be utilized in decision making support.

1.1 Artificial Intelligence Paradigms

Artificial Intelligence paradigms are used to mimic the behavior of humans in a limited way. These include tools such as symbolic logic, Artificial Neural Networks (ANNs), fuzzy systems, evolutionary computing, Intelligent Agents and probabilistic reasoning models [1, 2]. In conventional programming methodologies, explicit logic and numerical calculations are provided to solve a problem. In contrast, an ANN mimics some biological systems by solving problems using training and learning to generalize for new problems.

Uncertain and imprecise knowledge can be represented with fuzzy logic [3] and ANNs [4]. They are effective ways of describing complex behavior that is difficult to describe mathematically using conventional methods. Evolutionary computing

techniques [2] evolve a solution to a problem guided by algorithms such as optimization of a multi-dimensional problem. A widely reported type of evolutionary algorithm is a Genetic Algorithm (GA).

Artificial Intelligence paradigms have been used successfully to solve problems in many disciplines including business, management, engineering design, medical diagnosis, decision making and web-based systems [4, 5, 6, 7, 8]. One fruitful area of research appears to be the fusing of these paradigms using hybrid agents [5].

The following section describes briefly a selected number of Artificial Intelligence paradigms used in decision making.

1.2 Artificial Intelligence in Decision Making

The application of Artificial Intelligence to decision making is certainly not new. Recent advances have made Artificial Intelligence techniques accessible to a wider audience as seen by the increase in the number of applications in such areas as intelligent decision support systems. Artificial Intelligence is being used in decision support for tasks such as aiding the decision maker to select actions in real-time and stressful decision problems; reducing information overload, enabling up-to-date information, and providing a dynamic response with intelligent agents; enabling communication required for collaborative decisions; and dealing with uncertainty in decision problems. Leading Artificial Intelligence professional organizations recognize the current effort in “focusing on problems, not on hammers. Given that we (i.e. Artificial Intelligence researchers) do have a comprehensive toolbox, issues of architecture and integration emerge as central” [9]. Several recent examples from the literature are given demonstrating the pragmatic applications of various Artificial Intelligence techniques.

An expert system was implemented to automate the operations of petroleum production and separation facilities [10]. Such systems provide access to plants in remote areas by automatically collecting, transmitting and analyzing data for analysis. The system is able to monitor operations, detect abnormalities, and suggest actions to the human operator based on domain-specific expertise acquired during development of the system. A preliminary evaluation of the system showed satisfactory results.

Case Based reasoning (CBR) is being applied to health services in a variety of areas [11]. Current application of CBR is in bioinformatics, support to the elderly and people with disabilities, formalization of CBR in biomedicine, and feature and case mining. Recent advances are design of CBR systems to account for the complexity of biomedicine, to integrate into clinical settings and to communicate and interact with diverse systems and methods.

Collaborative decision making and knowledge exchange can be enabled with Artificial Intelligence even in difficult clinical health-care decisions by incorporating a social context [12]. In sophisticated neonatal intensive care units, parents, physicians, nurses and other parties must collaborate to decide whether to initiate, limit, continue or discontinue intensive treatment of an infant. The system integrates likely outcomes of the treatment with the physician's interpretation and parents' perspectives. It provides a method of communicating difficult information in a structured form that is still personalized and customized to facilitate decision making.

Fuzzy modeling incorporated into a decision support system has been used to enable forest fire prevention and protection policies in Greece, although the system can be applied on a global basis [13]. Existing approaches use specific geographic boundaries in determining long-term forest fire risk. An inference mechanism based on fuzzy sets has been demonstrated to estimate forest fire risk more successfully. Avineri¹ [6] presents a fuzzy decision support system for the selection of transportation projects. The selection procedure is a multiple objectives process, and projects are rated using linguistic variables on both on a quantitative and qualitative basis. Both fuzzy weighted average and noncompensatory fuzzy decision rules are used to describe a given transportation policy,

An ANN is used by Konar et al.² [6] to develop a scheme for criminal investigation using multi-sensory data including voice, fingerprint, facial image and incidental description. When matching results are poor, the speaker identification scheme RBF-BP Neural Net is invoked. When no conclusion about the suspect could be detected by voice, incidental description is used as the resource for criminal investigation. Kates et al.³ [8] present a decision support system for diagnosing breast cancer using neural networks. The authors took into account the time dependence of underlying risk structures in the formulation of the neural network.

Genetic programming has been used in a decision support system for a tactical air combat environment [7]. The system uses a combination of unsupervised learning for clustering the data and three well-known genetic programming techniques to classify the different decision regions accurately, namely, Linear Genetic Programming (LGP), Multi Expression Programming (MEP) and Gene Expression Programming (GEP). The clustered data is used as the inputs to the genetic programming algorithms.

Intelligent Agents (IA) are perhaps the mostly widely used applied Artificial Intelligence method in recent years. Their utilization has significantly advanced many applications, particularly Web-based systems (see for example [14]). Learning can be incorporated into agent characteristics to extend the capability of systems [15].

2 Session Papers

There are six papers in this session as described below.

2.1 Description of Session Papers

The first paper by Kunchev and Jain [16] is entitled “Path Planning and Obstacle Avoidance for Autonomous Mobile Robots: A Review.” Mobile robotics is of particular interest to the military in potentially life-threatening situations such as reconnaissance. Path planning and obstacle avoidance for autonomous systems can be accomplished using artificial intelligence techniques. The paper presents recent advances and cooperation issues for multiple mobile robots.

¹ Chapter 11.

² Chapter 10.

³ Chapter 2.

Fitch [17] provides a description of pulse processing and signal identification techniques in Radar Warning Receiver and Electronic Support Measures systems with the objective of describing their similarity to Artificial Intelligence techniques. The paper is entitled “Is there a role for Artificial Intelligence in future Electronic Support Measures?” It presents aspects for which future developments in Artificial Intelligence-based techniques could support the objectives of such systems, both during operation and during the more detailed analysis of data after operations and in counteracting future trends in radar developments.

The paper by Phillips-Wren [18] entitled “Agent-Enabled Decision Support for Information Retrieval in Technical Fields” describes an application of intelligent agents for the non-expert who seeks information in a technical field. Information retrieval tasks are compared to the decision making process, and the similarities suggest that intelligent decision support system design is relevant to information retrieval. The paper describes the implementation of intelligent agents to assist a non-technical user to locate the desired information in a technical medical field. Agents are autonomous, proactive, communicative, mobile and goal-driven in order to support the user.

“Twined Topographic Mapping for Decision Making” by Thatcher and Fyfe [19] discusses a new form of self-organising maps which is based on a non-linear projection of latent points into data space. The paper presents a review and comparison with related work in Generative Topographic Mapping. A second mapping based on harmonic averages is demonstrated and compared on real and artificial data sets.

The paper by Abraham and Liu [20] entitled “Job Scheduling on Computational Grids Using Fuzzy Particle Swarm Algorithm” focuses on grid computing, a computing framework to provide increased computational resources. Intelligent functions are needed for such decision tasks as resource management, security, grid service marketing, collaboration, and load sharing. This paper introduces a novel approach based on Particle Swarm Optimization (PSO) for scheduling jobs on computational grids by extending the representation of the position and velocity of the particles in the conventional PSO from real vectors to fuzzy matrices. A dynamic optimal schedule is generated to complete the tasks in a minimum period of time and utilize the resources in an efficient way. The performance of the proposed approach is evaluated with a direct Genetic Algorithm and Simulated Annealing approach.

Donegan and Majaranta [21] discuss eye control methodology in “Optimising the use of Eye Control Technology to meet the needs of people with complex disabilities ? what do users really need?” Eye control methods can be used to provide access to technology for people with complex disabilities. The paper presents issues in eye control, user perspectives, customization and challenges based on research in the COGAIN program. Future research and development needs are discussed.

Acknowledgements

We would like to thank the authors for their excellent contributions. The reviewers are gratefully acknowledged for their time and effort to enhance the quality of the papers.

References

1. Jain, L.C., De Wilde, P. (eds.): Practical Applications of Computational Intelligence Techniques. Kluwer Academic Publishers, Norwell, MA, USA (2001)
2. Jain, L.C., Martin, N.M. (eds.): Fusion of Neural Networks, Fuzzy Sets, and Genetic Algorithms. CRC Press LLC, Boca Raton, FL, USA (1999)
3. Jain, L.C. (ed.): Electronic Technology Directions Towards 2000 **1**, IEEE Computer Society Press, USA (1995)
4. Hammerstrom, D.: Neural networks at work. *Spectrum* **30** (1993) 26-32
5. Jain, L.C., Jain, R.K. (eds.): Hybrid Intelligent Engineering Systems. World Scientific Publishing Company, Singapore (1997)
6. Tonfoni, G., Jain, L.C. (eds.): Innovations in Decision Support Systems. Advanced Knowledge International, Australia (2003)
7. Abraham, A., Grosan, C., Tran, C., Jain, L.C.: A Concurrent Neural Network-Genetic Programming Model for Decision Support Systems. In Proceedings of the Second International Conference on Knowledge Management (2005) 377-384.
8. Jain, A. et al. (eds.): Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis. World Scientific Publishing Company, Singapore (2000)
9. Mackworth, A.: The Coevolution of AI and AAI. *AI Magazine* **26** (2005) 51-52
10. Chan, C.W.: An expert decision support system for monitoring and diagnosis of petroleum production and separation processes. *Expert Systems with Applications* **29** (2005) 131-143
11. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences: What's next? *Artificial Intelligence in Medicine* **36** (2006) 127-135
12. Frize, M., Yang, L., Walker, R.C., O'Connor, A.M.: Conceptual framework of knowledge management for ethical decision-making support in neonatal intensive care. *IEEE Transactions on Information Technology in Biomedicine* **9** (2005) 205-215
13. Iliadis, L.S.: A decision support system applying an integrated fuzzy model for long-term forest fire risk estimation. *Environmental Modelling and Software*. **20** (2005) 613-621
14. Phillips-Wren, G., Jain, L.C. (eds.): Intelligent Decision Support Systems in Agent-Mediated Environments. IOS Press, The Netherlands (2005)
15. Valluri, A., Croson, D.C.: Agent learning in supplier selection models. *Decision Support Systems* **39** (2005) 219-240
16. Kunchev, V., et al.: Path planning and obstacle avoidance for autonomous mobile robots: A review. In Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Verlag-Springer, Berlin (2006) in press
17. Fitch, P.: Is there a role for Artificial Intelligence in future Electronic Support Measures? In Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Verlag-Springer, Berlin (2006) in press
18. Phillips-Wren, G.: Agent-Enabled Decision Support for Information Retrieval in Technical Fields. In Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Verlag-Springer, Berlin (2006) in press
19. Thatcher, S., Fyfe, C.: Twined Topographic Mapping for Decision Making. In Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Verlag-Springer, Berlin (2006) in press

20. Abraham, A., Liu, H., Zhang, W.: Job Scheduling on Computational Grids Using Fuzzy Particle Swarm Algorithm. In Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Verlag-Springer, Berlin (2006) in press
21. Donegan, M., Majaranta, P.: Optimising the use of Eye Control Technology to meet the needs of people with complex disabilities ? what do users really need? In Proceedings of the 10th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Verlag-Springer, Berlin (2006) in press

Path Planning and Obstacle Avoidance for Autonomous Mobile Robots: A Review

Voemir Kunchev¹, Lakhmi Jain¹, Vladimir Ivancevic²,
and Anthony Finn²

¹ School of Electrical and Information Engineering,
Knowledge Based Intelligent Engineering Systems Centre,
University of South Australia, Mawson Lakes SA 5095, Australia
kunvy001@students.unisa.edu.au,
Lakhmi.Jain@unisa.edu.au

² Defence Science and Technology Organisation
{Vladimir.Ivancevic, Anthony.Finn}@dsto.defence.gov.au

Abstract. Recent advances in the area of mobile robotics caused growing attention of the armed forces, where the necessity for unmanned vehicles being able to carry out the “dull and dirty” operations, thus avoid endangering the life of the military personnel. UAV offers a great advantage in supplying reconnaissance data to the military personnel on the ground, thus lessening the life risk of the troops. In this paper we analyze various techniques for path planning and obstacle avoidance and cooperation issues for multiple mobile robots. We also present a generic dynamics and control model for steering a UAV along a collision free path from a start to a goal position.

1 Introduction

In the past few decades there has been a great interest in the problem of motion planning for autonomous mobile robots. To better define motion planning problem we can decompose it into path planning and trajectory planning. Path planning is taking care of the generation of obstacle free path taking into consideration geometric characteristics of obstacles and the kinematic constraints of the robot. Trajectory generation deals with the robot’s dynamics, moving obstacles or obstacles not known priori which are time dependent constraints [1]. The basic mobile robot navigation can be divided into the following tasks:

- Generate a model of the environment in the form of map.
- Compute a collision free path from a start to a goal position.
- Traverse the generated trajectory (with specified velocity and acceleration) and avoid collision with obstacles.

Among the mobile robot research society reactive behavior and planning behavior are often accepted as opposite approaches. The mobile robot must be able to act accordingly when unforeseen obstacles are found on the fly. If the robot rely only on pure path planning the robot is prone to physical collision with an unforeseen obstacle. On the other hand without path planning, with the use of reactive obstacle avoidance method only it will be impossible for the robot to reach its goal location.

Considering the robot environment motion planning can be either static or dynamic. We have static environment when the location of all the obstacles is known priori. Dynamic environment is when we have partial information about obstacles prior the robot motion. The path planning in a dynamic environment is done first. When the robot follows its path and locates new obstacles it updates its local map, and changes the trajectory of the path if necessary.

In his work Fox [2] divides the collision avoidance problem into “global” and “local”. The global techniques that involve path planning methods rely on availability of a topological map defining the robots workspace and obstacle location. He explains that the benefit from using path planning is that the entire path from start to goal can be planned, but this method is not suitable for fast collision avoidance due to its slowness caused by their complexity. On the other hand the local approaches of using pure obstacle avoidance methods suffer from the inability to generate an optimal solution. Another problem is that when using local approach only the robots often get ensnared into a local minimum. Because of these shortcomings, a reactive local approach representing obstacle avoidance cannot be considered feasible for dealing with robot navigation. Due to the reason that there is not a single universal method that can deal with both problems we need to combine both obstacle avoidance and path planning techniques to develop a hybrid system (combining reactive and deliberative approaches) overcoming the weakness of each of the methods.

The hybrid architecture unites the reaction with planning in a heterogeneous system by combining “low-level control” and “high-level reasoning” [3]. The most common hybrid systems are comprised of three layers [4]:

- Reactive layer uses low level sensor based decisions.
- Deliberative layer (planning layer) provides global planning. Its decisions can be based on predefined data (map) or data learned from sensors.
- Executive layer is the intermediate layer between the other two. It process commands from the planning to the reactive layer.

In that sense we can divide robot navigation problem into two sub tasks:

- Obstacle avoidance
- Path planning

2 Review of Obstacle Avoidance Techniques

During the last decades scientists working in AI have contributed to development of planning methods and algorithms for the purpose of navigation of mobile robots. Considerable work has been done in the development of new motion planning and path planning techniques. This section is surveying common obstacle avoidance algorithms.

The purpose of obstacle avoidance algorithms is to avoid collisions with obstacles. Obstacle avoidance algorithms deals with moving the robot based on the feedback information from its sensors. An obstacle avoidance algorithm is modifying the trajectory of the mobile robot in real time so the robot can avoid collisions with obstacles found on its path.

- **Virtual Force Field**

Bornstein's research [5] on real-time obstacle avoidance for a mobile robot is based on Virtual Force Field (VFF) method. This method involves the use of histogram grid for representing the robots work area and dividing it into cells forming the grid. Any of these cells have a "Certainty Value $C(i, j)$ " showing the measure of confidence that an obstacle is located in the cell. During its movement the robot maps the "range readings" into the Certainty Grid. In the same time the VVF method examines a frame area in the Certainty Grid for the occupied cells. Then the occupied cells repel the robot away. The extent of repellent force depends on the concentration of occupied cells in the examined frame, and it is "inversely proportional to the square of the distance between the cell and the robot".

- **Vector Field Histogram**

Even though the VFF method performs quite fast it has its shortcomings. The implemented test-bed shows that often the robot would not move between obstacles to close to each other due the repellent effect from both sides, causing the robots to repel away, a problem also experienced in the Potential Field method [6]. To solve the problems with VFF Borenstein and Koren [7] developed the Vector Field Histogram VFH technique. The method employs the use of 2D histogram grid to represent the environment, being reduced to single dimension "polar histogram" which is build around the position of the robot in a certain moment. The sectors presented in the polar histogram show the "polar obstacle density". The direction of the robot is computed by choosing the sector with least concentration of obstacles. The map represented by the histogram grid is renewed from the robot's sensors with data containing the distance between the robot and obstacles.

- **VFH+**

The VFH+ method [8] is similar to the VFH but introduces some novelties by employing "threshold hysteresis" to improve the shape of the trajectory, and the use of a cost function. The cost function is used to choose the best direction in between all candidate directions (which are free of obstacles) provided by the polar histogram. The selected direction is the one with the lowest cost. The new VFH+ method considers the vehicle width by enlarging the cells containing obstacles, which makes it easy to experiment with various vehicle dimensions. The trajectory of the vehicle is also considered with by "masking sectors that are blocked by obstacles in other sectors".

- **Dynamic Windows Approach**

The Dynamic Window Approach (DWA) [9] is another method for reactive obstacle avoidance dealing with the kinematical and dynamic constraints of the vehicle in contrast to VFF and VFH methods. The method might be described by a "search for commands" computing the velocities of the vehicle which are then passed to the velocity space. The robot's trajectory consists of a "sequence of circular arcs". The arcs are defined by a velocity vector (v_i, ω) , in which v_i denotes the translational velocity and ω stands for the rotational velocity, together they represent the search space. The

search space is being reduced to form a dynamic window, which takes into account the trajectory formed by the circular arcs and defined by the velocity vector.

$$V_r = V_s \cap V_a \cap V_d \quad (1)$$

The region V_r located in the dynamic window is intersected by the space of possible velocities represented by V_s , the area V_a in which the vehicle is able to stop and avoid collision, and the Dynamic window denoted by V_d .

• Nearness Diagram

The problem of obstacle avoidance in highly dense and troublesome environment is presented in [10]. The Nearness Diagram (ND) method uses “divide and conquer” approach splitting the environment into sectors to represent the location of obstacles. Experiments show that ND method can successfully avoid local minima trap only if it is completely visible to the sensors. The ND method utilizes the behavioral based “situated activity” paradigm. The concept uses predefined groups of condition states consisted of different problems and their corresponding actions. When algorithm is performed the current state based on information from sensors is defined and its corresponding action is executed as described in [11].

• Curvature Velocity Method

The Curvature Velocity Method (CVM) [12] takes into account the dynamic constraints of the vehicle allowing it to move fast in a dense environment. The velocity and acceleration constraints of the robot, and the presence of obstacles presented as circular objects are added to a velocity space. The velocity space consists of translational and rotational velocity. The presumption is that the robot’s trajectory is based along arcs of circles $c = \omega/v$. The velocity is selected on the base of objective function that corresponds to the part of the velocity space that realizes the physical constraints of the robot and the obstacles.

• Elastic Band Concept

The Elastic Band Concept [13] works by deforming the original obstacles free path supplied by a path planner. The reason for that is that often the path planner computes a path that has sharp turns, which makes it impossible for the robot to steer. The path modified using the Elastic Band concept is shorter and smoother than the original path. This method can adapt to dynamic changes in the environment modifying the path if new obstacles are detected, avoiding the need for a new path preplanning. There are two forces that modify the form of the new path. A force that mimics the stretching of an elastic band eliminating the “slack” called “contraction force”, and an opposite force called “repulsion force” providing more room by repelling the robot away from obstacles.

The modern obstacle avoidance and algorithms reviewed in this section represent the synthesis of vector-field based techniques and agent-based AI techniques. In our future work we plan to further develop this approach by including some rigorous Lie groups and Lie algebras based methods (refer Section 4).

3 Review of Path Planning Algorithms

- **A* heuristic search**

A* is one of the most common path finding algorithm [14]. For its map representation A* utilizes a grid based search area divided into squares. Each square can be either a free space or an obstacle. In order to find the shortest path a collision free trajectory is calculated comprised of free space squares (also called nodes). To find the shortest path to the goal the A* algorithm uses heuristic approach. A* first adds its starting node A to OPEN set of free space nodes comprising a possible path. The next step is to look for free space nodes around node A and add them to its list and set node A as their parent node. The next step is to add node A to a CLOSED set and delete it from the OPEN set. The next node to be processed is determined by its minimum cost F towards the goal. The lowest cost $F=G+H$, where G is the cost for getting to the next node, and H is the estimated distance to the goal point. A* provides efficient and complete path finding, but one of its major weakness when dealing with large environments is the vast memory usage caused by the use grid representation of the map.

- **Visibility Graph**

The visibility graph method [15] consists of straight line segments joining at the obstacles visible vertices (but not crossing the obstacle) to define a roadmap from a start to a goal position. In this method the shape of an obstacle is represented as a polygon. The task of the visibility graph method is to connect the start and goal positions with all the vertices of the polygons that are visible. Then connect every single vertex of a polygon with the visible vertex of another polygon. In the created visibility graph any straight line can be a part of the path. The shortest possible path is then calculated using simple graph search technique. This method is prone to let the mobile robot collide with the edge of an obstacle due to the very close distance with its path. A solution of this problem is to artificially increase the size of the polygons before the path is planned so the robot can pass it from a safe distance. Due to increase the number of vertices the visibility graph method performs rapidly in areas when the number of obstacles (polygons) is low.

- **Generalized Voronoi Diagram**

The Voronoi diagram [16] consists of arcs (lines) which are equidistant from the two nearest obstacles. The obstacles in the Voronoi diagram are presented as polygons. The maximized clearance between the Voronoi arc segments and the polygons helps the robot maintain safe distance away from the obstacles. Another advantage of the this method is that it provides a complete path solutions (if possible path exists) based on the fact that if there is a gap between two obstacles there will be a Voronoi line in between. After the completion of the graph, two straight lines are added to the diagram. One line connects the graph with the start position, and a second line is used to connect the goal position [17]. Then graph search technique is used to calculate the roadmap based on the generated Voronoi diagram where the arcs are equivalent to the edges of a graph.

The path finding algorithms reviewed in this section represent the synthesis of computational geometry and AI techniques. In our future work we plan to further

develop this approach by including some rigorous optimal control methods, like PMP (Pontryagin Maximum Principle).

4 Dynamics and Control of Steering a Generic UAV

Mechanically, every UAV represents a free-flying rigid body governed by the SE(3)-group, that is, the 6 DOF Euclidean group of 3D motions. Recall that the SE(3)-group couples 3D rotations with 3D translations; technically, it is defined as a non-commutative product of the rotational group SO(3) and the translational group \mathbf{R}^3 . The corresponding nonlinear dynamics problem (that had been resolved mainly for aircraft, spacecraft and submarine dynamics) is called “dynamics on SE(3) group,” while the associated nonlinear control problem (resolved mainly for general helicopter control) is called “control on SE(3)-group” [18].

We can represent a point in SE(3) by the 4x4 matrix g defined as:

$$g = \begin{bmatrix} R & x \\ 0 & 1 \end{bmatrix} \tag{2}$$

where $R \in \text{SO}(3)$ and $x \in \mathbf{R}^3$. The UAV-dynamics are now defined by:

$$\dot{g} = g \begin{bmatrix} 0 & -u_3 & u_2 & 1 \\ u_3 & 0 & -u_1 & 0 \\ -u_2 & u_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3}$$

where u_1, u_2 and u_3 – being 3 scalar inputs (UAV thrusts). The standard quadratic cost function for a particular trajectory is given by:

$$\frac{1}{2} \int_0^T \sum_{i=1}^3 c_i u_i^2(t) dt \tag{4}$$

where c_i are the cost weights and T is desired final time. The optimal inputs are:

$$u_i = \frac{P_i}{c_i} \quad (i=1,2,3) \tag{5}$$

where P_i 's are solutions of the Euler-like equations:

$$\begin{bmatrix} \dot{P}_1 \\ \dot{P}_2 \\ \dot{P}_3 \end{bmatrix} = \begin{bmatrix} \frac{c_2 - c_3}{c_2 c_3} P_2 P_3 \\ \frac{c_3 - c_1}{c_1 c_3} P_1 P_3 \\ \frac{c_1 - c_2}{c_1 c_2} P_1 P_2 \end{bmatrix} \tag{6}$$

If the cost weights are equal, then the system is (analytically) integrable and represents the so-called ‘‘Lagrange’s top.’’

In case of a fixed altitude, the above dynamics and control problem reduces to the so-called ‘‘Hilare robot car,’’ governed by the SE(2)-group of motions (which is a 2D subgroup of the SE(3) group). Given that this UAV always drives forward at a fixed speed, we need to find the steering controls so that the robot, starting from an initial position and orientation, arrives at some final goal position and orientation at a fixed time T . In this case the UAV-dynamics are simplified to:

$$\dot{g} = g \begin{bmatrix} 0 & -u & 1 \\ u & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{7}$$

with a single input u corresponding to the UAV’s turning velocity. The quadratic cost function is simplified to: $\frac{1}{2} \int_0^T cu^2(t)dt$. In this case, we have the constant of motion

given by $l^2 = P_1^2 + P_2^2$, and optimal trajectory given by:

$$P_1(t) = l \cos \theta(t), \tag{8}$$

$$P_2(t) = l \sin \theta(t), \tag{9}$$

where $\theta(t)$ denotes the orientation of the UAV at any given point in time.

The model presented in this section can be further developed by including both modern Lie-derivative based control techniques and Hamiltonian optimal control.

5 Conclusion and Future Work

In this paper we have presented various algorithms and techniques for efficient obstacle avoidance and path planning for mobile robots. We have also presented a generic dynamics and control model for steering a UAV along a collision free path from a start to a goal position. In the case of fixed altitude, this model reduces to the ‘‘Hilare robot car.’’ The major area of our future research involves motion planning for multiple mobile robots based on subsumption architecture. At the top level is a path planner which will work in tandem with a flocking behavior. The flocking action will be based on steering behaviors for a flock of mobile robots based on Boids [19] which will be overridden by reactive obstacle avoidance behavior. Our dynamical and control model will be further developed by including both modern Lie-derivative based control techniques and Hamiltonian optimal control.

References

1. Clean, S., L. ‘‘Path planning & high level control of an unmanned aerial vehicle’’, University of Sydney (2002)
2. Fox, D., W. Burgard and Thrun, S. ‘‘The Dynamic Window Approach to Collision Avoidance’’, IEEE Robotics and Automation Magazine, March (1997)

3. Coste-Manière, Ève; Simmons, Reid: "Architecture, the Backbone of Robotic Systems". In Proceedings of the 2000 IEEE International Conference on Robotics & Automation, San Francisco, CA, (April 2000)
4. Russell, S. & Norvig, P. *Artificial Intelligence: a Modern Approach*. Prentice-Hall, 1995.
5. Borenstein, J. and Koren, Y. "Real-time Obstacle Avoidance for Fast Mobile Robots." *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5) (Sept/Oct 1989)
6. Khatib, O. "Real-Time Obstacle Avoidance for Manipulators and Mobile Robots" *The International Journal of Robotics Research*, 5(1), (1986)
7. Borenstein, J. and Koren, Y. "The Vector Field Histogram- Fast obstacle avoidance for mobile robots." *IEEE Journal of Robotics and Automation* 7(3), (June 1991)
8. Ulrich, I., and Borenstein, J. "VFH+: Reliable Obstacle Avoidance for Fast Mobile Robots" *IEEE International Conference on Robotics and Automation*, p1572, Leuven, Belgium, (1998)
9. Brock, O., Khatib, O. "High-speed navigation using the global dynamic window approach." In *Proc. ICRA*, pages 341-346, (1999)
10. Minguez, J., Montano, L. "Nearness Diagram Navigation (ND): Collision Avoidance in Troublesome Scenarios". *IEEE Transactions on Robotics and Automation*, (2004)
11. Arkin, R. "Behavior-Based Robotics". Cambridge, MA: MIT Press, (1999)
12. Simmons R., "The Curvature Velocity Method for Local Obstacle Avoidance," *IEEE Int. Conf. on Robotics and Automation*, Minneapolis, USA, (1996)
13. Quinlan S., Khatib O. "Elastic Bands: Connecting Path Planning and Control," *IEEE Int. Conf. on Robotics and Automation*, Atlanta, USA, (1993)
14. Wilson, N. J. "Principles of Artificial Intelligence" Springer Verlag, Berlin (1982)
15. Nilsson, N. J. "A Mobile Automaton: An Application of Artificial Intelligence Techniques" *Proc. 1st Int. Joint Conf. on Artificial Intelligence*, Washington D.C., 509-520 (1969)
16. Latombe, J. C. "Robot motion planning" Kluwer Academic Publishers, (1991)
17. Eldershaw, C. "Transfer Report: Motion planning" (1998), Unpublished.
18. Ivancevic, V. and Ivancevic, T. "Natural Biodynamics," World Scientific Singapore, Series: Mathematical Biology, (2006)
19. Reynolds, C. "Steering Behaviors for Autonomous Characters," *Proceedings of Game Developers Conference*, (1999)

Adaptive Nonlinearity Compensation of Heterodyne Laser Interferometer

Minsuk Hong, Jaewook Jeon, Kiheon Park, and Kwanho You

Sungkyunkwan University, Suwon, 440-746, Korea
{uriadl, jwjeon, khpark, khyou}@ece.skku.ac.kr

Abstract. With its outstanding ultra-precise resolution, the heterodyne laser interferometer systems are commonly used in semiconductor manufacturing industry. However, the periodical nonlinearity error caused from frequency-mixed cross talks limits the accuracy in nanometer scale. In this paper to improve the accuracy of laser interferometer system, we propose an adaptive nonlinearity compensation algorithm using RLS (recursive least square) method. As a reference signal, the capacitance displacement sensor mounted on a linear piezo-electric transducer gives a feedback information on how to transform the elliptical phase into a circular one.

1 Introduction

The heterodyne laser interferometer is a frequency-detecting method to measure length in various industries. The application includes motion control in robotics, photo lithography in semiconductor manufacture, and some velocity sensors. Especially it covers from several meter to sub-nanometer (nm) scale [1-2]. For ultra-precise length measurement using the laser interferometer, however, there are some barriers to overcome. The first restriction resides in the precise alignment of a laser interferometer and stabilization of a laser source. The next thing to be considered is a measurement error from external noises. The external noise is related to environment such as vibration, temperature change, and air turbulence. The error caused by environment can be overcome through high quality sensors.

As one of the main factors in measurement errors, the nonlinearity is occurred from the imperfect optical equipment. In the use of heterodyne laser interferometer, the accuracy of nanometer-scale measurement is quite often restricted by nonlinearity. The nonlinearity errors happen mainly from two factors: polarization-mixing and frequency-mixing [3]. The polarization-mixing happens within an imperfect polarization beam-splitter (PBS). This is a primary factor of nonlinearity errors in the homodyne and one frequency interferometer. Meanwhile in case of laser interferometer with two different frequencies, the frequency-mixing problem, which arises from elliptical polarization, non-orthogonality between two frequencies, and imperfectly aligned PBS, etc., induces the nonlinearity errors.

To overcome the nonlinearity, many efforts have been done [4-7]. Under the assumption that the nonlinearity error is distributed as a white Gaussian noise,

Park [6] tried to increase the accuracy by using a Kalman filter. Using a phase-quadrature mixing technique, the phase of two output signals from demodulators are adjusted electrically [7]. This method is also too complicated and expensive to apply for real system.

In this paper, we compensate the nonlinearity error of a laser interferometer using a capacitance displacement sensor. The proposed algorithm is remarkable in that the compensation parameters are not fixed. Meanwhile, it modifies the compensation parameters in an adaptive way. Therefore the compensation process adjusts the input data to the optimized value.

The heterodyne laser interferometer is shown in Fig. 1. In Fig. 1, there are two orthogonally polarized beams with different frequencies (ω_1, ω_2). Passing through a PBS, one of the waves becomes a reference signal ($A\omega_1$) and the other is a measurement signal ($B\omega_2$). Two signals are reflected from fixed mirror and moving mirror separately. After being recombined through a PBS and $\lambda/4$ wave plate again, it is observed by a photo detector.

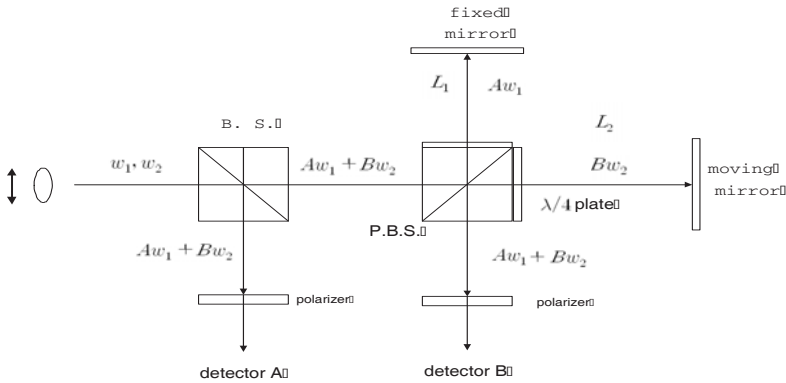


Fig. 1. Configuration of heterodyne laser interferometer

For the heterodyne laser interferometer without frequency-mixing problem as in Fig. 1, the intensity of each electric field can be expressed as follows from two detectors, A and B,

$$\begin{aligned}
 E_{A1} &= \frac{1}{\sqrt{2}} A e^{j(\omega_1 t + \Phi_A)} \\
 E_{A2} &= \frac{1}{\sqrt{2}} B e^{j(\omega_2 t + \Phi_B)} \\
 E_{B1} &= \frac{1}{\sqrt{2}} A e^{j(\omega_1 t + \Phi_A)} \\
 E_{B2} &= \frac{1}{\sqrt{2}} B e^{j(\omega_2 t + \Phi_B + \Delta\Phi)}
 \end{aligned} \tag{1}$$

where A and B are the amplitudes, Φ_A and Φ_B are the initial phase values, and $\Delta\Phi$ is a phase difference between fixed path and moving path.

The intensity of an input signal to the photo detectors A and B is

$$\begin{aligned}
 I_r &\propto (E_{A1} + E_{A2})(E_{A1} + E_{A2})^* \\
 &= \frac{1}{2}(A^2 + B^2) + AB \cos[\Delta\omega t + (\Phi_B - \Phi_A)] \\
 I_m &\propto (E_{B1} + E_{B2})(E_{B1} + E_{B2})^* \\
 &= \frac{1}{2}(A^2 + B^2) + AB \cos[\Delta\omega t + (\Phi_B - \Phi_A) - \Delta\Phi]
 \end{aligned} \tag{2}$$

where $\Delta\omega$ means the frequency difference of $\omega_2 - \omega_1$ and $\Delta\Phi$ stands for phase difference between fixed path and moving path as,

$$\Delta\Phi = \frac{4\pi n \Delta L}{\lambda} \tag{3}$$

Here λ is a mean wavelength between ω_1 and ω_2 , n stands for the refractive index and ΔL is a difference between fixed path and moving path ($\Delta L = L_2 - L_1$). The length of a movement (ΔL) can be expressed in terms of phase difference ($\Delta\Phi$) as in (3).

2 Nonlinearity in Heterodyne Laser Interferometer

Fig. 2 shows the configuration of heterodyne laser interferometer under the influence of frequency-mixed cross talk. As shown in Fig. 2, some portions of two orthogonal frequency beams are mixed up as a cross talk in terms of $\alpha\omega_1$ and $\beta\omega_2$. Under the effect of frequency-mixing as depicted in Fig. 2, the intensity of electric field from photo detector B is as follows.

$$\begin{aligned}
 E_{B1} &= Ae^{j(\omega_1 t + \Phi_A)} + \beta e^{j(\omega_2 t + \Phi_\beta)} \\
 E_{B2} &= Be^{j(\omega_2' t + \Phi_B)} + \alpha e^{j(\omega_1' t + \Phi_\alpha)}
 \end{aligned} \tag{4}$$

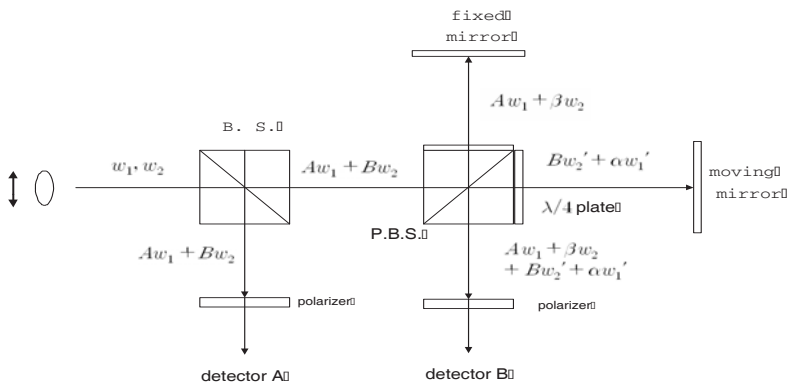


Fig. 2. Heterodyne laser interferometer with frequency-mixed cross talk

Here ω'_1 and ω'_2 are the frequency by Doppler effect of ω_1 and ω_2 respectively. The intensity I_m of the measured signal at photo detector B is

$$\begin{aligned}
 I_m &\propto (E_{B1} + E_{B2})(E_{B1} + E_{B2})^* \\
 &= \frac{1}{2}(A^2 + B^2 + \alpha^2 + \beta^2) + AB \cos[(\Delta\omega + \psi)t + (\Phi_B - \Phi_A)] \\
 &\quad + A\beta \cos[\Delta\omega t + (\Phi_\beta - \Phi_A)] + B\alpha \cos[\Delta\omega t + (\Phi_B - \Phi_\alpha)] \\
 &\quad + A\alpha \cos[\psi t + (\Phi_\alpha - \Phi_A)] + B\beta \cos[\psi t + (\Phi_B - \Phi_\beta)] \\
 &\quad + \alpha\beta \cos[(\Delta\omega - \psi)t + (\Phi_\beta - \Phi_\alpha)]
 \end{aligned} \tag{5}$$

Here ψ means the frequency difference of $\omega'_2 - \omega_2 = \omega'_1 - \omega_1$. If we suppose that the moving mirror does not change rapidly, the DC component and near-DC component(ψt) of I_m can be separated by using high pass filter. The AC component of I_m can be expressed as a sum of three harmonic functions.

$$I_{m,AC} \propto \cos(\Delta\omega t + \Phi) + \Gamma_1 \cos(\Delta\omega t) + \Gamma_2 \cos(\Delta\omega t - \Phi) \tag{6}$$

with $\Gamma_1 = (A\beta + B\alpha)/(AB)$, $\Gamma_2 = \alpha\beta/AB$, and $\Phi = \psi t$. The first term in the Eq. (6) is a base signal, the second term $\Gamma_1 \cos(\Delta\omega t)$ and third term $\Gamma_2 \cos(\Delta\omega t - \Phi)$ are the undesired nonlinearity components. These nonlinearity factors restrict the accuracy in nanometer scale length measurement. The nonlinearity terms are called as first-order and second-order phase error respectively.

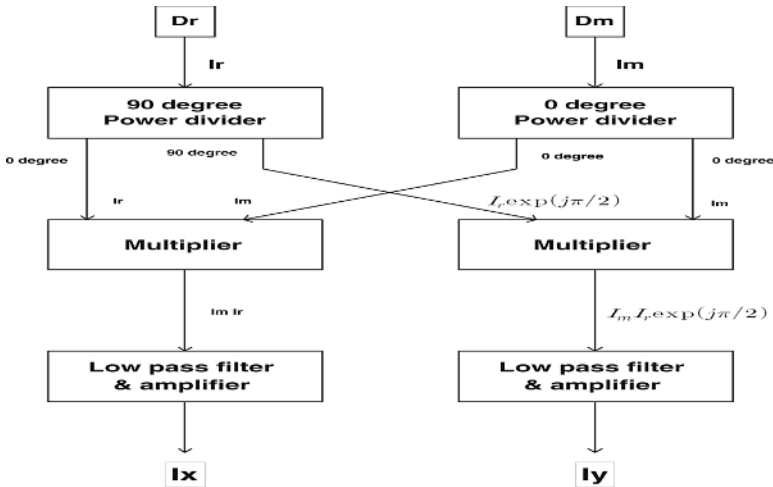


Fig. 3. Block diagram of a lock-in-amplifier

Fig. 3 shows a block diagram of a lock-in-amplifier. As depicted in Fig. 3, I_r of Eq. (2) and I_m of Eq. (6) are the input signals to the lock-in-amplifier. In this process, 90 degree delayed signal I_r is multiplied with signal I_m to extract the information of phase which is related to the length measurement. After passing

through a low pass filter, we can extract the intensities of I_x and I_y . As a result, the information of phase can be acquired as follows.

$$\begin{aligned} I_x &= \frac{AB + \alpha\beta}{2} \cos \Phi + \frac{A\beta + \alpha B}{2} = a \cos \Phi + h \\ I_y &= \frac{AB - \alpha\beta}{2} \sin \Phi = b \sin \Phi \end{aligned} \quad (7)$$

where α and β represent the extent of nonlinearity. As can be confirmed from (7), if there is no nonlinearity error (i.e., $\alpha = \beta = 0$), the phase can be expressed as $\Phi = \tan^{-1}(I_x/I_y)$ and the trajectory of intensities (I_x, I_y) forms a circle with a radius $AB/2$.

3 Adaptive Compensation for Nonlinearity

In this section we show how to compensate the laser interferometer using capacitance displacement sensor(CDS) in adaptive way. It is well known that the CDS maintains a high accuracy in nanometer scale. To get the phase signal from CDS as a reference signal, we change the length to phase using the relation of (3). Now we represent the phase Φ measured from CDS in intensity domain of I_x and I_y .

$$\begin{aligned} I_x &= \frac{AB}{2} \cos \Phi \\ I_y &= \frac{AB}{2} \sin \Phi \end{aligned} \quad (8)$$

Likewise we estimate the intensity of laser interferometer as \tilde{I}_x and \tilde{I}_y ,

$$\begin{aligned} \tilde{I}_x &= \frac{AB}{2} \cos \tilde{\Phi} \\ \tilde{I}_y &= \frac{AB}{2} \sin \tilde{\Phi} \end{aligned} \quad (9)$$

where $\tilde{\Phi}$ is the phase measured by laser interferometer. To compensate the nonlinearity appeared on \tilde{I}_x and \tilde{I}_y , we express the intensity as a function of phase Φ from CDS.

$$\begin{aligned} \tilde{I}_x &= \frac{AB + \alpha\beta}{2} \cos \Phi + \frac{A\beta + \alpha B}{2} = a \cos \Phi + h \\ \tilde{I}_y &= \frac{AB - \alpha\beta}{2} \sin \Phi = b \sin \Phi \end{aligned} \quad (10)$$

Here α and β represent the extent of nonlinearity. The optimal condition on \tilde{I}_x and \tilde{I}_y is that it should be on the circle with a radius of $AB/2$. However, as can be inferred from (10), the shape of circle is transformed into an ellipsoid from the effect of undesirable nonlinearity errors.

To find the optimal parameters of a , b , and h of \tilde{I}_x and \tilde{I}_y , we use the recursive least square (RLS) method [8]. The constraints on \tilde{I}_x and \tilde{I}_y can be formulated in a linear matrix form.

$$\begin{aligned}
 Mx &= H \\
 M &= \begin{bmatrix} \cos \Phi & 1 & 0 \\ 0 & 0 & \sin \Phi \end{bmatrix}, \quad H = \begin{bmatrix} \tilde{I}_x \\ \tilde{I}_y \end{bmatrix}
 \end{aligned} \tag{11}$$

where x means the compensation parameters defined as $x^T \equiv [a, b, h]$. In this paper, we focused on the adaptive compensation. This means that the compensation parameters are not fixed from the initial measurement. Instead the parameters are updated to follow the reference target value. To do so, the RLS method is suitable as an optimal parameter search method.

Let's represent P_k as $P_k^{-1} \equiv M^T M$. Applying the compensation parameters x to RLS method, we can find the optimal values in an iterative way as following,

$$\begin{aligned}
 P_{k+1} &= P_k - P_k M_{k+1}^T (I + M_{k+1} P_k M_{k+1}^T)^{-1} M_{k+1} P_k \\
 x^{(k+1)} &= x^{(k)} + P_{k+1} M_{k+1}^T (H^{(k+1)} - M_{k+1} x^{(k)})
 \end{aligned} \tag{12}$$

The new compensation parameters a , b , and h are modified through the Eq. (12). To find the optimal values of a^* , b^* , and h^* , we inserted an additional constraint as a boundary condition.

$$\sqrt{(a^+ - a^-)^2 + (b^+ - b^-)^2 + (h^+ - h^-)^2} < \varepsilon \tag{13}$$

where the superscript $+$, $-$ mean the prior and posterior values respectively. For example, if the boundary value of ε is too large, RLS algorithm will stop within several iterations. We choose an enough small ε value as to guarantee the optimized compensation parameters.

With the optimal values of a^* , b^* , and h^* , the intensities of laser interferometer can be transformed to an optimal one as,

$$\begin{aligned}
 I_x^* &= \frac{\tilde{I}_x - h}{2a} AB \\
 I_y^* &= \frac{\tilde{I}_y}{2b} AB
 \end{aligned} \tag{14}$$

These new values are on the circle with a radius of $AB/2$. Using I_x^* and I_y^* , the compensated phase Φ^* and length measurement can be derived as

$$\begin{aligned}
 \Phi^* &= \tan^{-1} \left(\frac{I_y^*}{I_x^*} \right) \\
 L^* &= \frac{\Phi^* \lambda}{4\pi n}
 \end{aligned} \tag{15}$$

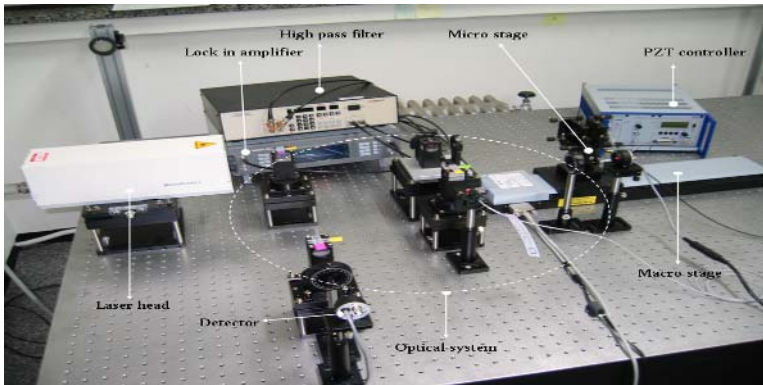


Fig. 4. Configuration of laser interferometer and micro stage

4 Experimental Results

In this section, we prove the effectiveness of our proposed compensation algorithm through some experimental results. The model of heterodyne laser interferometer in this experiment is Agilent:WT307B. The mean wavelength λ is given as $0.6329912 \mu\text{m}$. As a moving stage which can translate in nanometer-scale, we used the linear piezo-electric transducer (PI:P-621.1CL). Using an input command controller, the stage is moved to some fixed positions ($y=50 \text{ nm}$, 150 nm). To prove the efficiency of our proposed compensation algorithm, the capacitance displacement sensor (PI:D-100) is used as a reference. Fig. 4 shows the configuration of laser interferometer and micro stage used to implement displacement measurements.

Fig. 5 shows the result of heterodyne laser interferometer after applying the proposed adaptive compensation algorithm. In this figure, the thin solid line is

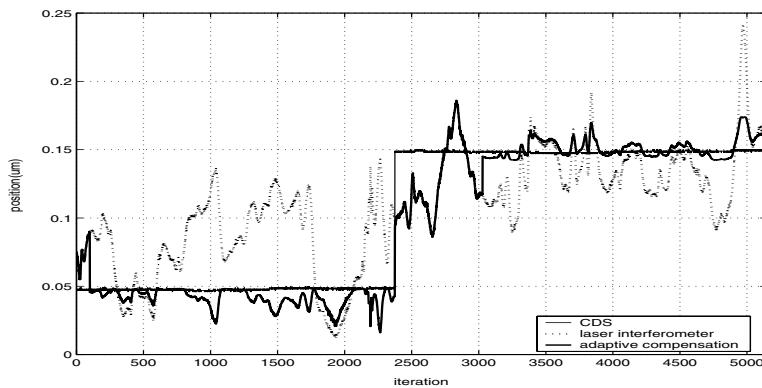


Fig. 5. Comparison of fixed length measurement ($y = 50 \text{ nm}$, 150 nm)

the length measurement with capacitance displacement sensor (CDS), the dotted line is that of laser interferometer, and the thick solid line is that of laser interferometer with adaptive compensation. As shown in Fig. 5, there is less chattering on each fixed position ($y=50$ nm, 150 nm) after adaptive compensation. We can confirm that the laser interferometer is now more robust to nonlinearity.

5 Conclusion

As an ultra-precise length measurement, the heterodyne interferometer is much affected by nonlinearity error from imperfect optical elements. To reduce the nonlinearity error, we proposed an adaptive compensation algorithm using recursive least square (RLS) method. The compensation parameters are modified adaptively according to the input data. Finally the compensation parameters approach to optimal values by using the data from capacitance displacement sensor as a reference signal.

Acknowledgement

The authors would like to thank the Korea Science and Engineering Foundation (KOSEF) for financially supporting this research under Contract No. R01-2004-000-10338-0 (2005).

References

1. Yeh, H.C., Ni, W.T., Pan, S.S.: Digital closed-loop nanopositioning using rectilinear flexure stage and laser interferometer. *Control Engineering Practice* 13 (2005) 559-566
2. Lawall, J., Kessler, E.: Michelson interferometry with 10 pm accuracy. *Review of Scientific Instruments* 71 (2000) 2669-2676
3. Freitas, J.M., Palyer, M.A.: Polarization effects in heterodyne interferometer. *Journal of Modern Optics* 42 (1995) 1875-1899
4. Wu, C.M., Su, C.S.: Nonlinearity in measurements of length by optical interferometer. *Measurement Science & Technology* 7 (1996) 62-68
5. Badami, V.G., Patterson, S.R.: A frequency domain method for the measurement of nonlinearity in heterodyne interferometer. *Precision Engineering* 24 (2004) 41-49
6. Park, T.J., Choi, H.S., Han, C.S., Lee, Y.W.: Real-time precision displacement measurement interferometer using the robust discrete time Kalman filter. *Optics & Laser Technology* 37 (2005) 229-234
7. Eom, T.B., Choi, T.Y., Lee, K.H.: A simple method for the compensation of the nonlinearity in the heterodyne interferometer. *Measurement Science & Technology* 13 (2002) 222-225
8. Chong, E.K., Zak, S.H.: *Introduction to optimization*. Wiley-interscience (2001)

Study on Safety Operation Support System by Using the Risk Management Information

Yukiyasu Shimada¹, Takashi Hamaguchi², Kazuhiro Takeda³, Teiji Kitajima⁴,
Atsushi Aoyama⁵, and Tetsuo Fuchino⁶

¹ Japan National Institute of Occupational Safety and Health, Chemical Safety Research Gr.,
1-4-6, Umezono, Kiyose, Tokyo 204-0024, Japan
shimada@s.jniosh.go.jp

² Nagoya Institute of Technology, Dept. of Engineering Physics, Electronics and Mechanics,
Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555, Japan
hamachan@nitech.ac.jp

³ Shizuoka University, Dept. of Materials Science & Chemical Engineering,
3-5-1, Johoku, Hamamatsu, Shizuoka 432-8561, Japan
tktaked@ipc.shizuoka.ac.jp

⁴ Tokyo University of Agriculture and Technology, Dept. of Chemical Engineering,
2-24-16, Naka-cho, Koganei, Tokyo, 184-8588, Japan
teiji@cc.tuat.ac.jp

⁵ Ritsumeikan University, Graduate School of Technology Management,
1-1-1, Nojihigashi, Kusatsu Shiga 525-8577, Japan
aoyama@mot.ritsumei.ac.jp

⁶ Tokyo Institute of Technology, Dept. of Chemical Engineering,
2-12-1, Oookayama, Meguro, Tokyo 152-8552, Japan
fuchino@chemeng.titech.ac.jp

Abstract. In case of abnormal situation of chemical process plant, it is required to judge the plant condition correctly and carry out the unerring response. Many operation support systems were proposed to help the plant operator with such judgment and operation decision making, but most of them could not indicate the rationale for operation, because they used only information on the result of plant design. Therefore, plant operator could not get the theoretical information for operation decision making from it. This paper proposes the safety operation support system which helps the operation decision making based on the risk management information (RMI). The RMI is useful one for the theoretical operation decision making, because it is considered and updated through the plant life cycle (PLC) and includes the information on the design rationale (DR) of the safety countermeasure, the intents of corresponding operation, etc. PVC batch process is used for verifying the effectiveness of proposed method.

1 Introduction

Safety of chemical process plant can be maintained by risk management through the plant life cycle (PLC). In the plant design activity, safety design is carried out through the risk assessment for the supported plant abnormal situation and the planning of countermeasure for it. In the plant operation and the maintenance activity, plant condition is monitored with response to the plant abnormal situation continually and

the equipments or the facilities are renewed depending on the plant condition. Such risk management activities require the safety management technique for various plant abnormal situations based on the risk concept and the information on each activity should be managed in an integrated fashion. On the other hand, recently, we had a large number of accidents in the chemical process plant. As a background problem of process safety management (PSM), it is pointed out that there is no technical environment to share the information on PSM techniques through the PLC. For example, plant operators and maintenance persons sometimes carry out each activity without having a clear understanding of the design intent of safety countermeasure or the required operation policy which are considered at plant design phase. This will entail the problem which plant operators carry out the unexpected operation and the expansion of damage is leaded.

Correspondingly, it is required to develop the computer-aided support system to enable plant operator to recognize the plant abnormal situation, infer the cause and the effect of process malfunction and decide the appropriated operation. Conventional operation support systems for plant abnormal situation selected the corresponding operation based on the process flow diagram (PFD) and the information on result of safety assessment, but they could not output the theoretical intent of the operation to the plant operator.

This paper proposes a safety operation support system to help the plant operator with the operation decision making in case of plant abnormal situation. This support system can provide plant operator with the clarified design intent of process plant and the purpose of operation procedure based on the risk management information (RMI) which is considered and updated through the PLC. PVC batch process is used to examine the practicability of operation decision making procedure.

2 Risk Management Information

Chemical process plants are designed to satisfy the operational conditions which are required for three operational phases; normal, abnormal and emergency operations [1, 2]. At the plant abnormal and emergency operational designs, safety problems of the plants are clarified through the risk assessment and best solutions to response to various abnormal situations are selected from some safety countermeasures by considering the requirements from the viewpoints of safety(S), cost(c), quality(Q), delivery(D) and environment(E). At this time, standard models which are defined by ISO, IEC, ISA, etc are also adopted as the design basis. Beside that, the stable and safety plant operation is improved gradually through the PLC based on above risk concepts. Consequently, it is important to maintain the records of each PSM activity as the RMI of the objective plant.

Table 1 shows the RMI which is considered at each activity phase; design, operation and maintenance. The RMI includes the information on the plant structure, the process behavior and the operation procedure, the result of risk assessment including the supported abnormal scenarios, the final results of design of safety countermeasures, etc. as the design information. Additionally, the operational information related to the log of operation, the record of plant maintenance of the objective plant, etc. is also included.

Table 1. Risk management information (RMI) and each example

Activity phase	RMI and it's examples
(1) Design	Plant information;
1) Normal operation	- Plant structure (Equipment, Plant layout, etc)
2) Abnormal and Emergency operation	- Process behavior (Mass balance, Heat balance, Material property, Reaction property, Risk data, etc)
a) Risk analysis; Abnormal scenario, Risk assessment	- Operation procedure (for Normal, Abnormal, Emergency) Risk analysis;
b) Safety design; Design of safety countermeasure to reduce the risk level	- Supported abnormal scenario - Result of risk assessment (HAZOP sheet, etc) Safety design; - Design of safety countermeasure and its rationale (DR)
(2) Operation	Logs of past operation (Near-miss information, etc)
(3) Maintenance	Records of plant maintenance (Facility update, etc)

Just as important is that the RMI includes not only the information on the result of each activity but also the logical information on PSM policy as background knowledge of it. For example, plant engineers design the safety countermeasures based on IPL (Independent Protection Layer) concepts by theoretical or empirical judgments, or in accordance with societal requirement and regulation. At this time, considered context, including their thinking process is design basis, i.e. why and how the safety facility or the operation are selected from some alternatives. Such reasoning embedded in the design can be explicitly expressed as design rationale (DR) [3, 4] and the recorded DR is an important information source for other engineering activities; safety operation, effective maintenance, management of change, etc [5].

Table 2 shows the example of specific DR for designing PVC process [6, 7]. For example, the reactor with exothermic reaction needs cooling system to control reactor

Table 2. Example of specific design rationale (DR) for PVC process design

Designed equipments or facilities	Specific design rationale (Viewpoints)
Reactor temperature sensor	To monitor reactor temperature (S, Q, D)
Reactor pressure sensor	To monitor reactor pressure (S, Q, D)
High reactor temperature alarm	To inform high reactor temperature (S, D)
High reactor pressure alarm	To inform high reactor pressure (S, D)
Redundancy pump	To maintain reactor temperature in case of pump failure (S) To enhance cooling ability (Q, D)
Back-up source of cooling	To maintain reactor temperature in case of some sort of troubles of cooling system (S, D)
Add shortstop chemical	To quench a runaway reaction (S)
Depressuring system	To vent reactor to a gas disposal system (S) To prevent explosion (S)
Manual activation of bottom discharge valve	To drop batch into a dump tank or to an emergency containment area (S, Q, D)

temperature and also requires the redundancy cooling system (pump) and the backup source of cooling in case of some sort of troubles of this system. These logics are also generic DR for designing this type of reactor.

On the other hand, on-site plant operators have their own operational policies and they effect their operation management in plant abnormal situation. These are also the theoretical RMI of the objective plant and can be used effectively for understanding the PSM policy of the objective plant theoretically.

3 Plant Safety Management Support System

3.1 Conventional Operation Support System

Operation support system is utilized to provide operator with proper instruction to judge the plant condition correctly and decide the most appropriate operation. Conventional operation support systems have used the plant information shown in the PFD and the P&ID (Piping and Instrumentation Diagram) and the information on the result of risk assessment, and indicated the only outline of the operation to plant operator. But they could not give any information on the technical reason why this operation is the best solution needed for current plant condition. Furthermore, it has been pointed out that operation support system makes no sense if it cannot identify the cause of the process malfunction, because they have tried to select the solution based on the only cause of process malfunction. To solve this problem, we have proposed a basic procedure of operation decision making to protect the fault propagation by using the plant design information effectively, even if the immediate cause cannot be identified [8]. Fig.1 shows the basic procedure of it. In this method, the operation can be selected based on the result of cause and effect analysis for the detected process malfunction. However, in this framework, the knowledgebase (KB) for the system included the only design result, but the information on the technical background and the engineer's logical idea could not be mentioned explicitly.

3.2 Safety Operation Support System by Using the Risk Management Information

This paper proposes the method of using the RMI as KB for supporting the operation decision making by plant operator in plant abnormal situation. Fig.2 shows a framework of proposed safety operation support system. More meaningful information for operation decision making can be given to the plant operator according to the basic procedure shown in Fig.1. Because the RMI includes the information on DR as described at chapter 2, the plant operators can understand the design intents of equipments and safety facilities and the reason why using them can be effective for the current plant abnormal situation. As a result of it, they can decide the appropriate operation with the risk-based and reasonable reason.

Step 1) Detection of process malfunction

Process malfunction is detected. Many fault detection methods have been proposed.

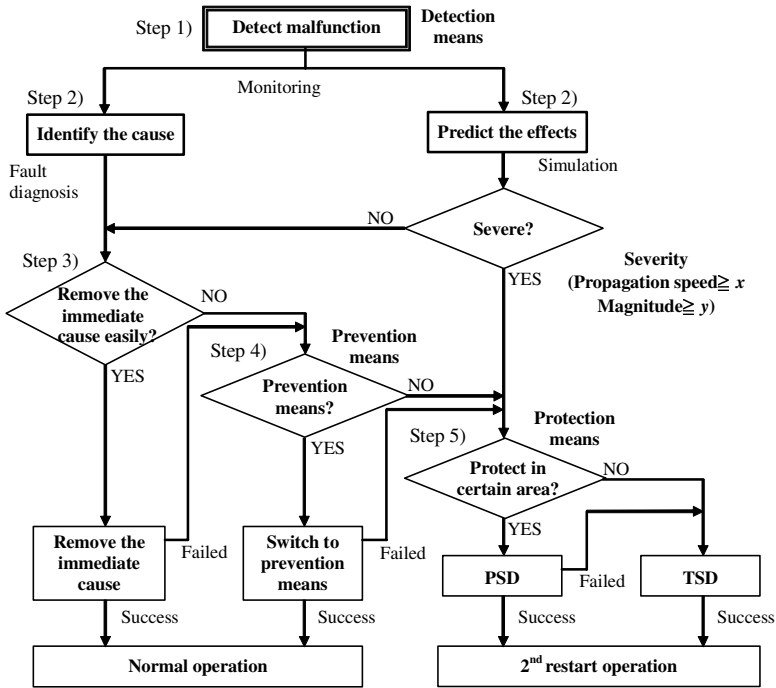


Fig. 1. Basic procedure of operation decision making

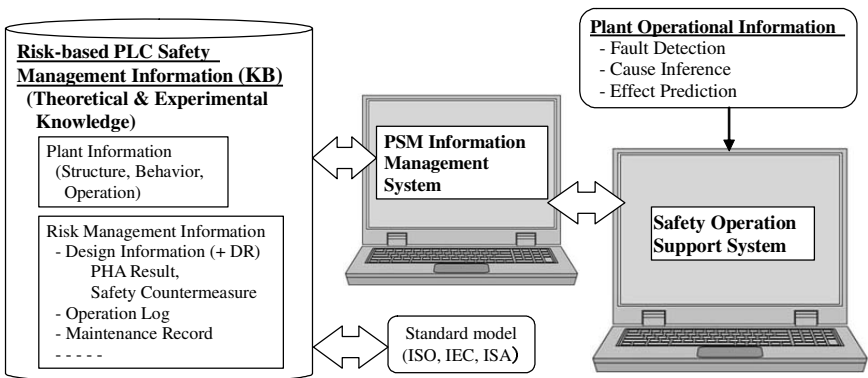


Fig. 2. Safety operation support system by using the risk management information

Step 2) Cause inference and effect prediction

For the detected process malfunction, the possible cause and the effects of it are identified by model-based or knowledge-based reasoning, and by using simulation techniques [9]. These methods in Steps 1) and 2) are out of scope in this paper.

Step 3) Removal of immediate cause

If the effects are not severe and the immediate cause can be removed easily, the support system decides the removal operation to recover the plant back to normal condition. Recovery operation for the causes is searched from the RMI in the KB.

Step 4) Switch to prevention means

When the immediate cause cannot be identified, or the response to it fails, switch to prevention means is considered. If the prevention means can be available, then the support system decides to switch to them as a recovery operation and outputs it with the reason why that operation is best solution for current plant condition.

Step 5) Plant shutdown operation

In case the recovery operation fails, or its effects on plant condition are severe, the shutdown operation by using the protection means is considered to protect the fault propagation. The severity is judged based on whether either the propagation speed, the magnitude of the effects, or both will exceed the preset value, or not. First the support system tries to select the partial shutdown (PSD) operation. If the speed of fault propagation is high and there is a possibility that the process malfunction may expand in the whole plant and leads to the accident, the support system selects the total shutdown (TSD) operation such as addition of shortstop chemical. After PSD or TSD operation, the plant will be recovered and restarted according to the documented manual etc.

In this proposed method, the support system can suggest the operation according to the PSM policy of the objective plant to plant operators. Then they can decide the appropriate operation based on the theoretical RMI.

4 Example for PVC Batch Process

PVC reactor unit is used to examine the effectiveness of proposed support system. This process is the polymerization of vinyl chloride monomer (VCM) to make polyvinyl chloride (PVC) and includes some process instrumentations, safety facilities and equipments as shown in Fig.3 [6]. A redundancy pump (P-R_CW) and a back-up source of cooling are installed based on the DR as shown in Table 2.

Steps 1) and 2)

In the PVC process, it is assumed that high reactor temperature, high reactor pressure and no CW flow are detected, its immediate cause is CW pump power failure or loss of cooling, and it leads to runaway reaction. The support system selects the appropriate operation for this plant conditions.

Steps 3) and 4)

It is judged that the CW pump failure or loss of cooling cannot be repaired easily. Then the support system selects the operation “switch to redundancy pump” to recover from the abnormal situation. By searching the DR of redundancy pump in the KB, the support system can show not only the operating procedure but also the reason why this operation is selected, i.e. the redundancy pump is designed to maintain the reactor temperature at designed value in case of CW pump failure. This is drawn from safety-related specific DR as shown in Table 2. If this DR information is not provided

to plant operators, they may confuse whether this operation should be done to keep the plant safely or to improve the productivity.

Step 5)

If the reactor temperature and pressure continue to increase, the support system can select following emergency shutdown procedure and their reasons.

- a) Add the shortstop chemical to quench the runaway reaction.
- b) Activate the depressuring system to avoid the reactor explosion.
- c) Open the bottom discharge valve manually to drop the batch into a dump tank.

The purposes of these operations are drawn from safety-related specific DR as shown in Table 2. To indicate the design intent based on the specific DR makes the purpose of each operation clear and plant operators can carry out the abnormal or the emergency operational responses according to engineer's design intents.

In previous works, plant operators had to decide the operation based on the only design result, but they could not understand the intent of the operation clearly. Using the RMI allows plant operator to decide the operation with reasoning selected according to the PSM policy through the PLC finally. Furthermore, operator can understand whether the operation is based on process specific idea or general concept with the viewpoints of operation selected (S, Q, C, D, E).

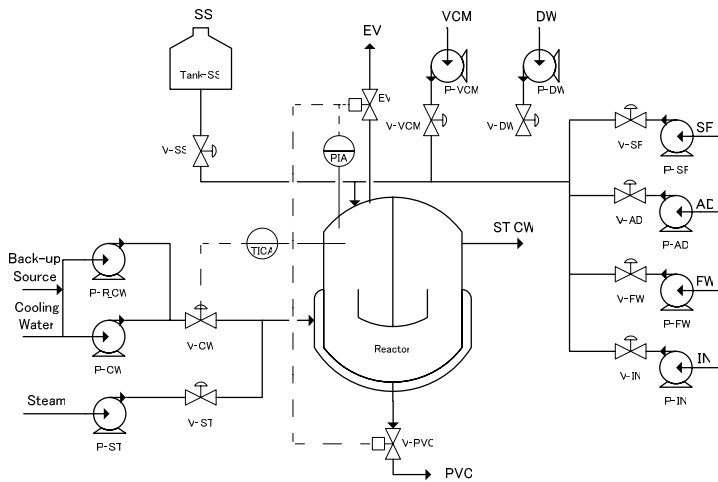


Fig. 3. PVC reactor unit

5 Conclusion and Future Works

This paper proposes the safety operation support system based on the risk management information (RMI) to help the plant operators with the decision making in case of plant abnormal situation. Proposed support system can suggest the appropriate operation and explain the reason of selecting such operation, because the RMI includes the theoretical process safety management (PSM) information and can

be used as knowledgebase for the safer operation support system. The ideas have been successfully simulated in the PVC batch process. This support system enable plant operator to understand the PSM policy of the objective plant through the plant life cycle in clearer way. As a result of it, it is expected that it makes safety operation according to the intent of integrated PSM a real possibility and human error such as false judgment and erroneous operation during the operation can be protected.

Future works includes the development of subsystems such as fault detection, fault diagnosis and on-line dynamic simulator and the integration of proposed support system with them. Furthermore, it is required that risk management method itself should be systematized so as to be managed in a logical process. Finally, the verification for actual plant is also needed.

Acknowledgement

This research was partially supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B), No.16310115, 2005.

References

1. Batres,R., Lu,M.L. and Naka,Y., Operation Planning Strategies for Concurrent Process Engineering, AIChE Annual Meeting 1997, (1997)
2. Fuchino,T. and Shimada,Y., IDEF0 Activity Model based Design Rationale Supporting Environment for Lifecycle Safety, Lecture Notes in Comput. Sci., 2773-1 (2003) 1281-1288
3. Regli,W.C., Hu,X., Atwood,M. and Sun,W., A Survey of Design Rationale Systems, Approach, Representation, Capture and Retrieval, Engng. Comput., 16 (2000) 209-235
4. Shimada,Y., Hamaguchi,T. and Fuchino,T., Study on the Development of Design Rationale Management System for Chemical Process Safety, Lecture Notes in Comput. Sci., 3681-1 (2005) 198-204
5. Shimada,Y. and Fuchino,T., Safer Plant Operation Support by reusing Process Design Rationale, Proc. of 11th Int. Symp. on Loss Prevention and Safety Promotion in the Process Industry, D (2004) 4352-4357
6. Drake,E.M., An Integrity Levels for Chemical Process Safety Interlock System, AIChE/CCPS, Proc. of Int. Symp. and Workshop on Process Safety Automation, (1999) 295-336
7. AIChE/CCPS, Guidelines for Process Safety in Batch Reaction Systems, AIChE/CCPS (1999)
8. Shimada,Y., A-Gabbar,H., Suzuki,K. and Naka,Y., Advanced Decision Method for Plant Operation Support System. Proc. of 10th Int. Symp. on Loss Prevention and Safety Promotion in the Process Industries, 1 (2001) 619-630
9. Vedam,H., Dash,S. and Venkatasubramanian,V., An Intelligent Operator Decision Support System for Abnormal Situation Management, Comp. & Chem. Engng., Suppl., (1999) S577-S580

Analysis of ANFIS Model for Polymerization Process

Hideyuki Matsumoto, Cheng Lin, and Chiaki Kuroda

Department of Chemical Engineering, Tokyo Institute of Technology
O-okayama, Meguro-ku, Tokyo 152-8552, Japan
{hmatsumo, clin, ckuroda}@chemeng.titech.ac.jp

Abstract. Adaptive-network-based Fuzzy Inference System (ANFIS), proposed by Jang, is applied to estimating characteristics of end products for a semibatch process of polyvinyl acetate. In modeling the process, it is found that an ANFIS model restructured in a way of cascade mode enhances predictive performance. And membership functions for temperature, solvent fraction, initiator concentration and monomer conversion, which are changed by training, are analyzed. Consequently, it is considered that the analysis of parameter adjustment in the membership functions can clarify effect of adding the conversion to an input variable of fuzzy sets on enhancement of robustness and improvement of local prediction accuracy in restructuring ANFIS model.

1 Introduction

“Soft sensor” is an attractive approach of offering alternative solutions for estimating dynamic behavior of chemical process, particularly when conventional hardware sensors are not available, or when the sensor’s high cost or technical limitation complicates its online use. It is well known that soft sensors are based on empirical models including Kalman filters, artificial neural networks, fuzzy logic, and hybrid methods. In our previous study [1], it has been reported that artificial neural networks are useful for modeling nonlinear behavior of chemical process, and then it has been claimed that it is important for the modeling to pick up key inputs from many process variables and parameters. So, it is considered that introduction of knowledge (qualitative information) about functional relationship of input-output pairs could make the modeling of chemical process simpler and more efficient.

It is also reported that fuzzy modeling, which is based on the pioneering idea of Zadeh, offers a powerful tool to describe a complex nonlinear system such as a biological system. A fuzzy logic system has ability to handle numerical data and linguistic information simultaneously, for a nonlinear mapping from an input data vector space into a scalar output space. The specific of the nonlinear mapping are determined by fuzzy set theory and fuzzy logic. Then, the fuzzy system models basically fall into two types [2]. One type includes linguistic models. Another type of fuzzy models is based on the Takagi-Sugeno-Kang (TSK) method of reasoning proposed by Sugeno et al. These models are based on a rule structure that has fuzzy antecedent and functional consequent parts, thereby qualifying them as mixed fuzzy and nonfuzzy models. It is considered that TSK fuzzy models are useful for modeling

chemical process with prior knowledge about cause-effect relationship among many process variables, because they have the ability to represent not only qualitative knowledge, but quantitative information as well.

In this study, we come up with applying the fuzzy modeling based on the TSK method to estimating characteristics of products from measurable process data in a polymerization process. According to the previous studies reported by Teymour [3], semibatch polymerization reactors have multiplicity of process dynamics and exhibit oscillatory behaviors of a steady state process in a parameter space. In the fuzzy modeling of such a semibatch polymerization process, there are two primary tasks: “*structure identification*” and “*parameter adjustment*”. The “*structure identification*” determines I/O space partition, rule premise and consequent variables, the number of IF-THEN rules, and the number and initial positions of membership functions. The “*parameter adjustment*” identifies a feasible set of parameters under given structure. Then, to deal with the problem of parameter adjustment, we propose to use a scheme of “Adaptive-Network-Based Fuzzy Inference System (ANFIS)”, proposed by Jang [4] in 1993.

ANFIS is based on the first-order TSK fuzzy model. Paradigm of multilayer feed-forward neural network is used in the fuzzy model. Parameters of premise and consequent parts are tuned by a hybrid learning algorithm based on the collection of input – output data, so that ANFIS possesses high accuracy of estimation and fast learning speed. In the past decades, many researchers have reported high applicability of ANFIS to process modeling. However, there are few studies that investigate applicability of the ANFIS to estimating characteristics of products in a polymerization process. Furthermore, there are a few studies that investigate hybrid effects of “*structure identification*” and “*parameter adjustment*” on predictive performance in the ANFIS model for nonlinear dynamics of chemical process.

Hence, in this article, we apply the ANFIS to estimating characteristics of end products for a semibatch polymerization process and analyze the ANFIS model. The purpose is to investigate on methods for analysis of the ANFIS model from the viewpoint of analyzing the process dynamics and improving structure of the model.

2 ANFIS

2.1 Architecture of ANFIS

For explanation simplicity, we assume that the ANFIS has two inputs x and y and output z . For the first-order TSK fuzzy model, a typical rule set with two fuzzy if-then rules is the following:

Rule1: If x is A_1 and y is B_1 , then $f_1 = p_1x + q_1y + r_1$

Rule2: If x is A_2 and y is B_2 , then $f_2 = p_2x + q_2y + r_2$

The corresponding equivalent ANFIS architecture is shown in Fig.1. The entire system architecture consists of five layers, namely, fuzzification layer, product layer, normalized layer, de-fuzzification layer and total output layer.

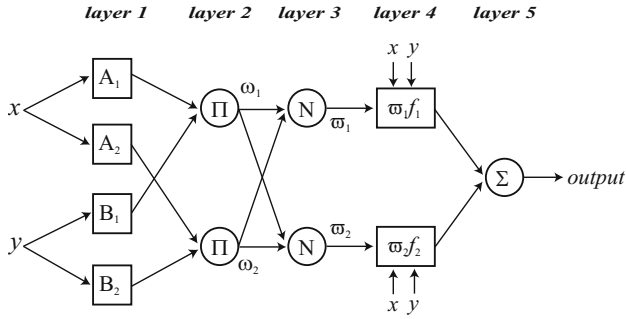


Fig. 1. Schematic diagram of ANFIS

[Layer1] Every node i in this layer is an adaptive node with a node function

$$O_{1,i} = \mu A_i(x), \quad i = 1,2 \tag{1,a}$$

$$O_{1,j} = \mu B_{j-2}(y), \quad j = 3,4 \tag{1,b}$$

where x (and y) is the input to node i and A_i (and B_{j-2}) is a linguistic label (such as “small” or “large”) associated with this node. $O_{1,i}$ is then the membership grade of a fuzzy set A ($= A_1, A_2, B_1$ and B_2) and a Gaussian parameterized membership function is used which guarantees a smooth transition between rule and rule:

$$\mu A(x) = \exp\left\{-1/2\left[(x - c_i)/\sigma_i\right]^2\right\} \tag{2}$$

where $\{\sigma_i, c_i\}$ is the parameter set. Nonlinear parameters in this layer are referred to premise parameters.

[Layer 2] The output of this layer is the product of all the incoming signals and represents the firing strength of a rule:

$$O_{2,i} = \omega_i = \mu A_i(x)\mu B_i(y), \quad i = 1,2 \tag{3}$$

[Layer 3] The outputs of this layer are the normalization of incoming firing strengths:

$$O_{3,i} = \bar{\omega}_i = \omega_i/(\omega_1 + \omega_2), \quad i = 1,2 \tag{4}$$

[Layer 4] Every node i in this layer is an adaptive node with a node function.

$$O_{4,i} = \bar{\omega}_i f_i = \bar{\omega}_i (p_i x + q_i y + r_i) \tag{5}$$

where $\bar{\omega}_i$ is a normalized firing strength from layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set of this node. Linear parameters in this layer are referred to consequent parameters.

[Layer 5] The single node in this layer computes the overall output as the summation of all incoming signals:

$$O_{5,i} = \sum_i \bar{\omega}_i f_i \tag{6}$$

2.2 Hybrid Learning Algorithm

The above-mentioned nonlinear and linear parameters in premise and consequent parts are adjusted by a hybrid learning algorithm, based on a collection of process data. One of the hybrid learning algorithms, which Jang [4] proposed, includes the *Gradient Descent Method* (GDM) and the *Least Squares Estimate* (LSE). Each epoch of the hybrid learning procedure is composed of a forward pass and a backward pass. In the forward pass, functional signals go forward till layer 4 and consequent parameters are optimized by the LSE, under condition that premise parameters are fixed. In the backward pass, the error rates propagate backward and the premise parameters are updated by the GDM.

Because the update algorithms of the premise and consequent parameters are decoupled in the hybrid learning procedure, further speed up of learning is possible by using other versions of the gradient method on the premise parameters. In this study, it is considered that simulation tests using the LSE and the GDM are still time-consuming. Thus, we apply the faster *Levenberg-Marquardt method* (LM) that is a nonlinear regression model method, which combines two gradient methods: the gradient descent method and the quadratic approximation method, to updating the premise parameters.

3 Process Simulation of Semibatch Polymerization Reactor

The case study in this article is the free-radical solution polymerization of vinyl acetate, for which Teymour et al. [3] analyzed the dynamic behavior in CSTRs. The process, which we deal with, is an exothermic polymerization in a semibatch mode. AIBN [2,2'-azobis (2-methylprppionitrile)] and t-butanol are used as an initiator and a solvent respectively in a feed to the semibatch reactor. Temperature of the feed, which contains the monomer, is 303 K. Volume of the reactor is 1000 l. And temperature of cooling water in the reactor jacket is set to 318 K.

The semibatch reactor model in this study is derived using the same approach as for the CSTR model reported by Teymour et al. The resulting ordinary differential equations (ODEs) display a structure similar to that of the CSTR model, except that no outflow term appears for the semibatch reactor case and that many ODEs explicitly depend on the volume of the reaction mixture. Thus, a dynamic total volume balance is necessary, and can be expressed as

$$\frac{dV}{dt} = q \frac{\rho_f}{\rho_{f(T)}} + V[MW]_m R_m \left[\frac{1}{\rho_p} - \frac{1}{\rho_m} \right] - V \sum_i \frac{v_i}{\rho_i} \frac{d\rho_i}{dt} \quad (7)$$

In addition, we set seven ODEs for temperature, monomer's fraction, solvent's fraction, initiator concentration, 0th moment, 1st moment and 2nd moment. And these ODEs are solved with kinetic equation and predictive equations for process parameters in Matlab[®] environment. Characteristics of end product in the process, e.g. number average molecular weight, weight average molecular weight, and polydispersity, can be calculated by using the above-mentioned moments.

In this study, an inlet flow rate scheduling strategy is introduced to process simulation of the semibatch reactor. In the inlet flow rate scheduling, the characteristic time constant $\hat{\theta}$ expressed by Eq. (8) is kept constant by increasing the flow rate of feed with increase of the reactor volume.

$$\hat{\theta} = V/q \quad (8)$$

In operating this strategy, $\hat{\theta}$ tends to the same limiting value of the equivalent CSTR residence time as the reactor fills up, so that the intensive states of the semibatch reactor will could the same dynamic behavior as those of the CSTR. Therefore, it is supposed that process simulations of the semibatch reactor with the flow rate scheduling may provide an efficient procedure for the startup phase of the CSTR, which is transient in nature and often results in off-specifications product. According to Teymour, it has been reported that dynamic behaviors of the semibatch reactor in the flow rate scheduling show multiplicity similar to those of the CSTR. For an example, it is shown that conversion changes periodically with time in the case when initiator concentration is 0.04 mol/l and $\hat{\theta}$ is 58 min. It is also known that forms of the time-variation for conversion and polydispersity vary in value of $\hat{\theta}$.

4 Results and Discussion

4.1 Identification of Characteristics of Polymer

In this study, we tried to construct an ANFIS model for estimating the characteristics of polymer from process variables that could be measured online. In order to determine an initial ANFIS model, it is necessary to determine four items: (1) input variables to the rules, (2) number of rules, (3) fuzzy partitioning of the I/O space, and (4) shapes and initial parameters of the membership functions. The first, four measurable variables, which were reactor temperature T , reactor volume V , initiator concentration I , and solvent fraction S , were picked as input variables of the models for estimating four variables; weight average molecular weight MW_w , number average molecular weight MW_n , monomer conversion X_m , and polydispersity PD .

To solve the problems of determining number of rules and fuzzy partitioning of the I/O space, *Subtractive Clustering Method* (SCM) was selected in the present case. This SCM is a modification of the *Mountain Clustering Method* proposed by Yager and Filev, developed by Chiu and provided by the Fuzzy Logic Toolbox for Matlab[®] [5]. Number of cluster centers represents number of fuzzy rules. In this study, a Gaussian function was used as a membership function in the ANFIS model. The width of the Gaussian function was determined by the cluster radius r_a .

In training of the above-mentioned initial ANFIS model using the above-mentioned hybrid learning algorithm, 6000 training data sets and 2500 validation data sets were prepared, which were collected using the process simulation of semibatch polymerization reactor, as described in chapter 3. Number of epochs for the training was set at 500. As the first step of training, sensitivity analysis was carried out to validate selection of the four input variables. In the sensitivity analysis, a value of 0.5

for cluster radius r_a for three input variables was selected except the one of interest, which would be assigned with a value of 0.1. When root mean square error (RMSE) for training was calculated, the highest value of RMSE in every output variables was estimated in the case of changing r_a for the reactor volume V . Since RMSE for the V increased with decrease of r_a (increase of clusters), it was considered that the reactor volume, which was an extensive property and an influential process variable as explained in the previous chapter, was not appropriate for constructing a robust model in the present case. Thus, we used three input variables (T, I, S) in the subsequent training and estimation for the ANFIS model.

If the ANFIS model has four input variables and three membership functions per an input, number of fuzzy rules is 3^4 (=81). However, by the SCM and the above-mentioned sensitivity analysis, 81 rules were reduced to 12 rules, so that long computation time and predictive performance for polydispersity could be improved. Moreover, we investigated robustness of the ANFIS model with the SCM. For analysis of the robustness, different groups of process data sets, which were collected in three different conditions (see Table 1), were prepared. 6000 data sets for training were collected in the process simulation of the condition 1, to which up to 0.5 % of disturbance was introduced in form of white noise. In the present case, process variables, of which signals included the disturbance, were cooling temperature, feed temperature, and initiator feed concentration. As a result of analyzing the robustness, good predictive performance for conversion was seen in simulation of every condition. However, as to polydispersity, the lower predictive performance was shown. For an example, RMSE was 0.085 in the case of condition 2. It was considered that the predictive performance came down when difference of $\hat{\theta}$ between training data and prediction data became larger.

Table 1. Simulations' conditions for analysis of robustness

	Condition 1	Condition 2	Condition 3
Cooling temp. (K)	318	317.5	319
Feed temp. (K)	304	303	302
Initiator feed conc. (mol/l)	0.04	0.038	0.042
Time const. ($\hat{\theta}$) (min)	58	53.5	60

4.2 Analysis of Parameter Adjustment

In order to improve predictability of the ANFIS model described in the previous section, we proposed to add monomer conversion X_m , which was an output for three inputs (T, I, S), as a new input variable for estimating PD, MW_w , and MW_n . Thus, the model was restructured in a way of "cascade mode". Before implementation of this scheme, sensitivity analysis for X_m was carried out by changing the cluster radius, as explained in the section 4.1. As a result of the sensitivity analysis, it was shown that predictive performance was increased by changing the value of r_a from 0.5 to 0.1, which encouraged us to use X_m as a new input variable. Then, cluster radius for X_m was determined to predict PD, MW_w , and MW_n , while the cluster radii for T, I, S were same as ones that were used in the previous modeling. As a consequence of simulations using the restructured ANFIS model, RMSE could be reduced from 0.085

to 0.027 for the case of predicting *PD* in the condition 2, which indicated better predictive performance as shown in Fig.2. Moreover, it was found that the restructured models had capability for rejecting disturbance in measurement of process data. Hence, effectiveness of adding a variable of conversion into inputs of the ANFIS model was verified in improvement of models for the semibatch polymerization process. It was supposed that conversion could be distinguished as a process state variable from other three output variables.

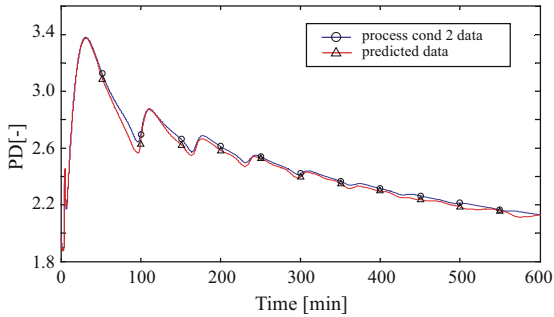


Fig. 2. Prediction of polydispersity in case of the condition 2

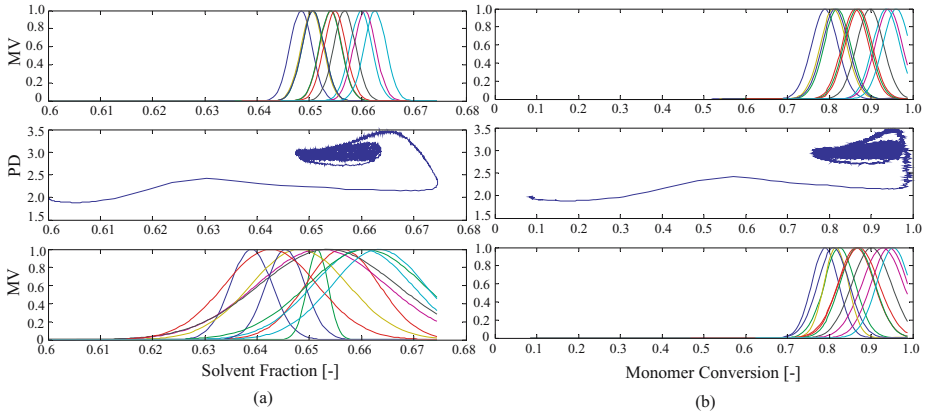


Fig. 3. Changes of membership functions by training: (a) Solvent fraction, (b) Monomer conversion

It was considered that one of advantages for using Neuro-Fuzzy modeling could be capability for facilitating analysis of cause-effect relationships among many process variables, as described in Introduction. Thus, we tried to analyze membership functions in the modeling, in order to make effects of fuzzy sets on predictive performance clearer. Figure 3 shows changes of parameters of Gaussian membership functions for predicting the polydispersity, before and after the training session. The top plots showed membership functions before training, and the bottom plots showed

ones after training. And the middle plots showed distribution of data points for polydispersity. As to initiator concentration and solvent fraction (see Fig. 3(a)), it was seen that the parameters were drastically adjusted by training. On the other hand, it was found that the parameters of membership functions for conversion (see Fig. 3(b)) and temperature were changed slightly. Therefore, it was considered that the functions for solution properties (I and S) contributed to adjusting parameters in the ANFIS model locally. It was also considered that the functions for process state variables (X_m and T) contributed to enhancing robustness of the model. Because the addition of X_m to the premise part in the cascade mode improved the predictive performance, we confirmed that X_m was a medium and key factor for determining characteristics of end product in the semibatch polymerization reactor.

5 Conclusions

ANFIS was applied to estimating characteristics of end products for a semibatch polymerization process, and membership functions in the ANFIS model were analyzed. It was confirmed that application of the SCM to determining the fuzzy sets improved predictive performance of the model. It was found that the analysis of parameter adjustment for the functions clarified that restructuring of the model in the cascade mode enhanced the robustness improved the predictive performance. Therefore, it seemed that the analysis of ANFIS model for polymerization process contributed to classifying key process variables, so as to facilitate validation of picking up input variables of fuzzy set.

References

1. Matsumoto, H., Kuroda, C., Palosaari, S., Ogawa, K.: Neural Network Modeling of Serum Protein Fraction using Gel Filtration Chromatography. *J. Chem. Eng. Japan*, 32 (1999) 1-7
2. Barada, S.: Generating Optimal Adaptive Fuzzy-Neural Models of Dynamical Systems with Applications to Control. *IEEE Trans. Syst. Man. Cybern.*, 28 (1998) 371-391
3. Teymour, F.: Dynamics of Semibatch Polymerization Reactors: I. Theoretical Analysis. *AIChE J.*, 43 (1997) 145-156
4. Jang, J-S. R.: ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Trans. Syst. Man. Cybern.*, 23 (1993) 665-685
5. Pra, A. L. D.: A Study about Dimensional Change of Industrial Parts using Fuzzy Rules. *Fuzzy Sets and System*, 139 (2003) 227-237

Semi-qualitative Encoding of Manifestations at Faults in Conductive Flow Systems

Viorel Ariton

“Danubius” University from Galati, Lunca Siretului no.3,
800416 – Galati, Romania
variton@univ-danubius.ro

Abstract. A complex system in industry is often a conductive flow system. Its abnormal behaviour is difficult to manage due to incomplete and imprecise knowledge on it, also due to propagated effects that appear at faults. Human experts use knowledge from practice to represent abnormal ranges as interval values but they have poor knowledge on variables with no direct link to target system’s goals. The paper proposes a new fuzzy arithmetic, suited to calculate abnormal ranges at test points located far deep in the conductive flow structure of the target system. It uses a semiqualitative encoding of manifestations at faults, and exploits the negative correlation of the power variables (pressure like and flow-rate like) in faulty cases. The method is compared to other approaches and it is tested on a practical case.

1 Introduction

Real technical systems are complex systems which normative model is hardly available as a whole. Instead, complexity is managed by modularization of the target system hence the normative model is actually a collection of independent models – each corresponding to a certain subsumed module. As expected, the abnormal behaviour of real systems is much more difficult to obtain as a whole model. The modularization does not offer premises for the independence of faulty models (one faulty module affects the others), while a certain fault provokes many effects and various faults show same effects. Consequently, no precise model of the faulty behaviour exists and the fault diagnosis relies on human experts. When a computational model is in concern, the imprecise values may get a fuzzy logic representation, while incomplete and qualitative knowledge become rules.

Almost any technical system (and not only) is a *Conductive Flow System* (CFS), i.e. it performs the ends by means of (matter and energy) flows, that pass through pipe-like conduits and observe the Kirchoff’s laws. For such a system, just from the design phase, each module is meant to achieve a definite end – carried out by specific functions upon the flows of the components of the module. A fault occurs at component level (located at module level, too), and the effects propagate throughout the system affecting the ends of the other modules. The effects’ propagation induce deviations of the flow power variables – flow rate like and pressure like, which are of greatest importance in the normal and abnormal behaviours’ assessment (concerning the control or the fault diagnosis).

Dedicated diagnosis expert systems may replace rare (and expensive) human experts in the field. [1] proposes a diagnosis expert system for complex CFSs, using fuzzy power variables and modularisation. However, when the human expert establishes fuzzy attributes related to the “abnormal” behaviour, he or she has refers to certain deviations of the end(s) of a module from the expected (“normal”) values. The variable in a test point directly linked to an end of a module (further denominated *primary test point*) benefits from the human expert’s knowledge when performing the fuzzy partition; a variable in a test points far from components achieving the ends (*secondary test point*) lacks such knowledge. For example, the flow rate in a test point located at the input of a hydraulic cylinder easily gets a fuzzy partition (with “normal” and “abnormal” attributes). Instead, the normal and the abnormal attributes of an ingoing flow rate in a Kirchoff’s node (e.g. supplying two hydraulic cylinders – see J_0 node in Fig. 3) gets the fuzzy partition from the sum of the respective attributes of the outgoing flow rates.

The paper proposes a new fuzzy arithmetic (shortly named OMSA), suited for fuzzy partitioning of power variables located in secondary test points, when using the fuzzy partitions of the “known” fuzzy variables – at their turn, originating from partitions of fuzzy variables in primary test points. It also presents advantages of OMSA. compared to Zadeh, Lukasiewicz and Probabilistic arithmetic, along with a practical example on fuzzy partitioning of power variables for a hydraulic installation in a rolling mill plant.

2 Fuzzy Encoding of Manifestations

Human diagnosticians use the term “manifestation” to describe one effect of a fault; usually, it is a linguistic value on the deviation of the numerical value observed for the variable when compared to the normal (expected) value. After that, the linguistic values enter qualitative relations used in manifestations-to-faults mapping, i.e. in the diagnosis.

As stated above, the manifestation is a pieces of knowledge simpler than a symptom (which usually is a combination of more manifestations), both associated to an abnormal behaviour and to a specific running context. The manifestation refers more often to a meaningful interval value than to a point-wise one, and offers a compact representation of experts’ incomplete and imprecise knowledge on the faulty behaviour of a target system.

2.1 Fuzzy Baricentric Encoding of Manifestations

The simplest computational model for a manifestation is the fuzzy attribute regarding normal and abnormal (“too low” or “too high”) interval values, as presented in Fig. 1. Kuipers [6] qualitative physics is a close approach to the common way human diagnostician acts when assessing manifestations as above: each attribute interval is delimited by *landmark* (Lm) values - as precise numeric values that separate meaningful ranges on the universe of discourse X of the variable x , in a given running context. For example, manifestation is the attribute *hi* (“too high”) attained by the variable x in the degree $\mu(v)$ for the instance value v observed in the current running context.

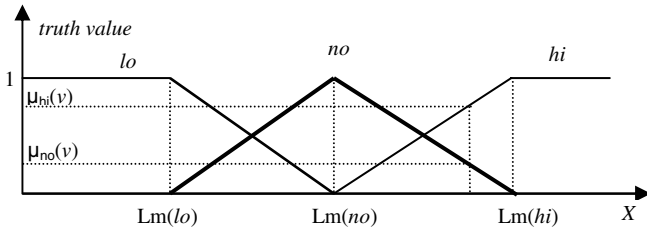


Fig. 1. Fuzzy partition of the observed variable x , over the universe of discourse X

The proposed encoding of manifestations is the same time qualitative – note the high level of abstraction for the three attributes no and lo, hi , also it is quantitative – given the truth values of an attribute when an instance manifestation occurs. The representation is hence semi-qualitative, closest to the qualitative way of thinking of human experts with a "linear" representation and including the "intensity" of the manifestation (as the truth value). Human expert should only define the no attribute – by its the landmarks $Lm(no), Lm(lo)$ and $Lm(hi)$, then automatically is generated the triangular "baricentric coding" membership function as in [3] – advantageous from many points of view.

Human experts can easily assess normal attributes no for the observed power variables in *primary test points*, using their knowledge on expected ends of the target system. However, the no attribute of a variable in a *secondary test point* results only from some processing – for example the fuzzy sum of no attributes of the already known variables (e.g. flow rates in a Kirchoff's node). Using Zadeh fuzzy arithmetic, the fuzzy sum leads to "ignorance" – i.e. the attribute tends to cover the entire universe of discourse of the fuzzy variable [5], hence an unrealistic no attribute. On the other hand, Zadeh fuzzy arithmetic only applies to "positive correlated" operands. So, it is necessary a fuzzy arithmetic suited to a realistic outcome in the case of power variables at faults in CFSs.

2.2 Fuzzy Arithmetic for Correlated Operands

The general Compositional Rule of Inference (CRI) applied to points of the two subset operands σ_i – each corresponding to respective instance membership values, is given by:

$$\mu_{no}(\sigma_k) = \bigcup_{i \leq k} \left(\bigcap_{i1 \neq i2} (\mu_{no}(\sigma_{i1}), \mu_{no}(\sigma_{i2})) \right) . \tag{1}$$

where \cap is the t-norm, and \cup is the t-conorm, relevant for the implication chosen in the given context [8]. In a certain inference rule $A \rightarrow C$, t-norm/conorm are expansion type operators – when given the fact A' the deduced consequence C' comply $C \subseteq C'$, or they are reduction type operators – when $C' \subseteq C$ holds [8]. In a fuzzy arithmetic, the t-norm is a numerical operation – e.g. the algebraic sum, and the t-conorm is an aggregation which complies the "logical correlation" between operand variables [7]; the correlation indicates the dependence of two variables' evolution (e.g. in a causal relation). When the t-norm is the sum, the t-conorm (the aggregation) should preserve the commutative property.

In the literature, there exist few commentaries on which fuzzy operator is suited to which specific correlation between operand variables; [7] proposes some specific logical aggregation in each of the three possible situations, as follows:

- Positive correlated variables – Zadeh AND, Zadeh OR:

$$A \text{ ZAND } B = \min(A, B), A \text{ ZOR } B = \max(A, B) . \tag{2}$$

- Negative correlated variables – Lukasiewicz (bold) operators:

$$A \text{ LAND } B = \max(0, A+B-1), A \text{ LOR } B = \min(1, A+B) . \tag{3}$$

- Zero correlated variables – probabilistic logic operators:

$$A \text{ PAND } B = A*B, A \text{ POR } B = A+B-A*B . \tag{4}$$

From the operators above, Zadeh operators lead to the expansion inference and Lukasiewicz operators lead to the reduction inference.

3 Opposite Maximum Support Arithmetic

The Kirchhoff's laws are balance equations for the two types of power variables: intensive (pressure like) and the extensive (flow rate like) – applied respectively to loops and nodes which are junctions in the bond graph approach of a CFS [4]. Each J bond graph junction exhibits a balance equation for one type of power variable (σ_j – e.g. the extensive one in the Kirchoff's node), and a unique value for the other type (κ_j – e.g. the intensive one):

$$\sum_i \sigma_{ji} = 0 \mid i \in J ; \kappa_j = \kappa_{j_i} = \kappa_j \mid i, j \in J . \tag{5}$$

In (5), the balance equation is applied to all variables σ_{ji} incident in the J junction (σ further denominated “summed variable”). If J is a loop, then all σ_{ji} are intensive variables in all test points along the loop but the extensive variables κ_{ji} have all the same value in the test points along the loop (κ further denominated “common variable”); if J is a node, σ_{ji} are all extensive variables and κ_{ji} are intensive variables.

3.1 Deviations of Power Variables at Faults

If a fault occurs, it induces a deviation of the σ_{jd} power variable (e.g. the flow rate) at the faulty component, so it becomes $\sigma'_{jd} = \sigma_{jd} + \Delta\sigma_{jd}$, where $\Delta\sigma_{jd}$ is the deviation from the optimal value, supposing it falls inside the normal *no* range. The others variables in the same junction get affected, so the balance equation becomes:

$$\sum_i \sigma'_{ji} = 0, \text{ i.e. } \sigma_{jd} + \Delta\sigma_{jd} + \sum_{i \neq d} (\sigma_{ji} + \Delta\sigma_{ji}) = 0, \tag{6}$$

and subtracting (5) it results:

$$\Delta\sigma_{jd} + \sum_{i \neq d} \Delta\sigma_{ji} = 0 . \tag{7}$$

Proposition 1: Deviations of the summed power variables, in the same bond-graph junction, have opposite signs to the faulty component comparing to the non-faulty ones.

Proof: Due to the fault, the common variable will also change $\kappa'_j = \kappa_j + \Delta\kappa_j$. However,

$$\text{sign}(\Delta\sigma_{j_i}) = \text{sign}(\Delta\kappa_j), \quad \forall i \neq d, \quad (8)$$

while they observe the Ohm's law: $\sigma_{j_i} = R \cdot \kappa_j$, (R resistance/conductance like parameter).

With (7) and (8) to be valid, it must:

$$\text{sign}(\Delta\sigma_{j_d}) \neq \text{sign}(\Delta\sigma_{j_i}), \quad \forall i \neq d. \quad (9)$$

Proposition 2: Deviations of the summed power variable at non-faulty components are smaller than the one at the faulty component, in the same bond-graph junction.

Proof: From (7) and (9) $\Delta\sigma_{j_d}$ is the sum of all the deviations at non-faulty components (all the same sign), it immediately results that any $\Delta\sigma_{j_i}$ ($\forall i \neq d$) is smaller than their sum $\Delta\sigma_{j_d}$.

The two propositions assert that: (1) the power variable at the faulty component is negative correlated to same variables at the non-faulty ones, (2) neither each deviation at non-faulty components nor their sum exceed the deviation at the faulty component.

Let us note that the two "wings" of the *no* membership function (toward *lo* and *hi*) are in fact the maximum deviations accepted by the human diagnostician for the faulty case; so, for example the interval $Lm(lo) - Lm(no)$ is the maximum deviation that may occur in a bond graph junction for the respective type of power variables.

3.2 Maximum Support Preserves Specificity and Commutativity

Consider two fuzzy power variables v_1 and v_2 – at a faulty and a non faulty component, and their sum $v_3 = v_1 + v_2$. The landmarks of *no* subset of v_3 results from:

$$Lm_3(no) = Lm_1(no) + Lm_2(no), \quad (10)$$

$$Lm_3'(hi) = Lm_1(hi) + Lm_2(lo) \text{ or } Lm_3''(hi) = Lm_1(lo) + Lm_2(hi), \quad (11)$$

$$Lm_3'(lo) = Lm_1(lo) + Lm_2(hi) \text{ or } Lm_3''(lo) = Lm_1(hi) + Lm_2(lo). \quad (12)$$

Each landmark $Lm_3(hi)$ and $Lm_3(lo)$ exhibits two values, as results of the sum of the opposite landmarks of the two negative correlated variables v_1 and v_2 (see Proposition 1). To assure the commutative property of the sum operation, the *no* attribute of v_3 should have the less specificity, which is obtained when the two wings of the *no* attribute have the maximum support (the same time observing Proposition 2); hence:

$$Lm_3(hi) = \max_{1,2}\{Lm_3'(hi), Lm_3''(hi)\}, \quad Lm_3(lo) = \min_{1,2}\{Lm_3'(lo), Lm_3''(lo)\}. \quad (13)$$

The *no* attribute of x_3 is obtained through *Opposite Maximum Support Arithmetic* – OMSA. In general, $\text{Supp}^{lo}(v_j(no))$ and $\text{Supp}^{hi}(v_j(no))$ denote left (low) and right (high) wings of the *no* attribute of a fuzzy variable v_j . The results *no* attribute of the v_j variable is the sum of v_i variables entering the same junction J , so for them it holds:

$$Lm_j(no) = \sum_i Lm_i(no) \mid i, j \in J, \quad (14)$$

$$Lm_j(hi) = Lm_j(no) + \max_i(\text{Supp}^{hi}(v_i(no))), \quad (15)$$

$$Lm_j(lo) = Lm_j(no) - \max_i(\text{Supp}^{lo}(v_i(no))). \quad (16)$$

OMSA applies to power variables which belong to the same bond graph junction, when calculating the *no* attribute of a variable in a secondary test point from the known *no* attributes of other variables (e.g. from primary test points), in faulty cases of CFSs.

3.3 Comparison to Other Fuzzy Arithmetic

As discussed above, the fuzzy aggregation depends on the correlation of the operand variables [7]. In the sequel the sum of two fuzzy variables v_1 and v_2 will be compared, regarding their *no* attributes (see Fig. 2), according to: Zadeh (2), Lukasiewicz (3), Probabilistic (4) and OMSA ((14) – (16)). The two operand variables are:

- $v_1(no)$ the fuzzy number 4 with symmetrical support $\text{Supp}(v_1(no)) = [2, 6]$,
- $v_2(no)$ the fuzzy number 7 with symmetrical support $\text{Supp}(v_2(no)) = [6, 8]$.

OMSA is more specific comparing to other arithmetic (e.g. $\mu_{\text{OMSA}} \subset \mu_{\text{Zadeh}}$). Table 1 shows the *similarity* measures (SM from (18)) based on Euclidean and Hamming distance measures (DM from (17)), also the *disconsistency* measure (D from (19)). OMSA fuzzy set is not disconsistent to no others. Comparing to Lukasiewicz arithmetic – which is also suited for negative correlated variables, OMSA is not only more specific but it preserves its point-wise kernel in all the inference steps; Lukasiewicz arithmetic rapidly raises the kernel to large and flat (i.e. crisp) membership shapes. OMSA is an “expansion” inference, but it is less expansive than Zadeh arithmetic (see the Hamming distance).

Table 1. Similarity measures of OMSA when compared to other approaches

Similarity Measure	Zadeh	Lukasiewicz	Probabilistic
Disconsistency	1	1	1
Euclidian Distance	0.924	0.862	0.815
Hamming Distance	1.032	1.05	1.092

Given a distance measure $DM_p(F,G)$ for two fuzzy sets F and G:

$$DM_p(F,G) = [\sum_{x \in X} |F(x) - G(x)|^p]^{1/p}, \tag{17}$$

the similarity measure is

$$SM = 1/(1+DM). \tag{18}$$

For DM in (18), $p = 2$ leads to Euclidean distance and $p = 1$ leads to Hamming distance. The disconsistency measure refers to intersection of the two fuzzy sets F and G:

$$D(F,G) = 1 - \sup \{ \min(F(x), G(x)) \mid x \in X \}. \tag{19}$$

Compared to other approaches OMSA is: (1) close to human diagnosticians’ way of thinking, (2) more specific, (3) preserves the triangular baricentric shape of membership functions, (4) less informative (maximum support is less than the sum of all supports).

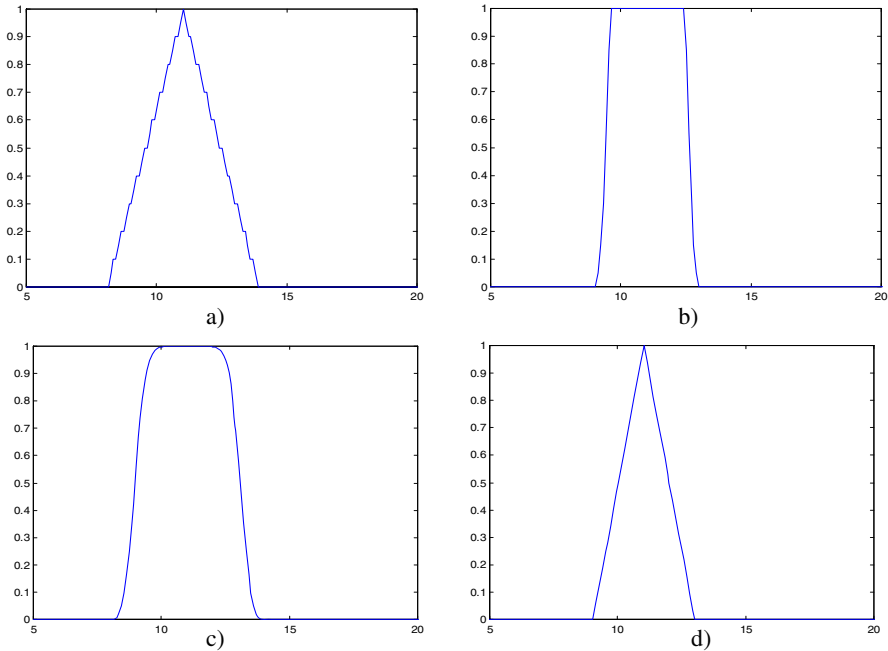


Fig. 2. Sum of the fuzzy numbers 4 (2-6) and 7 (6-8) with: a) Zadeh, b) Lukasiewicz, c) Probabilistic and d) OMSA fuzzy arithmetic

4 Using OMSA in the Fault Diagnosis

OMSA is meant for knowledge elicitation when building fuzzy-neural diagnosis systems as in [2]. The diagnosis system requires data about power variables (for *lo* or *hi* manifestations) from various test points along the flow paths in the target CFS. However, human experts can perform the fuzzy partition only for test points directly linked to goals of the target system (e.g. test points at actuators), while for test points far “inside” the target CFS they have few or no knowledge on the normal and abnormal ranges. For such points one can use Kirchoff’s laws to find out the intervals for normal and abnormal ranges. For example, using OMSA in Kirchoff’s nodes, one may get *lo* and *hi* ranges of the sum of flow rates getting in/out of the node.

OMSA is suited for negative correlated variables (as power variables are in CFSs), while Zadeh arithmetic is suited only for positive correlated ones. OMSA preserves the specificity of the result fuzzy set, and avoids ignorance. Zadeh fuzzy arithmetic rapidly leads to ignorance – i.e. after 2 or 3 sum operations the support of the result fuzzy set extends over the entire universe of discourse.

For the fault diagnosis systems presented in [2], OMSA is not directly used in the diagnosis; it is not meant for fuzzy inference on point-wise values but to ascertain fuzzy subsets at knowledge elicitation phase – eventually using a CASE tool as in [1]. Later on, during the diagnosis, manifestations get linguistic values (*lo*, *no*, *hi*) with

respective truth values – in all test points, and enter the neural blocks for discrimination of the faulty modules in the target CFS, then the neural blocks for fault isolation inside the module [2].

5 Case Study on a Simple Hydraulic Installation

OMSA is applied to indirectly assess the abnormal deviations (i.e. *lo* and *hi* ranges) when faults occur in a simple hydraulic installation of a rolling mill plant as in Fig. 3.

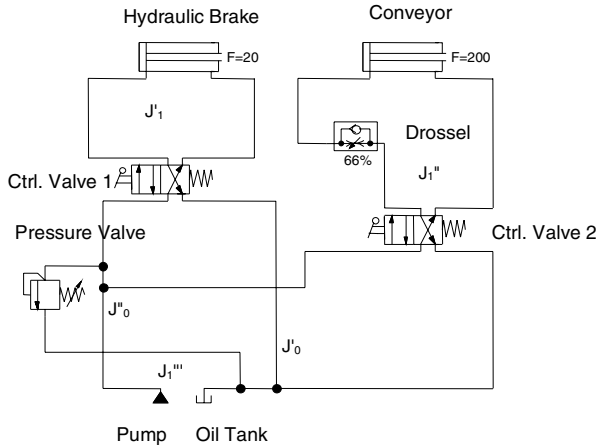


Fig. 3. Simple hydraulic installation under the fault diagnosis

The installation in Fig. 3 comprises three modules, each achieving a specific end: Supply Unit (pump, tank and pressure valve) – ensures the mineral oil flow at given rate and pressure, Hydraulic Brake (control valve, brake’s cylinder) – blocks the moving plate in the rolling mill, and Conveyor (control valve, self, the conveyor’s cylinder) – moves the plate carriage at a given speed. In the bond graph model, each module is in fact a loop (i.e. a 1 type junction), while the distribution points are nodes (i.e. 0 type junctions). To those junctions correspond the respective Kirchoff’s laws, and OMSA is applied in each junction for the fuzzy partitioning of unknown power variables required in diagnosis.

For example, when faults occur, the pistons’ movement deviates from the normal behaviour (e.g. it gets slower or non-uniform). Human diagnostician settles the normal range *no* for the mineral oil flow rate at test points located at actuators (cylinders) – i.e. along the J'_1 or J''_1 loops, given the expected normal movement of the respective piston. So, he or she settles the fuzzy partitions in *primary test points* located at cylinders. In points as from J'_0 and J''_0 junctions (which are *secondary test points*) the fuzzy partition is calculated using OMSA, based on the fuzzy partitions from primary test points.

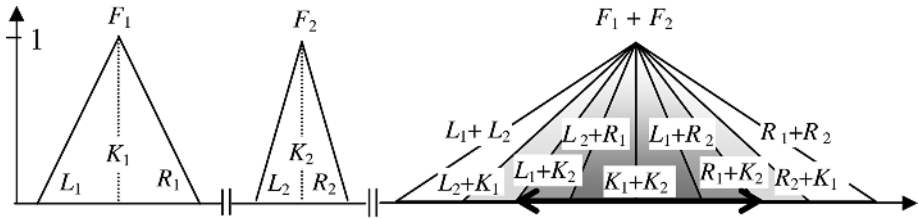


Fig. 4. Comparison of Zadeh and OMSA fuzzy arithmetic for the sum flow rate in J''_0 junction

The maximum support refers to the “worst case” (the faulty case) so, the deviations of power variables along non-faulty circuits (or in non-faulty cases) get smaller or equal support for the *no* range. Fig. 4 illustrates the specificity of the fuzzy sum of *no* subsets, for the emerging flow rates from J''_0 junction, using Zadeh arithmetic and OMSA.

Each point in the area of the sum flow rate $F_1 + F_2$ (in a secondary test point) is a sum of points from F_1 and F_2 flow rates (in primary test points). The lines $L_1 + L_2$, $L_1 + K_1$, etc. in Fig. 4 result from respective lines of F_1 and F_2 : left (L_1 , L_2), kernel (K_1 , K_2) and right (R_1 , R_2) lines. The shades indicate the specificity of sum points – the darker the bigger the density of summed points. The support of the sum fuzzy set obtained using OMSA is indicated by the arrowed segment – between $L_1 + K_2$ and $R_1 + L_2$ (F_1 exhibits the maximum support). Zadeh arithmetic extends the sum fuzzy set between $L_1 + L_2$ and $R_1 + R_2$ lines.

OMSA specificity is obviously better; however, OMSA may be used only because of the negative correlation between the two power variables at faults – flow rates F_1 and F_2 .

The fault diagnosis is more “suspicious” on a faulty case, when using OMSA for fuzzy partitioning, due to the smaller support of the *no* range, so, it reports abnormal values in secondary test points at smaller deviations than if using Zadeh fuzzy arithmetic.

6 Conclusion

The paper proposes a semi-qualitative representation of manifestations – close to human way of thinking during fault diagnosis, also a specific fuzzy arithmetic (OMSA) for negative correlated variables. OMSA is useful in the fuzzy partitioning of variables about which human diagnostician has poor knowledge, because they have no direct link to the ends of the target system. That is the case of the power variables far deep in the of target CFS’s circuitry, i.e. far from actuators. OMSA preserves better than Zadeh, Lukasiewicz and Probabilistic fuzzy arithmetics the specificity and the shape of the result fuzzy set.

OMSA was applied to automate the fuzzy partitioning of power variables along the circuits of a hydraulic installation from a rolling mill plant. Human diagnostician refers to knowledge from practice on establishing fuzzy partition for flow rates at the hydraulic cylinders, but in the flow distribution points (i.e. Kirchoff’s nodes) the fuzzy partition is calculated using OMSA. The approach is used in the knowledge

acquisition phase using a Computer Aided Knowledge Elicitation (CAKE) tool as in presented in [1], for building fault diagnosis expert systems as presented in [2].

References

1. Ariton V.: Handling Qualitative Aspects of Human Knowledge in Diagnosis. *Journal of Intelligent Manufacturing*, Springer USA, Vol. 16, No. 6, (2005) 615-634
2. Ariton V., Palade V.: Human-Like Fault Diagnosis Using a Neural Network Implementation of Plausibility and Relevance, *Neural Comput. & Applic.* (2005), 14: 149-165
3. Benzecri D.: La codage linear par morceaux. *Les Cahiers de l'Analyse des Donees*, XIV, (1989) 203-210
4. Cellier F.E.: Modeling from Physical Principles. In: W.S. Levine (ed.): *The Control Handbook*. CRC Press, Boca Raton, (1995) 98-108
5. Kruse, R. J. et. al.: *Foundations of fuzzy systems*. John Willey & Sons (1994).
6. Kuipers B. J.: *Qualitative reasoning: modelling and simulation with incomplete knowledge*. Cambridge, MA: MIT Press. (1994).
7. Siler W.: Fuzzy Reasoning - A New Software Technology. *PC AI Theme: Neural Networks and Fuzzy Logic*, Vol. 9 Issue 2, (1995) 22-38
8. Turksen B.: Fuzzy logic and the approximate reasoning. *Fuzzy sets and Artificial Intelligence*, 2 (1993) 3-32

Design Problems of Decision Making Support System for Operation in Abnormal State of Chemical Plant

Kazuhiro Takeda¹, Takashi Hamaguchi², Yukiyasu Shimada³, Yoshifumi Tsuge⁴,
and Hisayoshi Matsuyama⁵

¹ Shizuoka University Faculty of Engineering Department of Materials Science Chemical Engineering, Johoku 3-5-1, Hamamatsu, Shizuoka 432-8561, Japan
tktaked@ipc.shizuoka.ac.jp

² Nagoya Institute of Technology, Dept of Engineering Physics, Electronics and Mechanics, Gokiso-cho, Showa-ku, Nagoya, 466-8555, Japan
hamachan@nitech.ac.jp

³ Chemical Safety Research Group, National Institute of Industrial Safety, Tokyo, 204-0024, Japan
shimada@anken.go.jp

⁴ Department of Chemical Engineering, Kyushu University, Motooka 744, Nishi-ku, Fukuoka, 819-0395, Japan
tsuge@chem-eng.kyushu-u.ac.jp

⁵ School of Information, Production and Systems, Waseda University, Kitakyushu, 808-0135, Japan
matuyama@waseda.jp

Abstract. Any plants must achieve their activities, safely. Especially in chemical plants, there are possibilities of disasters or explosion, because these plants handle with many combustible or hazardous substances. Therefore, safety operation for the plants is required absolutely. To prevent the plant from abnormal state, plant maintenance techniques and human engineering has been proposed. And to prevent the plant from accidents in abnormal state, decision making support system for operation has been proposed. But the design of the system has some problems. In this paper, the design example using boiler plant simulator is illustrated, and the results and problems are proved.

1 Introduction

Any plants must achieve their activities, safely. Especially in chemical plants, there are possibilities of disasters or explosion, because these plants handle many combustible or hazardous substances. Furthermore, the harmful substances are used in the chemical plants. If the disaster or explosion is happened, it may have influence on not only the plant, but also surrounding people and environment. Therefore, safety operation for the plants is required absolutely.

To prevent the plant from abnormal state, plant maintenance techniques and human engineering has been proposed. And to prevent the plant from accidents in abnormal state, decision making support system [1], which is called DMSS, for operation has been proposed. But the design of the system has some problems.

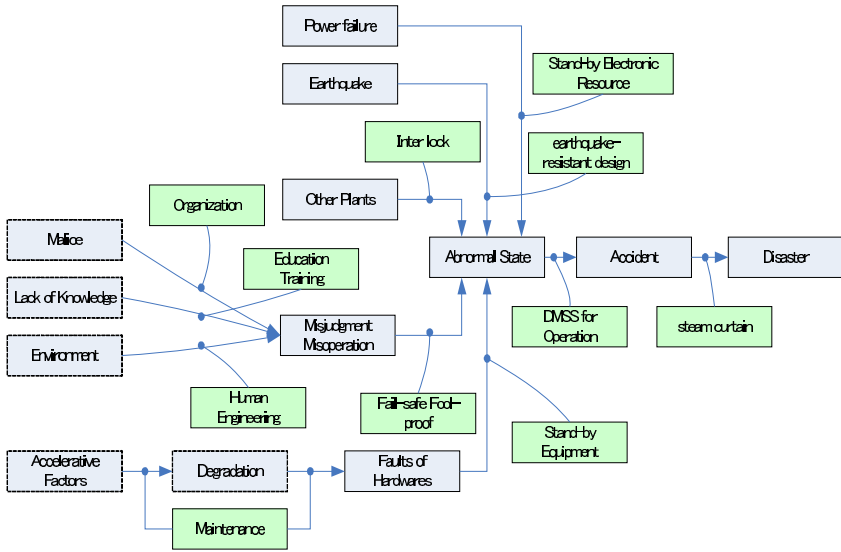


Fig. 1. Structure of Loss Prevention

In this paper, structure of loss prevention and the role of DMSS in the structure are shown. The structure and design procedure of DMSS are displayed. The design example using boiler plant simulator is illustrated, and the results and problems are proved.

2 Structure of Loss Prevention

The direct primary origins for plant failures can be categorized into three groups as follows;

- (1) External Origin: Abnormalities occurred at out side of the plant.
- (2) Human Origin: Human errors in judgment or operation
- (3) Hardware Origin: Failures of equipment.

Figure 1 shows the structure of loss prevention. The prevention of loss is nothing but to cut any of the paths between phenomenological origins and functional failures. Which is optimal to cut among paths depends on the characteristics of the origins.

Examples of external origin are abnormalities in utilities such as power failure, in flow rate or concentration of the feed, or in external signals to the control system. For encountering these abnormalities the diagnosis system should be so designed as to detect the abnormality at the phase the abnormal state appears in the plant.

Human errors are induced from lack of knowledge, poor skill, or inadequate operating environment. Failures of equipment are caused by deterioration such as wearing or corrosion which is accelerated by the operating factors such as improper lubrication or high humidity.

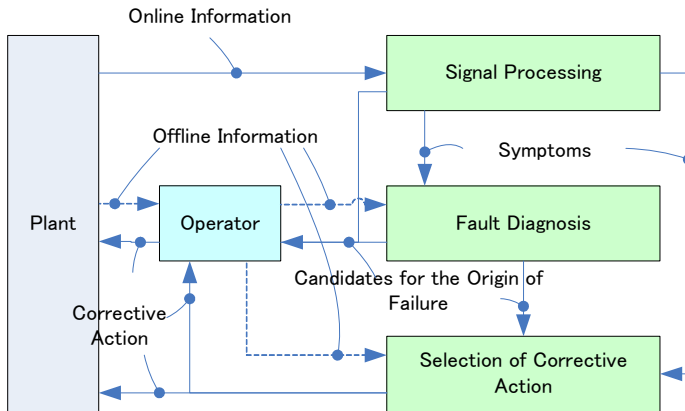


Fig. 2. Structure of DMSS for operation

The possibility of occurrence of the human errors in judgment and operation can be reduced to a large extent by means of education and training, improvement of environment, and organizational effort but cannot be completely eliminated. Therefore, it is necessary, in order to comprehensively prevent accidents, to prepare the devices of fail-safe or fool-proof or DMSS, in addition to the efforts in human origin prevention. In other words, misjudgment or misoperation which cannot be prevented by human origin prevention should be taken as the subject to DMSS.

For preventing accidents due to the equipment failures, such countermeasures as improvement of operating condition, preventive maintenance or improvement of equipment, stand-by equipment, and DMSS are effective.

3 DMSS (Decision Making Support System) for Operation

The DMSS for operation has been proposed to prevent the plant from accidents in abnormal state. In abnormal state, the system should be able to detect the abnormal state, estimate the origin of failure, and present the optimal corrective action. When candidates of the origin are serious origins, the DMSS should carry out corrective action without human operator.

Such the DMSS for operation is constructed by signal processing subsystem, fault diagnosis subsystem, selection of corrective action subsystem (Fig. 2 shows the structure).

3.1 Signal Processing Subsystem

This subsystem transforms the values of state variables to online signal by use of measurement, and extracts their features using signal processing technique. If the subsystem is able to identify the candidate of the origin of failure by extracted features, then the candidate is presented to operator and selection of corrective action

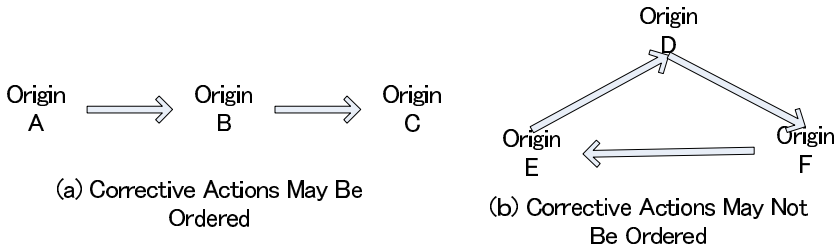


Fig. 3. Hazard Graph

subsystem. If not able to identify the candidate, the extracted features are brought to fault diagnosis subsystem.

3.2 Fault Diagnosis Subsystem

This subsystem owns cause-effect-feature relationships as base data. Using features given by signal processing subsystem and the relationships, the subsystem estimates the candidates of origins of failures.

3.3 Selection of Corrective Action Subsystem

This subsystem owns origins of failures and the related corrective actions as base data. Based on the data and information from signal processing and fault diagnosis subsystem, the optimal corrective actions are decided. When candidates of the origin are serious origins, the corrective action is executed without human operator.

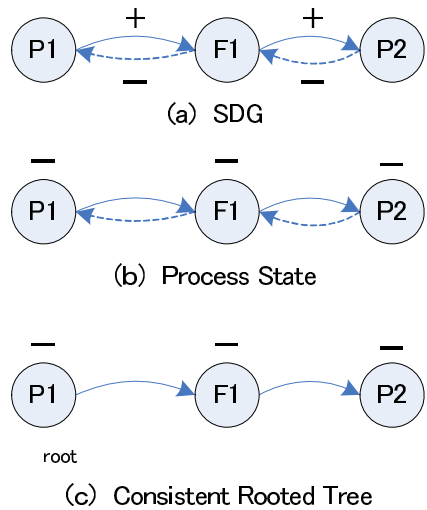


Fig. 4. Fault Diagnosis using SDG

4 Design of DMSS for Operation

The object of the DMSS for operation is to present or carried out the optimal corrective action in abnormal state. In abnormal state, two or more candidates of origin may be estimated, and two or more candidates of corrective actions may be suggested. Then, the order of executing the corrective actions should be decided. If the actions are executed in wrong order, then the situation will be worse, and will be extended to accident. To decide the order, the concept of hazard graph has been proposed [2].

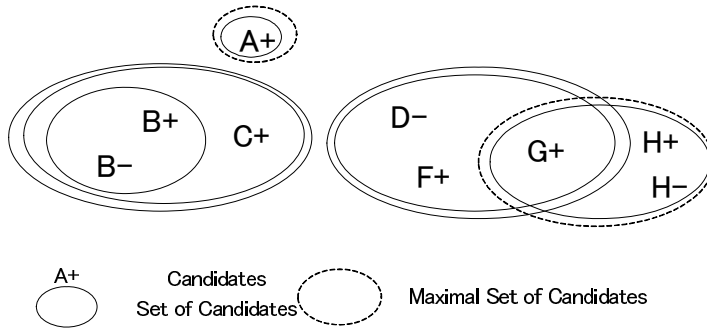


Fig. 5. Maximal Set of Candidates

4.1 Hazard Graph

The hazardous relationship is defined as relationship which the corrective action related with origin A in abnormal state caused by origin B must be extended to accident. The hazardous relationship is represented by $A \rightarrow B$. The graph represent hazardous relationship is called hazard graph. For example, when a sensor failure of inlet flow caused abnormal high fluid level of a tank, and a corrective action related with level sensor failure (cutting level control loop) caused overflow. The hazardous relationship is represented by (level sensor failure) \rightarrow (inlet flow sensor failure).

In case of Fig. 3(a), executing corrective action in order from downstream to upstream (at first, corrective action related with origin C, next corrective action related with origin B, and at last corrective action related with origin A) will prevent from accident for any candidates of origins.

In case of Fig. 3(b), the corrective action cannot be executed safely, because there is strong connected component. If the candidate of origin cannot be distinguished origin D and origin E, and the corrective action related with origin D in abnormal state caused by origin E, then accident will be caused. Therefore, the origins in strong connected component of hazard graph should not be included in one set of candidates of origin. That is to say, these origins should be distinguished in any abnormal state. Consequently, it is necessary to inspect whether the origins in strong connected component of hazard graph could be included in one set of candidates of origin or not. One of the inspection methods is maximal set of candidates (called MSC) [3] of signed digraph (called SDG) [4].

4.2 MSC (Maximal Set of Candidates) of SDG (Signed Digraph)

Iri et al. [4] have proposed SDG, which represent the cause-effect relationships in the objective plant qualitatively. As shown in Fig. 4 (a), each branch represents the immediate influence between state variables. Positive and negative influences, respectively, are distinguished by signs '+' (solid line) and '-' (dotted line) given to the branch. The combination of signs of the nodes corresponding to all the state variables in the system is defined as 'pattern'. The patterns using 3 range signs

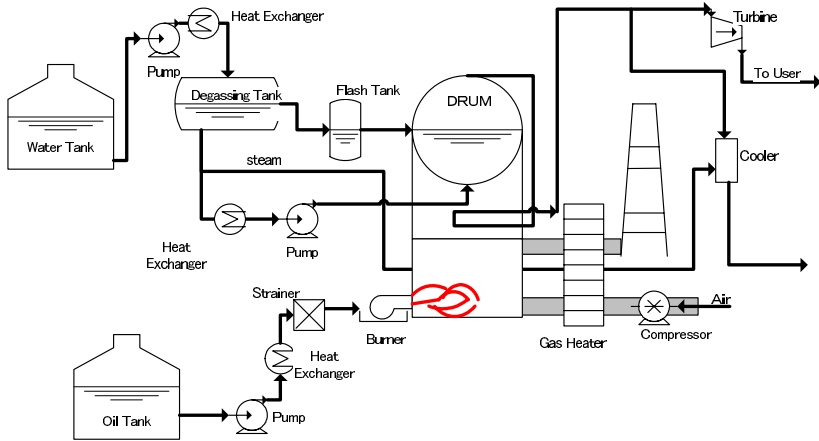


Fig. 6. Boiler Plant

Table 1. Candidates of the Origin

PV1102[-]	pressure control valve of feed water tank : close
SF1303[+]	flow meter for feed water : up
SL1301[-]	dram level sensor : down
PV1431b[-]	pressure control valve of turbine : close
TV1423[-]	temperature control valve of low pressure stream : close
PV1431c[-]	pressure control valve of turbine : close
TV1413[-]	temperature control valve of middle pressure steam : close

(abnormal high [+], normal [0] and abnormal low [-]) and 5 range signs (abnormal high [+], ambiguous high [+?], normal [0], ambiguous low [-?] and abnormal low [-]) [5] are proposed. In abnormal state, observed nodes have valid signs. Origin of failure is estimated using cause-effect relationships. Figure 4 (b) shows an abnormal process state. The branches from F1[-] to P1[-] and P2[-] to F1[-] represent negative influences, so the branches may not represent consistent relationships. These branches are disappeared (Fig. 4 (c)). The tree in Fig. 4 (c) are called consistent rooted tree and the root of the tree is the candidate of the origin of failure. The set of candidates of origin may be obtained.

A maximal set of candidates, which is called MSC, is defined as such a set of candidates that never become a partial set of any other set of candidates. The maximal set of candidates represents the diagnostic result for the worst case for all patterns. For example, set of candidates for one pattern is {B[+], B[-]}, and the other sets of candidates for the other patterns are {B[+], B[-], C[+]}, {A+}, {D[-], F[+], G[+]} and {G[+], H[+], H[-]}. The MSC are {B[+], B[-], C[+]}, {A+}, {D[-], F[+], G[+]} and {G[+], H[+], H[-]}. The problems to obtain the MSC are too many patterns will result in too large MSC and then too conservative estimation of accuracy. So, Physical analysis is needed to reduce the set of patterns, and branch and bound technique should be introduced to reduce the computation time.

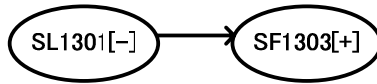


Fig. 7. A Part of Hazard Graph

If MSC include two or more origins in one strong connected component of hazard graph, then corrective actions may not be ordered in some abnormal state, else corrective actions can be ordered in voluntary abnormal state. Therefore, when MSC include these causes, it is necessary to rearrange the measurements or change process, and so on.

5 Experiments

5.1 Objective System

The objective system was the boiler plant simulator on dynamic simulator (Visual Modeler). Fig. 6 shows the system outline. This plant included 4 components, which were water feed subsystem, fuel feed subsystem, combustion subsystem, and steam supply system. Generated steam was supplied with 3 types of temperature and pressure for user's demand. The types were high, middle, or low temperature and pressure. This system was controlled with many controllers. The ratio of fuel and air was controlled.

5.2 Results and Discussions

The SDG for the objective system was constructed. The SDG included 362 nodes and 647 branches. The observed nodes were 114 nodes. The hazard graph for the system included 30 nodes and 55 branches. The MSC for the SDG was tried to calculate, but could not be gotten in practical time. So, all of the considerable failures were simulated on the simulator, and the observed patterns were gotten. The candidates of the origin for the patterns were diagnosed using the SDG.

For example, it is assumed that flow meter for feed water SF1303 was raised faulty and stuck when the corresponding flow rate was steady. The abnormal state was diagnosed for the observed pattern, and the candidates of the origin are shown in Table 1. A part of hazard graph related with the candidates is shown in Fig. 7. The hazard graph indicates that corrective action related with SF1303[+] should be done before corrective action related with SL1301[-]. The corrective actions related with the other candidates may be done in any order.

6 Conclusions

In this paper, structure of loss prevention and the role of DMSS in the structure were shown. The structure and design procedure of DMSS were displayed. The design example using boiler plant simulator was illustrated. As mentions above, it was not realistic to get MSC of SDG of large scale and complex plant. Therefore, all of the

considerable failures were simulated on the simulator, and the candidates of the origin for the patterns were diagnosed using the SDG. This method has problems. 1) Difficult to check all of the considerable failure comprehensively, 2) almost impossible to examine the abnormal state in real system, etc. In the future work, these problems will be solved.

Acknowledgements

This research was partially supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Science Research (B), No.16310115, 2005 and No.16360394, 2005.

References

1. Shibata, B., Tsuge Y., Matsuyama H.: A Consultant System for Operations during a Chemical Plant Emergency, Vol. 16. Kagaku Kogaku Ronbunshu (1990) 882-890
2. Tateno, S., Shibata B., Tsuge Y., Matsuyama H.: An Optimal Design Problem of Sensor Allocation for a Class of Fault Diagnosis Systems, Vol. 32. Transactions of the Society of Instrument and Control Engineers (1996) 577-586
3. Shibata B., Matsuyama H.: Evaluation of the Accuracy of the Fault Diagnosis System Based on the Signed Directed Graph, Vol.15. Kagaku Kogaku Ronbunshu (1989) 395-402
4. Iri, M., Aoki M., O'Shima E. and Matsuyama H.: A Graphical Approach to the Problem of Locating the Origin of the System Failure, Vol. 23. J. Oper. Res. Soc. Japan (1980) 295-312
5. Shiozaki J., Matsuyama H., Tano K., O'shima E.: Diagnosis of Chemical Processes by Use of Signed Directed Graphs –Extension to 5-Range Patterns of Abnormality-, Vol. 10. Kagaku Kogaku Ronbunshu (1984) 233-239

A Training System for Maintenance Personnel Based on Analogical Reasoning

Takashi Hamaguchi¹, Hu Meng¹, Kazuhiro Takeda², Yukiyasu Shimada³,
Yoshihiro Hashimoto¹, and Toshiaki Itoh¹

¹ Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
hamachan@nitech.ac.jp

² Shizuoka University, 3-5-1, Johoku, Hamamatsu, Shizuoka 432-8561, Japan

³ National Institute of Industrial Safety, 1-4-6, Umezono, Kiyose, Tokyo 204-0024, Japan

Abstract. This paper discusses the framework of a training system for maintenance personnel based on an analogical reasoning and a prototype system that is developed to check the effectiveness. It is difficult to get similar information on trouble and failure reports in short time. By combining trouble database and design database, and by searching using plant ontology, similar results can be searched.

1 Introduction

It is desired to accumulate knowledge on troubles and utilize them to have better production efficiency, because troubleshooting and the recovery maintenance toward unknown troubles tends to be a waste of time. It is quite important for manufacturers to increase MTBF (mean time between failures), because the trouble could lead to a halt in the production line. Therefore, highly skilled personnel are valuable assets to the manufacturers.

However, it is difficult to develop such specialists only by OJT (on the-job training). In the recent chemical industry, there are less opportunities of equipment replacement in a plant, plus the equipment has become relatively stable on stream. In this case, it is difficult for the operators and maintenance personnel to acquire enough experiences through real troubles. Also, even if an operator has experienced troubleshooting, the other operators cannot utilize the knowledge unless the operator disseminates the information throughout the company. Therefore, it is important to enhance information sharing.

In other industries such as the automobile or electronics industries, products and production facilities are frequently changed. In addition to the shortening of their life cycles, it takes longer to expose most of the troubles. Hence, many products and production facilities, on which few kinds of troubles have been reported, exist.

To support operators and maintain personnel, and to train them, it is very important to utilize helpful information of other products and/or production facilities. The concept, 'helpfulness', has many viewpoints. One fact is that it has many properties. Similarity between two facts depends on the viewpoints. The important properties might be different according to the aim of the search. For searching troubles, which might occur in the new facility, from the database; functions, components and

materials are important viewpoints. When the relationships between troubles and component makers want to be analyzed, functions might not be important. In searching similar reports, the viewpoints of similarity are expressed using properties of targets.

Although these properties are very important to search similar reports, they are not stored in troubleshooting databases. They are stored in design databases of the plants or equipments. Here, we assume that topology information can be recognized in the databases. Equipment is a part of a plant, a kind of facilities, and it is composed of some components. Components are made of some materials and are composed of sub-components. By using such ontology, it can be judged whether equipment or a plant contains some materials or other properties of their components. Even when reports are described using tag names of equipment, properties of its components can also be checked.

By combining trouble database and design database, and by searching using ontology, similar results can be searched.

In this paper, it is aimed to discuss an analogical reasoning system with multiple databases for training and support systems in the Japanese language.

2 System Framework

The following scenario is one of the usages of the proposed system. An instructor gives their trainees instructions, "Retrieve a trouble record reported on one year ago today and find its similar troubles in the database". The trainees select the viewpoint, from which similarity is judged, and find lists of trouble reports. The instructor asks why the viewpoint was selected to judge similarity and what kinds of information can be obtained from the searched lists. When the trainees can answer them, the instructor gives them another viewpoint to search similar reports. Such "question and answer" training is effective to improve maintenance skills.

When a relational database is adopted to store the information of troubleshooting and design, the data structure must be designed with utmost care. Because flexibility is poor, the data items and the primary key for search should be defined after activity analysis using such methods as IDEF0, UML. The flexible database structure like XML database is expected as one of the alternatives to conventional database structure, because it is difficult to define the suitable data structure in the first stage. The data used as samples in this study are the electrical documents for taking over shift operations in an automobile company. In order to accumulate data easily, the data structure is not strictly determined. Operators aren't bothered with the selection of the space to write information in. Therefore, the amount of information stored in documents becomes more than that where input format is strictly fixed. The data, which must be reported, are dates and writer names of reports. The information of detected phenomena, diagnosis result, and countermeasures is described in Japanese natural-language description. They might also contain information of dates and personnel for the activities. In some cases, only one of detected phenomena, diagnosis result, and countermeasures is described. The symbols to call the same equipment might be different because tag number, abbreviated names, equipment names, or

component names in the ill-conditioned equipment are allowed in the documents. Synonyms might appear to describe the same object, because natural-language description is allowed. Therefore, it is highly possible that lack of consistency in expression occurs in the documents.

A natural-language search system is aimed to be developed to enable the search of similar reports using one of the sentences in a report document as a search key. Both searched sentences and search key sentences have the problem of poor consistency.

In order to deal with the lack of consistency, Japanese morphological analysis, synonym dictionary and plant ontology are combined in this study. For Japanese morphological analysis, we used the free software “ChaSen”¹. “ChaSen” splits a natural-language sentence of a search key to sets of words in basic forms. For verbs and popular nouns, synonyms are extracted from a synonym dictionary. We used a reasonable dictionary for this search. The list of synonyms is utilized as the keyword list for a similar report search. The searched sentences in documents are also converted into lists of words in basic forms. In addition to this, the symbols corresponding to the plant, equipments, or components are extracted from the words generated by “ChaSen” using design databases of the plant.

The generation of a synonym dictionary based on the ontology in the design databases is important characteristic of this study. The relationships among plants, equipments, and components are various. In this study, the properties of the key object, which appears in the word lists in the search key, are considered as the options to express the viewpoint of similarity. A function such as “lift”, material such as “SUS201” and a maker name are examples of properties for the options. Plant, equipments, and components, which have the same properties, are dealt with as synonyms. If abbreviated names must be taken into account, they must be registered in some databases in addition to design databases.

The analogical reasoning tends to increase the number of searched data. Therefore, the method for narrowing down of the searched data is necessary. In this system, the frequency table of words in the selected sentences is generated to evaluate the adequacy of the search intent. The list of the words, which appear frequently in selected sentences, is shown to the maintenance personnel with the values of frequency. From this list, he/she can select the suitable keywords to narrow down the selected data. The new keywords are added to the keyword list for AND search. This refinement is useful to reduce the number of selected sentences.

3 Limits of Conventional Search System

In this section, we explain the search results using conventional search. Google desktop, which is the most popular conventional strong full text search system³, is used. In this work, real database, which has about 50,000 trouble and failure records, is utilized. The trouble records are collected from about 500 facilities. These records are not well organized for data search. All records were written in Japanese. Some of the stored records are shown in Fig.1.

	date	site	location	equipment	Trouble and failure	diagnosis	countermeasure
1	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
2	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
3	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
4	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
5	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
6	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
7	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
8	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
9	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
10	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
11	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
12	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
13	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
14	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
15	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
16	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
17	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
18	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
19	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
20	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
21	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
22	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
23	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
24	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
25	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
26	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
27	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
28	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
29	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
30	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
31	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
32	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
33	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
34	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
35	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
36	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
37	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
38	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
39	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。
40	2007/12/12	3011	3011A-3003A	エレベーター	エレベーターが停止した。	エレベーターの電源が切れた。	エレベーターの電源を再接続した。

Fig. 1. Example of failure and trouble reports

The example scenario in this section is as follows. A trouble was reported from a production line. The operator searched helpful information from his database. The report was “サブ1F設備でリフトが止まった。” It means “At the sub 1F equipment, the lift was suspended.” We used it as a query into the search system and wanted to find its similar records.

(Sentences cannot be utilized as queries)

When the sentence was entered into Google desktop, no records were found. Google desktop can search data which contain completely the same sentence as the query sentence. But, any difference between sentences cannot be allowed even if they have the same meaning. The sentence search is difficult by the conventional full text search system.

(Synonyms cannot be recognized)

Next, the keywords were extracted from the query sentence manually. The keywords are “サブ1F設備”, “リフト” and “止まった”. These keywords mean “equipment on sub floor 1”, “lift”, and “stopped” respectively. The keywords were used for AND search of Google desktop. This time, 216 records were found in the database.

However, there are other expressions having similar meaning as “stopped”. When “stuck”, “locked up”, and so on were added to the keyword list for OR search, more records were found. Unless the keyword list was selected in considering synonyms, useful information might be missed. However, if many synonyms were added to the search keywords, the number of found records might be too much.

More information should be added to clarify the intention of the search and to narrow down the number of found records. For example, more precise information of the search tag, such as the tag number, should be added. (Entering too precise target information makes the useful information lost)

In this example, we found the tag number of the stopped lift was “lift601”. When “lift601” was added to the keyword list, no record was found. The lift was one of the new machines and there were no trouble reported. (Information about useful similarity is selected manually)

In this case, helpful information has to be extracted from the reports about other machines. There must be useful similarity between “lift601” and them. The decision of target machines depends on the intention of the search. The selection of the machines is not easy without ontology of facilities.

4 Analogical Reasoning System Using Facility Ontology

In order to develop the analogical reasoning system, we used “Hidemaru Editor”⁴. The editor supports the regular expression and macro function as search engine. Procedure in an analogical reasoning system is shown in Fig.2. The query for trouble and failure reports database adopts Japanese natural sentences. Data of nouns and verbs are extracted from this query. If the data is a verb, it is converted into the basic form of the word. This data is used as keywords for AND search. To extract the data automatically, we used a Japanese Morphological Analysis System “ChaSen”.

If extracted keywords were only used for the search, the search result might lack the important records, which used synonymous terms. We entered “サブ1F設備でリフトが止まった。” into the search system as same as in the previous section.

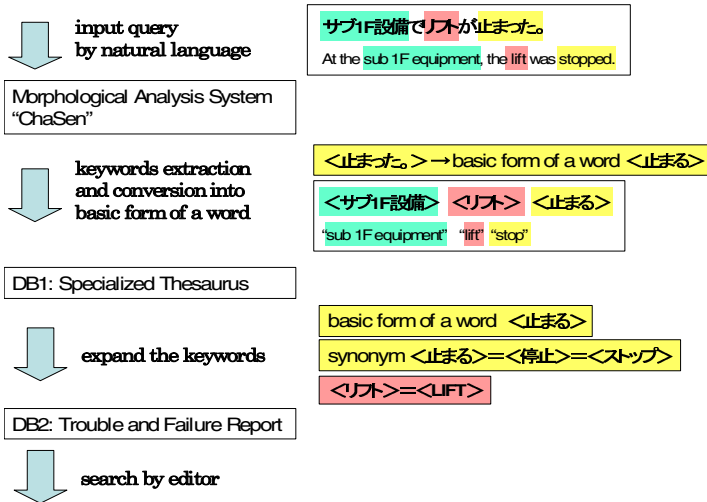


Fig. 2. The search procedure by the analogical search system

“ChaSen” extracted three keywords “サブ1F設備”, “リフト”, and “止まる”. These words mean “equipment on sub floor 1”, “lift”, and “stop” respectively. “止まる” is the basic form of “止まった” means “stopped”. The number of found records was 216.

By using a thesaurus, “停止” and “ストップ” were found as synonyms of “止まる”. These keywords were added for OR search. As the result, the number of the searched records was increased. In this example, the number of the founded records was increased from 216 to 385. In the query sentence”サブIF設備でリフトが止まった。”, the information about the place,サブIF設備” and the kind of machine “リフト” are included. The tag number of the lift was found as “lift 601” in equipment database using the combination of the information.

When the search was executed using the tag number as search key, no datum was found. As mentioned before, the lift 601 was a new facility, and no trouble records had been reported.

The other machines, which were similar to “lift601”, were searched by using ontology. For this purpose, the definition of the similarity in the viewpoints and intentions of the maintenance personnel is important. One machine has many kinds of properties. If two machines have some common properties, they are similar in the viewpoints of the common properties. In this study, the similarity is determined by selecting the properties to be common.

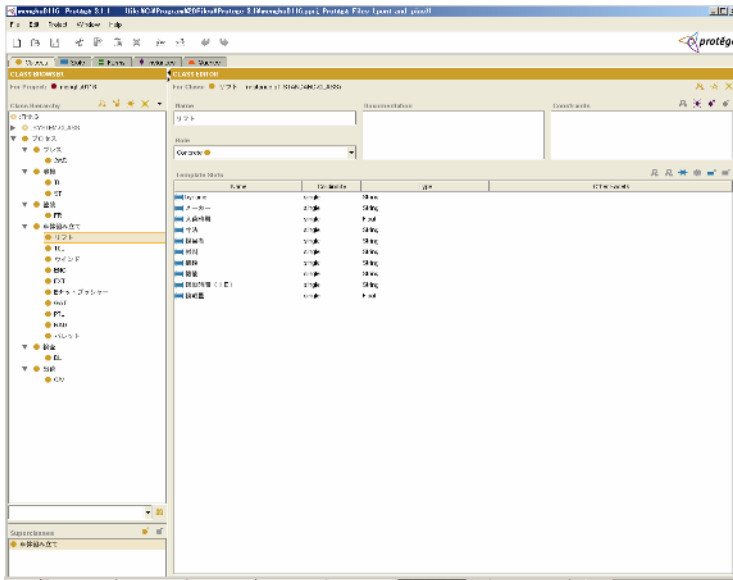


Fig. 3. The facilities database by Ontology in the Protégé

Fig. 3 shows an example of ontology of facilities. The database is designed on the Protégé, which is an ontology editor and knowledge-base framework⁵. In this example, a principal production process is divided by functional layers and the facilities class model is arranged in the layers. The object information such as type, maker, size, material, and function was registered in this database. The maintenance personnel focused on two attributes. They were the type and maker about lift601. They selected the data items to make a similar lift group for trouble reports search. The lift 601, lift 104R, and lift 501 are found by this search. It is shown in Fig.4.

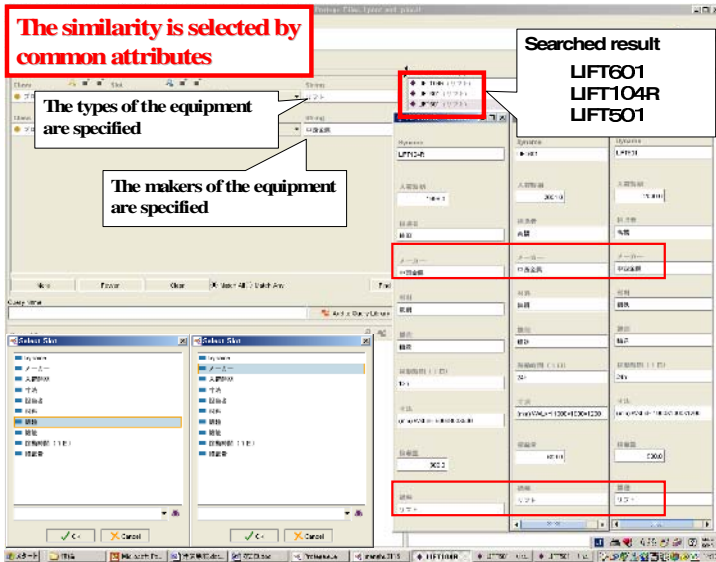


Fig. 4. Keywords selection from the facilities database

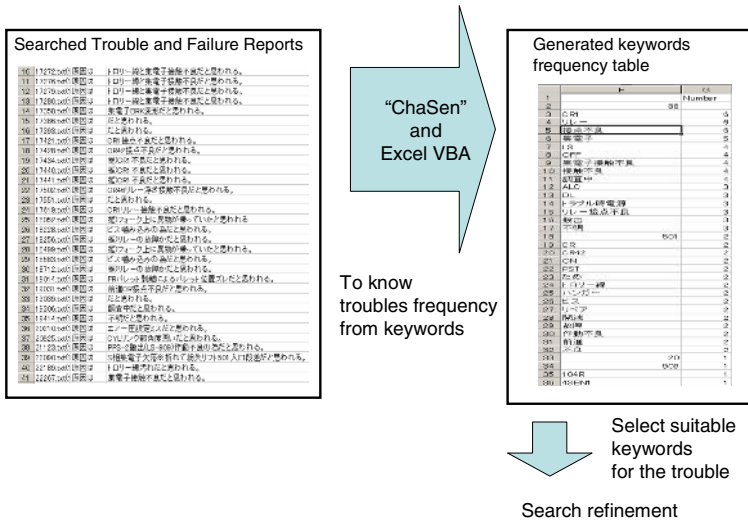


Fig. 5. The conversion from trouble records to frequency table

The three lift tags were added for OR search. In this example, the search succeeded in increasing the number of the search result from 0 to 104 records. It takes a long time to check them all. In order to search adequate keywords for further narrowing down the procedure, the list of words in the found sentences were generated. If any other keywords than these were given, the search result was zero. The frequency table of the words in the found sentences is shown in Fig. 5. In the

frequency table, some suitable keywords, such as "relay", and "poor contacting" could be chosen. After adding "poor contacting" as an AND search keyword, 104 records could be narrowed down to 14 records.

In the field of countermeasures in the selected records, information such as "connection cleaning" and "component replacement" was found. They could be considered as useful information. Although any troubles had not been reported about "lift601", it could be induced from the reports about similar facilities that "poor contacting of relay" might be the cause of the trouble and "connection cleaning or component replacement" might be effective countermeasures.

5 Conclusion

In this paper, we discussed an analogical reasoning system for multiple databases for the training and support system in the Japanese language. The databases used in this work had about 50,000 trouble and failure records. These records were collected from about 500 facilities. The conventional full text search system is not enough to search the necessary information from this database. The analogical reasoning system using multiple databases is effective for this purpose.

Acknowledgements

This research was partially supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Science Research (B), No.16310115, 2005 and No.16360394, 2005.

References

- [1] ChaSen web page: <http://chasen.naist.jp/hiki/ChaSen/>
- [2] Hamaguchi,T., Yamazaki,T., Hu,M., Sakamoto,M., Kawano,K., Kitajima,T., Shimada,Y., Hashimoto,Y., and Itoh,T., Analogical Reasoning based on Task Ontologies for On-line Support, KES 2005, LNAI 3681, 155-161, Springer (2005)
- [3] Google desktop web page: <http://desktop.google.co.jp/>
- [4] Hidemaru web page: <http://hide.maruo.co.jp/software/hidemaru.html>
- [5] Protégé web page: <http://protege.stanford.edu/>

On-Line Extraction of Qualitative Movements for Monitoring Process Plants

Yoshiyuki Yamashita

Tohoku University, Sendai 980-8579, Japan
yyama@pse.che.tohoku.ac.jp

Abstract. Qualitative representation of signal trend is useful in various applications to model temporal evolution. Monitoring and decision support of process plant operation is one of the typical application examples. Graphical representation of two temporal signals in a two variables plane often characterizes special features of data. This paper presents an effective online method to extract qualitative representation of movements in a two variables plane. The method is applied to a diagnosis of a control valve in an industrial plant.

1 Introduction

Many time series data are available in industrial systems but the extraction of meaningful information and interpretation are the problem. Trend analysis is useful in various applications such as the monitoring of process plants, understanding physical systems, and interpretation of economical systems.

Dash *et al.* proposed an interval-halving framework to automatically identify the qualitative shapes of sensor trends using a polynomial-fit based interval-halving technique [1]. But their method can not be used online. Charbonnier *et al.* proposed a methodology to extract online temporal trends from a time series [2]. They used the method to several applications such as alarm management in intensive care units, food process monitoring, and abnormal situation management in three-tanks system simulation.

In addition to the temporal representation of individual variables, graphical representation of two temporal signals in their variables plane is also often useful to characterize special features of trends. Qualitative representation of the behavior in two variables plane was proposed for the analysis of dynamic behavior of complex systems [3]. The method was successfully applied to the detection of valve stiction in industrial process plant [3]. In that study, conversion to qualitative representation was based on a simple differentiation and thresholding. This paper presents a new methodology to extract qualitative representation of system behavior in a two variables plane. The method is motivated by the idea of the Charbonnier's approach for an univariate time series [2] and extended to handle with qualitative movements in a two variables plane.

2 Trend Extraction

2.1 Temporal Segmentation

Consider splitting a univariate time series into successive line segments of the form:

$$\hat{y}(t) = (t - t_0)p + y_0, \quad (1)$$

where t_0 is the time when the segment begins, p is its slope and y_0 is the ordinate at time t_0 . Parameter identification is based on the linear least-square method. The detection whether the linear approximation is acceptable or not is based on the cumulative sum (cusum).

$$\text{cusum}(t_1 + k\Delta t) = \text{cusum}(t_1 + (k - 1)\Delta t) + e(t_1 + k\Delta t), \quad (2)$$

where

$$e(t_1 + k\Delta t) = \hat{y}(t_1 + k\Delta t) - (t_1 + k\Delta t - t_{o1})p_1 - y_{o1}. \quad (3)$$

The absolute value of the cusum is compared to two thresholds $th1$ and $th2$ at each sampling time. If the absolute value of the cusum is smaller than $th1$, the linear model is acceptable. If the absolute value of the cusum is greater than or equal to $th1$, the signal value and corresponding time are stored. If the absolute value of the cusum is greater than $th2$, the linear model is no longer acceptable and a new linear model is calculated on the stored values during the period $th1 \leq |\text{cusum}(t_1 + k\Delta t)| < th2$. Once the new linear model has been calculated, the cusum is reset to zero.

2.2 Classification and Aggregation of Segments

In this study, temporal patterns are classified into the following six classes: Steady, Increase, Decrease, Positive Step, Negative Step and Transient.

Let's consider two consecutive segments; present (i) and previous ($i - 1$). Let's define the following indices,

$$I(i) = y_e(i) - y_e(i - 1), \quad (4)$$

$$Id(i) = y_0(i) - y_e(i - 1), \quad (5)$$

$$Is(i) = y_e(i) - y_0(i). \quad (6)$$

Using these values, a shape of the segment can be classified as follows:

- If $|Id| \geq t_{hc}$ the shape is "Step" or "Transient". Else the shape is "Continuous".
- If the shape is "Continuous" and $|I| \leq t_{hs}$ then the shape is "Steady". Else it is "Increasing" or "Decreasing" depending on the sign of I .
- If the shape is not "Continuous" and if $Is < t_{hs}$ then the shape is "Positive Step" or "Negative Step" depending on the sign of Id . If $Is \geq t_{hs}$ and $\text{sign}(Id) = \text{sign}(Is)$ it is a "Increasing" or "Decreasing". If $Is \geq t_{hs}$ and $\text{sign}(Id) \neq \text{sign}(Is)$ it is a "Transient".

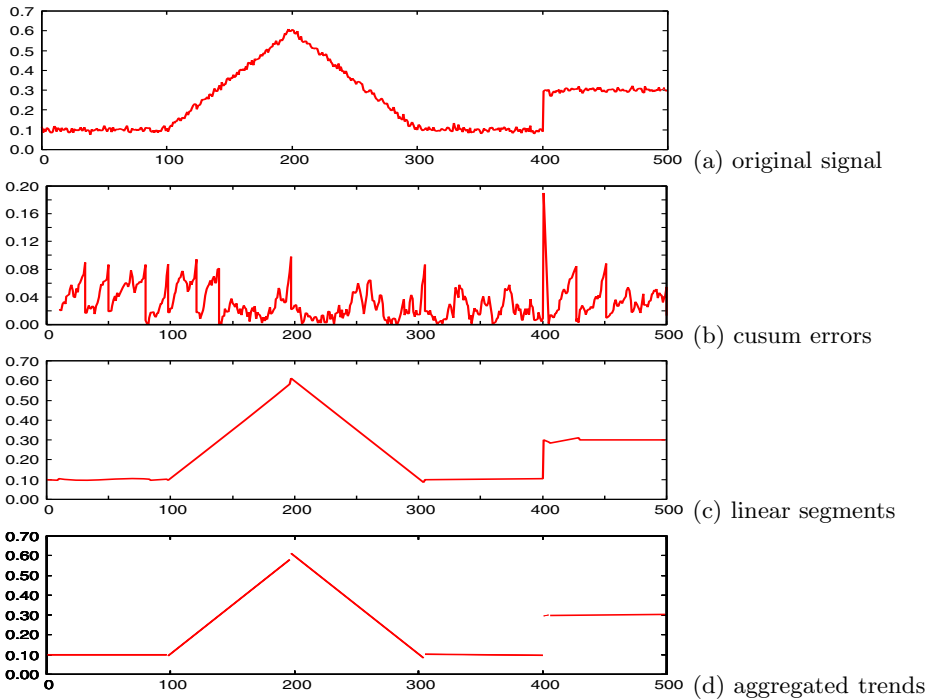


Fig. 1. Trend extraction process of a time series signal

After the classification of each segment, if possible, consecutive classes of the shapes are aggregated to abstract temporal patterns.

Figure 1 illustrate the methodology. The original signal includes Gaussian noise with 2% standard deviation of the range. Calculated cusums and extracted segments are shown in Figs. 1(b) and (c). In this example the threshold t_{h2} was 0.0897, which is the norm of the estimated residuals of the first 100 data. The value of t_{h1} was set to 3% of the t_{h2} . Each of the segment is classified into 6 classes based on the above algorithm, where the thresholds were set to 0.1 for t_{hc} and 0.04 for t_{hs} . Finally, the same consecutive classes are aggregated and obtain five segments of movement patterns as shown in Fig. 1(d).

3 Qualitative Movements

For the analysis of time series data, it is often useful to investigate the movement in a two variables plane. For example, graphical representation of the controller output and the valve position is known to be useful for diagnosis of industrial control valve [3]. In this section, qualitative representation of the movement in a two variables plane is considered and a new methodology to extract the qualitative movements is presented.

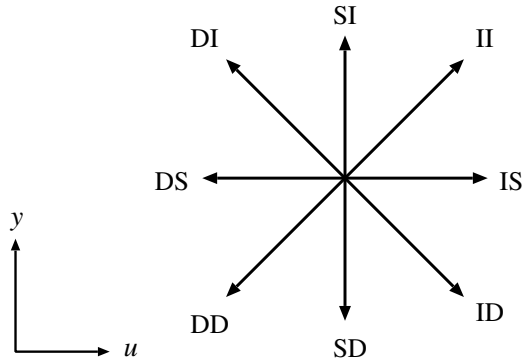


Fig. 2. Primitives for qualitative movement

3.1 Simple Approach

As shown in Fig. 2, nine primitives are used to represent qualitative movements in the plane. In this representation, ‘I’ stand for “Increase”, ‘D’ stand for “Decrease”, and ‘S’ stand for “Steady” for each variable. The movement SS is the center point, which represents no movement.

The simplest approach to represent qualitative movements is to use temporal differentiation and thresholding for individual variable on each sampling period [3]. Although the method is very simple and works fine for a signal without noise, it requires prefilters for noisy data. None of the denoising methods are perfect and sometimes amplify the noise, giving an unsatisfactory result through the differentiation process [4].

Another simple idea is to represent qualitative movements by combining two individual qualitative trends extracted by the method described in the previous section. Unfortunately, this method also sometimes fails to model the behavior as shown in the following example. Figure 3(a) shows an example of two signals in industrial plant data (dashed lines) [3]. Based on the method in previous section, these signals are transformed into qualitative trends (solid lines). Both of the signals seem to be modeled appropriately. Figure 3(b) shows qualitative movements in a two variables plane based on the qualitative trends of each signal. Although both of the individual signals modeled appropriately, the corresponding graphical representation in the two-variables plane includes some unexpected ragged lines. This awkwardness is mostly comes from that slant lines in the two variables plane often difficult to be expressed by the combination of quantized values of each variable.

3.2 Direct Approach

To overcome this problem, a direct approach to extract qualitative representation of movements in a two variables plane is considered in this subsection.

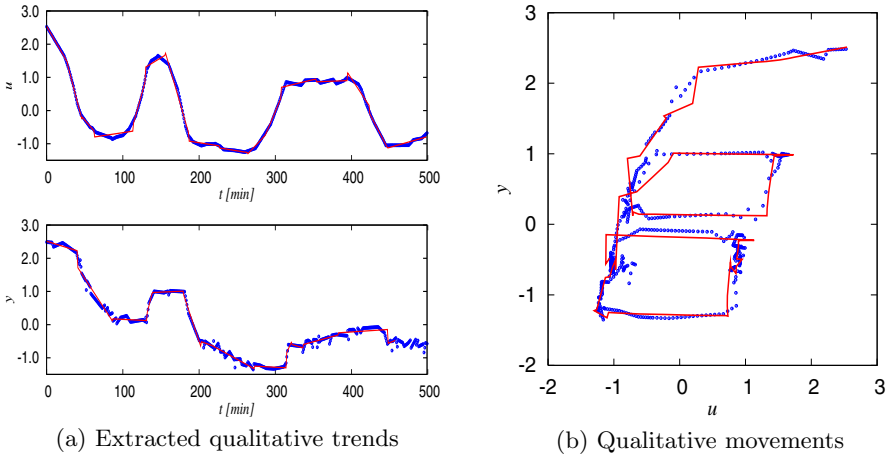


Fig. 3. Qualitative trends and movements (individual approach)

segmentation of movement

Consider a time series data of two variables $x(t)$ and $y(t)$ modeled as follows,

$$\hat{y}(t) = x(t)p_y + y_0, \tag{7}$$

where p_y is its slope and y_0 is the y ordinate at time t_0 . Parameter identification is based on the linear least-square method. The detection whether the linear approximation is acceptable or not is based on the cumulative sum (cusum).

$$\text{cusum}(t_1 + k\Delta t) = \text{cusum}(t_1 + (k - 1)\Delta t) + e(t_1 + k\Delta t), \tag{8}$$

where

$$e(t_1 + k\Delta t) = \hat{y}(t_1 + k\Delta t) - (t_1 + k\Delta t - t_{o1})p_{y01} - y_{o1}. \tag{9}$$

The absolute value of the cusum is compared to two thresholds $th1$ and $th2$ at each sampling time. If the absolute value of the cusum is smaller than $th1$, the linear model is acceptable. If the absolute value of the cusum is greater than or equal to $th1$, the signal value and corresponding time are stored. If the absolute value of the cusum is greater than $th2$, the linear model is no longer acceptable and a new linear model is calculated on the stored values during the period $th1 \leq |\text{cusum}(t_1 + k\Delta t)| < th2$. Once the new linear model has been calculated, the cusum is reset to zero.

classification and aggregation of segments

Let's consider two consecutive movements; present (i) and previous ($i - 1$). Let's define the following indices,

$$I_y(i) = y_e(i) - y_e(i - 1), \quad I_x(i) = x_e(i) - x_e(i - 1), \tag{10}$$

$$Id_y(i) = y_0(i) - y_e(i - 1), \quad Id_x(i) = x_0(i) - x_e(i - 1), \tag{11}$$

$$Is_y(i) = y_e(i) - y_0(i), \quad Is_x(i) = x_e(i) - x_0(i). \tag{12}$$

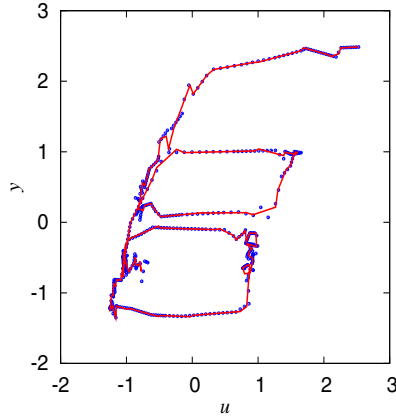


Fig. 4. Qualitative movements (direct approach)

Using these values, a movement pattern can be classified as follows:

- If $\max(|Id_x(i)|, |Id_y(i)|) \geq t_{hc}$ the movement is not “Continuous”. Else the movement is “Continuous”.
- If the movement is “Continuous” and $|I_x(i)| \leq t_{hs}$ then the movement is “SD” or “SS” or “SI” depending on the value of $I_{s_y}(i)$.
- If the movement is “Continuous” and $|I_y(i)| \leq t_{hs}$ then the movement is “DS” or “SS” or “IS” depending on the value of $I_{s_x}(i)$.
- If the movement is not “Continuous” insert a new continuous segment between i and $i - 1$.

After the classification of each segment, if needed, consecutive classes of the movements are aggregated to abstract movement patterns. If the movement is “Continuous” and if $|I_a(i)| < t_{ha}$ then the two movements i and $i - 1$ can be aggregated, where

$$I_a(i) = \frac{I_y(i)}{I_x(i)} - \frac{I_y(i-1)}{I_x(i-1)}. \tag{13}$$

Solid lines in Fig. 4 show extracted qualitative movements based on this method. In this calculation, thresholds were set to 0.03 for t_{h1} and 0.3 for t_{h2} , based on the norm in the preliminary analysis. Thresholds for the classification were set to 0.25 for both t_{hc} and t_{hs} . Aggregation was not used in this example.

The result shows good approximation of the original signal in the two variables plane and confirmed that this method gives better approximation than the method using qualitative trends of individual variables.

4 Application

The representation of qualitative movements in a two-variables plane can be applied to various decision making problems. In this section, an application

Table 1. Index values to detect valve stiction

	original data	noisy data
conventional method (finite difference)	0.69	0.32
proposed method (direct approach)	0.89	0.83

to valve diagnosis in industrial process plant is presented. Stiction of control valve often cause oscillations in controlled variables and the detection of stiction is highly demanded from industrial engineers. Among several methods for the diagnosis, graphical representation of the input and output signals of a valve is known to be useful for characterizing the behavior of a valve [3].

A data set shown in Fig. 3 represents controller output u and flow rate y of a control loop in an industrial plant [3]. The controller output corresponds to the valve input, and the flow rate corresponds to the valve output. Using the proposed extraction methodology of qualitative movements, the original signals are converted into qualitative movements as shown in Fig. 4. From this representation, one can easily acquire the information related to the diagnosis of valve.

When the valve position does not move against the changes of controller output the representation of the movement should be ‘IS’ or ‘DS’. So the first index for the diagnosis is defined based on the time duration of these two movement patterns:

$$\rho = (\tau_{IS} + \tau_{DS}) / \tau_{move}. \tag{14}$$

Where τ_{move} is time period during which the controller output moves. This index is expected to show larger value when the valve has the stiction problem.

The calculated value of the index is 0.89 as shown in Table 1. It is large enough to indicate stiction of the valve. For comparison, the value calculated by conventional method using the finite differences of each variable is also shown.

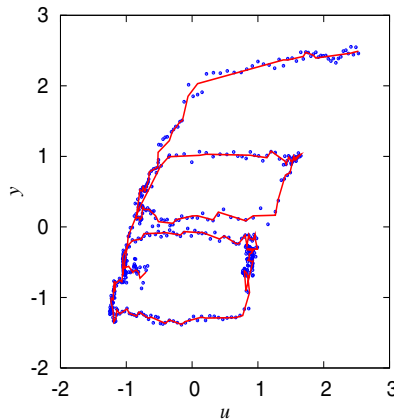


Fig. 5. Qualitative movements of noise added data

In the quantization of finite differences, standard deviations of the difference signals are used as thresholds. Although the value 0.69 is smaller than the value of proposed method, it is considered to indicate stiction for this case.

To validate the effect of noises in the signal, noisy signal was generated by adding 3% Gaussian noise to the flowrate. The generated noisy signal and its extracted qualitative movements are shown in Fig. 5. All the thresholds in the following calculation are the same to the values used in the calculation for the original data. The calculated value of the index for this data is 0.83, which is almost the same to the value for the original data. On the contrary, the calculated value using the finite differences of each variable becomes 0.32, which is too small to indicate stiction. This result shows that the proposed method is robust against the noise and provides stable calculation for this application.

5 Conclusion

A methodology to extract qualitative movement patterns of time series data in a two variables plane were presented. Nine primitives are used for the qualitative representation of a movement. The method contains the following three steps: segmentation of the movements of the two signals into line segments, classification of the segments into nine primitives, and aggregation of the primitives.

The usefulness of the method was demonstrated on an application to the diagnosis of valve stiction on industrial plant data. The result shows that the characteristic behavior of the signals is well represented by the extracted qualitative movements. The method can deal with noises without losing characteristic feature of the movements.

References

1. Dash, S., Maurya, M.R. and Venkatasubramanian, V.: A novel interval-halving framework for automated identification of process trends. *AIChE J.* **50** (2004) 149–162
2. Charbonnier, S., Garcia-Beltan, C., Cadet, C. and Gentil, S.: Trends extraction and analysis for complex system monitoring and decision support. *Eng. App. Artif. Intel.* **18** (2005) 21–36
3. Yamashita, Y.: An automatic method for detection of valve stiction in process control loops. *Cont. Eng. Pract.* **14** (2006) 503–510
4. Moussaoui, S., Brie, D. and Richard, A.: Regularization aspect in continuous-time model identification. *Automatica* **41** (2005) 197–208

An Expansion of Space Affordance by Sound Beams and Tactile Indicators

Taizo Miyachi¹, Jens J. Balvig¹, and Jun Moriyama²

¹Faculty of Electronic Engineering, Tokai University
1117 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan

²NISHINIPPON SYSTEM INSTALLATIONS AND CONSTRUCTION CO, LTD.
3-15-7 Kuhonji, Kumamoto, 862-0976, Japan
{miyachi, 4keem002}@keyaki.cc.u-tokai.ac.jp

Abstract. The information in mobile terminal should be connected to objects in an immediate space where a pedestrian walks. We propose a mobility assistance system by parametric speakers cooperating with the Braille blocks with embedded IC-tag so as to allow pedestrians to safely perform tasks in the artificial “space affordance.” Complicated structures in a town, perceptive information and cognitive information useful for various scenes in the town are also discussed.

1 Introduction

The Free Mobility Assistance (FMA) Project which embeds information in objects in the environment has been working to construct new environments of ubiquitous computing in Japan since 2004 [1]. A ubiquitous chip (IC tag) is embedded in tactile ground surface indicators called “Braille blocks” or “ubiquitous seals”, and the information in them can be acquired from a portable terminal, the Ubiquitous Communicator (UC), when the ubiquitous cane detects an IC tag in the block. The technique has been tested in Kobe, at Aichi Expo 2005, in Ueno Park and Ueno Zoo as free mobility support and sightseeing promotion. It is expected that the ubiquitous environment can provide lots of useful information in a town using the UC. However, it does not solve all problems met by pedestrians. For example, it is difficult for pedestrians to connect useful information provide by the UC with objects in the immediate space where he is.

An important concept in architecture of a city is the idea of “Imageability”, a way of structuring cities that allows pedestrians to effectively move about and enjoy the city [3]. Concerning safe movement of a vision impaired pedestrian, affordance [4] in city, and the perceptual and cognitive information [2] are proposed as the base in psychology. A training system of obstacle perception for the Blind was proposed [7].

In this paper we propose a mobility assistance system using parametric speakers together with IC-tag embedded Braille blocks. Pedestrians perform tasks in the artificial “space affordance” that is a kind of expansion of affordance in a town. We also examine how information in the UC can be supplied to pedestrians who have senses and reasoning capability from both physical and psychological point of views and

how this information in can be connected to objects in various scenes. The main features of the parametric speaker [5, 6] are also described.

2 Free Mobility Assistance System and Structure of a Town

Kevin Lynch [3], who had produced brilliant achievement in the field of image maps of a city and urban planning, analyzed the complicated structures of cities like Boston and how to enjoy the town. Movement to a destination can be made easy and can be enjoyed by using geographical features as a navigational aid. In Japan, as there are no street names, recognition of continuity is unclear and movement to the destination is difficult. FMA which is a Japanese national project has been developing a portable terminal: Ubiquitous Communicator (UC) for free mobility assistance in cities under the ubiquitous computing environment in Japan since 2004 [1].

Five major problems were discovered in the FMA Project experiment.

(1) **Speed of walking.** Pedestrians can not access information if he walks past the Braille block supplying the information. As he does not know where the block is he may not be able to go in the correct direction and can end up frustrated not knowing where to go.

(2) **Acquiring information requires stopping on the Braille block.** Since it is necessary to stop on a Braille block to hear the information, the pedestrian runs the risk of someone bumping in to him/her from behind.

(3) **Getting lost.** If a pedestrian does not hear the relevant information correctly it is very difficult for him to reconfirm the direction. He would have to imagine the route back and this can cause a great amount of stress.

(4) **Lack of connection between location of information and the pertaining object.** If the Braille block is separated from the entrance of a rest room, it can be very difficult to find the entrance of a toilet. As the information is purely audio-based, matching the information with the objects in real space is a difficult task. Moreover information automatically acquired from an information booth via infrared transmission does not have exact directivity anchored in the immediate space.

(5) **Technical errors.** Due to metal carried in a pocket (such as loose change or keys) the direction sensor would sometimes make mistakes. If the pedestrian is misinformed, there is a great risk of an accident occurring.

3 A Mobility Assistance System by Parametric Speakers and Braille Blocks

Mobility Assistance system should be useful in helping both sighted pedestrians and pedestrians with vision-impairment due to physical or mental reasons. It is important to understand that since the spatial information needed by vision impaired pedestrians is different than the spatial information used by sighted pedestrians the task performed by vision impaired pedestrians is different and more complex than that of sighted pedestrians. Vision impaired pedestrians can learn very little by observing buildings, trees, fences, and so forth. We propose a mobility assistance system using parametric speakers together with Braille blocks, as a part of a ubiquitous computing

environment, which provides perceptual and cognitive information about both immediate space and remote space and creates a space affordance in the spaces.

3.1 Perceptual Information and the Parametric Speaker System

Perceptual information is defined as contemporaneous information in an immediate space. It is information that is acquired directly from the space in which a task is being performed, and used while the task is in progress. Perceptual information causes a kind of intersection between a pattern of activities in human life and a pattern of changes of objects. Even if large amounts of useful information are available in the UC, if there is no clear physical correspondence with the object in the immediate space, moving pedestrians can not carry out their desired tasks. Two major subjects of mobility of vision impaired pedestrians are (1) discovery of the accurate position of target objects and direction to move, and (2) avoiding dangerous objects.

Parametric speaker systems (PS) supply pedestrians with the information of a target object by sound beam of ultrasonic wave with high directivity, which is anchored in the immediate space (Refer to Fig.1). PS can connect the information in UC to real objects in the immediate space with the following features.

- (a) Pedestrians can hear the sound only in the central part of a sound beam because of its high directivity. The sound of PS does not turn into noise outside the sound beam. There is no sound heard during transmission because of an ultrasonic wave.
- (b) Pedestrians can know a direction of sound beam and can find out the exact position of a target object by hearing the explanation from the sound beam (Refer to Fig.2).
- (c) The sound is heard from the reflected point.
- (d) The sound is reproduced close to the ears and the volume of it is low.

Useful functions and situations are described below (marked "FS").

FS1. Parametric speaker systems (PS) supply a walking pedestrian with the accurate position and direction of a telephone box. He risks colliding with the large transparent upper part of the box before noticing the legs of the box using his cane (Fig.2). Examples of other objects: a pole, signboard, flower bed box, pond.

FS2. PS supplies pedestrians in a dangerous situation with warnings via sound beams in order to inform about dangerous objects such as the gap between a train and the track.

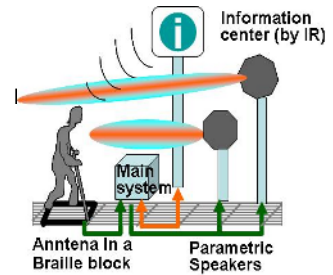


Fig. 1. A parametric speaker system cooperating with Braille blocks with IC-tags



Fig. 2. A telephone box is connected with the information in UC

3.2 Orientation for Movements Without Getting Lost

A pedestrian sometimes gets lost or walks the wrong way. A parametric speaker system used together with Braille blocks supplies pedestrians with accurate information on position, direction, and actions via a voice beam anchored in the immediate space. PS is mainly classified into (A) Direct type, (B) Long & Short range type, (C) Middle range & Reflected type. Since the sound beam of PS does not easily turn into noise,

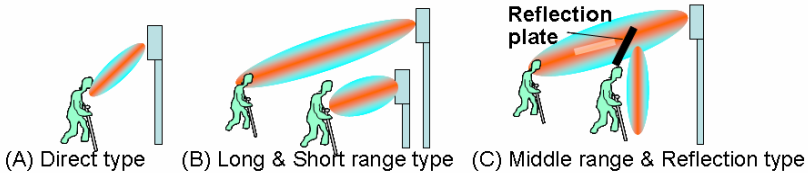


Fig. 3. Three types of parametric speaker systems

the PS can transmit a long explanation to the pedestrian. Long range type and middle range type of parametric speakers supply pedestrians with directions, predictions and longer explanations. Pedestrians can also perform actions more effectively by obtaining information concerning not only each object but the relation of objects and spatial relation in the immediate space.

FS3. Directive voice beam guides pedestrians in the right direction of elevators and escalators at a station or inside a department store, ticket office or restroom etc. Complicated crossings like five-forked roads or rotaries can also be assisted by the spotlight.

FS4. Directive voice spotlight above a target object informs of the accurate position and gives a detailed explanation of the item. Multiple voice spotlights supply pedestrians with explanations in multiple languages (separate from adjacent speakers).

FS5. Getting lost. A pedestrian sometimes changes suddenly direction when walking along a curved road. In the case that cautions will be taken by passing of large-sized vehicles in a big curve a pedestrian can not also recognize the exact change of direction. The pedestrian may lose his way if he changes direction at the exit of a curved way. A voice beam tells pedestrians the accurate direction anchored in the immediate space and how to safely walk along the curved way.

3.3 Safe Areas and Affordance in a Space of Town

By understanding geographical features of a city it is easy for a pedestrian to recognize the direction of movement, and the attainment to the destination becomes easy [3]. The affordance which Gibson [4] says is the view that the environment side has given us the message. To get where pedestrians want to go, they move through spaces that humans have furnished with familiar contents. In the constructed spaces through which pedestrians usually move, there are affordances. That is to say, movement along some courses is easy, whereas movement along other courses is difficult or impossible [2]. We call this affordance “space affordance” in a town.

The affordance usually forms a combination of straight lines in cities. Because (1) pedestrians want to move the shortest path, (2) he does not want to be disturbed his smooth moving by the law of inertia, (3) he does not want to lose direction to walk. Voice beams are suitable to this form of the affordance to move. We propose a multiple PS system which offers previous notice concerning a safe course and a safe area from the place distant for a while by voice beam of PS and offers detailed information in near place. The combination of sound beams organize affordance in a space and the combination of voice beams make embedded affordance in a space obvious.

The concept of safety path and safety area is very convenient for pedestrians to keep safety while moving. A city consists of three kinds of area: (a) safety area, (b) watch area and (c) dangerous area. An Area includes (i) safety paths, (ii) watch paths, and (iii) dangerous paths. Voice beams of PS guide a pedestrian to keep walking in a safety paths in a watch area with an artificial affordance by the sound beams. Although only the affordance in a place currently seen can be known by vision, the affordance in the whole space of 360 degrees is known by sound beams.

Safety area is so useful that pedestrians are made free from Braille blocks. Benefits of safety paths and safety area are (1) – (5).

- (1) In safe areas, free movement unrestrained by Braille blocks is allowed.
- (2) Should a pedestrian enter a dangerous area, he can immediately escape to an adjacent safe area.
- (3) Guidance at the exact place where a pedestrian changes direction or has to cross a road.
- (4) Accidents such as a fall from the platform of a station in a watch area can be easily noticed.
- (5) Countermeasures to cope with difficulties using safe areas are clarified and explained by the PS voice beam.



Fig. 4. A safe path between two rows of bicycles which makes Braille block not to work

FS6. Space affordance by sound beams of PS and Braille blocks.

■ **No sudden stop and no waiting on the Braille block:** A combination of voice beams and information at a point of the Braille block makes pedestrian's walking smooth without a sudden stop and without having to stay on the block.

■ **Affordance outside:** a voice beam is useful for navigation within 70 meters.

■ **Affordance underground:** a sound beam two meters wide transmits a guide more than 50 meters away in an underground passage. A subway station in an underground passage can thus easily be found.

■ **Affordance in an unobservable area in underground passages:** a pedestrian discovers the elevator of the target building in an underground passage. The buildings can not be seen in the underground and the entrance of elevator usually hides behind stairs and pillars in underground passages.

Since various kinds of space affordance is automatically heard and shared by everybody we expects that residents find a new meaning in good points and detect bad points in the town. Accumulation of small actions for the improvement will make the town better and create a good new life style in the town.

3.4 Perceptual Information, Cognitive Information and Structure of a City

A pedestrian acquires and uses useful perceptual information in immediate space. On the other hand, in moving to an invisible space and a remote space, a pedestrian uses the cognitive information. Cognitive information is obtained from memory and also includes information established by inference (e.g., relationships among remembered spatial features that were not observed on the same occasion), generalizations made by the reiterations that characterize constructed spaces, and communicated information received by way of spoken or written language [2]. Cognitive information is not individual perceptual information but the common information on common recognition of two or more pedestrians. It can be matched with the object in the immediate space and should be translated into useful tasks.

Since a vision impaired person can perceive only the restricted information, he acts safely making full use of the memorized experiences and the cognitive information gathered beforehand. However, it is not easy for pedestrians to remember cognitive information which is once abstracted using symbols and the relationships, and useful tasks concerning an invisible space and a remote space. Pedestrians may sometimes misunderstand such information. Voice spotlights of PS supplies pedestrians with the cognitive information and the tasks both in remote place and in an information center near the destination without noise disturbance. The perceptual information and the tasks, which are translated from the cognitive information, can be also supplied for the pedestrians again in the immediate space by voice beams of PS, as supplementary information.

3.5 Comfortable Space Without Uneasiness or a Feeling of Fear

Navigation which is a sequence of guides is useful for pedestrians in the invisible remote space and the unfamiliar space in order to meet neither with uneasiness nor a feeling of fear.

Time margin for inference and the countermeasures to cope with difficulties in order to get ready for upcoming situations can be prepared by predicting an image or motions of dangerous objects. Typical patterns of the movement, the stereotype of dangerous objects, and spatial relationship between the objects, which can be heard by vision impaired pedestrians, are explained by voice beam of PS.

FS7. Preliminary announcement to sudden dangerous objects

Even a sighted person may be surprised at sudden danger. Pedestrian does not easily notice a car leaving the exit of basement car parking which locates near the pedestrian crossing. Because, the beep sound of leaving the garage is scratched out by the sign sound of a pedestrian crossing. A directive voice beam by a parametric speaker is effective in such a scene.

FS8. Information offer to a risk of changing in time

Depending on time zone of people, danger occurs or danger changes with time. Since it is not decided when danger will occur, it needs to be information provided for showing and avoiding a dangerous place with a sound beam from a walk environment side. Natural conditions, such as the weather, may be related. The dangerous things are collection garbage in the morning, overflowing illegally parked bicycle, riverbank ways at the time of a typhoon, a way beside a building from which a snowy lump may fall, and a depressed ground which tends to be covered with a gas of sulfide.

FS9. Information offer to the quantitative limit of danger

Dangerous quantity may increase, so that passing is impossible. It is warned that a passerby does not stray by a parametric speaker. Immediate refuge is guided to the pedestrians who strayed. There is a road which becomes such many traffic that residents cannot pass, either for traffic congestion evasion. Since the dangerous section of a school zone is short even if more far, there are some children who change an elementary school to another school.

FS10. The obstacle in a safe node in a safe area

A node which is a rest open space is close to a crossing near a station and a bench, a telephone booth, a flower bed etc. are centering on roadside trees. If a pedestrian follows the Braille blocks, it will collide with the obstacle of the node in spite of safe area. Danger is latent in a visually impaired person also in safe area. The scene which feels fear by getting to know the feature of an open space beforehand is avoidable. A parametric speaker system supplies the person with such information by directed sound beam.

3.6 Extension of Space Affordance by Bi-directional Communication

The course which many pedestrians often use to go to the destination is generally in a safe path. Navigation on this course is reasonable for efficient passing. However, even if easy for those who already know well, there are some crossings, a shopping street, a slope, a curved way, etc. in a course, and that it is not always easy for visitors to find the way. The visitors need to avoid the obstacle of building construction and that of road repairing in the end of a fiscal year.

If the destination is recorded in the RFID card, as a pedestrian can acquire not only the direction to proceed but also the next intermediate destination at each corner.

Information concerning sequence of directions at the corners and directions of the destination, a sequence of intermediate target points, obstacles in real time, a way of avoiding them by voice beams of PS, he can effectively walk in safe paths.

By sharing the information of destination by two way communication between a pedestrian and an information spot, the pedestrian walks in safe considering the course situation at present based on the information by the voice beams.

3.7 Broadcasting Emergency Information to the Whole Underground Passages

Since a PS with strong reflective sound can broadcast emergency information to the whole underground passages without too much echo, this broadcasting is very useful for the emergency time (Refer to Fig.5). The broadcasting is also useful for a guidance of a public office in an underground passage with little traffic. The narrow underground passage which has little traffics after passing through noisy shopping quarter is a closed space, so a pedestrian becomes somewhat uneasy.

4 Experience of Basic Functions

Parametric speaker systems are newly productized within a few years by several makers. They are mutually quite different in features, materials, architecture, those are

trade secrets still. We examined main features of two maker's products and describe some results of them in section four.

A voice beam of PS with high directivity has less than 2 meter width is transmitted to the point 55 meters away in a under ground path. Setting the voice beam at a position of 3 meters high on the ground transmitting 70 meters away allows it to be clearly heard. A voice by sound beam at the point of 1.5 meter high was transmitted to the point only 15 meters away. The volume of PS sound ranges between 55 and 63 dB by sound level meter RION NL-26. The power of PS is expected to be strengthened up to about 80 dB. Pedestrians with practiced could recognize the sound of the PS better than beginners. Vision impaired pedestrians might be able to recognize better the sound than sighted pedestrians. Pedestrians could separately hear both direct sound and the reflected sound coming from a different direction. More than ten combinations of PSs were examined both outside on the ground and underground.

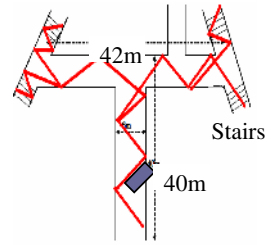


Fig. 5. A broadcasting of sound by reflective sound beam

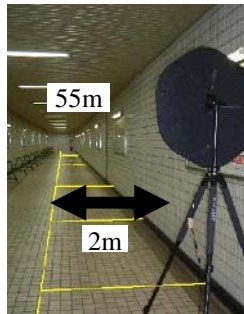


Fig. 6. A voice beam with a width of 2m

References

1. Ministry of Land, Infrastructure and Transport, <http://www.jiritsu-project.jp/> (2006).
2. Emerson Foulke: The Roles of Perception and cognition in Controlling the Mobility Task, International Symposium on Orientation and Mobility, Trondheim, Norway, (1996)
3. Kevin Lynch: The Image of the City, MIT press, (1960).
4. Gibson J.J.: The senses considered as a perceptual system, Boston, Houghton, Mifflin (1966).
5. Noboru KYOUNO: Technology Trends on Parametric Loudspeaker, JSME (2004-7).
6. Shinichi SAKAI, et al.: Effect of the Directivity of a Loudspeaker on the Walk of the Visually Handicapped, TECHNICAL REPORT OF IEICE, EA2004-6, (2004-9).
7. Yoshikazu SEKI, Kiyohide ITO, Study on Acoustical Training System of Obstacle Perception for the Blind, Assistive Technology Research Series 11, Assistive Technology - Shaping the Future (Proceedings of AAATE Dublin 2003), 461-465 (2003).

Automatic Discovery of Basic Motion Classification Rules

Satoshi Hori¹, Mizuho Sasaki¹, and Hirokazu Taki²

¹ Monotsukuri Institute of Technologists
333 Maeya, Gyoda 361-0038 Japan
hori@iot.ac.jp
<http://www.iot.ac.jp/manu/HORI.html>
² Wakayama University
930 Sakaedani, Wakayama 640-8510 Japan

Abstract. There is a keen demand for a method of sharing better work practices in a factory because better work practices are the key to improving productivity. We have developed a system that can measure a worker's motion and automatically generate a manual that describes his movements. This system employs motion study as used in Industrial Engineering to identify the important steps in a job, and it has proven to be effective especially in the fields of factory machine operation and maintenance. However, work procedures often include unique basic motions. The determination of basic motions and the creation of an algorithm that can classify these basic motions are time consuming and complex tasks. Therefore we have employed the C4.5 algorithm to discover rules that classify the basic motions. Experimental results prove that our method can successfully discover rules for various work procedures.

1 Introduction

Firstly we outline why better work practices need to be recorded and shared. The objective of automatic discovery of basic motion classification rules is then described after a brief introduction to our manual generation system.

1.1 Background: Better Work Practices Need to Be Shared

Better work practices contribute greatly to higher productivity in factories. Skilled workers have better work practices because they have learned through experience in the workplace. For example, a service technician or worker moves his arms and legs while he checks and repairs machinery. He walks, pushes switches, and looks at meters. The behavior of an experienced technician is more efficient than that of a novice, and we can learn a lot by observing the work practices of a well-trained and experienced worker. Efficient work processes need to be documented in manual form and shared with others. Good operation manuals improve work productivity. There is a keen demand for manual generation systems because documenting work processes is a time-consuming task. This is the motivation of our research.

1.2 Objectives: To Automatically Discover Basic Motions

We have developed a system that records a worker's actions in order to generate a manual. Our system employs motion study [1] used in Industrial Engineering (I.E.) to

identify important events in a work procedure. Why is the identification of important events required? An item of work may take from 30 minutes up to several hours to complete. Hence, it is necessary to pick up several important events in the work procedure in order to teach novices the key steps in the work.

Motion study shows that a work process is a series of basic motions. Most work is done using the two hands, and all manual work consists of relatively few fundamental motions which are performed over and over again. "Get" or "pick-up" and "place" or "put-down" are two of the most frequently used groups of motions in manual-assembly work.

Motion study has a history going back almost one hundred years, and it is still used in many production lines. However, basic motion recognition still depends on human I.E. engineers. On top of this, a work procedure often includes unique basic motions. It is a time-consuming and complex task to determine basic motions and create an algorithm that can classify the basic motions. Therefore we have employed the C4.5 algorithm to discover rules that classify the basic motions. Some experimental results prove our method can successfully discover the rules for various work procedures.

2 Skill Intelligence and Motion Study

Experienced workers' skill is the key to improved productivity in a manufacturing facility, and a method for sharing skill intelligence has been in demand. This section defines skill intelligence and briefly introduces motion study. Our manual generation system is also explained.

2.1 Definition from the Hierarchical Paradigm of Robotic Control

Knowledge and/or intelligence can be found not only in the form of language but also in the behaviors and skills of living creatures. With a motion capture system, ubiquitous sensors, and other electronic devices, we can electronically observe and record human motions. The skill intelligence elicited from such data will be useful if it is shared among novice workers.

The term 'skill' includes a wide range of meanings. When a very fast ball is pitched or a violin is played, skill is required. In this case, 'skill' involves the control of muscles. On a factory floor, manufacturing machinery has to be appropriately operated and maintained. In this case, 'skill' includes knowledge and experience.

Robotics provides a good domain for recognizing skills on a factory floor as compared to, for example, the skill required to play a violin. The hierarchical paradigm [2] is one method for organizing intelligence in mainstream robotics. This paradigm, as shown in Figure 1, is defined by the relationship between the three primitives (SENSE, PLAN, ACT). Considering a walking robot, first the robot senses the world, the robot plans all the directives needed to reach a goal, and the robot acts to carry out the first directive.

Human skills exist in Sense, Plan, and Act modules. Some experienced workers can sense $1\mu\text{m}$ noise on a steel sheet. On the other hand, the skill intelligence that

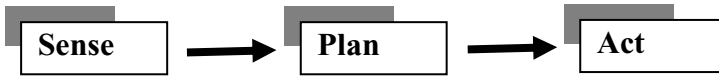


Fig. 1. Hierarchical Paradigm of Robot Controller

we will be storing and sharing is 'skill' in the Plan module. An experienced worker can make an appropriate plan for fabrication and repair because he has a lot of background knowledge and experience. This task can be realized as a planning and classification task in artificial intelligence. This skill intelligence is rarely documented in manual form. Therefore a system that can observe worker's actions and generate a manual automatically is required.

2.2 Industrial Engineering (I.E.) Motion Study

Industrial Engineering (I.E.) provides a strategy for finding a preferred method of performing work, which is referred to as Motion Study [1]. Frank and Lillian Gilbreth did pioneering research in this area and established the field of Motion Study in the beginning of the 20th century. Gilbreth noticed that most work is done with two hands, and all manual work consists of relatively few fundamental motions that are performed over and over again. "Get" or "pick up", and "place" or "put down" are two of the most frequently used groups of motions in a production line. Gilbreth developed subdivisions or events, which he considered common to all kinds of manual work. He coined the word *therblig* to refer to any of the seventeen elementary subdivisions.

- Motion sensing, motion labeling, and work recognition: an I.E. engineer observes what a worker is doing on a production line. He understands the worker's movements and writes them down as a series of basic motions using *therblig* symbols. He then establishes how the work is done. Table 1 describes an example.
- Use of observed motion: The record of basic motions is analyzed and used for designing a better work procedure.

The I.E. Motion Study is expensive because it manually analyzes the worker's movement.

Table 1. Example of Motion Study: Task of "signing a paper"

<i>Therblig</i>	Description of Motion
1. TE: Transport Empty	Reach for pen.
2. G: Grasp	Take hold of pen.
3. TL: Transport Loaded	Carry pen to paper.
4. P: Position	Position pen on paper for writing.
5. U: Use	Sign the paper
6. TL: Transport Loaded	Return pen.

2.3 Proposed System

Figure 2 depicts the outline of our proposed system. The system consists of two sub-systems: (1) a wearable sensor system that observes the body movement and vision range of a worker, (2) a manual-generator that generates a manual from the motion data recorded by the wearable sensor system. We employed IC accelerometers to measure a worker's motion. A camera is used to record what the worker sees while doing a job.

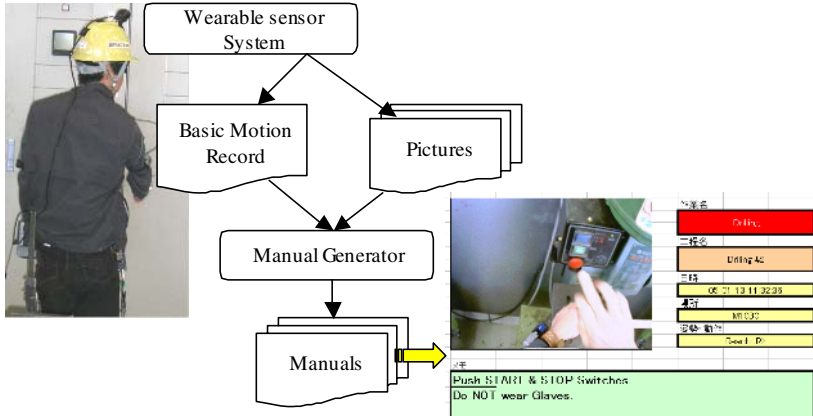


Fig. 2. Our Manual-Generation System

2.3.1 Motion Observation with IC Accelerometers

We employed IC accelerometers to measure motion [2]. The sensors are attached to the arms and legs of a service technician. This section describes IC accelerometers and their signal processing. Analog Device's ADXL202 is employed as an accelerometer. The ADXL202 is a low-cost, low-power, complete 2-axis accelerometer sensor with a measurement range of ± 2 g. The sensors attached to a right arm are shown in Figure 3. Figure 4 demonstrates how "reach" motion is detected intuitively from the output of a wrist sensor. "Reach" motion often occurs when a technician operates switches. The right hand is first extended close to the right leg. In the "reach" motion, the hand is raised and lowered.

To detect a basic motion, the output signal is processed as follows:

- Step-1. Measure raw X, Y acceleration data with a 100 msec sampling rate.
- Step-2. Eliminate noise by taking the moving average of each of five samples.
- Step-3. Data labeling: Classify each data item into 3 bins labeled -1, 0, and 1.

This data set is described as a gravity pattern.

- Step-4. Motion labeling: Prepare a table that maps a gravity pattern change to a basic motion. As an example, when an arm's gravity pattern changes from 0101 to 1010, we determine that "reach" motion is occurring (see Figure 4).

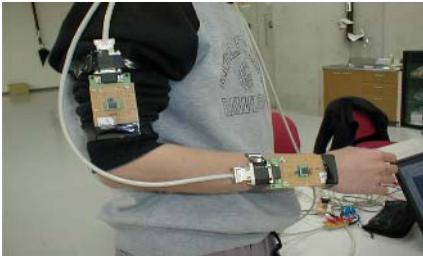


Fig. 3. Accelerometers on Right Arm

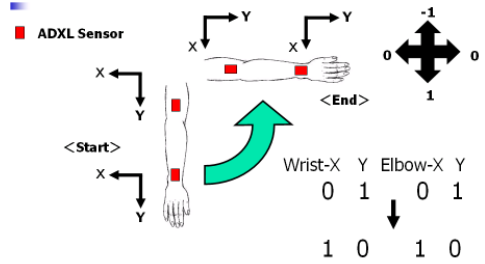


Fig. 4. Wrist Accelerometer's Output for Reach Motion

3 Basic Motion Discovery

A basic motion is a fundamental motion that occurs over and over again while performing a job. When a basic motion is observed, we recognize it as an important event worth recording. Our method can generate a set of rules that discover basic motions. The rules are generated from a set of supervised learning data with the C4.5 algorithm [3].

3.1 C4.5 Algorithm

The algorithm C4.5 uses top-down induction of decision trees. Given a set of classified examples, a decision tree is induced, biased by the information gain measure, which heuristically leads to small decision trees. These trees can be easily transformed into a set of production rules. Learning data, a set of examples, are given in attribute-value representation. The set of possible classes is finite.

Table 2 shows an example of classes. Part of the learning data is shown in Table 3. The goal of this example is to discover rules that decide whether people play golf, if weather outlook, temperature, etc., are given. Figure 5 lists the generated rules.

Table 2. Attributes and Values for Golf Problem

Attributes	Value
Outlook	sunny, overcast, rain
Temperature	cool, mild, hot
Humidity	high, normal
Windy	true, false

Table 3. Supervised Learning Data for Golf Problem

No.	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	false	no play
2	sunny	hot	high	true	no play
3	overcast	hot	high	false	play
:	:	:	:	:	:
14	rain	mild	high	true	no play

Rule 1:	IF Outlook = overcast	THEN play golf.
Rule 2:	IF Outlook = sunny AND Humidity = high	THEN do not play golf
:	:	:

Fig. 5. Generated Rules

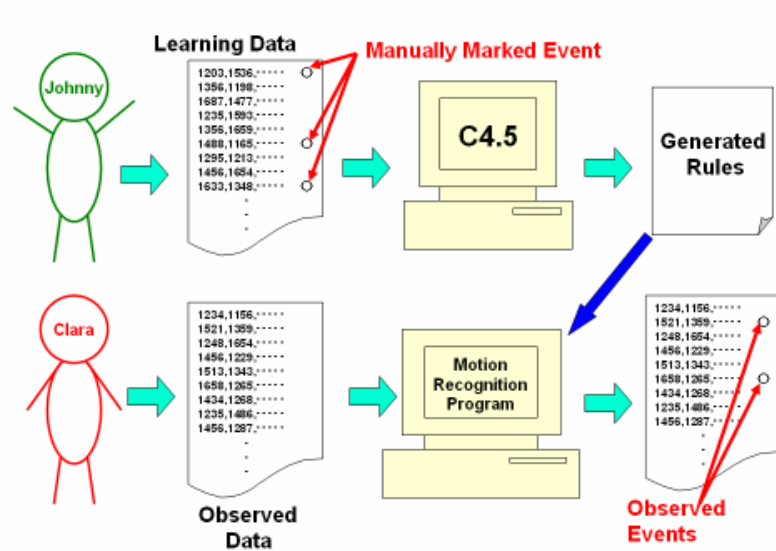


Fig. 6. Outline of Basic Motion Discovery

3.2 Outline of Classification Rule Discovery

The framework of rule discovery and basic motion detection is described in Figure 6. This procedure consists of two major parts, (1) Rule discovery, and (2) Basic motion classification using the rules discovered.

(1) Rule discovery

Step-1. Prepare supervised learning data.

Human motion data, Sensor outputs, are recorded while Johnny performs a certain target job. Important events are manually marked. Thus we obtain the supervised learning data.

Example: A mouse button is pushed when Johnny throws a dart, so this throwing event is marked on the data file.

Step-2. Prepare the data file.

Details of this preparation are described in the next section.

Step-3. Feed this data file to the C4.5 program. C4.5 generates the classification rules.

Example of generated rules:

IF RWY < -90[deg] THEN dart throw occurred.

(2) Basic motion classification

Step-1. The rules discovered above are implemented in a recognition program.

Step-2. Clara's motion is recorded while Clara performs the same job.

Step-3. The classification program determines what basic motion occurred using the rules.

Example: Events in which Clara threw a dart are marked in the output file of the recognition program.

3.3 Learning Data Preparation

This section describes how to prepare learning data from the observed motion data. We take dart-throwing as an example. In the case of dart-throwing, we want to identify events involved in throwing a dart.

The learning data is depicted in Figure 7. Two sensors are attached, one to the wrist and the other to the elbow of the right arm. Observed data consist of the voltage output by the sensor. RWX denotes the voltage of the Right arm's Wrist X-axis output. REY denotes the voltage of the Right arm's Elbow Y-axis output.

The sensor data are sampled with a 10 Hz sampling frequency. Each record in the observed data represents a right arm pose carried out every 0.1 second. The dart-thrower pushes a button when he throws a dart, and the "Throw" label is added to a record, as shown in Figure 7(B).

The records in a 0.5-second interval are assembled to give a record of the learning data, as shown in Figure 7(C). One record in Figure 7(C) denotes a move of 0.5 seconds. C4.5 processes this supervised learning data and generates rules that distinguish the dart-throwing and other movements.

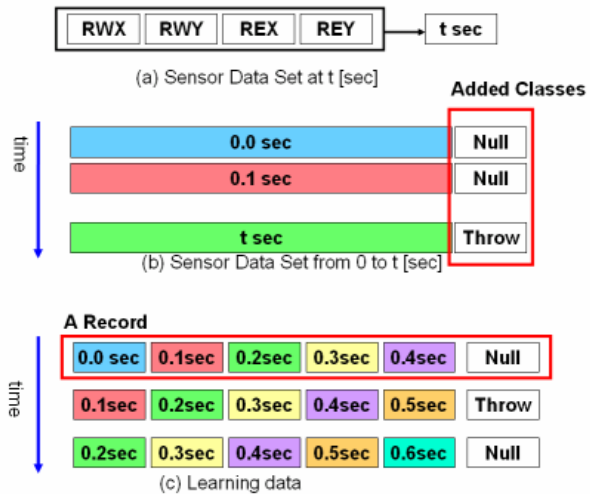
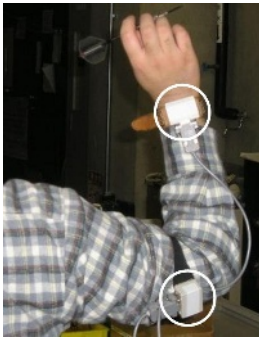


Fig. 2. Dart Throwing and Learning Data

3.4 Experimental Results

We conducted experiments for two types of motion, dart-throwing and confirmation-by-pointing, in order to evaluate how effectively the generated rules identify important events in the stream of motion.

- Dart-throwing

Firstly, we generated rules for the dart-throwing task. 10 dart-throwing motions were recorded as supervised-learning data. C4.5 then generated the following rules that recognize dart-throwing motion as distinct from other motion.

Rule-1: IF RWY < -90[deg] THEN dart-throw occurred.

**Rule-2: IF RWY at 0.2[sec] > 90[deg] AND RWY at 0.4 [sec] < 60
THEN dart-throw occurred.**

The rules successfully classified all 55 dart-throwing events. No false classifications occurred. 23 false accepts did occur but these were the motions of retrieving darts from a target. The reason the program misclassified these motions is because this retrieving-darts motion was not included in the learning data.

Table 4. Dart-throwing

No. of Events	Correct	FA	FR
55	55	23	0

FA: False Accept, FR: False Reject.

- Confirmation-by-pointing

Secondly, Confirmation-by-pointing was tested. Assumed work is as follows: A participant in the experiment counts red boxes in a warehouse. He ensures that all red boxes are counted by pointing at them one by one. His right arm moves in a manner similar to the "Reach" motion when he performs this confirmation-by-pointing motion.

We observed ten confirmation-by-pointing motions performed by the person. C4.5 generated a rule as follows:

**Rule-1: IF REY at 0.2[sec] < 70[deg] AND RWY at 0.2 [sec] > 40[deg]
AND REY at 0.4[sec] < 70 [deg]**

THEN Confirmation-by-pointing occurred.

Using this rule, the recognition program classifies the confirmation-by-pointing motions performed by another person. Table 5 shows the result. The participant performed the confirmation-by-pointing eight times. Seven out of the eight motions were successfully classified. One motion was missed.

Table 5. Confirmation-by-pointing

No. of Events	Correct	FA	FR
8	7	0	1

FA: False Accept, FR: False Reject.

4 Conclusion

There is a keen demand for a method of sharing better work practices in a factory because better work practices are the key to improving productivity. To analyze work practices, we need to recognize and record a human worker's motion. Motion study used in Industrial Engineering provides a good method for analyzing and designing better work practices. However we also need to develop algorithms to classify basic motions for the motion study. In this paper, we employed the C4.5 algorithm to automatically generate classification rules for basic motions. The experimental results of section 3.4 prove the effectiveness of our method.

References

- [1] Barnes, "Motion and Time Study", John Wiley & Sons, Inc., (1968).
- [2] Robin R. Murphy, "Introduction to AI Robotics", MIT Press, (2000).
- [3] J.R. Quinlan, "C4.5 an Induction System", Academic Press ().
- [4] J. R. Quinlan: "Decision Trees and Decision making", IEEE Trans. on Systems, Man and Cybernetics, Vol.20,No.2, pp.339-346 (1990).
- [5] S.Hori, K.Hirose, H.Taki, "Acquiring After-Sales Knowledge from Human Motions", KES2004, LNAI3214, pp.188-194 (2004).
- [6] T. Nakata, "Automatic Generation of Expressive Body Movement Based on Cohen-Kestenberg Lifelike Motion Stereotypes," Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.7 No.2, pp.124-129, (2003).
- [7] Song, Goncalves, Perona, : "Unsupervised learning of human motion", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.25, Iss.7, pp.814-827 (2003).

An Interpretation Method for Classification Trees in Bio-data Mining

Shigeki Kozakura¹, Hisashi Ogawa¹, Hirokazu Miura², Noriyuki Matsuda²,
Hirokazu Taki², Satoshi Hori³, and Norihiro Abe⁴

¹ Graduate School of Wakayama University, 930 Sakae-dani, Wakayama, Japan

² Faculty of Systems Engineering, Wakayama University,
930 Sakae-dani, Wakayama, Japan

`is-staff@sys.wakayama-u.ac.jp`

³ Monotukuri Institute of Technologist

⁴ Kyusyu Institute of Technology

Abstract. This research describes the analysis of a decision tree to interactively make rules from data that includes noise. A decision tree is often used in data mining technology because classification rules generated by the decision tree form new knowledge for the domain. However, it is difficult for a non-specialist user of data analysis to discover knowledge even if the decision tree is presented to the user. Moreover, the target data for mining may have both discrete values and continuous values, and these may include a lot of noise and exceptions. To understand the rules, the user needs the intermediate results of mining. Our system has interactive functions to realize interactive mining, and in this research, we propose a data mining technique with an interpretation function so that the user can understand analyzed data from which a required rule is derived.

Keywords: data mining, decision tree, mining corresponding to noise, continuous value mining.

1 Introduction

Data mining technology has recently been used to discover new rules in large amounts of data. The a priori algorithm[1] is a primary algorithm to find rules using certainty factors and support factors, calculated by statistical analysis. Many applications have been developed in various fields, such as finance, manufacturing, communication, chemistry, medicine and bio-technology. These data have various features and both discrete data and continuous ones can be treated. When the data mining system analyzes discrete data, it is easy to make logical rules. However when the set of data includes continuous values, pre-processing is needed to make discrete data from continuous values. There are many methods for separating continuous into discrete values, according to time intervals. The results of the data mining are dependent on the separation methods. Therefore, the system must have translation techniques for complex relation discovery. If the set of data has mixed data which include both discrete and continuous values, the number of rule candidates will be huge. The system must use background knowledge about the application field to remove unnecessary relations. Moreover, if a lot of noise is present in the data, the system gives incorrect

mining results. Normally, the system uses statistics to find important rules in a huge data set, but if the size of the data set is not large, the system cannot use statistics efficiently. Our target data set includes discrete and continuous data with noise and its size is not large. The data mining system discovers rules by analyzing the features and relations of each datum in the data set, and the probability of co-occurrence between data is used. In order to obtain correct rules, the system must have noise removal methods and interval making methods which generate discrete values from continuous values. We consider the decision tree making method, C4.5 [3] for mining noisy and small size data, using knowledge about the application field. To employ knowledge about genome analysis data which are used to find relations between genes and phenotypic expressions, we are developing a user interface so as to interactively obtain biological knowledge in data mining processes.

2 Problems in Mining Data

Data that are analyzed in data mining are not always accurate. Normally, they contain several kinds of noise, in addition to exceptions. There are two types of data, discrete data and continuous data. The continuous data cannot be treated in logical form, and in data mining, they have to be translated into discrete data according to values in separate time intervals. In our research, the mining system treats bio data which contain both discrete and continuous values: for example, the shape of botanic seeds is round or square, which are discrete values, while the height of a botanic body is expressed in continuous values.

2.1 Transposon m-ping

The data mining system has to analyze the relations between gene and phenotypic expression. The system needs data about many different types of genes. A mutation results in a new gene, and artificial mutation occurs by transposition. A sequence of DNA, known as a Transposon, or "Moving gene", can freely move on DNA, and can cause the mutation. In this research, we deal with plants in which there is a transposon "m-ping" [4]. The m-ping is a kind of transposon, and it actively moves, and frequently causes mutation.

2.2 Transposon Analysis

We explain the analysis method referring to Fig. 1, which consists of the following steps:

1. Extracting the gene from a plant.
2. Generating fragments using a digestive enzyme.
We know that the gene is fragmented according to a constant rule, and if an array of the same gene is obtained, the result of the fragmentation is also the same.
3. Electrophoresis
This involves sorting the fragmented gene by length.
4. Graphing
The horizontal axis of the graph represents length, and the vertical axis is the amount of the fragmented gene.

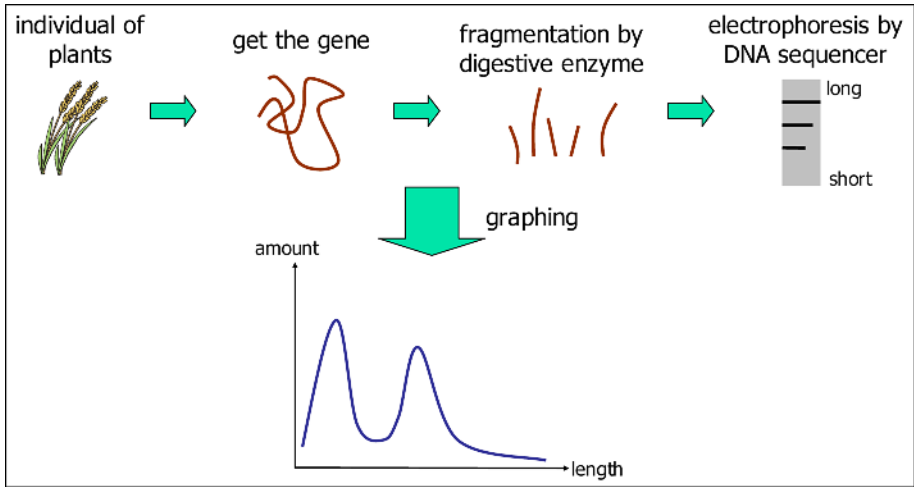


Fig. 1. Measurement of amounts of DNA according to type

2.3 Comparison of Transposon Analysis

We judge whether the transposon was inserted into the DNA by comparing the mutant with the amount of specific DNA in the reference plants. Graphs of reference data and of comparison data are collected and changes in the graphs are discovered. If the reference data have peaks although the comparison data have no peak, we call it "mutation(-)"; conversely, if the reference data have no peak although the comparison data have a peak, we call it a "mutation(+)"

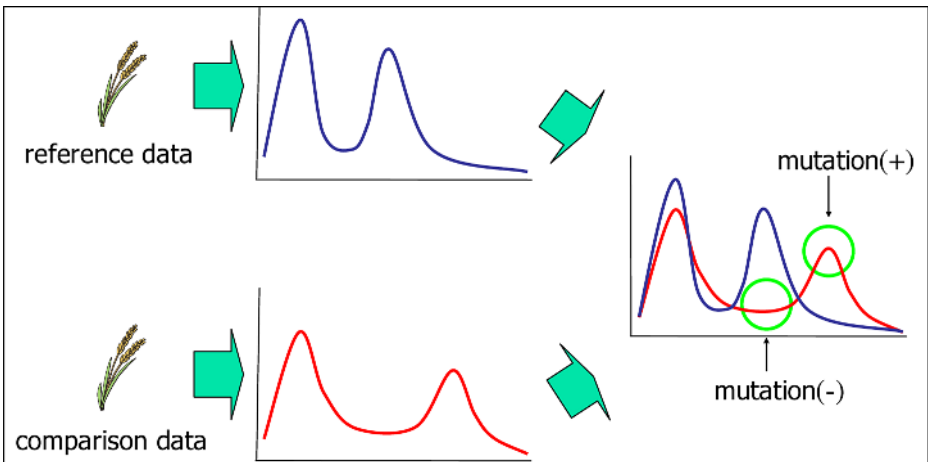


Fig. 2. Change in amount of specific DNA in mutant

2.4 Noise in the Data

Human analysis is needed to make target data from the experimentation. A biologist checks the original data and makes a set of gene changes and phenotypic expressions. He sometimes misunderstands and makes mistakes in the process, and there is much noise in the target data. Moreover, sometimes there is no relation between the change in the length of the gene and its features. For instance, as shown in the figure below, when an m-ping enters the part where there is no relation in a feature, change takes place in the length of a gene, but change does not take place in a feature, so we need to carry out the same treatment as in the case where change does not take place in the length of a gene.

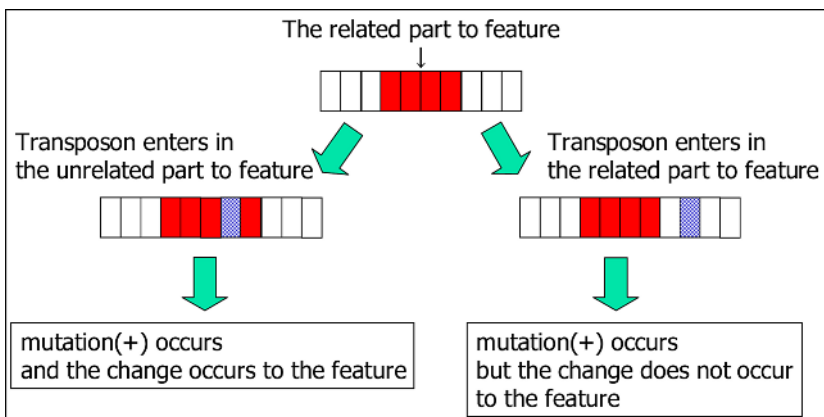


Fig. 3. Relation between insertion part and transposon gene

When the system cannot make complete rules by data mining, it provides a hypothesis for the correct data. The user, who supports the data analysis, uses an interactive function of data mining to select candidates for the hypothesis. In this research, we propose a data mining method that has an interpretation function which makes simple data from complex results.

3 Decision Tree

The decision tree making method generates classification rules to analyze sample data sets that have the same attributes and values. This method is used for selection of attributes to make classification rules in the data mining. This method not only makes logical rules but also decides the order of attributes to make efficient classification. Many classification methods and decision tree generation methods have been researched. ID3 [2] and C4.5 are well known decision tree making algorithms. ID3 uses entropy of classification results and makes orders of useful attributes to classify the correct selection. However, ID3 is not applicable to continuous data. C4.5 is an

algorithm which is an improved version of ID3, and it can treat discrete and continuous values to make the classification rules. In this research, we use an open source data mining tool, WEKA [5], which is developed by the University of Waikato (New Zealand). WEKA includes J48 which is an improved version of C4.5. J48 also uses entropy to make decision trees. It divides the continuous value into several sections.

J48 treats these sections as discrete data. For instance, when data (10, 20, 4, 7, 8, 16, 15, 6) in a continuous data expression are converted into 2 values, the values are defined according to whether the original values are bigger than 10 or not. Thus, the original data are expressed as ($\geq 10, \geq 10, < 10, < 10, < 10, \geq 10, \geq 10, < 10$). As a result, the continuous values of (10, 20, 4, 7, 8, 16, 15, 6) can be represented as two values which are " ≥ 10 " and " < 10 ". This separation method is useful for translating continuous values into discrete values. We currently are improving separation methods to make efficient translations in the data mining process, [6] and [7].

4 Features of the Target Bio Data

We consider the mining target to be the gene of a plant and the character of the plant. Table.1 shows our target bio data which include gene sets and phenotypic expression sets. Each gene length value is expressed as "=", "+", or "-". The length in a normal state (reference gene) is shown by "=". Thus, "=" expresses no change of length of the gene. "+" indicates that the gene has m-ping. "-" indicates that the gene loses m-ping. The features in the data express phenotypic expressions, for example the size of its seed.

Table 1. Example of evaluation data

Data name	gene1	gene2	gene3	gene4	gene5	feature1	feature2	Feature3
DATA1	+	-	=	+	-	10	A	1
DATA2	=	-	-	=	-	7	B	0
DATA3	+	=	-	=	=	6	C	1
DATA4	+	-	=	+	-	12	A	0
DATA5	=	=	-	+	=	14	C	1

5 Generation of Decision Tree with J48

In the example shown in Fig. 4, the data mining system classified the target data into two classes, "A" and "B". The value of each of the data is written in the tree leaf. This expression includes a feature value and the total number for the same data. For instance, if A(40) is written, it means the value of the feature is A, and its value is 40. However, if the decision tree is complex, the user may not be able to understand the result.

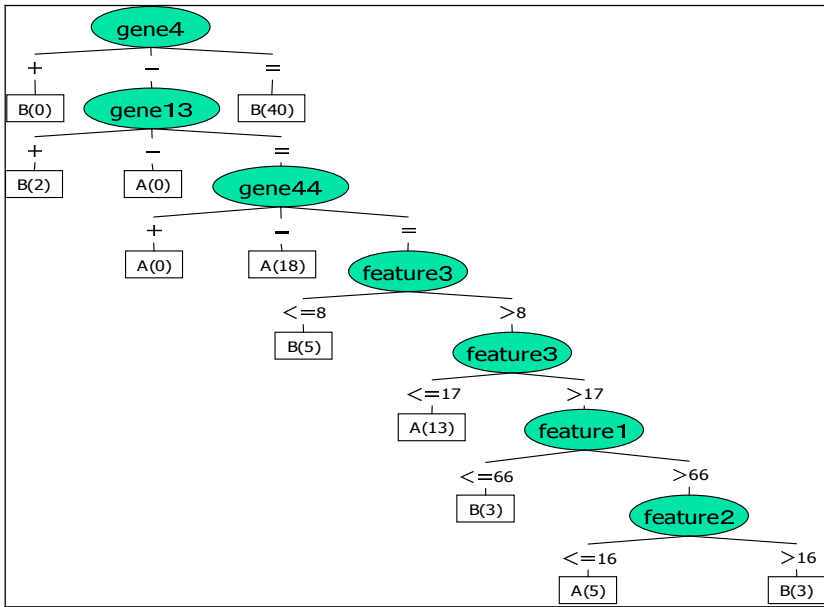


Fig. 4. Sample of decision tree

6 Interpretation of Decision Tree

The decision tree is not always simple and easily understandable. Therefore, an interpretation support function is important for the user. The steps are as follows:

1. Simplify the decision tree

In the decision tree, there are important and unimportant parts. The criteria for importance are frequency of data appearance. The display area of the tree is also useful for the user in understanding the result. The display area is controlled by the depth of the tree, certainty factor and support factor.

2. Explain the decision tree in natural language

Sometimes, the tree representation is very complex for the user to understand. In such cases, the system also provides natural language expressions to present new rules. The system translates part of the tree into a natural language sentence after simplifying the tree.

3. Interactive mining

When the user ignores noise from the original data, he defines a hypothesis, for example, that the gene value must be "+" even if now it is "=". The user considers that a continuous value should be separated into 3 sections even if the data mining system offers 4 sections. As a result, the system treats the hypothesis for the data and makes the results according to the hypothesis. The system supports the user making the hypothesis and applies it to the target data. It mines the rules from the data using the hypothesis. Thus, it supports an abduction function and truth maintenance function to manage the mining process.

7 Interactive Data Mining Methods

Based on the following decision tree, we explain the interaction sample in the decision tree analysis.

1. Depth control of the decision tree

When the system selects the part in which the depth of the tree is limited to four levels, the figure changes from Fig. 7-1 to Fig. 7-2.

2. Branch selection by the support factor

When the system selects the parts at which the support factors exceed 25%, the tree changes from Fig. 7-2 to Fig. 7-3.

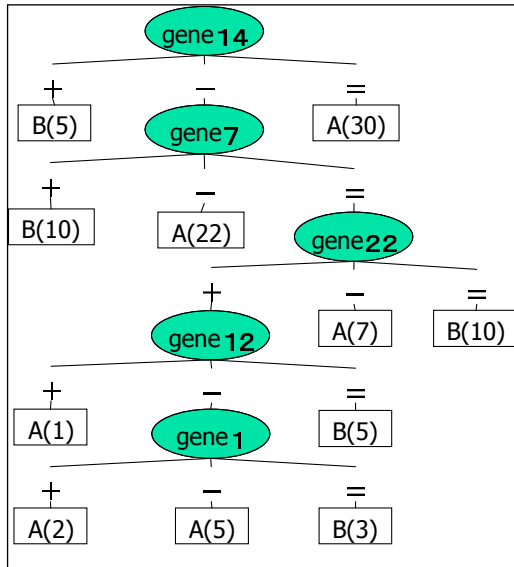


Fig. 7-1. Original decision tree

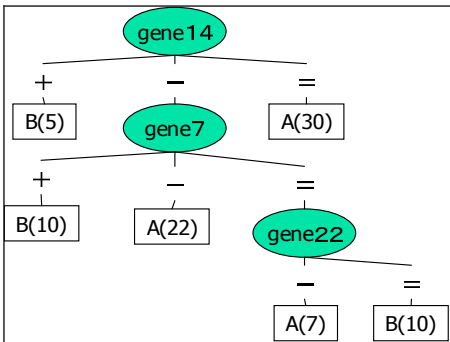


Fig. 7-2. Depth limited tree

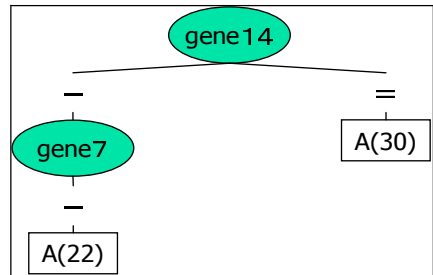


Fig. 7-3. Part of tree selected by its support factor

3. Composition of sentences from the decision tree

For example, by composing sentences from Fig.7-3, we obtain the sentence that "If attribute value of gene14 is "=", the feature is "A" and 30 data support this rule."

8 Conclusions

We constructed an interactive data mining system. This system can handle discrete values and continuous values. Data including noise has two or more interpretations, so we formed a method in which the system proposes a candidate for a correct interpretation of earlier data. Moreover, we developed a method for making a complex decision tree simple and forming sentences. It is planned to verify the effectiveness of the system using real data.

References

1. Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, in Bocca, J. B., Jarke, M., and Zaniolo, C. eds., Proc. of the 20th Very Large Data Bases Conference, Morgan Kaufmann (1994) 487-499.
2. Quinlan, J. R.: Discovering rules by induction from large collections of examples. In Expert Systems in the Micro-electronic Age, Michie, D., Editor, Edinburgh University Press, Edinburgh, Scotland, (1979) 168-201.
3. Quinlan, J. R.: C4.5: Programs for Machine Learning, Morgan Kaufmann Series in Machine Learning, 1993.
4. Kikuchi K, Terauchi K, Wada M, Hirano HY.: "The plant MITE mPing is mobilized in anther culture", Nature.;421(6919) (2003)167-170,
5. WEKA: <http://www.cs.waikato.ac.nz/ml/weka/>
6. Motoda, H.and Washio, T.: "Machine Learning and Data Mining", Journal of the Japanese Society for Artificial Intelligence, Vol.12, No.4 (1997) 505-512 (in Japanese)
7. Watanabe, Y., Komori, M., Abe, H.and Yamaguchi, T.: "Data pre-processing Based on the Integration of Factor Analysis and Feature Selection", The 17th Annual Conference of Japanese Society for Artificial Intelligence, (2003) (in Japanese)

Adequate RSSI Determination Method by Making Use of SVM for Indoor Localization

Hirokazu Miura¹, Junichi Sakamoto¹, Noriyuki Matsuda¹, Hirokazu Taki¹,
Noriyuki Abe², and Satoshi Hori³

¹ Graduate school of Systems Engineering, Wakayama University
930 Sakae-dani, Wakayama 650-8510, Japan
Is-staff@sys.wakayama-u.ac.jp

² Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology

³ Institute of Technologists

Abstract. Context-aware computing that recognizes the context in which a user performs a task is one of the most important techniques for supporting user activity in ubiquitous computing. To realize context-aware computing, a computer needs to recognize the user's location. This paper describes a technique for location detection inside a room using radio waves from a user's computer. The proposed technique has to be sufficiently robust to cater for dynamic environments and should require only ordinary network devices, such as radio signal emitters, without the need for special equipment. We propose performing localization by relative values of RSSI (Received Signal Strength Indicator) among wireless nodes. Furthermore, we use SVM (Support Vector Machine) to find the criteria for classification (whether a node is inside or outside a given area), in the case where absolute RSSI values are used for localization.

1 Introduction

Context-aware computing [1], which offers services based on the context in which a user is performing a task, is one of the most important techniques in supporting user activities in ubiquitous computing. In context-aware computing, a computer needs to recognize the user's environment, e.g., the user's current location, the time at which a service is requested and the user's current activities. To recognize activities, the computer has to know where the user/computer is located. In this paper, we focus on techniques for detecting the location of a user/computer.

Considerable work has been done in the research field of indoor location determination systems [2], [3]. The systems described need special hardware and communication interfaces such as infra-red radiation and ultrasonic wave signals, and the exact location, i.e., the coordinates of the user device, has to be calculated. To improve performance, it is necessary that a large number of devices be arranged in the space concerned (e.g., building, room, etc.). Thus, the achievement of good performance with these systems is very expensive. To reduce the costs involved, systems using ordinary network devices that are widely used for wireless communications between computers, such as radio, are proposed [5]-[7]. But these systems are weak where signal conditions change, due to shadowing, fading, etc.

APIT [10] has been proposed to solve these problems. In this approach, the system detects the user's location by the signal strength from the user's radio device, rather than by change in signal strength. The system can decide whether the user is inside a room or not, without calculating the coordinates. This means that APIT is robust in dynamic environments. However, in the case where the user is inside a building that has many floors, there may be some errors. This is because APIT does not take the 3-dimensional structure of rooms in a building into account, and, as a result, indoor location determination cannot be effectively carried out.

In this paper, we propose an indoor localization method, making use of the relative value of RSSI. Our localization method achieves good performance in unstable signal environments.

Furthermore, we use SVM (Support Vector Machine) to find criteria for classification (inside or outside), in the case where the absolute RSSI values are used for localization. The performance evaluation shows that it is more important to arrange and select anchors adequately rather than to use relative RSSI values from an anchor node.

2 Related Work

There have been several reports of work on location determination techniques. Active bat [2] and Cricket [3] determine location by the use of sensor devices such as infrared or ultrasonic wave signal devices. The performance of these systems depends on how many devices can be arranged in the space considered. To achieve good performance with these systems, a large number of devices is needed, e.g., 720 devices for every 1000 square meters. Thus it is very expensive to maintain good performance and practical implementation is difficult. We believe that what is really required is a method which relies only on ordinary network devices that are widely used for wireless communications between computers, e.g., Wireless LAN, Bluetooth, Zigbee and Mote [4].

Almost all the systems that use radio signals [5]-[7] calculate the location coordinates of the object concerned, which means that the system is very complicated. With indoor location, it is necessary to determine which room the computer is located in, i.e., it is more important to know the boundary between the rooms than coordinates. In these systems, however, the estimation error is large, leading to mistaken results. In the case shown in Figure 1, the computer is originally located in room Y, but the system may estimate that it is located in room X because of the large error. These systems use the formula given in [8] for a wireless model, but there are many cases where real world radio signals cannot be modeled with the formula.

Ogawa et al. [9] have proposed a location determination method using radio signals from base stations, with supervised learning. In their work, the assumption is made that change of environment does not occur, i.e., the device does not move. Thus, with their system it is difficult to determine the location of a mobile device.

The APIT (Approximation of the Perfect PIT Test) which is robust in dynamic environments has been proposed in [10]. In APIT, the wireless mobile node has a 2-dimensional map and determines its own location by communicating with anchor nodes. Anchor nodes are made aware of their own coordinates beforehand. The

mobile node looks for other nodes and anchor nodes to which it can communicate. If it finds them, it determines their ID, before plotting anchor coordinates in the map and drawing triangular areas connecting anchor to anchor. The mobile node determines that its own position is inside these triangular areas. This allows it to narrow down the area where it can potentially reside. An overview of APIT is shown in Figure 2. The center of gravity of the shaded area indicates the node's coordinates.

We describe how APIT determines whether a node is inside these triangular areas or not. In Figure 3 and 4, nodes M, 1, 2 and 3 are mobile nodes and nodes A, B and C are anchor nodes. Each mobile node measures the signal strength from each anchor node. Node M compares the signal strength which it receives from the anchor nodes, with the signal strength which neighboring mobile nodes receive. If no neighboring node of M is further from or closer to all three anchor nodes simultaneously, node M assumes that it is inside the ΔABC . Otherwise, M assumes that it is outside the ΔABC . Figure 3 presents a scenario where M will assume that it is inside the ΔABC . In Figure 4, node 1 will report to node M that it is further away from A, B and C than M. Figure 4 shows the case where it is outside the triangle ABC.

On the other hand, in the case where the user is inside a building that has many floors, there may be some error in the system, as we mentioned above. This is because the 3-dimensional structure of rooms is not taken into account. In some cases, the device may receive signals from lower or upper floors, leading to misdetection. Therefore, in this paper we propose an indoor location determination technique using a topological model representing the hierarchical structure of rooms in the building.

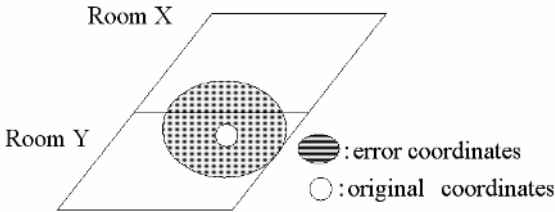


Fig. 1. Problem of indoor determination

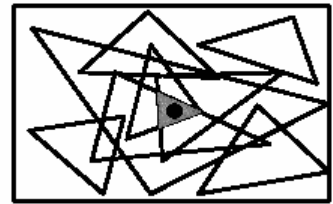


Fig. 2. APIT overview [10]

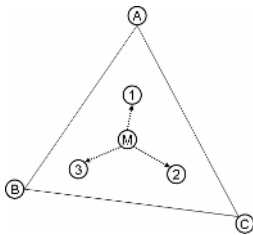


Fig. 3. APIT inside case

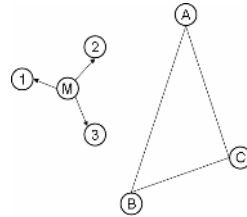


Fig. 4. APIT outside case

3 Indoor Localization

In [13], the indoor localization method using RSSI and a topological model has been proposed. However, it has not been implemented or tested. In our research, we have implemented our method and evaluated its performance. In this section, we describe the detailed operation of the proposed technique. Our proposal makes use of the relative values of RSSI between the anchor and mobile nodes. Our proposal uses ordinary network devices without special equipment and does not calculate location coordinates. Thus, our system is inexpensive and simple.

The values of RSSI measured indoors are unstable due to multi-path fading, compared to the outdoor values as shown in Figure 5. Therefore, it is difficult to determine the distance between the wireless nodes directly from the measured RSSI. We use the relative values of RSSI between the anchor nodes and mobile nodes rather than the absolute values. In other words, the values of RSSI between nodes are compared.

However, there are areas where the relative values of RSSI cannot be used for the comparison, e.g., the case where the distance between the wireless nodes is longer than 3.5 m. Therefore, it is necessary to initially investigate the areas where the relative values can work effectively.

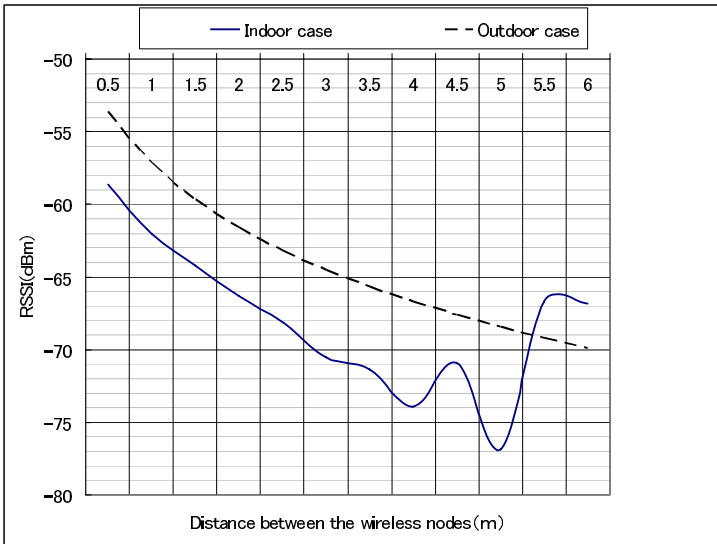


Fig. 5. RSSI vs. Distance

3.1 Localization Using Relative RSSI

The procedure is as follows. In Figure 6, nodes A, B, C and D are anchor nodes and node M is a mobile node. Nodes A-D are arranged at the corners of the area. The node M estimates its own location by the following steps.

- Step1. Each anchor node measures the RSSI between itself and the other anchor nodes.
 The mobile node M receives signals from the anchor nodes and obtains information such as the anchor IDs and signal strengths ($RSSI_{MA}$, $RSSI_{MB}$, $RSSI_{MC}$, $RSSI_{MD}$ in Figure 6).
- Step2. Node M collects the information (obtained in Step 1) from the anchor nodes ($RSSI_{AC}$, $RSSI_{BD}$, $RSSI_{CA}$, $RSSI_{DB}$).
- Step3. Node M compares $RSSI_{MA}$ and $RSSI_{AC}$, and determines whether its own location is inside or outside circle C_{AC} . The node M repeats the same operation for the other anchor nodes. Finally, the node M decides whether its location is inside or outside the area ABCD.

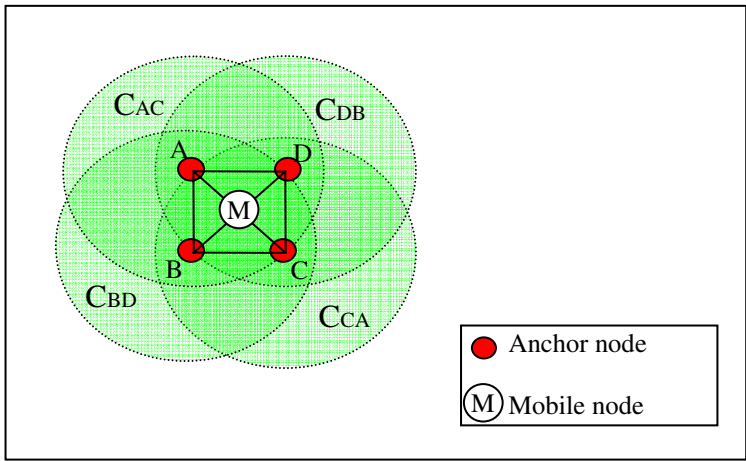


Fig. 6. Node placement

3.2 Overview of Our Localization System

We implement our localization system using an ad-hoc wireless sensor device, Mote. The overview of our system is described in Figure 6.

In our system, the anchor nodes and mobile node measure RSSI between each anchor node and between the anchor node and mobile node, respectively. The mobile node compares the values of RSSI and determines whether its own location is inside or outside an area surrounded by the anchor nodes.

3.3 Experimental Results

We investigated the performance of our system experimentally, as follows. Figure 7 shows our experimental environment. Our localization system decides whether the position of the mobile node is inside or outside an area surrounded by anchor nodes. We investigated whether the localization was correctly achieved at 21 measurement points, as in Figure 8.

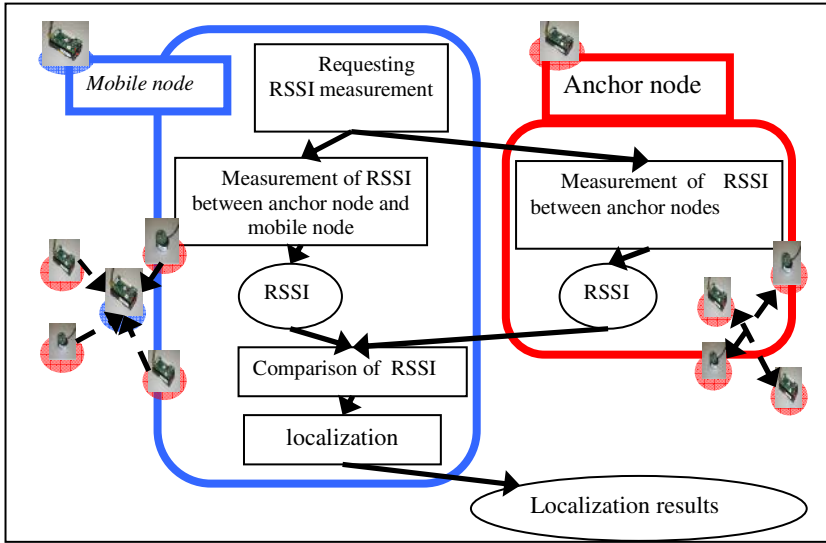


Fig. 7. System Overview

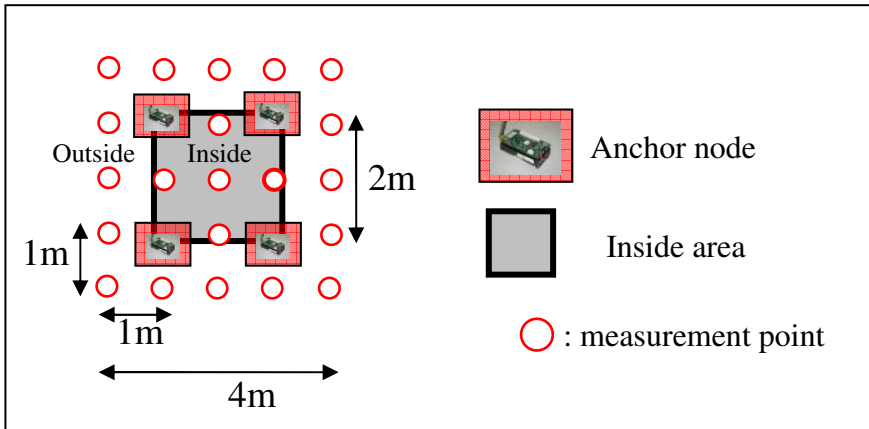


Fig. 8. Experimental Environment

The performance of our system is compared with localization by making use of the absolute value of RSSI. The comparison method calculates the difference of signal strength and RSSI value from the propagation model, and the coordinates of the mobile node by multilateration.

Table 1 shows the results of the localization. In the case where an obstacle does not exist in the area, the comparison method localized the mobile node in 57% of the measurement points. In contrast to this, the percentage of measurement points where our system could localize the mobile node was 95%. From these results, our system is shown to have good performance.

Table 1. Localization Results

	No obstacle			Obstacle		
	Total	Inside	Outside	Total	Inside	Outside
Comparison method	57 %	80%	50%	76 %	0%	100%
Our system	95 %	80%	100%	100 %	100%	100%

3.4 Localization Using SVM

As mentioned above, the relation between the RSSI and distance among wireless nodes varies according to the signal condition. In other words, it is difficult to calculate the distance from the absolute values of the RSSI. However, the combination of the RSSI among nodes whose location can be decided, can be found by SVM. Therefore, we implemented the SVM (support vector machine) [12] in our system for indoor localization as an adequate RSSI learning mechanism.

Furthermore, we evaluated the performance of this system. We classified the 21 results in the previous section into an inside class or an outside class. From these results, 414 combinations of RSSI data were classified by the SVM. In this evaluation, the value of the RSSI between the mobile node and the 4 anchor nodes is used as 4 properties. As the result, 93% of the RSSI combinations are classified successfully. The result shows almost the same performance as in the previous section. These results show that it is important to arrange and select the anchors carefully.

4 Applications

This section describes applications using our proposed technique. There are three kinds of applications. The first is where the user wishes to use nearby computer peripherals, such as a printer in an office. A simple message is sent to printers connected in a wireless network as follows: "Print this document using the printer in the location where I am now". Printers receiving this message sense their own location and that of the user. A printer is found in the location of the user and the user's document is printed. Another example would be if the user requests information concerning shops in a shopping complex, a computer provides relevant information arranged according to shop proximity.

The second application we consider is where a location administrator communicates with users inside its location area. For instance, since our system enables judgment as to whether a user is inside or outside a designated location, user input of working hours in an office or school can be automated. When the user enters his/her office, the system starts recording, and when the user leaves the office, the system stops recording. Another example would be where only those persons inside a specific location, such as a library, would be able to browse an e-book free of charge. This is based on the premise that only people inside the library have the right to consult the library books.

The third application is where other people wish to know the location of a user. In cooperative work using mobile computers, it is often not possible to know where others are located – building, car, train, etc. This problem is solved by using our proposed technique so that a user can be made aware of another user's location. Additionally, it is possible to provide a user's location while taking privacy into account by applying inclusion relations in the topological model.

Thus, our proposed technique can be used in a varied range of applications.

5 Conclusions

We have described an indoor localization technique using relative values of RSSI among nodes. Furthermore, we have implemented SVM in our localization system to achieve robustness for unstable signal conditions. The advantage of our technique is that it does not use coordinates, special hardware is not required, and it is simple and robust in dynamic environments. Future work will investigate required signal strengths and how many nodes and anchors are needed. For this purpose, we are preparing a simulation program that will utilize various parameters.

References

1. B. Schilit, N. Adams, and R. Want, "Context-Aware Computing Applications" in IEEE Workshop on Mobile Computing Systems and Applications, Santa Cruz, CA, US, 1994.
2. Andy Harter, Andy Hopper, Pete Steggles, Andy Ward, Paul Webster, The Anatomy of a Context-Aware Application, Proceedings of the Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 1999), pp. 59-68, Seattle, Washington, USA, August 1999.
3. N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, The Cricket location-support system. in the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 2000), pp. 32.43, Boston, MA, USA, Aug. 2000.
4. <http://www.tinyos.net/>
5. Akiko Iwaya, Nobuhiko Nishio, Masana Murase, and Hideyuki Tokuda, GOMASHIO: Proximity Based Localization In Wireless Ad-Hoc Sensor Networks, IPSJ SIG Mobile Computing and Ubiquitous Networking, Vol. 2001 108 pp. 23-30, Nov. 2001 (in Japanese).
6. Teruaki Kitasuka, Tsuneo Nakanishi, and Akira Fukuda, "Wireless LAN based Indoor Positioning System WiPS and Its Simulation," 2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM'03), pp. 272-275, Aug. 2003..
7. N. Bulusu, J. Heidemann and D. Estrin, GPS-less Low Cost Outdoor Localization for Very Small Devices, IEEE Personal Communications Magazine, Vol. 7, No. 5, pp. 28-34, October 2000.
8. Deepak Ganesan, Deborah Estrin, Alec Woo, David Culler, "Complex Behavior at Scale: An Experimental Study of Low-Power Wireless Sensor Networks", Technical Report UCLA/CSD-TR 02-0013, 2002.
9. Tomoaki Ogawa, Shuichi Yoshino and Masashi Shimizu, The In-door Location Determination Method Using Learning Algorithms with Wireless Active Tags, IPSJ Ubiquitous Computing Systems (UBI), Vol. 2004, No. 66, pp. 31-38, 2004 (in Japanese).

10. Tian He, Chengdu Huang, B. M. Blum, John A. Stankovic, and Tarek F. Abdelzaher, Range-Free Localization Schemes in Large Scale Sensor Networks, the Ninth Annual International Conference on Mobile Computing and Networking (MobiCom 2003), San Diego, CA, September 2003.
11. B. Brumitt, and S. Shafer, Topological World Modeling Using Semantic Spaces, Workshop Proceedings, UbiComp 2001, pp. 55-62, 2001.
12. V. Vapnik, Statistical learning theory, John Wiley & Sons, New York, 1998.
13. J. Sakamoto, H. Miura, N. Matsuda, H. Taki, N. Abe and S. Hori, "Indoor Location Determination using a Topological Model", Proc. of 9th International Conference Knowledge-based Intelligent Information and Engineering Systems (KES2005), LNAI 3684, pp. 143-149 (2005).

Ontia iJADE: An Intelligent Ontology-Based Agent Framework for Semantic Web Service

Toby H.W. Lam and Raymond S.T. Lee

Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{cshwlam, csstlee}@comp.polyu.edu.hk

Abstract. The Internet becomes part of our lives. Users can access different kinds of information through the Internet. However, the exponential growth of the Internet led to a great problem: difficult to locate relevant information. The Semantic Web is an extension of the current web system in which information is organized in a well-defined manner and for enabling computers and people to work in cooperation. Agent technology gives a new way to this intelligent web system. In this paper, we proposed an innovative agent platform – Ontia iJADE which supports the Semantic Web Service. Ontia iJADE contains an ontology Server known as intelligent Java-based ontology Server (IJAOS), which stores and manages the ontology information for agents to share and reuse.

1 Introduction

In modern AI, scientists try to build intelligent software objects that can mimic human intellectual behavior for the purposes of problem solving, scheduling, data mining and to generally assist humans in all of their activities. In the past years, the developers of agents have implemented various agent systems, ranging from AuctionBot [1] for e-auction to our own iJADE Web Miner for intelligent web mining. Most of these agents, however, are either task-specific or “rigidly” defined in response to specific environments. As yet, it has not been possible to develop agents that are “truly” adaptable to their environments and capable of autonomous learning.

Recently, the semantic web has received substantial attention from the research community. The semantic web, the next generation of World Wide Web, aims to provide a new framework that can enable knowledge sharing and reusing. Related to the development of the semantic web, new research and findings have been done in various areas such as Knowledge Engineering, ontology-based Information Retrieval and ontology-based Agent.

To enable knowledge interoperability in the semantic web, ontologies is one of the main components to provide this service. From the philosophical point of view, ontology means the study of entities and their relationship. In AI point of view, ontology is the explicit specification of concepts. In fact, ontology is usually defined as an explicit specification of conceptualization [2]. In general, it refers to the description of the concepts and relationships that can exist for an agent or a community of agents.

Ontology also plays a vital role in knowledge sharing and exploration, particularly in multiagent-based communication where the content of messages can exchange among different agents. In summary, agents which integrated with ontology have the

ability to i) categorize logical (and linguistic) concept items; and/or ii) translate ontologies which are different either in terms of their degree of categorization or the languages being used. As a result, such agent will be more adaptive and platform independent in different environment as it could communicate with other agents in different language and culture.

Ontia iJADE is an intelligent ontology-based agent framework for semantic web service. It is the extension of our current research work – intelligent Java-based Agent Development Environment (iJADE). Ontia iJADE integrated the intelligent Java-based Agent ontology Server (iJAOS) into the Intelligent Layer of the iJADE Framework. It further increases the functionalities of the agent and enables the knowledge reuse and sharing. In this paper, we describe and explain the design of our new generation of agent platform namely Ontia iJADE.

The rest of this document is organized as follows: Section 2 shows the related works on semantic web and ontology. Section 3 describes the methodology and system design of Ontia iJADE. Section 4 presents the main architecture of the Ontia iJADE Section 5 presents the conclusion and future work.

2 Related Work

Nowadays, the Internet becomes part of our lives. Users can gain access to all kinds of information over the Internet. But, such exponential growth has led to a great problem: much time is wasted in locating relevant information [3]. In order to help users to search useful information on the web, many well-known web portals such as Yahoo! have organized web documents into some predefined categories including Arts & Humanities, Computers & Internet, and Entertainment [4]. Users can search relevant information by browsing a topic hierarchy.

However, current topic-based directory systems suffer from the bottleneck of the manual classification of newly collected documents. For example, Yahoo!, the largest directory system on the Internet, contains roughly 1.2 million links in its topic hierarchy and more than 150 editors are needed to classify web documents [5][6], a process that is costly and slow [7]. In addition, the total number of documents in the directory systems is much lower than the database being used by search engines. The topic directories are often incomplete and generally out of date [8]. Therefore, automatic classification has become both an interesting and vital topic for research and application [9].

Precise classification is the main concern in semantic web because the results of automatic classification are usually not as precise as those arising from human decisions [10]. As Tim-Berners Lee, the founder of the World Wide Web, stated “*The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.*” The next generation of the World Wide Web has been referred as the “Semantic Web,” where information will be and must be machine-processable in ways that support intelligent network services such as search agents [11].

Ontology enables among different agents to share knowledge, reuse knowledge and provide fundamental knowledge in a specific domain. Ontology in AI and agent technology is well represented by the conceptualization theory of Gurarino and

Giaretta [12] and furthers work by Guarino [13]; the design of ontologies for knowledge sharing by Guber [14]; the Representation ontologies proposed by Van Heijst et al. [15].

Current applications on ontology includes the Semantic Web proposed by Maedche and Staab [16] using ontology learning techniques, and the application of Core ontology technology to enable scalable assimilation of information from diverse multimedia sources on MPEG-21 proposed by Hunter [17]. For agent technology, current ontology standards include OKBC (Open Knowledge Base Connectivity) jointly developed by the Artificial Intelligence Center of SRI International and the Knowledge Systems Laboratory of Stanford University and K-ontologies which developed by INM of the Institute for New Media in Germany. INM also acts as information brokers, helping users to exploit the ontologies of the Web. For applying agents to intelligent inhabited environments, Khedr and Karmouch using ontology model in multi-agent system under context-aware environments [18]. Susperregi et. al develop a location-aware intelligent laboratory which used AI technique with agents technology [19]. Hargas et al. has a similar research but they developed an intelligent hotel instead [20].

In this paper, we present the design of our new generation of agent platform namely Ontia iJADE. Ontia iJADE has an ontology server, iJAOS, which enable to share and reuse knowledge. Section 3 shows the details of Ontia iJADE.

3 Ontia iJADE Framework

The Ontia iJADE system framework is shown in Figure 1. It is an extension of iJADE. The aim of Ontia iJADE is to provide comprehensive APIs for intelligent agents and applications. In our current Ontia iJADE platform, it consists of four layers:

i) **Application Layer** – this is the uppermost layer that consists of different intelligent agent-based applications. This layer accepts the data result from the intelligent layer and is connected to external application.

ii) **Intelligent Layer** – this is an intelligent layer includes a Sensory Area, Logic Reasoning Area, Analytical Area and intelligent Java-based Ontology Server.

iii) **Technology Layer** – this layer provides all the necessary mobile agent implementation APIs for the development of intelligent agent components in the Intelligent Layer.

iv) **Supporting Layer** – This layer provides a programming language and protocols to support the development of the Technology Layer.

In the intelligent layer of Ontia iJADE, there is an intelligent Java Agent-based ontology Server (iJAOS), which aims at the implementation of a truly autonomous and adaptive agent (See Figure 2). By using the iJAOS, it can be used to:

- share knowledge – automatic knowledge sharing among different agents
- reuse knowledge – reuse knowledge to other similar domain
- priori knowledge – provide fundamental knowledge in a specific domain

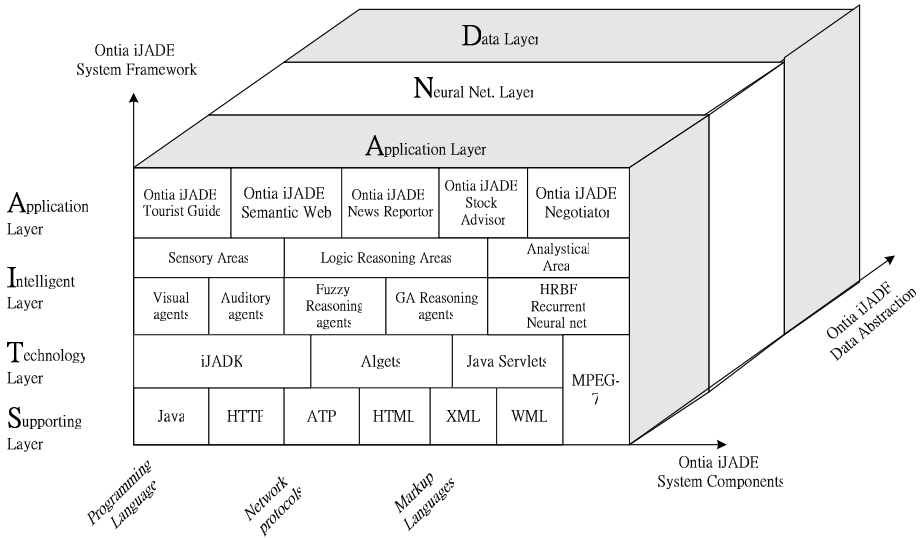


Fig. 1. System architecture of Ontia iJADE model

The intelligent Java Agent-based Ontology Server (iJAOS), which consists of five functional modules:-

- iJASC (intelligent Java Agent-based Sensation Center)
- iJAMC (intelligent Java Agent-based Memory Center)
- iJAKC (intelligent Java Agent-based Knowledge Center)
- iJALC (intelligent Java Agent-based Language Center)
- iJAEC (intelligent Java Agent-based Ethics Center)

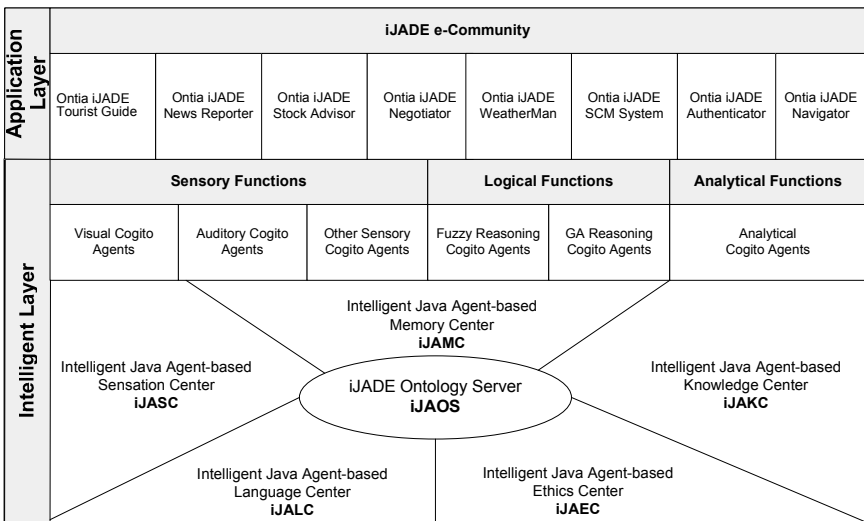


Fig. 2. System architecture of the iJAOS in the Intelligent Layer of Ontia iJADE Framework

From the application point of view, this new generation of iJADE platform could develop applications which can be categorized into two types: i) Ontology-based Location-aware Agent Applications and ii) Ontology-based Context-aware Agent Applications.

i) Ontology-based Location-aware Agent Applications

- Ontia iJADE Tourist Guide – a highly proactive and adaptive iJADE Tourist Agent which helps travelers to navigate for sightseeing, to locate particular landmark, to find-out the closest restaurant and other related information with respect to his/her location [21];
- Ontia iJADE Navigator - iJADE Navigator is an intelligent route finder and itinerary planner for drivers travelling around the globe. It integrates Global Positioning System (GPS), Genetic Algorithms (GA) and Shortest Path Algorithm (SPF) on intelligent agent helping user to navigate around the globe and to seek for the optimal route.

ii) Ontology-based Context-aware Agent Applications

- Ontia iJADE Semantic Web – can intelligently construct an effective semantic web system.
- Ontia iJADE News Reporter - can intelligently select useful news for users.
- Ontia iJADE Negotiator - intelligently switch its negotiation strategy by assessing the opponent's negotiation patterns automatically
- Ontia iJADE SCM – dynamic cooperation agreement in Collaborative Planning, Forecasting and Replenishment for SCM strategy.

Ontia iJADE data level, DNA model, provides a comprehensive data manipulation framework that is based on neural-network technology. The “Data Layer” corresponds to the raw data and input “stimuli” (such as the facial images captured from a Web camera and the product information in a cyber store) from the environment. The “Neural-Network Layer” provides the “clustering” of different types of neural networks for the purpose of organization, interpretation, analysis and forecasting operations that are based on the inputs from the Data Layer”. The neural networks are used by the iJADE applications in the “Application Layer”. Another innovative feature of the iJADE system is the ACTS model, which provides a comprehensive layered architecture for the implementation of intelligent agent systems.

4 Conclusion and Future Work

In this paper, we introduced the details of our next generation agent platform – Ontia iJADE. Ontia iJADE is the extension of our iJADE (intelligent Java-based Agent Development Environment). In the Intelligent Layer of Ontia iJADE, there is an intelligent Java-based ontology Server (iJAOS).

Ontia iJADE is the new generation of agent platform. It is able to share knowledge, reuse knowledge and provide fundamental knowledge in a specific domain. In the future, we would create other applications based on the Ontia iJADE platform. Besides application development, we would like to include different kinds of domain knowledge in the iJAOS. It would further increase the functionality of Ontia iJADE and the

interoperability between different kinds of agents. We showed some potential applications which can be developed under such framework. In the future, we would like to implement a prototype of this framework. Besides, we would like to implement application, such as Ontia iJADE Tourist Guide and Ontia iJADE Semantic Web, under the framework. After developed the prototype, we would like to do some experiments to evaluate the performance of the framework.

Acknowledgements

This work was partially supported by the iJADE projects B-Q569, A-PF74 and Cogito iJADE project PG50 of the Hong Kong Polytechnic University.

References

- [1] AuctionBot: <http://auction2.eecs.umich.edu/auction/>
- [2] M. R. Grenesereth, and N. J. Nilsson, "Logical Foundation of Artificial Intelligence", Morgan Kaufmann, California, 1987.
- [3] D. Fensel and M. Musen, "The Semantic Web: A Brain for Humankind," IEEE Intelligent Systems, March/April 2001
- [4] C. Haruechaiyasak, M.-L. Shyu, S.-C. Chen, "Web Document Classification Based on Fuzzy Association", Proceedings. 26th Annual International Computer Software and Applications Conference, 2002, pp. 487 – 492, 2002
- [5] Web Directory Sizes: URL : <http://searchenginewatch.com/reports/directories.html>
- [6] S.-H. Lin, M. C. Chen, J.-M. Ho, Y.-M. Huang, "ACIRD: Intelligent Internet Document Organization and Retrieval", IEEE Transactions on Knowledge and Data Engineering, Vol. 14, no. 3, pp. 599 – 614, 2002
- [7] Z. Peng and B. Choi, "Automatic Web Page Classification in a Dynamic and Hierarchical Way," Proceeding of IEEE International Conference on Data Mining, 2002. ICDM 2002, pp. 386 – 393, 2002
- [8] C. Jenkins and D. Inman, "Adaptive Automatic Classification on the Web", Proceeding of the 11th International Workshop on Database and Expert Systems Applications, pp.504 – 511, 2000
- [9] R. Prabowo, M. Jackson, P. Burden, H.-D. Knoell, "ontology-Based Automatic Classification for the Web Pages: Design, Implementation and Evaluation", Proceedings of the 3rd International Conference on Web Information System Engineering (WISE'02), pp. 182-191, 2002
- [10] A. Rotshtein and D. Katelnikov, "Design and Tuning of Fuzzy If – Then Rules for Automatic Classification", Conference of the North American on Fuzzy Information Processing Society - NAFIPS, pp. 50 – 54, 1998
- [11] L. Weihua, "Ontology Supported Intelligent Information Agent," Proceedings on the First International IEEE Symposium on Intelligent Systems, vol. 1, pp. 383 – 387, 2002
- [12] N. Guarino, M. Carrara, and P. Giaretta, "An ontology of Meta-Level Categories", Proceedings of the Fourth International Conference of Knowledge Representation, pp. 270-280, 1994.
- [13] N. Guarino, "Formal ontology in Information Systems", Proceedings of Formal Ontology in Information Systems, pp. 3-15, 1998.

- [14] T. R. Gruber, "Toward Principles for the Design of ontologies used for Knowledge Sharing", *International Journal of Human and Computer Studies*, vol. 43 no. 5, pp. 907-928, 1993
- [15] G. van Heijst, A.Th. Schreiber, B.J. Wielinga, "Using explicit ontologies in KBS development," *International Journal of Human and Computer Studies*, vol. 46 (2/3), pp.183-292, 1997
- [16] A. Maedche, and S.Staab, , "ontology Learning for Semantic Web, " *IEEE Intelligent Systems*, vol. 16, no. 2, pp. 72-79, 2001
- [17] J. Hunter, "Enhancing the Semantic Interoperability of Multimedia Through a Core Ontology", *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 49-58, 2003
- [18] M. Khedr and A. Karmouch, "Negotiating Context Information in Context Aware Systems," *IEEE Intelligent Systems*, vol.4, pp. 21-29, 2004
- [19] L. Susperregi, I. Maurtua, C. Tubio, I. Segovia. M.A, Perez and B. Sierra, "Context aware agents for Ambient Intelligence in Manufacturing at Tekniker", *AgentLink Newsletter*, vol. 18, pp. 28-30, 2005
- [20] H. Hagra, V. Callaghan, M. Colley, G. Clarke, A. Pounds-Cornish, and H. Duman, "Creating an Ambient-Intelligence Environment Using Embedded Agents," *IEEE Intelligent Systems*, vol.4, pp.12-20, 2004
- [21] T.W.H. Ao Jeong, T.H.W. Lam, A.C.M. Lee and R.S.T. Lee, "iJADE Tourist Guide: A Mobile Location-Awareness Agent-Based System for Tourist Guiding," *Proceedings of the 9th International Conference Knowledge-Based Intelligent Information and Engineering Systems (Part I)*, pp. 671-677, 2005

iJADE FreeWalker: An Ontology-Based Tourist Guiding System

Toby H.W. Lam and Raymond S.T. Lee

Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
{cshwlam, csstlee}@comp.polyu.edu.hk

Abstract. In this paper, an ontology-based tourist guiding system, iJADE FreeWalker, is presented. The system integrated with agent and Semantic Web technologies to provide tourist information for the tourist. The tourist's geographical information is gathered by GPS receiver. The GPS Agent migrates to the tourist information center and retrieves the related tourist information with respect to the user's location. We defined an upper-level travel ontology by using Web Ontology Language (OWL). By using such system, tourists could browse the relevant information based on their interest and location.

1 Introduction

There are a number of tourists love traveling with their backpack, guidebook and map. As these backpackers do not have any tourist guide (human), they would get lost easily and miss some attractions. It is necessary to develop a system which travelers can access the latest travel information according to their geographic information at anytime and in anywhere. In this paper, we proposed an ontology-based tourist guiding system, called iJADE FreeWalker, for tourists. As there are many wireless devices such as PDA, mobile phone in the market, they are usually small in size, light in weight and long run time. We developed a prototype system in a pocket PC. The main aim of this context aware tourist guiding system is to provide the tourists more fruitful tourist information.

iJADE FreeWalker was developed under iJADE which is an intelligent Java Agent-based Development Environment [1]. The system is integrated with GPS (Global Positioning System) receiver, to gather the user's location. The system would give the tourists' nearby tourist information such as shopping, sightseeing, entertainment and restaurant. To enable knowledge sharing and reuse, we adopted the latest Semantic Web technology to develop an upper-level *travel* ontology.

The rest of this document is organized as follows: Section 2 shows the related works about location-aware applications. Section 3 describes the system design of iJADE FreeWalker. Section 4 describes the ontology of *travel*. Section 5 offers our conclusion and outlines some directions for future work.

2 Related Work

GPS is used to gather geographical information. Since the price for GPS receiver is not expensive at present, these equipments are widely used and adapted today. There are

some consumer products developed for outdoor usage such as hiking, flying and sailing. Furthermore, the receivers can also be used for route guiding.

There is a research project called CRUMPET, which is funded by European Union, for developing a tourist information system with personalized nomadic platform [2]. The system is integrated with GPS and agent technology for robust, scalable and seamlessly accessible services. CRUMPET collects two different types of information for tourists: dynamic and static information. The dynamic information is the information that retrieved according to the user's moving route. The static information is the information that retrieved according to the user's profile and request. Besides, the system would learn the user's interest and preferences automatically. The system could filter the irrelevant information to the user.

Georgia Institute of Technology developed a tourist guiding system called Cyber-guide [3]. This guiding system provides information to the users based on the data of users' position and orientation. There are two different kinds of Cyberguide: indoor and outdoor. The indoor cyberguide uses infrared (IR) as the positioning component and the outdoor cyberguide uses GPS to capture the user location.

Cheverst et al. developed a context aware tourist guide system called GUIDE [4]. GUIDE would show tourist information in display of the user's mobile device. The system could construct a tour plan for the user by using user's preference and interest. Similar to CRUMPET system, GUIDE also employs wireless communications and GPS to ensure the mobile connectivity and context-awareness. Experiments showed that there is a high level of acceptability for the system in different types of user.

A number of location aware tourist guide systems have been developed. The main objective of these systems is to select or filter the relevant information to the users based on user's preference and location. However, the existing applications not fully utilized the AI technologies. To further improve the functionality, we propose a mobile location-aware agent-based system for tourist guiding. The detail of iJADE FreeWalker is shown in next section.

3 System Architecture

iJADE FreeWalker is composed of three major components, 1) iJADE FreeWalker Client, 2) GPS Agent, and 3) iJADE Tourist Information Center

3.1 iJADE FreeWalker Client

The iJADE FreeWalker Client is a graphical user interface to display the map and tourist information for the user. The client gathers the user's location information by using the GPS receiver. The GPS receiver is employed to receive simultaneous GPS data and ascertain the user's location. The GPS receiver is connected with the pocket PC via Bluetooth. Figure 2 shows the screenshot of the iJADE FreeWalker Client.

3.2 GPS Agent

As we all know, mobile device has a narrow bandwidth in wireless connection which is a critical problem in developing a mobile information retrieval system. To overcome

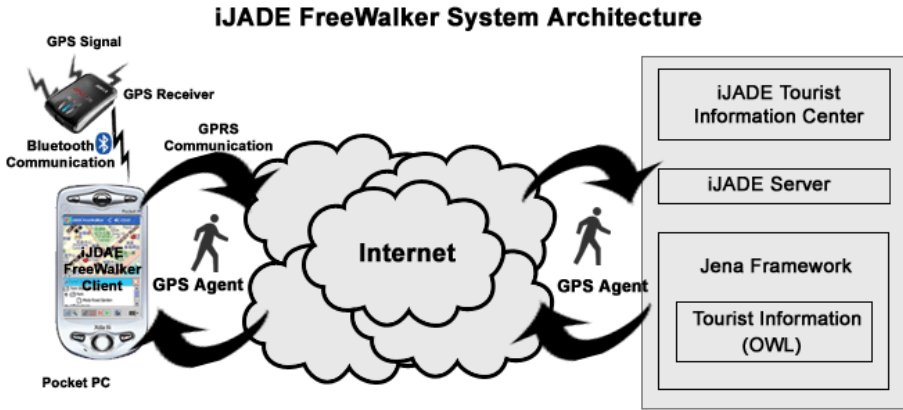


Fig. 1. iJADE FreeWalker System Architecture



Fig. 2. Screenshot of iJADE FreeWalker

such problem, we utilize mobile agent technology in iJADE FreeWalker. In a single request, the agent could conduct multiple interactions with different information database systems. The results are then sent back to the device so as to reduce the network loading.

The GPS Agent is a software agent which can freely migrate from one host to another. First, the GPS Agent captures the user location information from GPS receiver. Then, it migrates from the client (end-user handheld device) to a remote iJADE Tourist Information Center through GPRS communication. When the GPS Agent reaches the information center, it will query the server to collect the tourist information with

respect to the user's geographical information. At last, the GPS Agent returns to client with related tourist information. The client will collect the information from the GPS Agent and show the context information to the user.

3.3 iJADE Tourist Information Center

The Tourist Information Centre has two components: iJADE Server and Jena Framework [5]. Jena is an open source Java framework for building Semantic Web applications. It offers a number of APIs for handling RDF, RDFS and OWL. The main purpose of using Jena is to parse and query about the travel ontology. Jena searches the data from the travel ontology by using the SPARQL Protocol And RDF Query Language (SPARQL). Figure 3 shows an example of SPARQL that is used to get related information about Guesthouse. The search returns as an RDF graph and send back to the tourist information server.

iJADE Server acts as a communication platform. It is a container to receive and send GPS Agent through GPRS. After the GPS Agent arrived at the information center, it uses SPARQL statement to parse RDF. The related tourist information will send back to client.

```
# Select related information about Guesthouse
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX travel: <http://www.comp.polyu.edu.hk/~cshwlam/ontology/travel.owl#>
SELECT ?URI ?NAME ?ADDRESS ?DISTRICT ?ROOM ?URL ?EMAIL ?PRICE
?FAX ?PHONE
WHERE { ?URI rdf:type travel:Guesthouse .
        ?URI travel:hasName ?NAME .
          ?URI travel:hasAddress ?ADDRESS .
        ?URI travel:hasDistrict ?DISTRICTURI .
        ?DISTRICTURI travel:districtName ?DISTRICT .
        ?URI travel:hasRoom ?ROOM .
        OPTIONAL { ?URI travel:hasURL ?URL } .
        OPTIONAL { ?URI travel:hasEmailAddress ?EMAIL } .
        OPTIONAL { ?URI travel:hasStandardRoomPrice ?PRICE } .
        OPTIONAL { ?URI travel:hasFaxNumber ?FAX } .
        OPTIONAL { ?URI travel:hasPhoneNumber ?PHONE }
}
```

Fig. 3. A SPARQL example used in iJADE FreeWalker

4 Semantic Web Technologies for Tourist Information

Artificial Intelligence (AI) deals with reasoning about models of the world. In AI, ontology is usually defined as an explicit specification of conceptualization [2, 6]. In general, ontology refers to the description of the concepts and relationships that can exist for an agent or a community of agents. The defined ontology is usually in some formal and preferably machine-readable manner. Since the ontology could be reused, so it could save more time and cost for the development.

In this paper, the class or concept of the ontology is *travel*. In general, a class is a collection of elements with similar properties. Classes constitute a taxonomic hierarchy (a subclass-super class hierarchy). A class hierarchy is usually an IS-A hierarchy. For example, Chinese Cuisine IS-A subclass of Food (super class). We employed Protégé, an ontology editor, to create the ontology [7]. We defined an upper-level *travel* ontology which expressed in OWL (Web Ontology Language) [8]. OWL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. Figure 4 shows a simple definition of the concept of *travel* in OWL. Figure 5 shows an instance of *travel*.

```

...
<owl:Class rdf:ID="Custom_Tailor">
  <rdfs:subClassOf rdf:resource="#Shop_Type"/>
</owl:Class>
<owl:Class rdf:ID="Useful_Telephone_Number">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="General_Information"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="Handbags_Shoes_and_Leather_Goods">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="Shopping"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="Hotel">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="Accommodation"/>
  </rdfs:subClassOf>
</owl:Class>
<owl:ObjectProperty rdf:ID="hasCuisine">
  <rdfs:range rdf:resource="#Cuisine_Type"/>

```

Fig. 4. Code snippet of travel ontology

```

<rdfs:domain>
  <owl:Class>
    <owl:unionOf rdf:parseType="Collection">
      <owl:Class rdf:about="#Chinese_Food"/>
      <owl:Class rdf:about="#Other_Asian_Food"/>
      <owl:Class rdf:about="#Western_Food"/>
      <owl:Class rdf:about="#Other"/>
    </owl:unionOf>
  </owl:Class>
</rdfs:domain>
</owl:ObjectProperty>

...

```

Fig. 4. (continued)

```

<Hotel rdf:ID="THE_MARCO_POLO_HONGKONG_HOTEL">
  <hasTelephoneNumber rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >2113 0088</hasTelephoneNumber>
  <hasFaxNumber rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >2113 0011</hasFaxNumber>
  <hasEmailAddress rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >hongkong@marcopolohotels.com </hasEmailAddress>
  <hasURL rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >www.marcopolohotels.com </hasURL>
  <hasDistrict rdf:resource="#Tsim_Sha_Tsui"/>
  <hasRoom rdf:datatype="http://www.w3.org/2001/XMLSchema#int"
  >710</hasRoom>
  <hasStar rdf:datatype="http://www.w3.org/2001/XMLSchema#float"
  >5.0</hasStar>
  <hasStreet rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >Canton Road</hasStreet>
  <hasName rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
  >The Marco Polo Hongkong Hotel</hasName>
  <hasAddress rdf:datatype="http://www.w3.org/2001/XMLSchema#string"

```

Fig. 5. An instance of a hotel

```

>Harbour City, 3 Canton Road, Tsimshatsui, Kowloon</hasAddress>
<hasStandardRoomPrice rdf:datatype="http://www.w3.org/2001/XMLSchema#float"
>1050.0</hasStandardRoomPrice>

</Hotel>

```

Fig. 5. (continued)

5 Conclusion and Future Work

In summary, we proposed an ontology-based tourist guiding system, iJADE FreeWalker. It integrates agent, GPS and Semantic Web technologies to provide location-aware tourist information. Tourists can access the nearby tourist information by a small handheld device with limited computation power and limited network bandwidth. We showed the details of the iJADE FreeWalker and we developed the prototype system. By using the mobile agent technology, it increased the efficiency and scalability of the system. The system could run in the handheld device such as PDA and pocket PC. It would provide the tourist information such as sightseeing, entertainment and restaurants for the tourists according to his/her geographical information. By using ontology context modeling, it would retrieve the relevant tourist information and further filter the irrelevant information.

In the future, we would like to extend iJADE FreeWalker to include:

- **Voice interface** - To extend the usability and interaction, we would like to use voice interface as the input for the system. The system uses Natural Language Processing (NLP) as the communication tool to mimic the human-like tourist guide. User can chat with the tourist guide and query the relevant tourist information by using voice interface.
- **Learn user preference** – To further increase the usability and functionality of the system, we would add a Neural Networks (NN) module to learn the user preferences on the tourist information.

Acknowledgement

This work was partially supported by the iJADE Projects B-Q569, A-PF74 and A-PG50 from the Hong Kong Polytechnic University.

References

- [1] iJADE, <http://www.ijadk.org>
- [2] S. Poslad, H. Laamanen, R. Malaka, A. Nick, P. Buckle, A. Zipl, "CRUMPET: creation of user-friendly mobile services personalised for tourism," The Second International Conference on 3G Mobile Communication Technologies, pp. 28-32, 2001

- [3] G. Abowd, C. Atkeson, J. Hong, S. Long, R. Kooper and Pinkerton, M, "Cyberguide: A mobile context-aware tour guide," ACM Wireless Networks, vol. 3, pp. 421-433, 1997.
- [4] K. Cheverst, N. Davies, K. Mitchell and A. Friday, "Experiences of developing and deploying a context-aware tourist guide: The GUIDE project," The sixth International Conference on Mobile Computing and Networking, pp. 20-31, 2000
- [5] J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne and K. Wilkinson, 'Jena: Implementing the Semantic Web Recommendations, ' Proceedings of the 13th international World Wide Web conference (WWW 2004), pp. 74-83, 2004
- [6] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Laboratory Technical Report KSL-01-05, 2001
- [7] Protégé, <http://protege.stanford.edu/>
- [8] OWL Web Ontology Language, <http://www.w3.org/TR/owl-ref/>

***i*JADE Content Management System (CMS) – An Intelligent Multi-agent Based Content Management System with Chaotic Copyright Protection Scheme**

Raymond S.T. Lee, Eddie C.L. Chan, and Raymond Y.W. Mak

The Department of Computing, The Hong Kong Polytechnic University,
Hung Hong, Kowloon, Hong Kong
{csstlee, cscclchan, csywmak}@comp.polyu.edu.hk

Abstract. In this paper, an Intelligent Content Management System with Chaotic Copyright Protection scheme called *i*JADE CMS is presented. *i*JADE CMS focuses on how web mining techniques can be effectively applied on Chinese content, with the integration of various AI techniques including intelligent agents, agent ontology and fuzzy logic based data mining scheme. Through the adoption of chaotic encryption scheme, *i*JADE CMS demonstrates how agent-based copyright protection can be successfully applied to digital media publishing industry. From the application perspective, *i*JADE™ CMS provides the state-of-the-art content management function by the integration of *i*JADE™ Technology with the Ontological Agent Technology. *i*JADE™ CMS assists user to organize the content in the most semantic way. Moreover, the web mining information retrieval method such as Term Frequency Times Inverse Document Frequency (TFIDF) scheme is adopted to mine the linguistic meaning of the information content.

1 Introduction

One of the most challenging problems in retrieving, categorizing, managing and reporting useful content to user is to find the most suitable articles for user to read by machines. The common interest among researchers working in diverse fields is motivated by our remarkable innate ability to study and report information (e.g. news, articles) as human. The current search engines provided by Google or Yahoo! on news retrieval do not have flexible categorization and every news article is non-standard which is difficult for reading. In this paper, an intelligent Content Management System with Chaotic Copyright Protection called *i*JADE CMS is presented. *i*JADE [1] (intelligent Java Agent Development Environment) provides an intelligent agent-based platform to support the implementation of various AI functionalities.

2 Literature Review

2.1 Chaos System

Chaos Theory holds a belief which is totally different from classical probability. Although chaos theory is the study of the world of dynamics – which is highly dynamical

and non-linear, it is not uncertain or a matter of chance. Strictly speaking, chaos theory believes in the world of determinism rather than probabilism. Chaos theory maintains that all the dynamics in the world no matter how complex they are; they can be (and must be) somehow modeled by certain chaotic motions deterministically. The fact is, whether we can find out these chaotic motions is another question. [14]

2.2 Agent Applications in Web Content Mining

Agents such as Harvest [8], FAQ-Finder [9], Information Manifold [10], OCCAM [11], and Parasite [12] rely either on pre-specified domain specific information about particular types of documents, or on hard-coded models of the information sources to retrieve and interpret documents. The Harvest system [13] relies on semi-structured documents to improve its ability to extract information. For example, it knows how to find author and title information in Latex documents and how to strip position information from postscript files. Harvest neither discovers new documents nor learns new models of document structure. Similarly, FAQ-Finder [9] extracts answers to frequently asked questions (FAQs) from FAQ files available on the web.

2.3 Web Page Ontology

Web page ontology can be defined in different ways depending on the objective of the ontology. Most of the web sources have its semantic meaning. Techniques for Ontology Generation, Ontology Mediation, Ontology Population and Reasoning from the Semantic Web have all been major areas of focus. Most web documents are organized in a content hierarchy, with more general nodes placed closer to the root of hierarchy. Each node is labeled by a set of keywords describing the content of documents that are placed in the node. Each document is described by a one-sentence summary including a hyperlink that points to the actual Web document located somewhere on the Web.

3 iJADE CMS Implementation

iJADE™ CMS consists of two main modules – the Front End System Module and Back End System Module. Figure 1 shows the system overview of the iJADE™ CMS System.

3.1 iJADE CMS – Front-End System Module

Basically, iJADE CMS front-end system consists of four kinds of agent. They are Search Agent, Category Agent, Update Agent and Report Agent. In the following section, we will focus on the Category Agent.

3.1.1 iJADE Category Agent

A stationary iJADE agent categorizes articles into different regions and categories by calculating similarity between web documents (TFIDF method) and article clusters. This agent has high degree of intelligence, as it increases the accuracy by calculating

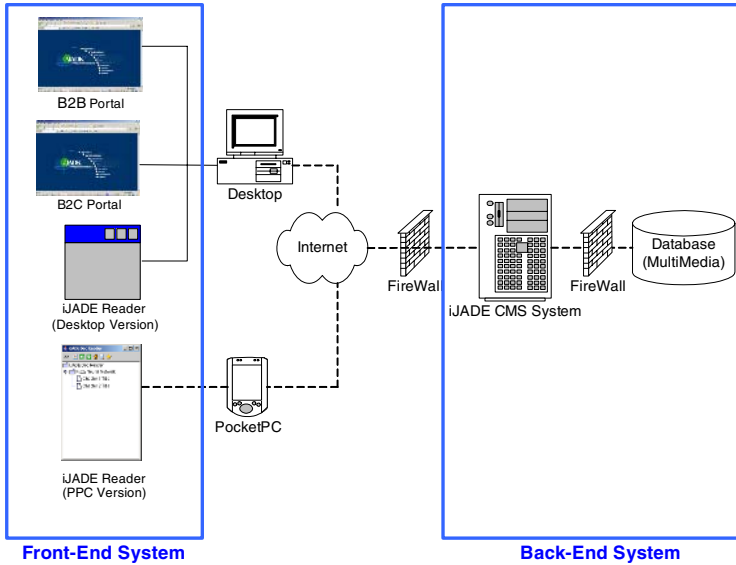


Fig. 1. iJADE CMS System Overview

similarity with many articles. Term Frequency times the Inverse Document Frequency (TFIDF) algorithm is adopted to query searching in web documents. TFIDF is a simple and powerful algorithm for machine to understand semantic document. TFIDF exhibits strong characteristics of word frequencies presented in a document. Vector Space Model (VSM) will be used to represent a document. Finally, by calculating the Euclidean distance of two vectors of two documents, the similarity can be computed.

Term Frequency times Inverse Document Frequency (TFIDF) Algorithm

TFIDF[2] is an information retrieval algorithm commonly used in query searching. The theory behind algorithm is to calculate a specific value of the semantic meaning among words and documents. TFIDF is very useful as it is simple and powerful to express the abstract idea of semantic meaning.

Vector Space Model (VSM) is typically used to represent Web documents. The documents constitute the whole vector space. **TFIDF** is being used as a weight of term in document. If a term t occurs in document d ,

$$w_{di} = tf_{di} \times \log(N / idf_{di}), [2]$$

where t_i is a word (or a term) in document collection, w_{di} is the weight of t_i , tf_{di} is term frequency (term count of each word in a document) of t_i , N is the number of total documents in the collection and idf_{di} is the number of document in which t_i appears.

Modified TFIDF by Ontology

In this section, we will introduce ontology-based term frequency (otf). Assume each word is related to other word, a relationship graph can be constructed to represent the relationship of all the words. [3]

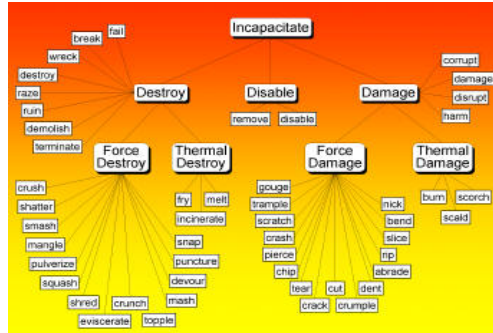


Fig. 2. The word graph example of “Destroy” and “Damage”

In Fig.2., “destroy” and “damage” are at the same level of hierarchal structure and they have relation between each other, so they are very similar. The similarity of two words can be measured by distance of tree structure. With comparison of meanings of two terms, the ontology-based term frequency can be obtained,

$$otf_i = tf_i \times (1 + (1/D(t_1, t_2))^{tf_2}),$$

where t_1, t_2 , are different terms; otf_i is ontology-based term frequency of t_i ; tf_1, tf_2 are term frequency respectively to t_1 and t_2 ; $D(t_1, t_2)$ is the depth between t_1 and t_2 . $D(t_1, t_2)$ can be calculated by WordNet[7].

For example, assume the term frequencies of “destroy” and “damage” are 3 and 2 respectively and depth between “destroy” and “damage” is 3. The ontology-based term frequency of “destroy” will be $otf_1 = 3 \times (1 + (1/3))^2 = 5.33$. The ontology-based term frequency of “damage” will be $otf_2 = 2 \times (1 + (1/3))^3 = 4.67$. After adjustment of each term frequency, their term frequency value could be increased, so that two terms will become more significant after computing TFIDF.

3.2 iJADE CMS – Back-End System Module

iJADE™ CMS back-end system consists of two main sub-modules – iJADE™ Content Management Module (iJCMM) and iJADE™ Copyright Protection Module (iJCPM). iJCMS and iJCPM not only allow user to read and search through all the digital media contents, they also provide a service portal for the management and copyright protection of the digital media contents such as news and magazines digital articles. The iJCMM consists of two main components - iJADE™ ContentOrganizer and iJADE™ ContentSeeker. The iJCPM consists of two main components - iJADE™ ContentReader and iJADE™ ContentProtector. Fig. 3 shows the System Model of iJCMM and iJCPM.

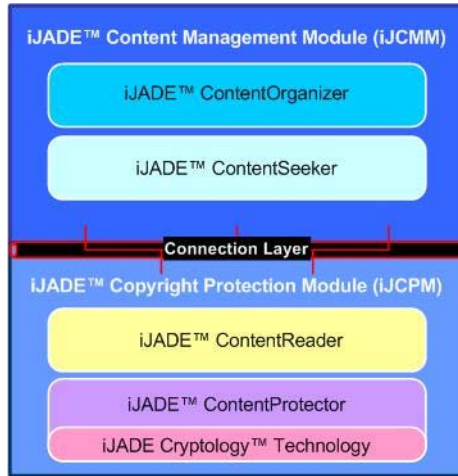


Fig. 3. The System Model of iJCMM and iJCPM

3.3 System Architecture of iJCMM and iJCPM

The System Architecture consists of six components. They are 1) Desktop Client, 2) Mobile Client, 3) iJADE™ Certification Authority – a center (iJADE Cryptology™ Server) to encrypt digital media contents, authorize agent and issue agent certificate, 4) iJADE™ Server – an Agent Pool that store different working agents, 5) iJCMM & iJCPM and 6) Internet. This System Architecture illustrates the total picture of how iJCMM & iJCPM work with users and the digital media contents.

The iJADE™ CMS allows desktop and mobile clients (installed iJADE™ ContentReader) to access digital media contents. Once the user request for the content, the

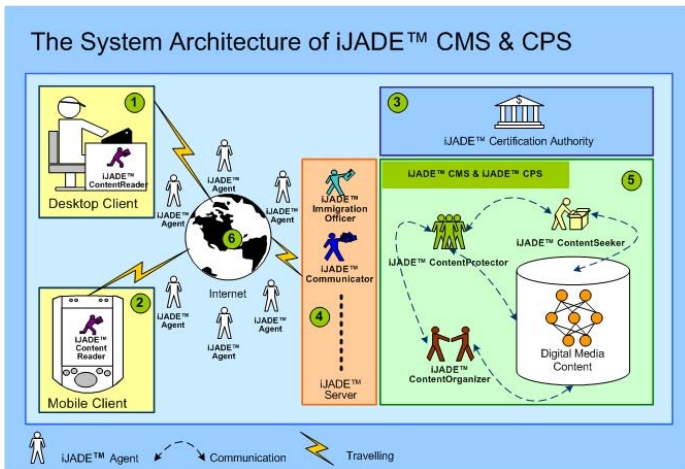


Fig. 4. The System Architecture of iJADE™ CMS & CPS

iJADE™ ContentReader Agent will dispatch itself to the iJCMM and collaborate with all those related agents (i.e. iJADE™ Transporter, iJADE™ Communicator and etc.) for retrieving the content. In order to avoid any kind of unauthorized access and ensure iJADE™ as a secure environment for the protection of the valuable digital media contents, the iJADE™ ImmigrationOfficer cooperates with the iJADE™ Certification Authority to authorize and certify all those incoming and outgoing agents before any kind of interactions and/or transactions are done. After that iJADE™ ContentSeeker will help to search for the contents and the iJADE™ ContentProtector will help to encrypt the contents by the adoption of iJADE Cryptology™ Technology. Finally, the iJADE™ ContentReader agent will carry the encrypted contents to the user via the Internet Channel.

4 Experimental Results

In this section, the precision rate of categorizing news is tested by the methods for categorizing news. The article database is used into comparison contains 500 records, in which the news is subdivided into six categories. There are 83 records for Business, Health, Education and Science/Nature each; while 84 records for Technology and Entertainment.

A test set of 100 articles without categorization is input to evaluate the performance of the models. There are 400 words in each article inside the content. To determine whether the article is categorized correctly or not, human judgment is used.

In Table 1 and 2, we find that by using iJADE CMS, the precision rate is 95.68% with clustering time 1185 seconds. As compared the iJADE CMS with other methods, such as neural network as well as TFIDF and hierarchical clustering, iJADE CMS gives a better precision rate for categorizing article and reasonable time taken for clustering.

Table 1. Comparison of the Precision Rate (%)

class/category	NN (3-Layer)	TFIDF+hierarchical clustering	iJADE CMS
Business	68.40%	90.30%	95.13%
Health	58.45%	93.37%	96.72%
Education	73.43%	93.88%	95.14%
Science/Nature	67.00%	94.33%	97.44%
Technology	70.22%	93.78%	94.38%
Entertainment	50.23%	91.67%	95.24%
Average	64.62%	93.72%	95.68%

Table 2. Comparison of time taken for clustering 3 different sets of total 100 news (seconds)

class/category	NN (3-Layer)	TFIDF+hierarchical clustering	iJADE CMS
Set 1	5277 seconds	1114 seconds	1175 seconds
Set 2	5432 seconds	1123 seconds	1186 seconds
Set 3	5175 seconds	1176 seconds	1195 seconds
Average	5295 seconds	1138 seconds	1185 seconds

5 Conclusion and Future Work

In this paper, an agent-based intelligent content management system – iJADE CMS is proposed. Experiments show that iJADE CMS produces very high accuracy result to categorize article in an effective and secure manner. It also provides a convenient and attractive environment for user to read digital articles. Future work can be focused on the user reading preference, query reconstruction and 3D Reporter broadcasting, which can be adopted to produce better quality of digital broadcasting.

Acknowledgment

This work was partially supported by the iJADE projects PG50 of the iJADE Project Development Group of the Hong Kong Polytechnic University.

References

- [1] [Online] iJADE, <http://www.ijadk.org>
- [2] Catherine W.Wen, Huan Liu, Wilson X. Wen and Jeffery Zheng; *A Distributed Hierarchical Clustering System for Web Mining*. WAIM2002, LNCS2118, pp. 103-113, Springer-Verlag Berlin Heidelberg, 2001
- [3] Everett, J. O.; Bobrow, D. G.; Stolle, R.; Crouch, R. S.; de Paiva, V.; Condoravdi, C.; van den Berg, M.; Polanyi, L. *Making ontologies work for resolving redundancies across documents*. *Communications of the ACM*. 2002 February; 45 (2): 55-60.
- [4] Jianwei Ham and Micheline Kamber: *Data Mining Concepts and Techniques*; Morgan Kaufmann Publishers; 2002
- [5] [Online] BBC News Website, <http://www.bbc.co.uk>
- [6] [Online] CNN News Website, <http://www.cnn.com>
- [7] [Online] WordNet, <http://www.wordnet.com>
- [8] C.M.Brown, B.B Danzig, "The harvest information discovery and access system.", In Proc. 2nd International World Wide Web Conference; 1994
- [9] R. B. Doorenbos, O. Etzioni, and D.S. Weld, "A scalable comparison shopping agent for the world wide web", Technical Report TR 97-027, University of Minnesota; 1997
- [10] A. Y. Levy, T. Kirk, and Y. Sagiv, *The information manifold*, presented at the AAAI Spring Symposium on Information Gathering From Heterogeneous Distributed Environments, P14, 1995.
- [11] C. Kwok and D. Weld, "Planning to gather information," in Proc. 14th Nat. Conf. AI, P15, 1996.
- [12] E. Spertus, "Parasite: Mining structural information on the web," presented at the Proc. 6th WWW Conf., P16, 1997.
- [13] O. Etzioni, D. S. Weld, and R. B. Doorenbos, "A Scalable Comparison - Shopping Agent for the World Wide Web," Univ. Washington, Dept. Comput. Sci., Seattle, Tech. Rep. TR, P17, 96-01-03, 1996.
- [14] Raymond S.T. Lee, "Advanced Paradigms in Artificial Intelligence", International Series on Natural and Artificial Intelligence Volume 5, P219, 2006.

Agent-Controlled Distributed Resource Sharing to Improve P2P File Exchanges in User Networks

J.C. Burguillo-Rial, E. Costa-Montenegro, F.J. González-Castaño,
and J. Vales-Alonso*

Departamento de Ingeniería Telemática, Universidad de Vigo, Spain
{jrial, kike, javier}@det.uvigo.es

*Departamento de Tecnología de la Información y las Comunicaciones,
Universidad Politécnica de Cartagena, Spain
Javier.Vales@upct.es

Abstract. In this paper, we evaluate the feasibility of distributed control of shared resources in user-managed networks. This paradigm has become possible with the advent of broadband wireless networking technologies such as IEEE 802.11. One of the most popular applications in these networks is peer-to-peer (P2P) file exchange. Node cooperation can optimize the usage of shared “external” accesses to the Internet (set of links between the user network and the Internet). In a previous paper, we evaluated different agent-oriented distributed control schema, based on the concept of *credit limits*, on ideal mesh networks subject to uniform traffic. Each node in the mesh network chooses to behave as a cooperator or a defector. Cooperators may assist in file exchange, whereas defectors try to get advantage of network resources without providing help in return. Before this paper was completed, we observed that popular P2P protocols like eMule, Kazaa and BitTorrent were evolving towards the same credit-oriented strategies we previously proposed. Now, we realistically model both user network traffic and topology, and evaluate a new advanced agent-based distributed control scheme. The simulation results in this paper confirm on realistic networks the main conclusion in our previous research: autonomous node agents become cooperators in their permanent state when they take decisions from local information, checking that file exchange services offered to neighbor nodes do not exceed appropriate credit limits.

Keywords: P2P, agents, network management, mesh networks.

1 Introduction

User-controlled networks have become possible with the advent of broadband wireless networking technologies such as IEEE 802.11. For applications like peer-to-peer (P2P) file exchange [9], it may be useful to consider the “external” access to the Internet (set of links between the user network and the Internet) as a shared resource whose usage can be optimized by node cooperation (i.e. if a node cannot serve its demand with its own external link, it requests help from another node via the high-bandwidth internal user network). The nodes decide if they grant cooperation requests

or not, from their own experience or from limited information on their neighbors' states.

It can be argued that sharing Internet accesses may be illegal. However, parasitism of operator resources is widespread and deserves academic attention. Consider, for example, the case of web caching [14], a P2P-boosting technique that exploits operator infrastructure, by mimicking non-P2P network services at port level.

In this paper, we analyze the conditions that enable cooperation in agent-controlled user networks. This is not evident in reality. In principle, it trivially holds that cooperation enhances overall network performance. However, a small population of intrinsically non-cooperative users may lead to massive defection, since node agents decide their course of action from *local* information (as real users do). Thus, the local decision algorithm of cooperation-prone nodes is the key issue to optimize the user network. Therefore, we have developed a framework to model the real problem and test agent algorithms allowing cooperator nodes with realistic behaviors to become a majority, which increases resource sharing and improves network performance.

We model resource sharing as a *game*, where nodes choose between two basic actions: cooperation and defection. Cooperators assist neighbor nodes in Internet file exchanges, while defectors do not (although they request help). Thus, our defectors correspond to “free-riding” P2P users [25].

This paper continues a previous work [13], where we showed that cooperation may emerge as the most popular strategy in ideal *mesh-like* cellular automata networks, even in case of autonomous distributed node operation.

Popular P2P protocols tend to apply the same technique we proposed in [13] as a safeguard against defectors (eMule [20], Kazaa [15], BitTorrent [16]): each node sets credit limits to its peers. Thus, our assumption in [13] is realistic.

The paper is organized as follows. Section 2 introduces user networks. Section 3 compares user networks and traditional P2P systems. Section 4 briefly introduces some basic concepts of game theory. Section 5 describes our models for network topology and traffic, which support the simulation model in section 6. Section 7 presents some simulation results, extended in section 8 to the scenario with adaptive credit limits. Finally, section 9 concludes.

2 User-Managed Networks

User-managed networks have become possible with the advent of wireless technologies such as IEEE 802.11 [4]. They represent an advanced stage in network control evolution [5]. This evolution started with telephone networks. In them, Telcos controlled both transport and applications. Later, the Internet allowed users to control applications. Although Telcos still control transport in most of the Internet, in some scenarios carrier technology is affordable to end-users, and user-managed “islands” appear in a natural way [6,7,8]. For example, these infrastructures are currently being used to provide broadband access in Spanish rural areas, from satellite IP gateways.

A typical node in a wireless user-managed network is composed by a router and an IEEE 802.11 access point (AP) and/or some IEEE 802.11 cards to set links to other nodes. Nodes may provide service to a multi-user LAN (covering a building, for example).

A subset of the nodes will have cable or DSL access, providing “external” connection to the Internet. We will assume that all nodes are “externally connected”. Additionally, we assume that user network capacity is larger than external access capacity (this holds for reasonable internal and external networking technologies, for example IEEE 802.11 and 1-Mbps DSL respectively), so that the internal network always has spare capacity. In a user network, nodes easily share contents, due to the large internal bandwidth. The bottleneck is the set of “external” connections. By optimizing their usage, overall performance (and, as a consequence, user satisfaction) can be greatly improved.

By *network stability* we refer to the condition such that external demands (e.g., downloads or uploads at the edges of the user network for P2P file exchanges) can be satisfied with external capacity, on average. This certainly holds if:

1. The external capacity of each node can satisfy its own demand, on average.
2. All nodes cooperate via the user network and their combined external demand can be satisfied with their combined external capacity, on average.

Even if the first condition holds (and therefore cooperation is not strictly necessary to guarantee network stability), cooperation minimizes demand service time (nodes with temporarily idle external connections can help neighbors with demand peaks). However, there is no central authority, and selfish nodes act from limited information on their neighbors' state, based on their own past experience. Cooperation grant/denial with limited information can be modeled as a *game*. The main goals of this paper are to demonstrate (1) that cooperator node players can be implemented as distributed node agents and (2) that they may become dominant in *real* user networks, thus improving resource sharing and, as a consequence, network performance. In our terminology, an *agent* is the program that implements the node behavior, adapting itself to its surrounding context, i.e. other node agents (we follow the same approach as in [19]).

3 User-Managed Networks and P2P

P2P applications, which include file-sharing protocols (such as eMule [20], Kazaa [15] or BitTorrent [21]), are interesting from many perspectives [22][23]. In user-controlled networks, cooperation can be a useful strategy to improve P2P performance. However, rational selfish peers refuse to serve others when they do not have clear incentives. As a consequence, the *tragedy of commons* [24] may arise, leading to generalized defection. Some of the difficulties to discover the resulting *free-riding* users [26] in P2P systems are:

- **Large population and high turnover:** P2P protocols have many clients, often short-lived.
- **Asymmetry of interest:** transactions lack reciprocity.
- **Zero-cost identity:** nodes can freely switch identities.
- **Lack of history:** it is unfeasible for a node to store the identities of all its past peers.
- **Unawareness of others:** it is impossible to know the transactions of all peers.

These conditions are the main cause of free-riding in P2P systems. The incentive mechanisms to fight them belong to two categories: token-based schemes [27] and reputation-based schemes [28]. In the former, the management of system-specific currencies becomes complex as the network grows and it is necessary to adopt a centralized secure credit authority. The later is based on shared transaction history, but it is vulnerable to collusion attacks (when several defectors claim to have received service from other defectors). The MaxFlow algorithm in [26] solves this problem but it is quite expensive and requires extra information to construct the graph.

On one hand, P2P user networks share some problems with general P2P systems: large population, asymmetry of interests and unawareness of others. On the other, fixed nodes in user networks may not experiment high turnover, zero-cost identity and lack of history. However, these problems may appear in *mobile* user networks with dynamic connection establishment. We will not consider mobile networks in this paper and leave them for future work.

4 Game Theory Basics

Game Theory [10] provides useful mathematical tools to understand the strategies of selfish agents. The simplest type of game is the single-shot simultaneous-move game. In it, all agents must choose one action and all actions are effectively simultaneous. Each agent receives a utility that is a function of the combined set of actions. In an extended-form game, agents participate in turns and receive a payoff at the end of a cycle of actions. In general, a single-shot game is a good model for many distributed systems where encounters require coordination.

Cooperative games and cooperation evolution have been extensively studied in biological, social and ecological contexts [1], seeking general theoretical frameworks like the Prisoner's Dilemma (PD). In his seminal work, Axelrod showed that cooperation can emerge in a society of individuals with selfish motivations [2]. For a review of related work in the last twenty years see [3].

Game Theory and the Generalized Prisoner's Dilemma have been applied to solve incentive problems in P2P systems. Examples can be found in [23] and BitTorrent [16] itself considers an alternative of the Tit-for-Tat strategy [2]. BitTorrent proposes a Bit-for-Bit incentive mechanism, where peers receive as much as they contribute. In a simulation environment with many repeated games, persistent identities and no collusion, Axelrod [2] shows that the Tit-for-Tat strategy is dominant. This strategy is simple, but it degrades the performance of the whole system if peers with asymmetric network bandwidth are present.

The approach we follow in this paper is a composite spatial game where actions are effectively simultaneous but each agent may interact with several neighbors at a time. Every agent receives a data throughput payoff every turn. The better the strategy the better the payoff is, i.e. the higher the throughput is. Note that the payoff of a given agent depends on the choices the rest take. At each turn (24 hours = one turn), agent i chooses a strategy $s_i \in S$, where S is the set of strategies available. The agent follows that strategy to interact with its neighbors until the next day. We consider two basic strategies: defection and cooperation.

5 Considerations on User Network Topology and Traffic

Once we have checked that our assumption on P2P client interaction based on credit limits is realistic (from a review of well-known P2P protocols), we describe our model of a realistic user network topology and its P2P traffic. For the former, we employ the IEEE 802.11 user network deployment algorithm in [17]. Figure 1 shows a resulting user network in Vigo (Spain), with 46 node locations and the corresponding wireless links for near-maximum network bandwidth (AP channel cellular planning as in [12], for 50×50 m² cells). We use it in the simulations in section 7.

Regarding traffic, we model the elapsed times between node demands of P2P chunk transfers (either incoming or outgoing) by means of Pareto distributions, following the well-known result in [18]. The exact settings are provided in section 7.

6 Distributed Control Strategies

If we let every agent in the system to interact with the other $N-1$ agents, we have a panmictic population. However, in this paper we are mainly interested in the spatial effects of the game, because in real user networks each node interacts with a few neighbors. Therefore, we consider that each agent i only interacts with the K agents in its immediate neighborhood.

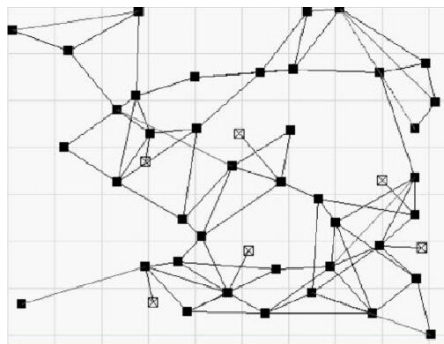


Fig. 1. “Vigo” user network

Interaction is driven by demand service times of externally inbound or outbound data, i.e. when the external traffic queue of a node is larger than a particular threshold, the agent controlling the node contacts its neighbors requesting help to handle the file chunks involved. Our time unit to generate new traffic demands and to interact with the neighbors is an hour. Thus, the total number of interactions per hour is up to $K \times N$.

We assume that agents keep the same strategy during a 24-hour timeframe, since hourly traffic patterns are similar across different days. Once a day passes, s_i can change as described next for two different control scenarios:

- *Shared information scenario (SI)*. This scenario assumes that every agent knows the state of its K neighbors. It can be considered as an *ideal* reference for fully distributed scenarios, such as the next one (in other words, it helps to determine if cooperation is feasible at all or not). The agent in node i mimics the strategy of its “best” neighbor k whose agent provided i the best worst-case hourly throughput during the previous day (out of 24 measurements). We decided to compare hourly rather than daily throughput for user satisfaction to be based on a compromise between worst case and average experience.

At the end of every day d , agent i calculates $x^* = \operatorname{argmax}_{x \in K} (\min_{h \in d} (\operatorname{th}(h, x)))$, where $\operatorname{th}(h, x)$ is the throughput of neighbor x during hour h . Then, agent i mimics the strategy of agent x^* . Obviously, it keeps its previous behavior if the most successful agent also followed it.

- *Fully distributed scenario (FD)*. **This is the real scenario we are interested in.** Our approach is inspired by [19]. Agent i keeps a state vector whose component $NT_i(s)$ is the number of times the agent followed strategy s in the past. We define the daily actualization of the efficiency estimator of strategy s as:

$$EE_i(s) := W f(i, d) + (1 - W) EE_i(s),$$

where $f(i, d) = \min_{h \in d} (\operatorname{th}(h, i))$ represents the minimum throughput achieved by agent i during day d . Parameter W results from:

$$W = Z + (1 - Z) / NT_i(s),$$

where Z is a real-valued constant. We set ($Z=0.3$) from [19]. The term $(1-Z)/NT_i(s)$ is a correcting factor, which is only significant for low $NT_i(s)$ values. As $NT_i(s)$ grows, this term becomes negligible compared to Z . To select the strategy for the next day we need a probability distribution. Initially, we force every agent to test every possible strategy at least once. Then, to determine the probability to choose strategy s the following day, we perform:

$$\operatorname{Prob}_i(s) = EE_i(s)^\alpha / \sum_s EE_i(s)^\alpha,$$

where α is a positive real constant. This biases the selection towards good past strategies. The weight of the bias depends on α ; the larger its value, the stronger the bias is. For a high α (e.g. $\alpha > 20$), the agent will always choose the strategy with the best record. But, as explained in [26], this option does not let the agent explore other strategies in changing contexts. Therefore we set $\alpha = 5$.

Each hour, if the queue length of a node exceeds a defined threshold, the agent ruling that node requests help from its neighbors for every pending file chunk (incoming or outgoing). Neighbor agents may grant their external connections or not, depending on their present state and strategy. We implement help transactions using the Contract Net Protocol: neighbors may answer with offers or refusals. The requesting agent selects the offering node that provided the best average throughput in the past.

We model the two strategies in both scenarios, *defection* and *cooperation*, as follows:

- *Defection*: a defector never helps, so it will never grant its external connection. Nevertheless, defectors ask their neighbors for help when they need it. Thus, they use shared resources opportunistically as many users do in P2P networks.

- *Cooperation*: a cooperator helps neighbors up to a certain credit limit. If node j reaches its credit limit, node i does not help j again unless j returns its debt by helping i when requested.

Finally, concerning node demand distribution, we define two node types, A and B . A nodes are busy during the day and quiet during the night. In B nodes activity is the opposite.

7 Simulation Results

We set system parameters as follows:

- Type A nodes generate maximum-activity traffic between 12 PM and 12 AM, and minimum-activity traffic otherwise. Type B nodes have opposite maximum and minimum activity timeframes.
- All nodes have an external 512 Kbps DSL access, i.e. 18 Gbph.
- P2P chunk size is 512 KB (4 Mb).
- During maximum and minimum activity timeframes, there are respective average node demands of 2880 Mbph and 360 Mbph. Thus, the aggregated external access is highly loaded (~90%).
- Pareto distribution settings for elapsed times between successive chunk demands (inbound or outbound) are $a=2.5$ $b=3.0$ $c=0$ for maximum activity timeframes (average=5 s, standard deviation= 4.5 s) and $a=4.0$ $b=30.0$ $c=0$ (average=40 s, standard deviation=14.14 s) for minimum activity timeframes.

From these parameter values, table 1 shows samples of the percentages of cooperators in permanent state vs. credit limit values (in Mb). The plots are concave in all cases.

We observe the following:

- In ideal networks, SI and FD behave as in [13], i.e. a concave plot that reaches a peak for a certain credit limit setting. Also, the more the information on neighbor nodes (SI), the higher the resource sharing.
- In both scenarios, the permanent-state percentages of both strategies are similar for near-zero credit limits. This is logical, since a cooperator with zero credit limit is equivalent to a defector.
- As credit limit grows, defectors tend to win.

8 Simulation Results with Adaptive Credit Limit

In the previous section all cooperators apply the same credit limit. However, in a realistic multiagent approach every node learns the best credit limit from its context, i.e. its surrounding neighbors and its interaction with them.

Agents cannot exhaustively explore the space of credit limit values, since a delay in choosing a reasonable value could degrade throughput. We think that genetic algorithms are simpler and perform much better in this case than more sophisticated optimization techniques [29]. We employ the following genetic algorithm, given the

Table 1. Cooperators in permanent state

Scenario	Credit limit	Cooperators in permanent state (%)
SI	0	48%
SI	180	61%
SI	1800	72%
SI	7200	49%
SI	36000	42%
FD	0	48%
FD	180	57%
FD	1800	53%
FD	7200	53%
FD	36000	46%

fact that, after some simulation time, the best credit limit interval was [1000, 5000] in all cases:

0. Each agent takes one random sample in each of the following intervals: [1000,2000], [2000,4000], [4000,5000]
1. Initially, the agent chooses the pair of best credit limit values ($CL1$, $CL2$) within the samples in step 0. They become the *parents*.
2. The resulting newborn credit limit $CL3$ is $CL1 + (1 - \text{rand}(0,1)) CL2$.
3. Mutation: IF ($\text{rand}(0,1) < 12 / \text{elapsed hour}$) THEN $CL3 = CL3 + \text{rand}(-5,5)$
4. If $CL3$ is better than $CL1$ and $CL2$ then the worst parent is replaced by $CL3$.
5. Return to step 2.

Note that we apply mutation with a decreasing probability depending on the simulated hour (the first check takes place after 24 simulated hours, so the probability is always lower than 1).

Using this credit limit adaptation in cooperator nodes, we obtain the results in table 2, which shows the mean and standard deviation of the credit limits after 10 simulation runs. The cooperators learn credit limits that confirm the results in table 1, as well as our previous results with cellular automata in [13].

We observe the following:

- SI gets better results than FD, but SI is unrealistic because nodes share state information.
- In the realistic distributed FD scenario, results improve by letting nodes learn their credit limits. The poor results in table 1 seem due to the fact that a global credit limit may be inadequate for some cooperator nodes.

Figure 2 shows the evolution of cooperators and defectors in the SI scenario. It represents a temporal window of 100 days. The number of cooperators increases and the percentages become stable after 23 days.

Figure 3 shows the evolution of cooperators and defectors in the FD scenario. We observe initial oscillations until the nodes determine their best strategy and credit limit values. After some time (58 days), cooperation emerges as the preferred strategy.

Table 2. Cooperators in permanent state with adaptive credit limit

Scenario	Avg. credit limit	Std. dev. of credit limit	Cooperators in permanent state (%)
SI	1698	309.1	82%
FD	2818	1084.2	71%



Fig. 2. SI scenario with adaptive credit limit. Cooperators win. The x axis represents time in seconds. Percentages become stable after 23 days.

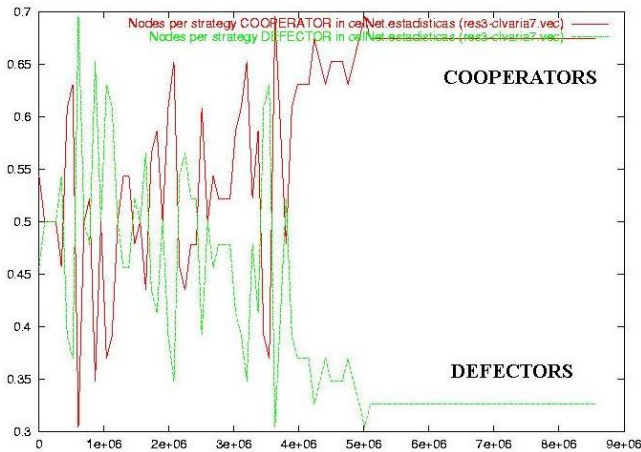


Fig. 3. FD scenario with adaptive credit limit. Cooperators win. The x axis represents time in seconds. Percentages become stable after 58 days.

9 Conclusions

The approach in this paper is an abstraction of the complex problem of negotiation and resource sharing in user networks. We consider it reflects the main features in a real user network, and we show that, if nodes learn their credit limits and adapt their strategy to their context, cooperation emerges as the preferred strategy even in case of fully distributed node operation. We also show that node behavior in fully distributed scenarios varies with credit limit values as in shared information scenarios, although they do not necessarily reach the same level of cooperation.

We observe that, in user networks, node cooperation arises if traffic demands are strongly variable during the day, for example in case of several node types with complementary activity timeframes.

Now that cooperation can be considered a stable strategy in realistic agent-managed user networks, we plan to evaluate the distribution of cooperators and defectors under different workload conditions and more complex control scenarios. We also plan to extend the model to mobile user networks with dynamic connection establishment.

References

1. F. Schweitzer, J. Zimmermann, H. Muhlenbein: Coordination of decisions in a spatial agent model. *Physica A*, 303(1-2), pp. 189-216, 2002.
2. R. Axelrod: *The evolution of Cooperation*. Basic Books, New York, 1984.
3. R. Hoffmann: Twenty years on: The evolution of cooperation revisited. *Journal of Artificial Societies and Social Simulation*, 3(2), 2000.
4. IEEE 802.11. <http://grouper.ieee.org/groups/802/11/>
5. J.P. Hubaux, T. Gross, J.Y.L. Boudec, M. Vetterli: Towards self-organized mobile ad-hoc networks: the terminodes project. *IEEE Commun. Mag.*, 1, pp. 118-124, 2001.
6. Madrid Wireless. 2004. <http://madridwireless.net>.
7. Wireless Athens Group. 2004. <http://www.nmi.uga.edu/research>.
8. N. Negroponte: *Being Wireless*. *Wired Magazine*, 10.10, Oct. 2002.
9. Kazaa news. 2004. <http://www.kazaa.com/us/news/index.htm>.
10. Ken Binmore: *Game Theory*. Mc Graw Hill, 1994.
11. Narendra: *Learning Automata*. Prentice Hall, 1989.
12. F. Box: A heuristic technique for assigning frequencies to mobile radio nets, *IEEE Trans. Veh. Technol.*, VT-27, pp. 57-74, 1978.
13. J.C. Burguillo-Rial, F.J. González-Castaño, E. Costa-Montenegro, J. Vales-Alonso: Agent-Driven Resource Optimization in User Networks: a Game Theoretical Approach. *Lecture Notes in Computer Science (LNCS)*, 3305, pp. 335-344, 2004. *Proc. 6th Intl. Conf. on Cellular Automata for Research and Industry, ACRI 2004*.
14. eMule-Board. <http://forum.emule-project.net/index.php?showtopic=61326&hl=webcache>.
15. Kazaa participation ratio. http://www.kazaa.com/us/help/glossary/participation_ratio.htm.
16. K. Tamilmani, V. Pai, A. Mohr: SWIFT: A System With Incentives For Trading. *Proc. Second Workshop of Economics in Peer-to-Peer Systems*, 2003.
17. F. J. González-Castaño, E. Costa-Montenegro, U. García-Palomares, M. Vilas Paz, P. S. Rodríguez Hernández: Distributed and Centralized Algorithms for Large-Scale IEEE 802.11b Infrastructure Planning. *Proc. Ninth IEEE International Symposium on Computers & Communications (ISCC 2004)*.

18. W.E. Leland et al: On the Self-Similar Nature of Ethernet Traffic. *IEEE/ACM Transactions on Networking*, 1994.
19. A. Schaerf, Y. Shoham, M. Tennenholtz: Adaptive Load Balancing: A Study in Multi-Agent Learning. *Journal of Artificial Intelligence Research*, 2, pp. 475-500, 1995.
20. Y. Kulbak, D. Bickson: The eMule Protocol Specification. http://leibniz.cs.huji.ac.il/tr/acc/2005/HUJI-CSE-LTR-2005-3_emule.pdf
21. The official BitTorrent page. <http://www.bittorrent.com>
22. B. Gu, S. Jarvenpaa: Are Contributions to P2P Technical Forums Private or Public Goods?- An Empirical Investigation. In 1st Workshop on Economics of Peer-to-Peer Systems, 2003.
23. M. Castro, P. Druschel, A. Ganesh, A. Rowstron, D.S. Wallach: Security for Structured Peer-to-Peer Overlay Networks. In *Proc. of Multimedia Computing and Networking*, 2002.
24. G. Hardin: The Tragedy of the Commons. *Science*, 162, pp. 1243-1248, 1968.
25. E. Adar, B. A. Huberman: Free riding on Gnutella. Technical report, Xerox PARC, Aug. 10 2002.
26. M. Feldman, K. Lai, I. Stoica, J. Chuang: Robust Incentive Techniques for Peer-to-Peer Networks. *ACM E-Commerce Conference (EC'04)*, 2004.
27. P. Golle, K. Leyton-Brown, I. Mironov: Incentives for Sharing in Peer-to-Peer Networks. In *ACM Conference on Electronic Commerce*, 2001.
28. Y.-H. Chu, J. Chuang, H. Zhang: A Case for Taxation in Peer-to-Peer Streaming Broadcast. *ACM SIGCOMM Workshop on Practice and Theory of Incentives and Game Theory in Networked Systems (PINS)*, August, 2004.
29. U.M. García-Palomares, F.J. González-Castaño, J.C. Burguillo-Rial: A Combined Global & Local Search (CGLS) Approach to Global Optimization. *Journal of Global Optimization* (in press).

An Agent-Based System Supporting Collaborative Product Design

Jian Xun Wang and Ming Xi Tang

Design Technology Research Centre, School of Design, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
{Jianxun.wang, sdtang}@Polyu.edu.hk

Abstract. Collaborative design can create added value in the design and production process by bringing the benefit of team work and cooperation in a concurrent and coordinated manner. However, the difficulties arising from the differences between heterogeneous system architectures and information structures undermine the effectiveness and the success of collaborative design. This paper presents a new framework of collaborative design which adopts an agent-based approach and relocates designers, managers, systems, and supporting agents in a unified knowledge representation scheme for product design. An agent-based system capable of assisting designers in the management and coordination of collaborative design process via the cooperation of a network of agents including mainly management agents, communication agents, and design agents etc is described.

1 Introduction

Increasing worldwide competition and rapidly changing customer's demands are forcing manufacturers to produce high quality and innovative products with shortened Time-to-Market and reduced production cost. An ideal product design environment which is both collaborative and intelligent must enable designers and manufacturers to respond quickly to these commercial market pressures [1].

Recently, agent technology has been recognized by more and more researchers as a promising approach to analyzing, designing, and implementing industrial distributed systems. An intelligent agent consists of self-contained knowledge-based systems capable of perceiving, reasoning, adapting, learning, cooperating, and delegating in a dynamic environment to tackle specialist problems. The way in which intelligent software agents residing in a multi-agent system interact and cooperate with one another to achieve a common goal is similar to the way that human designers collaborate with each other to carry out a product design project. Thus, we believe that a collaborative product design environment implemented by taking an agent-based approach will be capable of assisting human designers or design teams effectively and efficiently in product design.

This paper presents a new framework of collaborative design which relocates designers, managers, systems, and supporting agents in a unified knowledge representation scheme for product design, and describes an agent-based system capable of assisting designers in the management and coordination of collaborative design process

via the cooperation of a network of agents including mainly management agents, communication agents, and design agents etc.

This paper is organized as follows: Section 2 gives an overview of the work related to our research. The characteristics of a collaborative product design process and the requirements for a collaborative design system are examined in Section 3. Then, we give an overview of the system architecture of our proposed agent-based system in Section 4. Section 5 introduces the implementation strategy of the agent-based system. A number of concluding remarks are made in Section 6.

2 Related Works

The development of computational agents with unified data structures and software protocols can contribute to the establishment of a new way of working in collaborative design, which is increasingly becoming an international practice. Parunak examined the areas related to design activities where agent technology can be best employed and argued that agents were uniquely suited to addressing the problems characterized by modularity, decentralization, changeability, poor structure, and high complexity [2]. Shen et al. stated that the appropriate use of agents could result in desired modularity, allowing flexible simulations, and in better response and improved software reusability which an ideal collaborative design environment should exhibit [3].

As an emergent approach to developing distributed systems, agent technology has been employed to develop collaborative product design systems by a number of researchers. Their researches were generally concerned with three aspects: product modelling, consistency maintenance, and system architecture [4]. Intelligent software agents have mostly been used to enable cooperation among designers, to provide wrappers for integrating legacy software tools, or to allow better simulations [5]. One of the earliest projects in this area is PACT which demonstrates the use of agents to combine pre-existing engineering systems to constitute a common framework, using facilitators and wrappers, by adopting a federated architecture [6]. DIDE is developed to achieve collaboration and openness in an engineering design environment through the integration of CAD/CAM tools, databases, knowledge base systems, and etc [7]. SHARE project conducted at Stanford University, USA, including First-Link [8], Next-Link [9], and Process-Link [10], aimed at using agents to help multidisciplinary design engineers track and coordinate their design decisions with each other in the concurrent design of aircraft cable harnesses with the support of a Redux agent. Liu and Tang presented a multi-agent design environment that supported cooperative and evolutionary design by cooperation of a group of agents [11].

Although agent technology has been considered very promising for developing collaborative design systems, and most of the systems that have been implemented so far are domain dependent, and were intended for integrating legacy design tools. Such systems are still at a proof-of-the-concept prototype development stage. Furthermore, only a few literatures mentioned design process model for supporting dynamic design project management in a multi-agent collaborative design system. A detailed discussion on the issues and challenges in developing multi-agent design systems can be found in literature [12].

3 Requirements for a Collaborative Product Design System

Design is a complex knowledge discovery process in which information and knowledge of diverse sources are processed simultaneously by a team of designers involved in the life phases of a product (Tang 1991). The investigation into the development of large complex product which requires close cooperation among multidisciplinary designers reveals that a collaborative product design process is characterised by the following features:

- People working on the same product design project often appear to be temporally and spatially distributed.
- Collaborative design processes usually last for a long time. The requirements on the product may need to be changed during the product development process.
- Sharing and reusing various kinds of knowledge involved in the product life-cycle is common.
- Hardware and software systems for collaborative design are often heterogeneous (they may have different system architectures or information structures).
- During product design process, a lot of decisions are taken by each participant involved. These decisions influence each other and are interdependent of each other.

These features result in the identification of certain requirements for a collaborative product design system:

- The product design management functionality is required to facilitate design process planning, implementation, and coordination.
- Multi-modal communication tools, such as message, video conference, white-board, application sharing, and etc., should be provided to facilitate the cooperation among distributed designers.
- Mechanisms for design coordination and constraint propagation are required and the designers should be automatically notified with any constraint conflict as well as any design change being made by others.
- The system should support the easy integration of heterogeneous software used by designers and support the flow of information in the distributed environment.
- Popular CAD software should be integrated and Human Computer Interaction (HCI) should be user-friendly, so that the extra burden imposed on designers to get familiar with the collaborative design environment can be kept as low as possible.
- Product data management tools are needed to accelerate product development cycles by facilitating product data reuse and sharing, and protecting it from inadvertent changes.

4 The Agent-Based Collaborative Product Design System Architecture

Based on the analysis of the requirements for a collaborative product design environment above, we propose an agent-based system to facilitate, rather than automate,

teamwork in a collaborative product design project. It is organised as a network of intelligent agents which interact with each other and participating team members to facilitate collaborative design work. These agents include Project Management Agent, Design Agents, Product Data Management Agent, Design Coordination Agent, Design Communication Agent, Computational Support Agent, and Agent Manager.

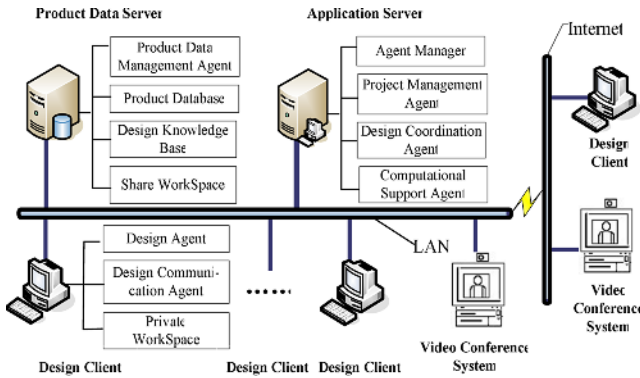


Fig. 1. General system architecture and hardware configuration

The system architecture and hardware configuration are illustrated in Fig. 1. The architecture integrates all the agents with design and engineering tools/services including product database, design knowledge base, share/private workspace, and video conference system, to interact with human designers in an open environment. Next, the main agents being developed are described briefly.

- **Project Management Agent** is responsible of helping project manager to describe the design project, specify the initial parameters and design constraints, decompose a complex design project into sub tasks, assign them to designers, schedule them and track the whole design process. It has a user interface for assisting project manager in managing the project plan and schedule, keeping track of progress of the project, and more importantly, making necessary changes while the plan gets more detailed and the higher levels of the plan have to be updated. Also, it serves to remind the project manager and involved designers of the outstanding issues related to schedule execution.
- **Design Agents** encapsulate some traditional popular CAD systems and are used by designers to fulfil design tasks through cooperation with other agents. There have been some standards, such as CORBA, DOM, allowing for legacy applications to be encapsulated using wrapper classes and to behave as distributed components.
- **Product Data Management Agent** is responsible for managing the product database and ensuring the consistency of the product data in the product design process, and informing concerned design agents of the product data change event (such as submission, modification and so on) made by other design agents.

- **Design Coordination Agent** serves to coordinate the design process through design constraint propagation, notifying involved designers of constraint conflicts and their respective reason, and helping designers address the conflicts.
- **Design Communication Agent** provides support for interaction among designers by services, such as email, video/audio conferencing, file transfer service, application sharing, whiteboard, and etc or coordination message transporting service among other agents and human designers.
- **Computational Support Agent** accounts for engineering calculations or analysis upon request. It may be a Finite Element Analysis tool, Optimiser of some other engineering computation tools.
- **Agent Manager** is a mandatory component of the system and acts as a directory facilitator to all other agents in the system. It is responsible for controlling the utilization and availability of all agents by maintaining an accurate, complete and timely list of all active agents through which agents residing in the system are capable of cooperating and communicating with each other.

All the agents are connected by a local network (LAN) via which they communicate with each other. Also, the external design agents residing on the internet carrying out specific design tasks can communicate with agents located in the local network via the Internet.

5 Implementation of the System Prototype

Our agent-based system supporting collaborative product design is now being implemented on a network of PCs with Windows 2000/XP and Linux operating systems. Java is chosen as the primary programming language for the system implementation. C++ is also used as the programming language for the integration of legacy systems. JNI (Java Native Interface) serves as the glue between Java-written applications and C++-wrapped native applications. JADE¹ (Java Agent DEvelopment Framework) which is a software framework fully implemented in Java language is utilised to develop the agent-based system. Currently, Inventor[®], as one of the most popular CAD systems, is being encapsulated into our design agents through Inventor[®] COM API. FIPA ACL serves as the agent communication language. MySQL[™] is used by the product data management agent as the database system for storing product data and design knowledge. The entire prototype of the system will be implemented before mid 2006. Here, we only brief the agent manager, and the design project manager which has been implemented so far.

The main user interface of our agent manager (JADE agent runtime environment) is shown in Fig. 2. All the agents geographically distributed in the collaborative design environment can register themselves in the agent manager and their life cycles can be controlled through this interface. When an agent comes on-line, a message is sent to the agent manager, where the name, location, a description of the skills and capabilities of the agent are registered and monitored. Then, on any request from any other agent in the system, the agent manager makes this information available to other

¹ JADE is a framework to develop multi-agent systems in compliance with the FIPA (Foundation for Intelligent Physical Agents) specifications. (<http://jade.cselt.it>)

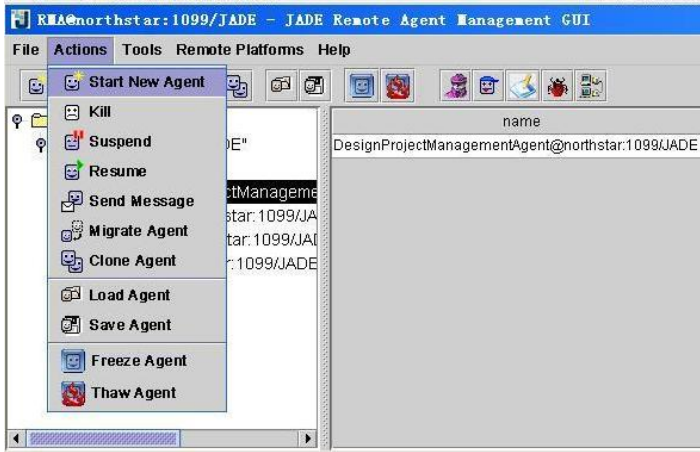


Fig. 2. The main user interface of the agent manager

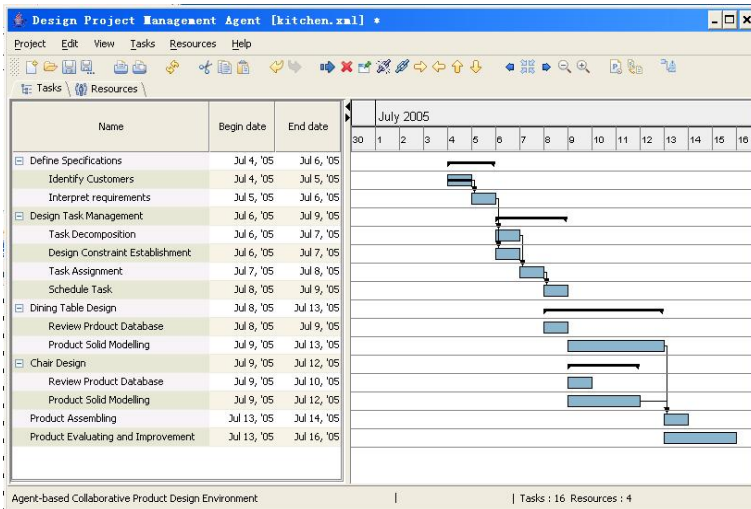


Fig. 3. The main user interface of the design project management agent

agents. When an agent becomes off-line whether or not unexpectedly, the agent manager must deregister it and then inform all other agents concerned.

The main user interface of the design project management agent is shown in Fig. 3. Through interaction with project management agent, a human design manager can describe and decompose a design project into a number of design tasks. Each design task is associated with a set of attributes including duration, start time, end time, assigned agent, inputs, and expected outputs, and etc. Then, these design tasks can be planned and scheduled by assigning them to design team members. A design task tree can be easily established and the relationship among design tasks can be visually

managed with the help of design project management agent. As shown in Fig. 3, a design project to design a dining table and chairs has been planned and scheduled.

6 Conclusions

In this paper, on the basis of the problem identification and the analysis of the requirements for a collaborative product design system, an agent-based system supporting collaborative product design to facilitate the management and coordination of collaborative product design process via the cooperation of a network of intelligent agents is presented. Current and future work focuses on the implementation of the proposed system. As the system is still being fully implemented, more experiments are required to be carried out in order to test and improve our system. In the future testing of our approach and software, we intend to involve designers in the process with real design examples since we believe that a design-oriented approach needs to be taken in order to identify those key tasks that need collaboration and support by software agents.

Acknowledgements

This project is supported by a PhD studentship from The Hong Kong Polytechnic University.

References

1. Frazer, J.H.: Design Workstation on the Future, Proceedings of the Fourth International Conference of Computer-Aided Industrial Design and Conceptual Design (CAID & CD 2001), International Academic Publishers, Beijing (2001) 17–23.
2. Parunak, H.V.D.: What can agents do in industry, and why? An overview of industrially-oriented R&D at CEC, In M. Klusch, and G. Weiss, (Eds.): Cooperative information agents II: Learning, mobility and electronic commerce for information discovery on the Internet: Second International Workshop, CIA'98, Springer, Paris, France (1998) 1–18.
3. Shen, W. and Wang, L.: Web-Based and Agent-Based Approaches for Collaborative Product Design: an Overview, International Journal of Computer Applications in Technology, Vol. 16(2/3) (2003) 103–112.
4. Rosenman, M.A., Wang, F.: A Component Agent Based Open CAD System for Collaborative Design, Automation in Construction, Vol. 10(4) (2001) 383–397.
5. Hao, Q., Shen, W., Park, S.-W., Lee, J.-K., Zhang, Z., Shin, B.-C.: An Agent-Based Engineering Services Framework for Engineering Design and Optimization, In Orchard R., et al. (Eds.): IEA/AIE 2004, LNAI 3029 (2004) 1016–1022.
6. Cutkosky, M.R., Engelmores, R.S., Fikes, R.E., Genesereth, M.R., Gruber, T.R., Mark, W.S., Tenenbaum, J.M., and Weber, J.C.: PACT: An experiment in integrating concurrent engineering systems, IEEE Computer, Vol. 26(1) (1993) 28–37.
7. Shen, W., Barthes, J.P.: An experimental environment for exchanging engineering design knowledge by cognitive agents, In Mantyla, M., Finger S., and Tomiyama T. (Eds.): Knowledge intensive CAD-2, Chapman and Hall (1997) 19–38.

8. Park, H., Cutkosky, M., Conru, A., Lee, S.H.: An Agent-Based Approach to Concurrent Cable Harness Design, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, Vol. 8(1) (1994) 45–62.
9. Petrie, C., Cutkosky, M., Webster, T., Conru, A., Park, H.: Next-Link: An Experiment in Coordination of Distributed Agents, *AID-94 Workshop on Conflict Resolution*, Lausanne (1994).
10. Goldmann, S.: Procura: A Project Management Model of Concurrent Planning and Design, *Proceeding of WET ICE'96*, Stanford, CA (1996).
11. Liu, H., Tang, M.X., Frazer, J.H.: Supporting evolution in a multi-agent cooperative design environment, *Advances in Engineering Software*, Vol. 33 (6) (2002) 319–328.
12. Lander, S.E.: Issues in Multiagent Design Systems, *IEEE Expert*, Vol. 12(2) (1997) 18–26.
13. Tang, M.X.: A Representation of Context for Computer Supported Collaborative Design, *Automation in Construction*, Vol. 10(6) (2001) 715–729.
14. Wang, J.X., Tang M.X.: Knowledge Representation in an Agent-Based Collaborative Product Design Environment, *Proceedings of the 9th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2005)*, Vol. 1, IEEE Computer Society Press, (2005) 423–428.

Application Presence Fingerprinting for NAT-Aware Router

Jun Bi, Lei Zhao, and Miao Zhang

Network Research Center, Tsinghua University
Beijing, P.R. China, 100084
junbi@cernet.edu.cn

Abstract. NAT-aware routers are required by ISPs to administrate network address translation and enhance the management and security of access network. In this paper, we propose the new fingerprinting for NAT-aware router based on application-level presence information, which is usually not easily modified by network address translation gateways or the fingerprinted hosts behind gateways.

1 Introduction

NAT (Network Address Translation) is currently widely used in the Internet to provide private address space. Network address translation brings flexibility of access network, and resolves the problem of IPv4 address space limitation. However, from the viewpoint of Internet Service Providers (ISP), unauthorized NAT also brings management and security issues for an access network:

(1) Breaks the end-to-end model of Internet.

ISPs need to know the real topology, which is hidden by NAT gateways. The deployment of p2p applications will be accelerated and the performance will be enhanced without NAT gateways.

(2) Brings security issues.

IP source address, which is an important identifier to trace end users, is changed by NAT gateways. Therefore it's hard to trace the real end user.

(3) Breaks the address based accounting model.

Some ISPs charge fixed monthly fee for each IP address, but group of users can pay for only one NAT gateway address.

Especially, when IPv6 provides enough address space for end users, the address space is no longer a critical issue to be resolved by network address translation. The NAT-aware router is an intelligent system in an edge router that can be aware of the unauthorized NAT in an access network, based on NAT fingerprinting. It passively filters packets from target hosts and quickly matches results against a database of known NAT fingerprinting. The NAT-aware router for detecting and monitoring of network address translation is becoming an important requirement for ISPs, such as China Education and Research Network (CERNET).

NAT gateway in some ways behaves similarly as a normal host, so it is hard for ISPs to accurately detect it. Some existing NAT fingerprinting approaches are based

on collection and analysis on network layer (such as IP layer) fingerprints. However, network layer fingerprinting might be easily defeated by modifying NAT gateways. Some NAT fingerprinting approaches are based on transport layer (such as TCP) fingerprints, which can be defeated by hosts behind NAT gateway in the way of modifying their TCP/IP stacks.

In this paper, we propose a new detection system based on application layer fingerprinting, which is usually not easily modified by NAT gateways or hosts masquerading behind NAT. The advantages of application layer fingerprinting are:

- (1) It's hard for NAT gateway to modify application layer related information to avoid detection.
- (2) Users of fingerprinted hosts are hard to acquire and modify the source code of network application to avoid detection.
- (3) Applications developers don't have direct incentive to design features to help avoiding NAT detections.

The remainder of the paper is organized as follows. Based on the discussion on problem statement of exiting NAT fingerprinting technologies in section 2, we present new application presence fingerprints in section 3. Section 4 presents the detection algorithm in NAT-aware router. Section 5 summarizes the paper.

2 Problem Statement

Almost every system connected to the internet is identifiable to fingerprinting [1]. NAT Fingerprinting is the process of determining identities and the number of remote physical hosts on one IP address by analyzing packets from that address. The existing fingerprinting for NAT detection proposed by researchers in this area can be summarized as the following sections.

2.1 MAC Fingerprinting

This method checks whether a packet coming from a target has a MAC address that belongs to a known NAT device vendor.

This method requires the detection device is located in the same link layer network with the target NAT gateway. This method can be defeated by changing the MAC address of NAT gateway to a normal address.

2.2 IP ID Fingerprinting

This method counts the number of hosts by building up IP ID arrays for packets coming from the same source IP address [2].

There are two possible ways to defeat this method: make modifications on NAT gateway to change IP ID as a simple counter or use constant IP ID in those packets without fragment flag; or make modification on hosts to use other IP ID generation algorithms.

2.3 Operating System Fingerprinting

Each operating system has a special implementation of TCP/IP stacking, such as the initial value of TTL (Time to Live), MSS (Maximum Segment Size), and DF (Don't Fragment Bit). Table 1 shows some example of operating system fingerprints. If two or more fingerprints can be found in packets coming from the same source address, then it means there are multiple hosts behind this address. The related detection methods can be found in [3][4][5].

Table 1. Example of operating system fingerprint database

Operating System	TTL	DF
Linux 2.4	64	1
FreeBSD 5.1	64	1
Windows XP(SP1)	128	1

Users can defeat this method by choosing the same operating system for fingerprinted hosts or by making modifications on TCP/IP stack parameters. For example, Linux kernel 2.4 users can use IP personality [6] to change characteristics of the packets.

2.4 Clock Skew Fingerprinting

This method counts the number of hosts behind NAT by partitioning packets into sets corresponding to different sequences of time-dependent TCP or ICMP timestamps and applying a clock skew estimation technique on the sets [7].

One possible way to defeat this method is to make modification on NAT gateway to delete the timestamp option in TCP SYN packets. Then both side of the TCP would not use TCP timestamp option any more and thus this method fails.

3 Application Presence Fingerprints

Some network applications are user-oriented and designed to be used by an individual on one host. Therefore, normally users run only one application instance on one host. If there is more than one presence information of such application detected on one source IP address, it is likely that there exists a NAT gateway on this IP address. Hereafter we use the following terms:

Definition 1: Presence Packet. Presence Packet denotes the packets which carry the application presence information.

Definition 2: Presence Channel. Presence Channel denotes the data channel between the client and server of an application, which transfers presence packets.

From the prevalent network applications, such as Web, Email, FTP, IM (Instant Messaging), etc., we choose IM as the application for detection, for the following reasons:

(1) Usually, only one instance of one type of IM application runs on one host. Some IM applications (e.g., Microsoft MSN Messenger) have the limitation that only one instance can run on one desktop simultaneously. It is also reasonable that people usually do not run two or more instances of each type of IM at the same time. Therefore, the number of instances of one application denotes the number of hosts.

(2) Popular IM applications (e.g., MSN Messenger, Yahoo Messenger and Google Talk) have a large population of users. IM users often keep IM clients running for a relatively long period. It makes the detector have more chances to detect presence fingerprints.

IM provides a presence service which allows users to subscribe to each other and to be notified the changes in presence state [8]. Most IM applications use one stable data channel (e.g., a TCP connection) between the IM client and the IM server to transfer notifications. Thus this stable data channel can be a sign of an instance of IM client. If we get the number of such data channels on one source IP address, we can deduce the number of IM instances (in another word, the number of hosts) behind this IP address, and it is probably that there is a NAT gateway on this IP address.

MSN Messenger uses MSN Messenger Service Protocol [9]. In a typical session, the client of MSN Messenger will connect to three different kinds of servers: Dispatch Server (DS), Notification Server (NS) and Switchboard Server (SS). All of the three types of servers use a registered port number 1863. The IM client sets up a TCP connection to the DS; then DS dispatches a NS to serve the client. Then the IM client sets up a TCP connection to the NS which is mainly used to transfer presence information. The TCP connection between the client and the NS is the presence channel. The client periodically sends a “PNG” command to NS. This command is used to ensure that the TCP connection to be alive. The command is in the payload field and the format is:

PNG\r\n

This presence channel will last for the whole log-on session. When the IM client needs to send instant messages or transfer files to other IM clients, it asks NS to dispatch a SS to it and then sets up a TCP connection to the assigned SS. The connection between IM client and SS will be closed when it is no longer needed.

Google Talk uses XMPP [10] protocol suite. Google Talk client sets up one TCP connection to the IM server to transfer instant messages and presence information. This TCP connection is started when a client logs in, and continues to exist throughout the whole session. Google Talk uses port 5222 on IM server as its service port number, which is compliant with the assignment in XMPP. The connection is the presence channel.

Based on observations, we find that IM applications usually have three characteristics that can be used for fingerprinting:

(1) The IP addresses of IM servers are relatively stable. This is because IM service is often provided by companies and IP addresses hold by the company are used for IM servers. Thus the server address can be used to find out presence packets.

(2) The TCP/UDP port number of IM servers is often a registered port number, which can be used to find out presence packets.

(3) Certain presence packets of some IMs have special fixed formats and are sent periodically. A typical example is the packet that used for keeping alive purpose between a client and a server.

One can combine these three characteristics to trade-off between accuracy and efficiency of fingerprinting. Different IM applications have different design in its presence channels; therefore we have to use different fingerprints.

We use service port number and payload characteristic as fingerprints of multiple MSN Messenger clients. The presence packets of MSN Messenger can be find out by checking whether the port number is 1863 and whether there is a string "PNG" in the payload. The reason we don't use server address is the relative large number of NS servers.

We use server IP address and port number of its server as fingerprints of multiple Google Talk clients. Google Talk client connects to only one server in the whole log-on process. The total number of Google Talk servers is not large. We collected IP addresses of the servers by the domain name "talk.google.com". When a packet passes through the detecting point, we check whether the destination address in this packet is one of the Google Talk servers and whether the destination port is 5222, to judge whether it is a presence packet for Google Talk.

4 Algorithm in NAT-Aware Router

The scenario of IM presence fingerprinting is shown in figure 1. Multiple hosts share the same gateway address to access the Internet. A NAT detector, which a part of edge router, runs the fingerprinting algorithm and passively collects and analyzes IP packets passing through the edge router.

To reduce the heavy burden of filtering packets sent from every address in this access network, fingerprinting method is used in the second phase of the two-phase NAT detection methods proposed in [11]. In the first phase, the NAT-aware router roughly discovers suspicious target IP address of NAT gateway by light-weight methods presented in [11], and then verifies the suspicious target address in the second phase by application presence fingerprinting.

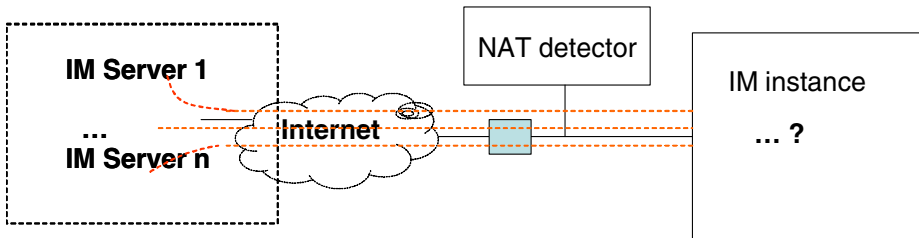


Fig. 1. NAT detection scenarios

A TCP connection is determined by its 4-tuple (source IP, destination IP, source port, destination port). Given the source IP (the suspicious target address i we observing) and the destination port number (service port number of a specific IM application IM_j), then the presence channel c_{ijkl} can be determined by destination IP k and source port number l .

A timer t_{ijkl} is set for each c_{ijkl} to denote the final updated time of that presence channel.

$TMAX_j$ is set the the maximum idle time of the presence channel for each IM application IM_j .

A threshold TH_j is set for the maximum number of allowed concurrent presence channels for each IM application IM_j .

The algorithm in NAT-aware router is described as follows:

Step 1: For each target IP address i in the monitored access network, NAT-aware router maintains a presence channel list C_{ij} for each IM_j .

Step 2: When a presence packet of IM_j coming from target IP address i is captured, NAT-aware router checks destination IP address k and source port number l to get the presence channel c_{ijkl} and determines whether is belongs to an existing channel. if $c_{ijkl} \in C_{ij}$, then $t_{ijkl} = 0$ (reset the update time); else, create a new presence channel c_{ijkl} , $C_{ij} = C_{ij} \cup \{c_{ijkl}\}$.

Step 3: NAT-aware router counts the number of current presence channels of target i and IM_j to make the verdict.

n_{ij} = number of members in C_{ij} .

if $n_{ij} \geq TH_{ij}$, then makes a verdict that this IP address i is a NAT gateway address.

Step 4: NAT-aware router checks each t_{ijkl} to remove expires presence channel c_{ijkl} . if $t_{ijkl} \geq TMAX_j$, then $C_{ij} = C_{ij} - \{c_{ijkl}\}$.

The NAT-aware router is used by network administrator to detect NAT gateways in their access network. We implemented the algorithm as a part of an edge router. As shown in figure 2, the system is composed by three main modules.

(1) Packet Filtering Module

This module is used to find out presence packets of different IM applications. Libpcap is used to do an initial filtering by the service port number of MSN Messenger and Google Talk. Then specific filter function for different IM is called to check whether a packet is really a presence packet.

(2) Presence Channel Analysis Module

This module is used to maintain the table for presence channels and run related algorithm.

(3) Result Reporting Module.

When the existence of a NAT gateway is detected, this module will report to the management system for further actions on the discovered NAT address.

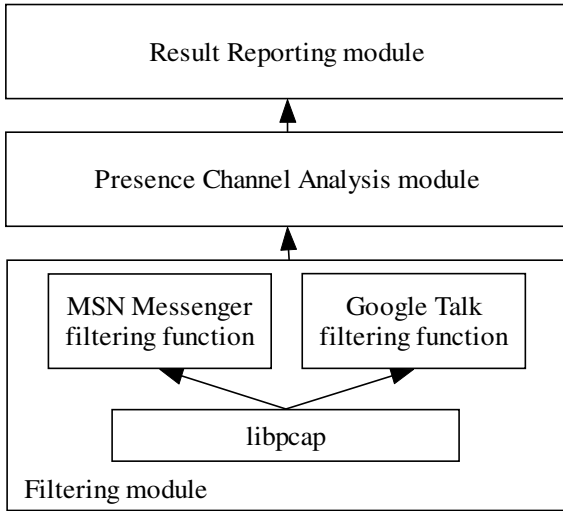


Fig. 2. Architecture of detection module

We did experiment with the system in Tsinghua campus network and CERNET. According to our experience, the parameters are set as table 2 shows.

Table 2. Experimental parameters setting

IM_j	$TMAX_j$	TH_j
MSN Messenger	100 seconds	2
Google Talk	50 seconds	2

5 Conclusions

This paper presents a new application presence fingerprinting for the NAT-aware router to detect NAT gateways. The main contributions of our work are:

- (1) Proposed novel application-level presence fingerprints for detection on network address translation.
- (2) Choose Instant Messaging as the applications and designed related algorithm for NAT-aware router.
- (3) A NAT-aware router is designed based on presence fingerprints of Google Talk and MSN Messenger.

The algorithm and system proposed by this paper provide a new way for administrators of ISPs to discover unauthorized NATs in their access network and enhance the network management and security. We will find more application-level

fingerprints in our future work and perform more experiments with the NAT-aware router.

References

1. M. Smart, G.R. Malan, F. Jahanian, Defeating TCP/IP Stack Fingerprinting, proc. of 9th USENIX Security Symposium, 2000, pp. 229-240.
2. S.M. Bellovin, A Technique for Counting NATted Hosts. proc. of 2nd Internet Measurement Workshop, 2002, pp. 267-272.
3. M. Zalewski, Passive OS Fingerprinting Tool, <http://lcamtuf.coredump.cx/p0f.shtml>, 2003
4. G. Taleck, Ambiguity Resolution via Passive OS Fingerprinting, proc. of 6th International Symposium Recent Advances in Intrusion Detection, 2003
5. R. Beverly, A Robust Classifier for Passive TCP/IP Fingerprinting, proc. of 5th Passive & Active Measurement Workshop, April 2004.
6. G. Roualland and J.M. Saffroy, Linux IP Personality, <http://ippersonality.sourceforge.net/>.
7. T. Kohno, A. Broido, and K.C. Claffy, Remote Physical Device Fingerprinting, IEEE Transactions on Dependable and Secure Computing, 2(2), 2005.
8. M. Day, J. Rosenberg and H. Sugano, A Model for Presence and Instant Messaging, RFC2778, Feb 2000.
9. R. Movva, MSN Messenger Service 1.0 Protocol, draft-movva-msn-messenger-protocol-00, Aug 1999.
10. P. Saint-Andre, Extensible Messaging and Presence Protocol (XMPP): Core, RFC 3920, Oct 2004.
11. J. Bi, M. Zhang, L. Zhao, and J. Wu, New Approaches to NAT Detection for Edge Network Management, proc. of IEEE 6th International Symposium and School on Advance Distributed Systems, Jan 2006.

Interaction for Intelligent Mobile Systems

G.M.P. O'Hare, S. Keegan, and M.J. O'Grady

Adaptive Information Cluster (AIC), School of Computer of Computer Science & Informatics, University College Dublin (UCD), Belfield, Dublin 4, Ireland
{gregory.ohare, Stephen.keegan, michael.j.ogrady}@ucd.ie

Abstract. Mobile computing poses significant new challenges due the disparity of the environments in which it may be deployed and the difficulties in realizing effective software solutions within the computational constraints of the average mobile device. Likewise, enabling seamless and intuitive interaction is a process fraught with difficulty. Embedding intelligence into the mobile application or the physical environment as articulated by the AmI vision is one potential strategy that software engineers could adopt. In this paper, some pertinent issues concerning the deployment of intelligent agents on mobile devices for certain interaction paradigms are discussed and illustrated in the context of an m-commerce application.

1 Introduction

Ambient Intelligence (AmI) is motivated by an awareness that the increasing proliferation of embedded computing within the environment may become a source of frustration to users if appropriate interaction modalities are not identified. Technically, AmI is closely related to the ubiquitous computing initiative articulated by Weiser over decade earlier. Both envisage computing artifacts being augmented with computational technologies. Both acknowledge the need for seamless and intuitive interaction. However, as developments in the necessary hardware and software continue, the issue of interaction has become increasingly important; hence the AmI initiative. This advocates the development of intelligent user interfaces that would mediate between the user and the embedded artifact.

Historically, text entry has been the traditional modality of interaction with computers starting with the QWERTY style keyboards back in the 1970s. However, developments in mobile telecommunications led to alternative interaction modalities being considered. At present, the default layout of mobile phone keyboards conforms to an ISO standard [1]. Numeric keys are overloaded with alphabetic characters and other symbols, and though the interface is not at first sight intuitive, the success of SMS suggests that significant numbers of people use keypad text entry without a thought. However, any difficulties that arise pale into insignificance when it is considered that sensors are considered to have sophisticated user interface if they support three LEDS! Thus, significant obstacles must be overcome if intelligent user interfaces are to be realistically deployed in AmI environments.

This paper is structured as follows: Section 2 considers interaction from a traditional mobile computing perspective. In Section 3, the intelligent agent paradigm

is examined. In Section 4, the use of intelligent agents for interaction monitoring is illustrated through a brief discussion of EasiShop, an m-commerce application. Some related research is presented in Section 5 after which the paper is concluded.

2 Observations on Interaction

In recent years, a number of different modalities of interaction have been researched and described in the literature. Multi-modal interaction where a number of different modalities of interaction, for example, voice, gesture and so on, are used for both for input and output, is a case in point. For the purposes of this discussion, however, the input modality is of particular interest and is viewed as ranging from explicit to implicit (Fig. 1). One could also classify the input interaction modality as ranging from event-based to streaming-based [2].

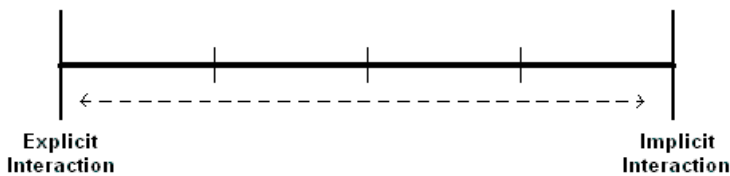


Fig. 1. The Interaction Continuum

Explicit interaction occurs when a user tells a computer quite clearly what it is to do. This may be accomplished by manipulating a GUI, running a command in a command window or using a voice recognition system, to mention but a few. In short, the user performs some action, thus unleashing a series of events, resulting in an expected outcome. Most interaction with computers is of the explicit variety.

Implicit interaction [3] is, from a computation perspective, a more recent development. In itself, it is not a new concept as people also communicate implicitly, for example by subconsciously smiling or frowning. Interaction is considered as being implicit when an action is performed that a user would not associate as being input but which a computer interprets as such. A classic example is that of a museum visitor. By moving towards certain exhibits and viewing them for a significant time span would be interpreted by a mobile electronic tourist guide as an indication of a strong interest in the exhibit. Therefore the guide could deduce with a reasonably high degree of certainty that the visitor might welcome some additional information about that exhibit.

Implicit interaction is closely associated with a user's context and some knowledge of this is almost essential if designers want to incorporate it into their applications. A potentially serious problem arises when an incomplete model of the prevailing context is available. At some stage, a decision has to be made as to whether to initiate an explicit interaction with the user. However, when to do this depends on the nature of the application in question. Using techniques such as machine learning, bayesian probability and so on, a threshold can be identified that, if exceeded, would prompt an explicit interaction. The value of this threshold would essentially be a judgment call

by the software engineer and may even change dynamically. Suffice to say that a health monitoring application would have a lower threshold than one that recommends local hostelrys. Thus, the need for a degree of intelligence in the application. In the next section, one solution, intelligent agents, is considered.

3 Intelligent Agent Architectures

Intelligent agents continue to be the subject of intense research in academia. Though their uses are varied and many, agents are perceived as offering alternative strategies for software development in areas that traditional techniques have not proved effective. Examples include domains that are inherently dynamic and complex. Three broad categories of agent architecture have been identified:

1. Reactive agents act in a simple stimulus-response fashion and are characterized by a tight coupling between event perception and subsequent actions. Such agents may be modeled and implemented quite easily. The Subsumption Architecture [4] is classic example of a Reactive Architecture.
2. Deliberative agents can reason about their actions. Fundamental to such agents is the maintenance of symbolic model of their environment. One popular implementation of the deliberative stance is the belief-desire-intention (BDI) model [5], which has found its way into commercial products, for example JACK [6]. In the BDI scheme, agents maintain a model of their environment through a set of beliefs. Each agent has set of objective or tasks that it seeks to fulfill, referred to as desires. By continuously monitoring its environment, the agent will detect opportunities when it is appropriate to carry out some of its desires. Such desires are formulated as intentions which agent proceeds to realize.
3. Hybrid architectures seek to adopt the best aspects of each approach. A strategy that might be adopted would be to use the reactive component for event handling, and the deliberative components for longer term goals.

A number of characteristics are traditionally associated with agents. Reactivity, proactivity, autonomy and societal are characteristics of so called weak agents while strong agents augment these further with rationality, benevolence, veracity and mobility [7]. Of course, not all of these will be possessed by individual agents.

4 Capturing Interaction Through Agents

On completing requirements analysis and, as part of the initial design stage, the software designer, possibly in conjunction with a HCI engineer, must decide on the necessity for supporting the implicit interaction modality for the proposed application. Should they decide favorably, then a mechanism for monitoring and capturing both explicit and implicit interaction must be identified. A number of options exist; and the final decision will be influenced by a number of factors that are application domain dependent. However, one viable strategy concerns intelligent agents. From the

previous discussion on interaction and intelligent agents, it can be seen that certain characteristics of the agent paradigm are particularly suited for capturing interaction.

Practically all applications must support explicit interaction. The reactive nature of agents ensures that they can handle this common scenario. As to whether it is prudent to use the computational overhead of intelligent agents just to capture explicit user input is debatable, particularly in a mobile computing scenario. Of more interest is a situation where implicit interaction needs to be captured and interpreted.

Implicit interaction calls for continuous observation of the end-user. As agents are autonomous, this does not present any particular difficulty. Though identifying some particular aspect or combinations of the user's context may be quite straightforward technically, interpreting what constitutes an implicit interaction, and the appropriateness of explicitly responding to it in a timely manner may be quite difficult. Hence, the need for a deliberative component.

As an illustration of the issues involved, interaction modalities supported by EasiShop, a system based on the agent paradigm are now considered. EasiShop [8] is a functioning prototype mobile computing application, developed to illustrate the validity of the m-commerce paradigm. By augmenting m-commerce with intelligent and autonomous components, the significant benefits of convenience and added value may be realized for the average shopper as they wander their local shopping mall or high street. In the rest of this section, the synergy between agents and interaction is demonstrated through an illustration of an archetypical EasiShop usage scenario.

4.1 EasiShop Usage Scenario

EasiShop is a suite of software deployed in a tripartite architecture to support m-commerce transactions. There are three distinct stages of interaction within EasiShop.

1. List construction and profile construction (explicit interaction)

EasiShop is initiated when the user constructs a seed user profile. This is comprised of a set of bounded personal information such as age and gender. A further set of generalized information, alluding to the type of product classes which are of interest, is then obtained from the user. This last type of data is obtained when the user constructs a shopping list. The list details what products are sought by the user and, to a certain extent, under what terms and context this acquisition is permissible.

2. Movement (implicit interaction)

EasiShop is primarily an automated shopping system. Once the user has specified their profile information and has constructed a shopping list (fig. 2a), the various components collaborate in a transparent and unobtrusive manner to satisfy the requirements of the user. To manage this process, a certain degree of coordination is required, hence the use of mobile agents.

As the user wanders their local high-street, a proxy agent migrates transparently from the user's device into a proximal store. From here, this entity may migrate to an open marketplace where representative agents from a myriad of stores (including the current proximal store) may vie for the user's agent's custom. This process entails a reverse auction whereby the user's requirements (and profile) are presented to the marketplace. Interested parties request to enter the ensuing auction and a set of the most appropriate selling candidates is chosen by the user's agent upon completion of



Fig. 2. Shoppers must explicitly inform EasiShop about their requirements (a) EasiShop implicitly monitors shopper behavior and (b) autonomously negotiates with nearby stores

that auction (if any). At this point the user's agent is ready to return to the user's device and will attempt to do so, though it may return from a different point.

3. Decision (explicit interaction)

When the agent has returned to the user's device, it is laden with the set of product offerings resulting from the auction process. This set is presented to the user from whom an explicit decision is requested as to which offering (if any) is the most acceptable (fig. 2b). Once this indication has been made, the user is free to collect the item from the relevant shop. This decision is used as reinforcement feedback in that selection data is garnered to determine what kind of choice is likely in the future. Should the result of the auction not meet the agent's initial requirements, the user is not informed of what has taken place. However, the process will be repeated until such time as the shopping list is empty.

4.2 Architecture of EasiShop

EasiShop functions on a three-tiered distributed architecture (fig. 3). From a center-out perspective, the first tier is a special centralized server called the Marketplace. The design of the marketplace permits trade (in the form of reverse auctions) to occur. The second tier is termed the Hotspot. This is a hardware and software hybrid suite, situated at each participating retail outlet, which is permanently connected to the Marketplace and which allows the process of moving (migrating) representative (selling) agents from the retailers together with (buying) agents representing shoppers

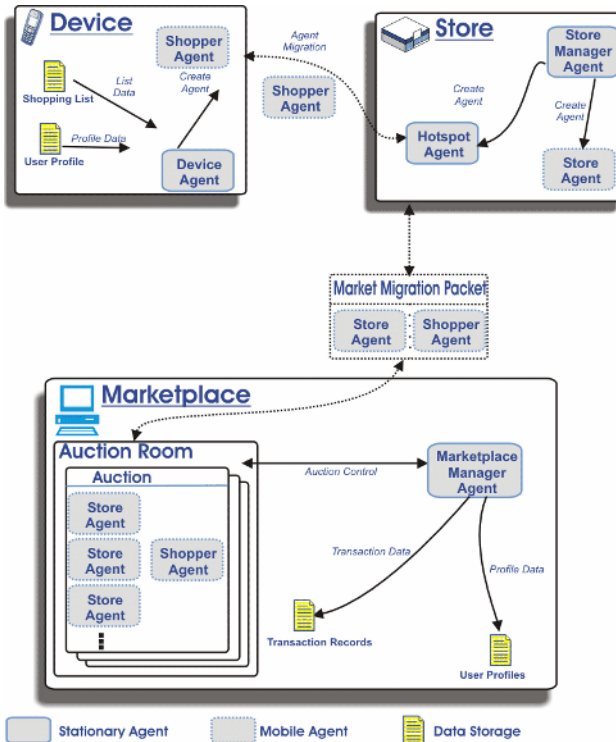


Fig. 3. Architecture of EasiShop

to the Marketplace. The final and outermost tier in this schema is collection of device nodes, usually a smart mobile phones or PDA.

4.3 Utilization of Mobile Intelligent Agents

The principal challenge when devising appropriate agents in a preference-based auctioneering domain is to deliver *personality embodiment*. These personalities represent traits of buyers (users) and sellers (retailers) and are required to encapsulate *dynamism* (in that traits change over time) and *persistency* (in that a record of the traits needs to be accessible to the agents). These requirements are satisfied by the utilization of file-based xml data storage which contain sets of values upon which traits depend. For example, the characteristics of a store are encapsulated in an xml file containing rules which can formalize attributes like preferences for certain types of shopper (of a certain age or gender), temporal constraints (like pricing products according to time of day) and stock restrictions (like pricing products according to current stock levels). The mechanism to deliver agent mobility is implemented in a separate subsystem and uses both fixed (LAN) and wireless (Bluetooth) carriers.

5 Related Research

Intelligent agents encompass a broad and dynamic research area. However, deploying agents on mobile devices has, until recently, been unrealistic primarily due to hardware limitations. However, ongoing developments are increasingly rendering these limitations obsolete and a number of agent environments have been described in the literature. In some cases, existing platforms have been extended. For example, LEAP [9] has evolved from the JADE [10] platform. Likewise microFIPA-OS [11] is an extension of the well-known open source platform FIPA-OS [12]. In the case of BDI agents, Agent Factory Lite [13] is one environment that supports such agents.

One classification of agents that is closely associated with interfaces are the aptly named Interface Agents. Maes [14] has done pioneering work in this area and regards interface agents as potential collaborators with users in their everyday work and to whom certain tasks could be delegated. Ideally, interface agents would take the form of conversational characters which would interact with the user in a social manner.

Recently, the use of intelligent agents for ambient intelligent applications has been explored. Satoh [15] describes a framework, based on mobile agents, that allows personalized access to services for mobile users. Grill et al [16] describe an environment that supports the transfer of agent functionality into everyday objects. Finally, Hagrais et al describe iDorm [17], an intelligent dormitory that uses embedded agents to realize an AmI environment. The system uses fuzzy techniques to derive models of user behaviors.

6 Conclusion

Mobile computing scenarios offer a different set of challenges from those traditionally experienced in networked workstation environments. Implicit interaction offers software designers an alternative model of capturing user intent and significant opportunities to proactively aid the user in the fulfillment of their tasks. However, capturing and interpreting implicit user interaction require that some intelligence either be hosted on the user's device or embedded in their surroundings.

In this paper, the use of intelligent agents has been discussed as a promising approach to managing both the traditional explicit interaction modality and, where necessary, the implicit interaction modality. Such an approach offers software designers an intuitive method of incorporating the implicit interaction modality into their designs. Attention is frequently at a premium in mobile computing environments thus the use of autonomous agents for managing user interaction in an intelligent manner offers a significant opportunity for enhancing the user experience – a fundamental prerequisite if mobile computing is to fulfill its potential.

References

1. ISO/IEC 9995-8:1994, Information technology -- Keyboard layouts for text and office systems -- Part 8: Allocation of letters to the keys of a numeric keypad.
2. Obrenovic, Z., Starcevic, D., Modeling Multimodal Human-Computer Interaction, IEEE Computer, vol. 37, no. 9, 2004, pp. 65-72.

3. Schmidt, A. Implicit Human Computer Interaction through Context. *Personal Technologies*, Volume 4(2&3), June 2000. Springer-Verlag. pp. 191-199.
4. Brooks, RA, Intelligence without representation, *Artificial Intelligence* 47, 139–159, 1991.
5. Rao, A.S., Georgeff, M.P., Modelling Rational Agents within a BDI Architecture. In: *Principles of Knowledge Representation. & Reasoning*, San Mateo, CA. 1991.
6. JACK - The Agent Oriented Software Group, <http://www.agent-software.com>.
7. Wooldridge, M., Jennings, N.R., Intelligent Agents: Theory and Practice, *The Knowledge Engineering Review*, vol.10, no.2, 1995, pp. 115-152.
8. Keegan, S., O'Hare, G.M.P., EasiShop: Enabling uCommerce through Intelligent Mobile Agent Technologies, *Proceedings of 5th International Workshop on Mobile Agents for Telecommunication Applications (MATA'03)*, Marrakesh, Morocco, 2003.
9. Bergenti, F., Poggi, A., LEAP: A FIPA Platform for Handheld and Mobile Devices, *Proceedings of the 8th International Workshop on Agent Theories, Architectures and Languages (ATAL-2001)*, Seattle, WA, USA, August 2001.
10. Bellifemine, F., Rimassa, G., Poggi, A., JADE - A FIPA compliant Agent Framework, *Proceedings of the 4th International Conference and Exhibition on the Practical Application of Intelligent Agents and Multi-Agents*, London, 1999.
11. Tarkoma, S., Laukkanen, M., Supporting Software Agents on Small Devices, *Proceedings of AAMAS*, Bologna, Italy, July 2002.
12. Foundation for Intelligent Physical Agents (FIPA), <http://www.fipa.org>.
13. Muldoon, C., O Hare, G.M.P., Collier, R.W., O Grady, M.J., Agent Factory Micro Edition: A Framework for Ambient Applications, *Proceedings of: Intelligent Agents in Computing Systems*, Reading, UK, May, 2006.
14. Maes, P., Agents that reduce work and information overload, *Communications of the ACM* 37(7), 30-40, 1994.
15. Satoh, I., Software Agents for Ambient Intelligence, *Proceedings of IEEE International Conference on Systems, Man and Cybernetics (SMC'2004)*, pp.1147-1150, 2004.
16. Grill, T., Ibrahim, I.K., Kotsis, G., Agents Visualization in Smart Environments, *Proc. of the 2nd International Conference on Mobile Multimedia (MOMM2004)*, Bali – Indonesia.
17. Hagraas, H., Callaghan, V., Colley, M., Clarke, G., Pounds-Cornish, A., Duman, H., Creating an ambient-intelligence environment using embedded agents, *Intelligent Systems*, 19(6), 12-20, 2004.

Design of a Intelligent SOAP Message Service Processor for Enhancing Mobile Web Service*

Gil-Cheol Park and Seoksoo Kim**

Department of Multimedia Engineering, Hannam University
133 Ojung-Dong, Daeduk-Gu, Daejeon 306-791, Korea
{gcpark, sskim}@mail.hannam.ac.kr

Abstract. As the mobile internet becomes one of the main methods for information delivery, mobile Web Services are regarded as a critical aspect of e-business architecture. In this paper, we proposed a intelligent mobile Web Services middleware that enhances SOAP processing performance by eliminating the Servlet container (Tomcat), a required component of typical Web services implementation. Our main contributions are to overcome the latency problem of current Web Services and to provide an easy mobile Web service implementation. Our system can completely support standard Web Services protocol, minimizing communication overhead, message processing time, and server overload. Finally we compare our empirical results with those of typical Web Services.

1 Introduction

Mobile internet services enable users to access the internet from any location at any time providing flexible personalized information according to users' location and their information needs [1]. The mobile internet can provide various value added services in addition to basic communication services. As the internet capabilities are widely understood and wireless technologies advance, mobile internet services will soon be a major mediator in information delivery and in business transactions [2].

Mobile internet services, however, still have physical devices, network and content limitations. Firstly, mobile devices are limited by system resources such as smaller screens and less convenient input devices. Secondly, wireless networks have less bandwidth, less connection stability, less predictability and a lack of standardized and higher costs [1, 3]. Lastly, mobile internet services also have content limitations because the amounts of available mobile content are still smaller than that of wired internet services, and the consistency between wired and wireless internet services is very critical. Physical device and network limitations make supporting common internet standards such as HTML, HTTP, and TCP/IP difficult because they are inefficient over mobile networks.

* "This work was supported by a grant No. (R12-2003-004-03003-0) from Ministry of Commerce, Industry and Energy".

** Corresponding author.

Therefore, new protocols such as WAP (Wireless Application Protocol) and WML (Wireless Markup Language) are proposed to address these issues. Content limitations encourage researchers to find a method that can support reusing current wired Web information. Some researchers focus on the conversion of HTML documents to mobile internet serviceable WML documents and direct access to databases, to provide efficient information delivery in the wireless environment [4-8].

However, these researchers do not focus on the capability that allows applications to interact over the internet in an open and flexible way, but on the capability that provides dynamic wireless internet service according to different network and device environments. In fact, the former goal can be achieved by implementing Web Services. If the implementation is successful, interactions between applications are expected to be independent from the platform, programming language, middleware, and implementation of the applications involved. Nowadays Web Services become key applications in business-to-business, business-to-customer, and enterprise applications integration solutions [9].

Web Services require specific messaging protocols known as SOAP for interaction. One main issue of SOAP implementation is the latency of SOAP execution [10-14]. We view that the latency problem is caused by the current SOAP message processing system architecture. The current SOAP processing system requires the Web Servlet container (e.g. Tomcat) to execute SOAP. Our hypothesis is that if a system processes the SOAP message directly, without help from Web Servlet container, the SOAP performance improves. To examine this hypothesis we implemented a SOAP message processing system, called SOAPProc.

The paper is organized as follows: Section 2 summarizes relevant research results, Section 3 illustrates our SOAPProc system implementation. In Section 4 we compare our system's performance with the typical Web Services implementation approach. Finally, conclusions and recommendations for further work are described in Section 5.

2 Mobile Web Service Implementation with Axis and Tomcat

There are several Web Services implementation methods, which differ in their support for class binding, ease of use, and performance [14]. Among them Apache Axis (Apache eXtensible Interaction System) with Tomcat is a popular implementation method. The Apache Axis project is a follow-on to the Apache SOAP project and currently has reached version 1.2.1, but it's not part of the Apache SOAP project. Axis is a completely new rewrite with emphasis on flexibility and performance. It supports HTTP SOAP request/response generation, SOAP message monitoring, dynamic invocation, Web service deployment, and automatic WSDL generation for Web services.

Fig. 1. illustrates this standard mobile Web service implementation. A Web Servlet container, like Tomcat, is required to provide mobile Web services with Axis. For wireless internet service, the server administrator should write MML (Made Markup Language) to parse Web contents by using the administrative tool. A MML is used to generate service request forms or service results by dynamically parsing the existing Web contents and sending them to relevant model and clients. When a client, whether it is wireless or wired client, requests Web service via SOAP request, Apache Tomcat transfers it to Axis. Axis interfaces the SOAP request message into a relevant service

by using the service management function. Service providing models are interfaced by using a WSDL module is provided by Axis. By implementing SOAP and distributed computing service, the system architecture can have a lightweight thin client structure and the service can be provided in a flexible way.

However, this implementation is not efficient because it requires additional process for Web Servlet engine (Tomcat) and communication port. For this reason, we propose an alternative system that can process SOAP messages without using Web Servlet engine. This implementation will be discussed in Section 4.

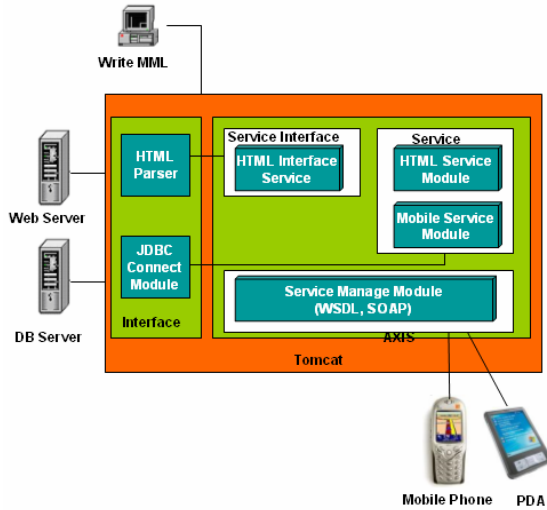


Fig. 1. Standard Mobile Web Service Implementation with AXIS and Tomcat

3 Implementation of SOAP Message Service Processor

In the Web Services, XML based SOAP messages are used when the clients request Web Services from the server or when the server sends Web Service response messages to the clients. In the standard Web Services implementation this is supported by Tomcat and AXIS. This architecture causes in efficiency as explained in Section 2.3. For this reason, we developed a SOAP message processing system, called SOAProc which directly processes the SOAP request and response messages without using Servlet engine.

Fig. 2. illustrates our Web Services system implementation architecture, in which the SOAProc and the WSDL builder are used. The most significant difference between the standard system (see Figure 1) and our implementation (see Figure 2) is that our system does not include Tomcat. Instead of using Tomcat’s WSDL and SOAP supporting function, WSDL files are directly generated by the WSDL builder and SOAP messages are processed by the SOAProc system.

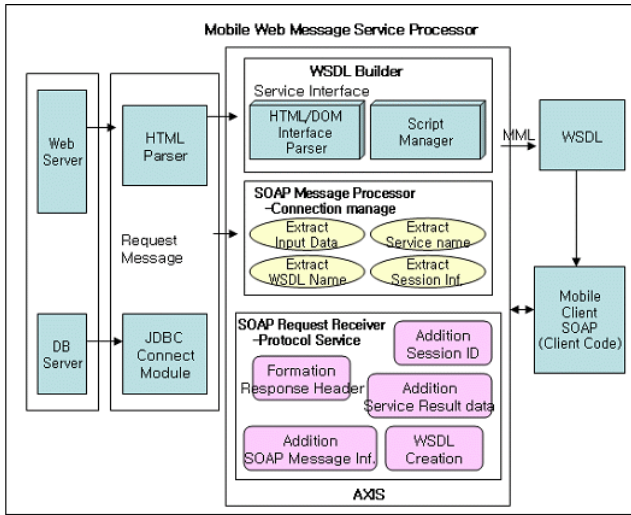


Fig. 2. A New Mobile Web Service Implementation by using the SOAPProc and the WSDL builder

Additional Web Services module is implemented to the Server. Functions for message handling, chain, WSDL processing, and SOAP request receiving are required to implement Web Services. To this end, following systems are implemented:

- SOAP message processor
- WSDL builder
- SOAP request receiver(Protocol Process)

SOAP request receiver receives SOAP request messages through port number 80 and sends them to the Web Server. We implemented SOAP message analyzer and generator (SOAP message processor) because the requested messages from clients consist of standardized SOAP message and the receiving messages from the Web server are required to be changed into SOAP message. In addition, the WSDL builder generates WSDL files, which will be used in the client. Though Java2WSDL supports WSDL generation by using Java Class, the WSDL builder supports WSDL generation by implementing XML parser. If a client requests Web services by sending SOAP message through HTTP, the server permits the client's connection. And we develop SOAP request receiver. It consists of SOAP analyzer, Service processor and SOAP generator. After the SOAP analyzer extracts service name and input data from the client's request, it requests service processing to the service processor module. The results that are returned from the service processor module are sent to the SOAP generator module. After creating response header and response message body, the SOAP generator module checks whether the service processing result is successful or not. If the result is successful, the results and session ID is added. If not, the failure information is attached. The service connector sent the response SOAP message to the client. When the client receives the SOAP response message, the Web service request is

complete. The access URL that the client uses are acquired from WSDL and its form is as follows: `http://host address : port/webservice?WSDL name`

The next Section describes how the SOAPProc system processes SOAP request and response messages in this implementation.

4 Experiment

4.1 Method

The experiment is focused on the performance evaluation of our mobile Web service system. Two sets of systems are prepared for our experiment. The first system is implemented with standard Web Services architecture as explained in Section 2. This implementation requires Tomcat Servlet container with AXIS. The second implementation is based on our approach. Where there is no Servlet container with the SOAP message processing performed by the SOAPProc system and WSDL created by the WSDL builder.

We conducted a simulated performance comparison experiment. Fig. 3. illustrates the experiment process. If a client requests Web services by submitting a SOAP request, the experiment system analyses the SOAP message and sends a HTTP request to the content Web servers. If the experiment system receives a HTTP response message from the Web server, it generates WSDL and sends a SOAP response message to the clients' mobile device.

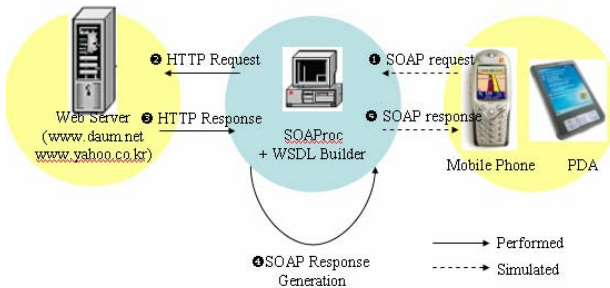


Fig. 3. Experiment System Procedure

SOAP requests are simulated by the mobile client simulation program, which connects to the experiment system and sends several SOAP request messages. There are time intervals, from 1 to 10 seconds between SOAP requests. If the connection is closed, the simulation program continually tries to connect to the experiment system. We assumed that there were 200 users at the same time. SOAP requests were created by four client programs and each program generated 50 threads at the same time. We chose two public Web sites [www.daum.net (dictionary) and www.yahoo.co.kr (stock)], which role as the content Web servers in our experiment. We assumed two kinds of specific information - dictionary and stock - are required by the user from these Web servers. Each service's timeout is 30 seconds.

4.2 Results

The following results were collected to compare two experiment systems:

- Test time: how many seconds were consumed for the test.
- Number of Requests: how many requests were generated within test time.
- Connection Timeout: the connection numbers that were not connected within the request timeout.
- Connection Refuse: the request numbers that could not be connected because the server was busy.
- Connection Handshake Error: the number of session configuration failures after connection
- Connection Trial Time: How many times the client could not connect to the server.
- Request Timeout: how many times the timeout was exceeded.

Table 1. summaries the experiment as results, which illustrate an enhanced performance in all categories. Though the test time of our system is shorter than that of the standard system, the total number of requests is greater than that of standard system and the timeout number is less than that of the standard system. For example, whist the average request per second of our system is 17.94, that of standard system is 9.64. There are many connection errors in the standard system. Only some portion of 200 requests is successfully connected to the server while the others get a “refused” message from the server. However, those kinds of connection failures do not happen in our system.

Table 1. Experiment Results

	SOAProc System	Standard System
Total Request	400,000	400,000
Connection Timeout	0	413
Connection Refused	0	18,250
Connection Handshake Error	0	483
Connection Trials	233	21,433
Request Timeout	745	113,782

5 Conclusions

In Internet service environment, the Mobile Web services are critical solutions in the internet service integration architecture. In this research we proposed a new Web Service architecture by implementing two significant systems. Firstly, we proposed a new SOAP message processing system to diminish SOAP latency problems by

eliminating the Tomcat Servlet container in the Web Services implementation. The SOAP request and response messages are directly processed by the SOAPProc system. Secondly, we proposed SOAP request receiver to analysis and generate SOAP message.

We can implement an alternative mobile Web Services system by using these two systems without violating standard Web Services protocols. Our experiment results demonstrate that the SOAP request processing performance of our approach is significantly better than that of the standard Web service implementation. Our system can process more service request about doubly efficient than that of typical Web service implantation with very small connection errors.

References

1. Siau, K., E.P. Lim, and Z. Shen, *Mobile commerce: promises, challenges, and research agenda*. Journal of Database Management, 2001. vol.12, no.3: p. 4-13.
2. Senn, J.A., *The emergence of m-commerce*. Computer, 2000. 33(12): p. 148-150.
3. Kim, H., et al. *An Empirical Study of the Use Contexts and Usability Problems in Mobile Internet*. in *35th Annual Hawaii International Conference on System Sciences (HICSS'02)*. 2002.
4. Kaasinen, E., et al., *Two approaches to bringing Internet services to WAP devices*. Computer Networks, 2000. 33(1-6): p. 231-246.
5. Kurbel, K. and A. Dabkowski. *Dynamic WAP content Generation with the use of Java Server Pages*. in *Web Databases/Java and Databases: Persistence Options (Web&DB/JaDa)*. 2002. Erfurt, Germany.
6. Metter, M. and R. Colomb. *WAP Enabling Existing HTML Applications*. in *First Australasian User Interface Conference*. 2000.
7. Saha, S., M. Jamtgaard, and J. Villasenor, *Bringing the wireless Internet to mobile devices*. Computer, 2001. vol.34, no.6: p. 54-58.
8. Pashtan, A., S. Kollipara, and M. Pearce, *Adapting content for wireless Web services*. IEEE Internet Computing, 2003. 7(5): p. 79-85.
9. Farrell, J.A. and H. Kreger, *Web services management approaches*. IBM Systems Journal, 2002. vol.41, no.2: p. 212-227.
10. Seshasayee, B., K. Schwan, and P. Widener, *SOAP-binQ: high-performance SOAP with continuous quality management*. Proceedings. The 2nd IEEE International Conference on Distributed Computing Systems, 2004: p. 158-165.
11. Kohlhoff, C. and R. Steele, *Evaluating SOAP for high performance applications in capital markets*. Computer Systems Science and Engineering, 2004. 19(4): p. 241-251.
12. Chiu, K., M. Govindaraju, and R. Bramley, *SOAP for High Performance Computing*, in *11th IEEE International Symposium on High Performance Distributed Computing HPDC-11 20002 (HPDC'02)*. 2002, Indiana University. p. 246.
13. Chiu, K., M. Govindaraju, and R. Bramley. *Investigating the Limits of SOAP Performance for Scientific Computing*. in *11th IEEE International Symposium on High Performance Distributed Computing (HPDC-11 '02)*. 2002.
14. Davis, D. and M. Parashar. *Latency Performance of SOAP Implementations*. in *2nd IEEE/ACM International Symposium on Cluster Computing and the Grid*. 2002.

Convergence Rate in Intelligent Self-organizing Feature Map Using Dynamic Gaussian Function*

Geuk Lee¹, Seoksoo Kim², Tai Hoon Kim³, and Min Wook Kil⁴

¹ Department of Computer Eng., Hannam University, 133 O-Jung Dong,
DaeJeon, South Korea
leegeuk@hannam.ac.kr

² Department of Multimedia., Hannam University, 133 O-Jung Dong, DaeJeon, South Korea
sskim@hannam.ac.kr

³ Security Engineering Research Group, DaeJeon, South Korea
taihoonn@paran.com

⁴ Dept. of Medical Inform., Mun Kyung College, HoGyeMyun, Mun Kyung, South Korea
mwkil@mkc.ac.kr

Abstract. The existing self-organizing feature map has weak points when it trains. It needs too many input patterns, and a learning time is increased to handle them. In this paper, we propose a method improving the convergence speed and the convergence rate of the intelligent self-organizing feature map by adapting Dynamic Gaussian Function instead of using a Neighbor Interaction Set whose learning rate is steady during the training of the self-organizing feature map.

1 Introduction

Among unsupervised learning models of the neural network, Self-organizing feature map proposed by Kohonen[1,2,3] is a neural network has a vector quantization function. Its strong points are to construct features of multi-dimensional input vectors into a feature map in 1-dimension or 2-dimension, and to automatically resolve classification problem associated with those input vectors. But there are weak points too. Speed of learning is slow since learning rate of the self-organizing feature map is steady regardless of vector distances between a winner neuron and other neurons in neighbor set, and it requires too many input vectors to converge into an equilibrium state.

Recent related researches are mainly focused on improving performance at the stage of learning. Lenne[4] proposed Weight alignment after grouping input vectors. Behis[5] suggested to learn by defining the geometric hash function. B. Bavarian[9,10] proposed the method providing different learning rates depending on topological positions by defining Gaussian Function as the neighbor interaction function. Even though it could obtain higher convergence rate than existing Self-organizing feature maps did, but inaccurate self-organization were also occurred.

* This work was supported by a grand No.R12-2003-004-02003-0 from Korea Ministry of Commerce Industry and Energy.

Therefore this paper proposes the method using Dynamic Gaussian Function to increase convergence speed of a neural network, to train the neural network with only a few input vectors, and to get more accurate self-organization.

2 Self-organizing Feature Map

The self-organizing feature map, which has different form with other neural networks, consists of input and output layers only, takes a simple form in which all neurons of the output layer takes all input pattern, which are interactively compete with each other and fully connected.

The rule determining the winner neuron is to choose a neuron which has the shortest Euclid distance. And the rule of learning is, as shown in Equation (1), to multiply a learning constant to the distance between the input vector and the weight vector so that weights of neurons within a radius are gradually closed to the input pattern.

$$\begin{aligned}
 M_i(t_{k+1}) &= M_i(t_k) + \alpha(t_k)[X(t_k) - M_i(t_k)], \text{ for } i \in N_f(t_k) \\
 M_i(t_{k+1}) &= M_i(t_k), \text{ for } i \notin N_f(t_k)
 \end{aligned}
 \tag{1}$$

where $X(t_k)$ is the input, $M_i(t_k)$ is the weight value of i -th neuron in the output layer, $N_f(t_k)$ is the neighbor set, and $\alpha(t_k)$ is the learning constant.

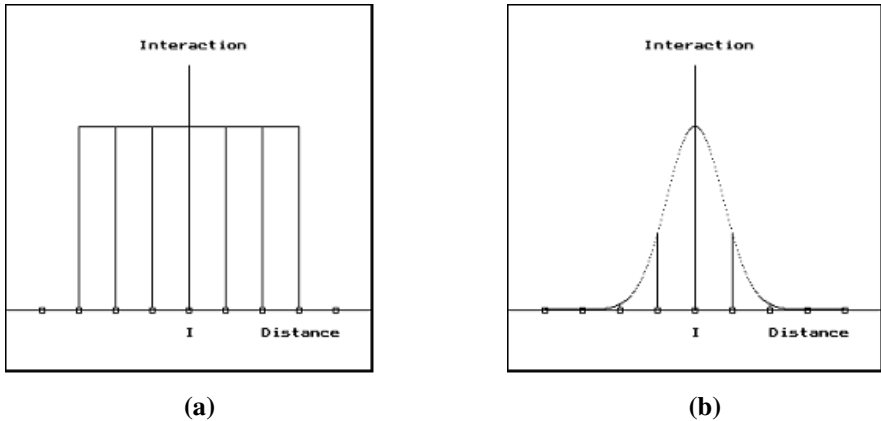


Fig. 1. (a) Neighbor Interaction Set (b) Neighbor Interaction Function

2.1 Kohonen's Neighbor Interaction Set

Kohonen's self-organizing feature map allows only neurons in neighbor sets, which is adjacent to the winner neuron as a center, to learn the input vector once after the winner neuron is determined.

But, as shown in Figure 1(a), all the neurons in the neighbor set learn with a same learning rate regardless of distances to the winner neuron. Therefore, many input vectors are required to be converged into equilibrium state, and speed of learning is slow.

Kohonen's self-organizing feature map has strong points such as fast learning time, simple structure, and self-organization of ambiguous relations among ambiguous features. But it cannot perform well if the size of the neural network is not large enough, requires many input vectors, and has to initialize weight values with random values.

2.2 Bavarian's Neighbor interaction function

B. Bavarian's method, which is different from existing neural networks, uses different learning rates depending on distances between the winner neuron and other neurons in the neighbor set by defining Neighbor interaction function. The neighbor interaction function AI uses Gaussian function as in Figure 1(b), and the expression is shown in Equation (2)

$$A_i(i, t_k) = c + d \cdot e^{-\frac{H(i-n)^2}{2\sigma^2}}, \quad \text{for } i \in N_f(t_k) \tag{2}$$

$\sigma = |N_f(t_k)|$

where c, d are constants, i is the winner neuron, and n is a size of the neighbor set. σ is decreased as time goes by.

Learning rate is in Equation (3) where the interaction function is appended into the learning rule in the self-organizing feature map.

$$M_i(t_{k+1}) = M_i(t_k) + \eta A_i(i, t_k) (x_i - M_i(t_k)), \quad \text{for } i \in N_f(t_k) \tag{3}$$

B. Bavarian's neural network can provide faster convergence speed than other self-organizing feature map do, and reduce a number of input vectors. But self-organization may not be completed because some parts of the codebook are overlapped when a size of the neural network becomes too larger.

3 Dynamic Neighbor Interaction Function

3.1 Analysis of the Neighbor Interaction Function

Analyzing Gaussian function, which is used as the interaction function, is a process to improve performance of the neural network, and can check variations of parameters and corresponding performance of the neural network. Figure 2. is a graph of Gaussian function related to the parameter H: the higher a value of H is, the lesser the width of Gaussian function is. And Figure 3 represents performance of the neural network corresponding to changes of the parameter H against the time domain using Equation (4). When the parameter H is increased, the error rate is decreased. But the error rate itself is not monotonically decreased when the parameter H is monotonically increased.

$$error(t_k) = \sum_j [(M_{i_1}(t_k) - \mu_{i_1})^2 + (M_{i_2}(t_k) - \mu_{i_2})^2] \tag{4}$$

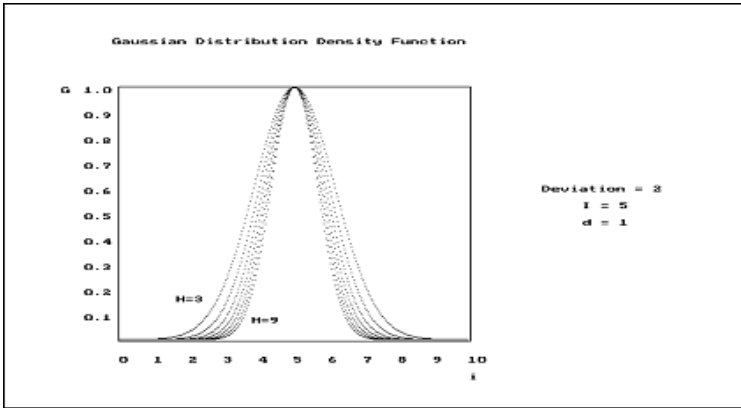


Fig. 2. Width of Gaussian function depending on Parameter H

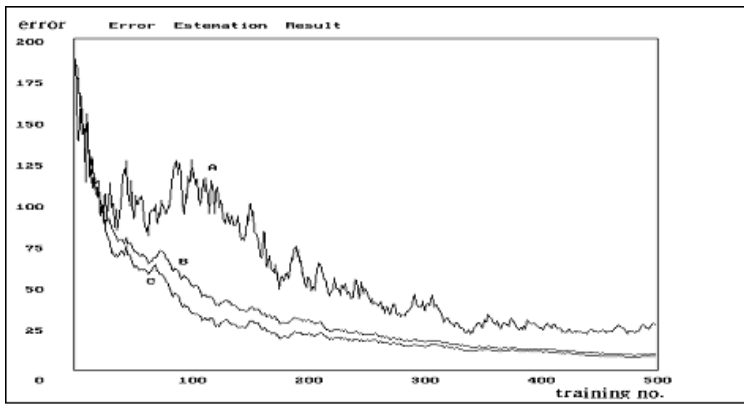


Fig. 3. Error rate of the neural network corresponding to changes of H (A : H=5 , B : H=8 , C : H=7)

3.2 Dynamic Neighbor Interaction Function

This paper proposes a new method to improve the convergence rate of the self-organizing feature map and increase the convergence speed by applying two characteristics to the neighbor interaction function. First one is Dynamic property of the width decrease according to increase in number of learning times and the other one is the different learning rate depending on topological position from the winner neuron.

As shown in Figure 4, the proposed method make Gaussian function have Dynamic property of the width decrease by time spent and different learning rates depending on topological position from the winner neuron to improve the convergence rate of the self-organizing feature map and increase the convergence speed.

Gaussian function, which is used in B. Bavarian's method, is modified as Equation (5) to decrease the width of the function according to the increase in number of

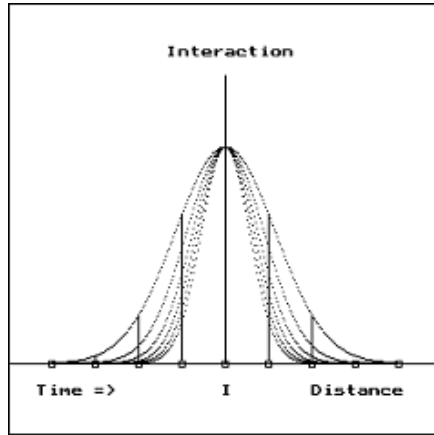


Fig. 4. Dynamic Gaussian function

learning times and to use different learning rates depending on the topological position of the winner neuron. Also, in order to resolve the situation in which weights are concentrated on the winner neuron at initial time period, the variance is adjusted larger than a size of neighbor set so that it will be gradually applied to all neurons in the neighbor set as time passes. The proposed neighbor interaction function is expressed as in Equation (5) and the rule of learning is in Equation (6).

$$A_f(i, t_k) = c + d \cdot e^{-\frac{(H + \frac{t_k}{T})(i-D)^2}{2\sigma^2}}, \quad \text{for } i \in N_f(t_k)$$

$$\sigma = |N_f(t_k)| + \alpha \tag{5}$$

where $c, d, f, \alpha,$ and H are constant.

$$M_i(t_{k+1}) = M_i(t_k) + \alpha(t_k)A_f(i, t_k)[X(t_k) - M_i(t_k)], \quad \text{for } i \in N_f(t_k)$$

$$M_i(t_{k+1}) = M_i(t_k), \quad \text{for } i \notin N_f(t_k) \tag{6}$$

Equation (6) increases the level of weight update and gradually applies to all neurons for time and efficiently constructs the codebook by uniformly distributing master neurons to get accurate self-organization in the self-organizing feature map.

4 Experiment Result

4.1 Application to 2-Dimension Topology

In order to estimate Topological ordering and to analyze level of self-organization, we construct a 32 X 32 neural network and apply to the 2-dimension topology. A 2-dimension input vector and a rectangular topology are chosen, and an input vector has Uniform distribution in the rectangular topology.

A learning constant $\alpha(t_k)$ is monotonically decreased as time passes, and the equation used in this experiment is below.

$$a(t_k) = Z(1 - \frac{t_k}{T_1}) \tag{7}$$

where Z is initial learning constant, and T₁ is a maximum number of learning times. The equation used in this experiment is Equation (8) which is Gaussian function where c=0 and d=1.

$$A_j(i, t_k) = e^{-\frac{(H + \frac{t_k}{T_1})(i - H)^2}{2\sigma^2}}, \text{ for } i \in N_j(t_k)$$

$$\sigma = |N_j(t_k)| + 1 \tag{8}$$

The initial value of H is set by 7, which resulted the best performance during analysis process, and the initial learning constant Z is set by 0.3. Figure 5 is a result using Equation (4) with above conditions against the time domain. It represents a convergence rate using a number of training times as a dependent variable, and shows the proposed method converges with the least error.

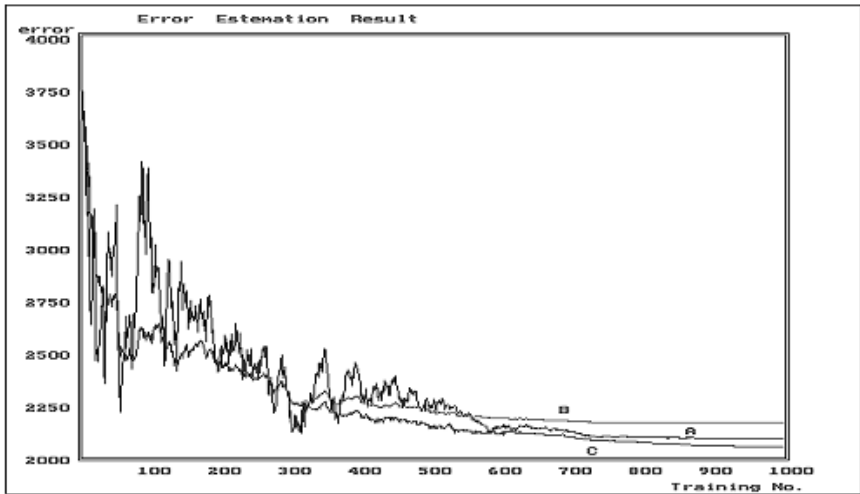
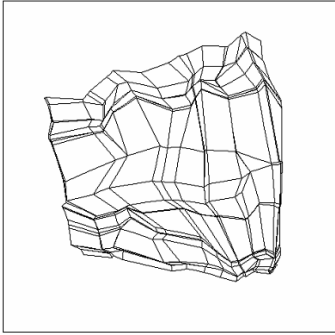
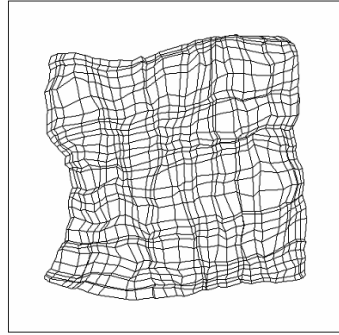


Fig. 5. Performance comparison according to a number of training times (A: Kohonen, B: Bavarian, C: Proposed method)

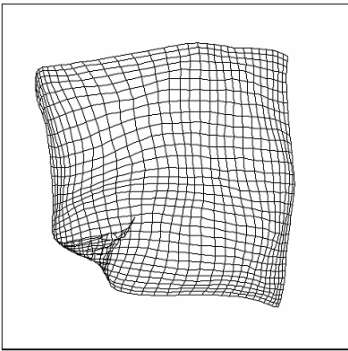
Fig. 6 displays neural networks trained 200 times and 1000 times on 2-dimension space. Coordinates of weight vectors of each neuron are marked on same coordinates of an input vector, and weight vectors of the nearest neurons are connected. That is, a cross point on the grid represents coordinates of a corresponding weight vector, and a line connecting two cross points means neurons, having weight vectors of those cross points, are adjacent. As a result of training the neural network, the proposed method resulted more accurate self-organization rather than Kohonen's method and Bavarian's method did.



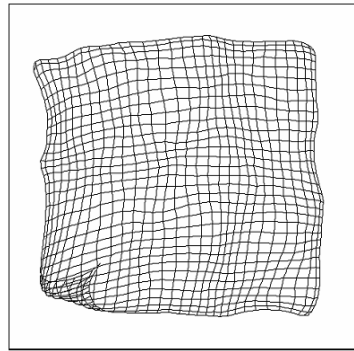
(a) Kohonen's method(200 times)



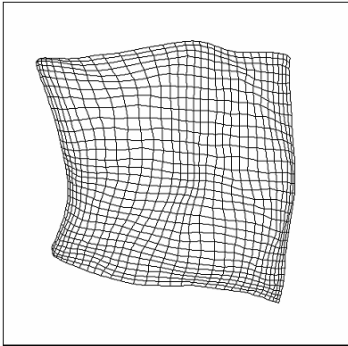
(b) Kohonen's method(1000times)



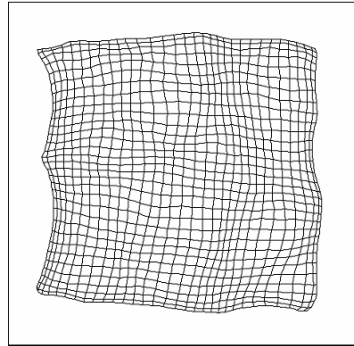
(a) Bavarian's method(200 times)



(b) Bavarian's method(1000times)



(a) Proposed method(200 times)



(b) Proposed method(1000times)

Fig. 6. 32 X 32 neural network trained 200 times and 1000 times

5 Conclusion

Among neural network models, Kohonen's self-organizing feature map sets up competition relationship according to Euclidean distance between one neuron and other

neurons in competitive layer, and these neurons perform unsupervised learning by self-organization. Also, this method converts intrinsic features of many input vectors into a feature map in 1-dimension or 2-dimension so that it can automatically solve classification problem of input vectors.

But this learning method using self-organizing feature map has weak points. A learning rate is slow and too many input patterns are required to converge into equilibrium state. In order to solve these limitations, this paper proposes Dynamic Gaussian function which can improve level of self-organization and convergence speed of the neural network by modifying, according to a number of learning time, a variance and a width of Static Gaussian function which is used as the neighbor interaction function in B. Bavarian's self-organizing feature map.

As a result of the experiment, the proposed method is better than most existing methods since the error curve is formed in almost linear instead of non-linear, and provides fast convergence speed into equilibrium state because the gradient of the curve itself is low. Therefore the proposed method can make a neural network learn with a fewer input patterns than Kohonen's method does if an input vector has Uniform distribution. And, as a result of applying to the voice recognition for performance analysis and estimation of proposed method, Kohonen's method, and B. Bavarian's method, the proposed method was resulted better than Kohonen's method and B. Bavarian's method.

References

1. Teuvo Kohonen, "The Neural Phonetic Typewriter", computer, Vol.21, No. 3, pp.11-22, 1988
2. Teuvo Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, 1989.
3. Teuvo Kohonen, "The Self-Organizing Map", Proc. of the IEEE, Vol.78, pp.1464-1480, Sep. 1990.
4. L.Paolo, T.Patrick, V. Nikolaos, " Modified self-organizing feature map algorithms for efficient digital hardware implementation", IEEE trans. on neural network, Vol.8, pp. 315-330, 1997.
5. B.George, Georgiopoulous, Michael, "Using self-organizing maps to learn geometric hash function for model-based recognition", IEEE trans. on neural network, Vol.9, pp. 560-570, 1998.
6. Teuvo Kohonen, E.Oja, A.Visa, O.Simula, "Engineering applications of self-organizing map", Proc. of the IEEE, pp.1358-1384, 1996.
7. P.Demartines and J.Herault, "Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets", IEEE trans. on neural network, Vol.8, pp.148-154, 1997.
8. Teuvo Kohonen, "Generalizations of the Self-Organizing Map", Proc. of IJCNN, Vol.1, pp.457-461, 1993.
9. Z.P.Lo and B. Bavarian, "Two Theorem for the Kohonen Mapping Neural Network", Proc. of IJCNN, Vol.4, pp.755-760, 1992.
10. Z.P.Lo Y.Yu and B. Bavarian, "Convergence Properties of Topology Preserving Neural Network", IEEE Trans. Neural Networks Vol.4, No. 2, March 1993.

Development of an Attack Packet Generator Applying an NP to the Intelligent APS

Wankyung Kim and Wooyoung Soh

Department of Computer Engineering, Hannam University,
Daejeon, S. Korea
{wankk12, wsoh}@hannam.ac.kr

Abstract. Security systems need to be tested on the network, when they are developed, for their security test and performance evaluation. Even though the security tests have to be done on the real network but, it is usually tested in a virtual test environment. APS (Attack Packet Simulator) is one of tools for performance test of security system on the virtual environment. In this paper, the development of an attack packet generator extracts the attack information from Snort rule and creates attack information in the Database using the extracted information applying intelligent APS. Also, the proposed generator generates high speed network attack packets to closely assimilate the real network for security system tests using an NP (Network Processor).

1 Introduction

Security systems are in general operating on a network involving a number of connected information systems. Therefore, such security systems need to be tested on the network, when they are developed, for their security test and performance evaluation. It is desired that the security test is done on the real network environment. However, it is usually tested in a closed network environment, due to the serious damages possibly occurred during the test and possibly propagated through the network. It is specially the case when the real network environment is too sensitive or important to allow any corruption of network or system during the test or when building a real-like test environment is too expensive. When the virtual test network is used, the problem is how to simulate the real network with various factors such as network speed, packet processing capacity, network load and etc. It is, for instance, very difficult to have enough packet amount and/or packet speed to assimilate the real network environment. Therefore, for the efficient and accurate test of the security system, it is necessary to have a pertinent method of providing the test environment with a high packet speed to assimilate the real attack environment.

This paper presents an attack packet simulator which performs test of information security to closely assimilate the real network for security system tests using a NP (a network processor equipped with an Intel chip). This paper applies intelligent APS [1] to H/W and creates attack packets.

Attack packets are constructed based on the rule set of Snort [2] which is an open source intrusion detection system.

The ASP's performance is better than others (such as SNOT [3], Mucus [4]) about intrusion detection rate and parsing success rate)

This paper is organized as follows. Section 2 describes the existing evaluation methods and the test conditions of security systems and APS. After describing, the design and implementation of the proposed attack packet simulator for security system tests in section 3, the results of performance evaluation by comparing the proposed packet generator with the existing one (Excalibur [5]) are discussed in section 5. Finally section 6 describes the conclusion and the future works.

2 Related Works

To develop and select a proper security system, it is necessary to have a proper evaluation method. Among other things, the security test and the performance test are the ones generally used [6], [7], [8], [9]. The security test is a test to observe vulnerabilities of a target system equipped with a security system using a vulnerability analysis tool and a well-known test is the 'penetration test' of CC (Common Criteria) [10], [11]. The performance test of information security system is a test which measures general traffic and throughput on the stress factors for information security system.

This chapter discusses the performance test and the test-bed in NSS [12] to derive the factors for the performance test and the configuration of test-bed for the proposed attack packet generator.

2.1 Factors of Performance Test for Security Systems

Common factors of performance test for security systems can be summarized as follows from NSS[12]:

- Limitation of authorized traffic for reliability
- Detection rate of defined intrusions
- Impact on network performance during process of received packets

2.2 Test Environment for Security Systems

The performance test environment for security systems can be divided into two categories: traffic environment or non-traffic environment, although it varies according to the test target. Traffic environment is configured as a real-network environment or a like. In case of testing on a real network environment, it is possible to test including the compatibility with OS. However, there can be a deviation of network traffic in accordance with when and how long the test is done, and further no guarantee that the necessary attacks happen during the test. Besides, if any security system without verification is installed on a real-network environment, it can be risky. On the other hand, a test on a network environment similarly configured to the real one can exclude the risk, but it bears the size and cost problem of network.

Non-traffic environment is a environment such that the packets including attacks are generated then the test is done with them using switch hub to target system. In this case, there is no risk from installing any non-verified security system and no cost of configuring a real-like network environment. However, it is difficult to test any stress factors as well as non-attack packets, because only the attack packets are used.

The traffic and non-traffic environments have its own merits and demerits. From the above discussion, the requirements of a test environment for security systems can be derived as follows:

- Can construct not only a single attack, but a complex set of attack
- Can construct non-attack packets as well as attack packets
- Can generate real-like traffic environment
- Can generate traffic over the processing ability of a target security system
- Should not be affected by the performance of the hardware installed with packet generator

The above 5 requirements can be summarized into 2 categories: capability of constructing complicated attack scenario, and capability of generating packets assimilating real network environment.

This paper intends to develop a network attack packet generator using a NP that can satisfy the above requirements and also provide a real network traffic capability and a real-like test environment for the security system test and development.

2.3 APS (Attack Packet Simulator) [1]

There are two kinds of packet providing system. One is to create intrusion related packets based on the snort rule set and then transmit them to a target system such as Snort and Mucus. The other is to gather intrusion related packets using TCP Dump data and then transmit them to a target system such as Packet Excalibur.

Most of the packet simulators provide useful intrusion packets but they must parse each time the packets are needed because they are not saving in a packet database. Furthermore, they have a low detection rate and a pure rule parsing success rate since they do not support the rule options of the latest Snort version.

APS has a different mechanism with those systems. The APS consists of 4 modules. First, an Intrusion Information Database Creating module collects necessary intrusion information by parsing Snort rules. Second, an Intrusion Information Database Management module updates the created intrusion information database. Third, a simulated intrusion Creation/ Transmission module creates the simulated intrusion and transmits to target IDS. Lastly, a Result Report module shows the transmitted results. This system works on Linux and uses MySQL for its database. Fig. 1. shows structure of APS.

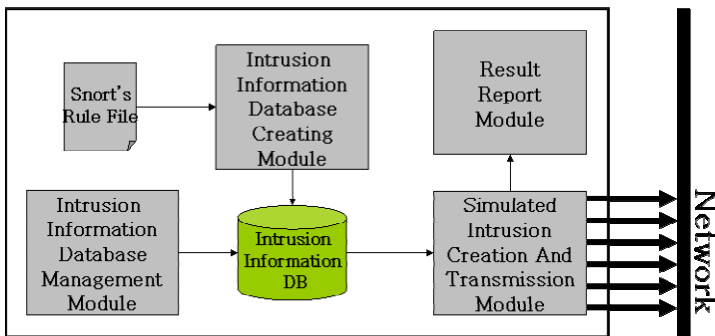


Fig. 1. Structure of Intrusion Simulator

This paper applies APS to H/W and creates attack packets.

3 Design and Implementation of the Proposed Attack Packet Generator

3.1 An Attack Packet Generator Using NP

The packet generator consists of two parts: Host part and NP Part. Host part mainly does the function of constructing the necessary packets, while NP Part mainly does the function of transmitting the constructed packets.

In this paper, the attack packet generator is designed and implemented such that the functions and performances can be accomplished according to the following conditions.

- An attack packet generator should be able to minimize the time of constructing and transmitting packets
- An attack packet generator should be able to transmit the correct packets of information to the tested system.

The above two conditions are set to meet the requirements derived from the objectives of the real network environment for testing the security systems.

3.2 Host Part

Host Part includes the user interface that serves the ease of constructing necessary attack packets, and also the DB of attack packet information derived by parsing the Snort rule set.

The host part uses RedHat 7.2, Kernel Ver. 2.4.17 and PCI Bus for communication between Host Part and NP Part [13].

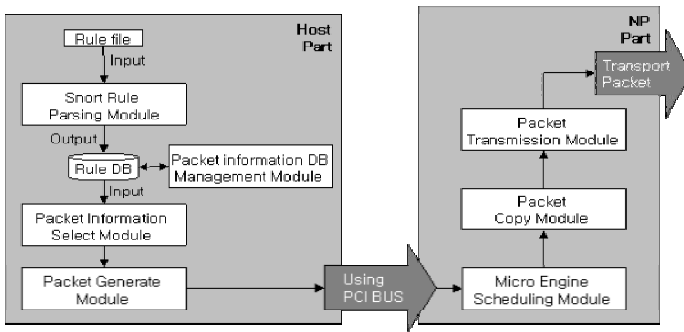


Fig. 2. Architecture of the Packet generator

Host Part is consisted of 4 modules: snort rule parsing module, packet information DB management module, packet information selection module and packet construction module. It provides the ease of managing DB and also the update function of rules with additional information through web GUI.

3.3 NP Part

NP Part is in fact a separate NP based system board and working with Host Part. In other words, NP Part receives the packets from Host Part through the PCI Bus and then transmits them after coping and scheduling. In this paper ENP-2506 by Radisys is used as NP. ENP-2506 uses Intel IXP1200 Network Processor, and supports two multi-optic ethernet port [14]. IXP1200 consists of Strong Arm Core Processor, six Micro Engines, SDRAM, SRAM, PCI BUS and IX BUS [15].

NP Part can be equipped with a separate OS, and in this paper Embedded Linux Kernel Ver. 2.3.99 is used. NP Part consists of Micro Engine, scheduling module, packet copy module and packet transmit module. It communicates with Host Part through Primary PCI Bus, and IXP1200 chip set communicates through Secondary PCI Bus to provide an external interface between IXP 1200 and a target system.

3.4 Construction and Management of Packet Information DB

The packet information DB is constructed with data gathered by parsing snort rule set. DB consists of a number of tables each of which has the attack information according to the attack type, and a key table which indices the corresponding attack information. The whole packet information is managed through web GUI.

3.5 Packet Construction

Through the key table, users extract the necessary information from DB and construct the desired packet. In other words, users input the transmission and destination IP addresses, and selects a desired attack type and an amount of packets. Then Packet construction module constructs the necessary packets, and then pushes the packets, the amount of packets, and the IP address information into the copy buffer of NP Part.

3.6 Copy and Transmission of Packets

Once the packets are pushed into the copy buffer, Micro Engine of NP Part executes Packet Copy Module which copies the packets by putting the IP address information to the packet header and pushes the results into the transmission buffer. Then Packet Transmission Module actually transmits the packets to the target system.

4 Performance of the Proposed Attack Packet Generator

4.1 Proposed Attack Packet Generator Using NP

To evaluate the performance of the proposed attack pack generator, the two requirements discussed in the previous chapter are evaluated.

Network attacks are generally consisted with more than two complicated attacks and/or sequential attack steps. Furthermore in a real network environment, there exist a great number of background data which may be attack related packets or normal packets to the target system. Therefore, a attack packet generator should be able to generate attack scenario including background data.

The proposed attack packet generator constructs attack packets based on the Snort rule set. Snort has a self-defined rule set of intrusion detection and it contains various attack packet information as a comparing factor with received packets. Therefore the proposed attack packet generator can generate various patterns of attack including complicated attack scenario.

4.2 Result of Test

4.2.1 Result of Simulation

This paper compares Snot and Mucus with Intrusion Simulator to verify the performance. The objects to compare are the parsing success rate of Snort's rule and the simulated intrusion detection rate of IDS.

It must conduct parsing to extract any information from Snort's rule. It was impossible to conduct parsing of rule options that have been added on current 2.X version because the existing programs conducted parsing according to the standard of Snort's rule 1.X version in parsing progress. It shows no differences in each rate of packet transmission in the study of the above.

In addition, when the simulated intrusion is created, the completion rate of the simulated intrusion depends on the accuracy of rule file. The existing programs have low completion rate because these programs use the rule file of Snort 1.X version. There is a striking contrast when we are using Snort 2.X version. Table 1 shows the simulated intrusion detection rate and the parsing success rate on each Snort version

The performance test was conducted by comparing the existing systems in different aspects. Table 2. compares various characteristics of the existing programs and the proposed generator.

Table 1. Simulated intrusion detection rate and parsing success rate on each snort version

	Snort 1.8.6 (Total : 1,212)	Snort 2.2.0 (Total : 1,838)
Snot	396/528(32.7%)	421/582(22.9%)
Mucus	684/725(56.4%)	769/831(41.8%)
Intrusion Simulator	967/1,100(78.8%)	1,097/1,272(59.7%)

Table 2. Characteristics of existing systems and proposed generator

	Snot	Mucus	Proposed Generator
Composition of Scenario	No	No	Yes
Creation of DOS attack packet	No	Yes	Yes
Exploit	Yes	Yes	Yes
Insert dummy code	No	Yes	Yes
Possession of Rule DB	Yes	Yes	Yes
Open-Source	Yes	Yes	Yes

4.2.2 Result of Performance Test Packet Excalibur and Proposed Generator

An open source packet generator, Packet Excalibur, has been used for performance evaluation of security systems. It basically reproduces pre-dumped packets for a pre-defined time period from a target network area to the tested system. The proposed attack packet generator is compare to evaluate the performance. The hardware specification of the proposed generator and the Packet Excalibur is listed in Table 3.

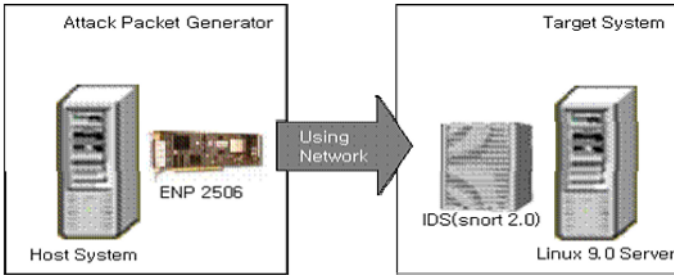


Fig. 3. Test Environment

Table 3. Hardware Specification

Hardware Specification	
CPU	P4-2.6Ghz
RAM	512Mbyte
HDD	120Gbyte
O.S	RedHat Linux 7.1

Packet Excalibur is a script-based network packet generator. It generates packets including background data according to the pre-defined script. For comparison of the two generators were compared by generating 1 GByte of packets including DoS attack packets and background data. The time required generating and transmitting and other factors are listed in Table 4.

Table 4. Comparison of the proposed generator and Packet Excalibur

	Generating packets	Packet Construction and Transmit time
Packet Excalibur	1 Giga byte	4min 40sec
Attack Packet generator	1 Giga byte	2min 50sec

The proposed attack packet generator was at least 40% faster than Packet Excalibur in generation and transmission of packets with the different amount of packets. The packets generated by the two were detected by Snort more than 95%, which means they generated packets and transmitted them at the reasonably acceptable rate.

5 Conclusion

The primary purpose of this study is to develop a way of providing a test environment for security systems with several desired characteristics including the ease of assimilating a distributed attack scenario in a single system and the high transmission rate of attack packets in a given short time period with a reasonable packet loss.

The proposed attack packet generator can generate not only attack packets but also background data, thus it can assimilate real network traffic. The proposed generator can minimize the effect of hardware environment, since it uses a NP. It means that with this generator one can configure a real-like network environment in a single Linux system. Thus one can reduce the cost and time of configuring a virtual network environment for testing security systems.

As the result of comparing the performance, processing time of the proposed generator is faster than software based packet generator, Packet Excalibur.

It is expected that the proposed system helps provide a way of developing an efficient test environment closely assimilating a real network environment for the precise test of security systems.

References

- [1] Junsang Jeon, Wooyoung Soh: Design and Implementation of An Attack Packet Simulator for Performance test of Information Security System, ICCMSE (2005)
- [2] Martin Roesch, Chris Green, SourceFire, INC.: Snort Users Manual", <http://www.snort.org>
- [3] Sniph, Snot, <http://www.sec33.com/sniph/> (2001)
- [4] Darren Mutz, Giovanni Vigna, Richard Kemmerer: An Experience Developing an IDS Simulator for the Black-Box Testing of Network Intrusion Detection Systems (2003)
- [5] <http://www.securitybugware.org/excalibur/>
- [6] Nicholas J. Puketza, Kui Zhang, Mandy Chung, BisWansth Mukherjee and Ronald A.Olsson: A Methodology for Testing Intrusion Detection Systems, IEEE Transactions on Software Engineering, Vol.22, No.10 (1996) 719–729
- [7] H.Debar, M.Dacier, A.Wespi and S.Lampart: An Experimentation Workbench for Intrusion Detection Systems, IBM Zurich Lab, Research Report (1998)
- [8] Richard P. Lippmann, David J. Fried, Isaac Graf, Joshua W. Haines, Kristopher R. Kendall, David McClung, Dan Weber, Seth E. Webster, Dan Wyschogrod, Robert K. Cunningham, and Marc A. Zissman: Evaluation Intrusion Detection Systems : the 1998 DARPA Off-Line Intrusion Detection Evaluation, Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (2000)
- [9] Robert durst, Terrence champion, Brian written, Eric miller, and Luigi spagnuolo: Testing and Evaluating Computer Intrusion Detection Systems, Communication of the ACM, Vol.42, No.7 (1999) 53–61
- [10] CCRA(Arrangement on the Recognition of Common Criteria Certificates), <http://www.commoncriteria.org>
- [11] CC: Common Criteria for Information Technology Security Evaluation, Version 2.1, CCIMB-99-031 (1999)

- [12] An NSS Group Report V 1.0, "Intrusion Prevention Systems(IPS)", Group Test, NSS, Jan 2004.
- [13] RadiSys Corporation, "Linux Setup guide for ENP-XXXX", <http://www.radisys.com>
- [14] RadiSys Corporation, "ENP-2506 Hardware Reference Manual", <http://www.radisys.com>
- [15] Intel Corporation, "IXP1200 Hardware Reference Manual", <http://www.intel.com>

Class Based Intelligent Community System for Ubiquitous Education and Medical Information System*

Seoksoo Kim**

Hannam University, Department of Multimedia Engineering, Postfach , 306 791
133 Ojeong-Dong, Daedeok-Gu, Daejeon, Korea
sskim@hannam.ac.kr

Abstract. Various kinds of community method is being researched to design effective systems within the ubiquitous environment. The research aimed to provide smooth mutual interlock among systems by grouping classified class based intelligent systems and focused on easy scalability and modification. The basic model of the research was; each class contains its own applications and these services mutually interlock in complex platforms. The community considered in this research is a basis for providing combined management of medical equipments and applications within environment where compatibility is causing problems during migration, and is suitable for designing medical information system within the intelligent ubiquitous environment.

1 Introduction

Generally, when the ubiquitous computing environment is designed using each detailed technologies, array of these devices or services search cooperative target they require according to the need and provide temporary cooperation services. That is many of devices and services only process given number of situations and consideration for repetition of these phenomenon has been omitted. These phenomenons are problems often seen from all systems using ubiquitous system. In particular, there is an immediate need to solve these problems for medical services where lives of people are depended to these services. In the case of medical services, various medical accidents are occurring during offline status and the possibility of much more problems occurring during application of online medical diagnosis is high. This suggests that, there is a need to equip mutual security system structure to reduce conflicts between various medical systems that have been introduced and to maintain integrity of information. Therefore, the researcher will design a class based community system that will categorize systems, as a means to design systematic medical information system in ubiquitous environment.

* “This work was supported by a grant No. (A02-2006-001-003-5) from Ministry of Commerce, Industry and Energy”.

** Corresponding author.

2 Related Literatures

2.1 Definition of Community

“Community” here means interactive relationships with all devices and services in a ubiquitous computing environment that has common goal application. That is, one community is made up of “goal, members, and functions of the members.” A community can be a member of other community and one device or service can be included in various communities. In other words, following the particular goal of community, it can form inclusive relationships and shared devices can be used as a service.

With the above definition or metaphor of community as basis, “community-based computing” aims to develop service and devices and unify those developments into creating new services. Additionally, instead of devising new development to meet the service demand arising dynamically, necessary services can be constructed from combining services and devices from already established ubiquitous computing environment [1].

2.2 Objective of the Community

The objective of engineers designing overall ubiquitous computing environment is for the users within the ubiquitous computing environment to comfortably and safely enjoy their lives. Therefore, objective of community composing of ubiquitous computing environment will correspond to above goal. As an extreme, whole community will desire to achieve an ideal objective and tasks involved in transition to this ideal situation will be carried out by the community. That means, a community having larger extent of objectives than a community with large objective may exist [2]. Specially, in the medical information community, approaching the system as a community might be very practical, as a means to accept various extended technologies and to connect them together.

2.3 Ubiquitous Education

From computer-based learning (CBI) to Web-based learning (WBI), the development of classes in computer environment demands huge investments of time and money. In order to overcome the inefficiency in development, e-learning researchers are looking for the reuse of developed contents and the sharing of contents developed by third parties. By developing a system for reusing a part or the whole of existing contents or sharing contents created by third parties, we can save a lot of time and money. Such efforts have been integrated into the establishment of e-learning technology standard.

2.4 Medical Information

The medical information tries to raise efficiency of introducing medical technologies by combining medicine with the information technology. This can be classified in to, medical information that will systematically manage various information required to provide patient diagnosis, medical education, medical research and medical management, the

hospital information that introduces hospital information system to design digital hospitals, e-Health where information system is designed to support decisions of doctors efficiently and rationally during patient diagnosis and personal health management by providing medical knowledge and patient information by utilizing information technology, and u-Health which utilizes ubiquitous technologies for the health management system. In particular, services such as Slipless, Paperless, Chartless and Filmless digital hospitals, mobile medical environment extending from within the hospital to living space of patient and at home medical services/remote medical technologies are being provided as the ubiquitous health care infrastructure is actualized[3].

In addition, EHR (Electronic Health Records) such as national personal electronic health record, sharing of information among medical facilities through information standardization, etc. and cooperation between organizations are being extended.

Figure 1 is showing schematic of e-Hospital constructed by combining existing cooperative system as the basis. Aside from the fundamental components such as patients, cooperative hospitals, pharmacies and medical logistic suppliers, we can realize various management systems such as ERM that stores all information related to clinical diagnosis of patients and supporting memory of clinical doctors by storing clinical records electronically, PACS which collects various clinical images occurred within the medical environment in digital data, saves these in the storage devices of the computer then transmits these information to various other computers connected to the network for utilization, POC used to process clinical information of patients efficiently right at the scene of diagnosis without limitation of time and space, ERP used to enhance hospital management efficiently by combining the diagnosis system and management tasks, Groupware System Electronic Cooperative Systems(Electronic Mail/ Transaction, Office Management System), DW modeling, construction of Data Mart, etc. are being combined.

It can be seen that various systems are connected together according to necessity creating a system that is combined and extended. It is predicted that these union in

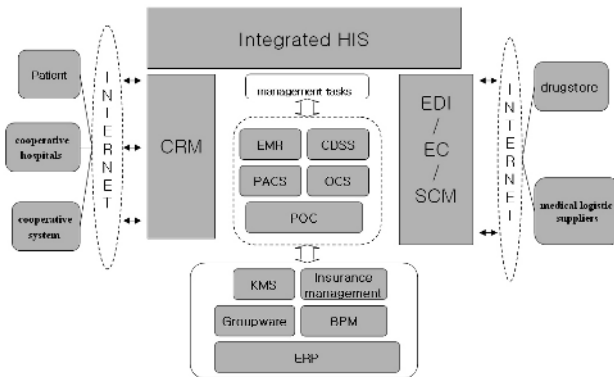


Fig. 1. e-Hospital constructed

medical information will be actively progressed even during the ubiquitous environment. However, extensive costs are required for modification of large network systems and for introducing new technologies. Furthermore, there exist difficulties in coping with fast changes in the fundamental technologies.

2.5 Ubiquitous Medical System

Ubiquitous Medical System has already been adopted by hospitals together with the introduction of POC. These technologies will support various types of medical system as the technology involved in basic infrastructure develops. A representing example is the u-Hospital, which utilizes RFID into medical system by using sensors. Massachusetts General Hospital of United States has already adopted RFID and GPS technology to identify location of medical equipments, doctors, nurses and patients. Also, diagnosis can be made remotely within normal households as “at home” medical services utilizing “at home” medical devices and internet develop[4].

The ubiquitous medical system utilizes existing systems that had been developed during the medical information process and has adopted sensors for diagnosing patients effectively[5]. However, countermeasures are being demanded for solving problems arising from compatibilities among equipments and data exchanges.

3 Designing and Evaluating MIC System

In this chapter, the author will design a community required to introduce systems reliably during the procedure of medical information transition. Various types of medical information system is being introduced as the ubiquitous environment develops, causing various compatibility problems. Therefore, class will be constructed between compatible systems and these classes will be combined to provide stability of medical information system. Suggested model has basic design as shown in Figure 2.

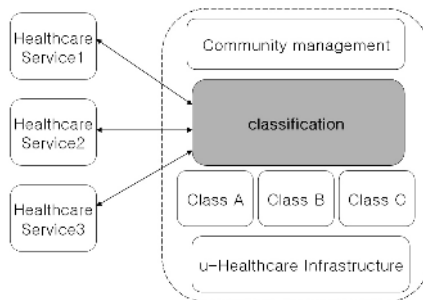


Fig. 2. The suggested medical community system

The suggested medical community system will classify healthcare services basing on the classification module. At this time, the classified healthcare service will be registered under the concerning class. Community management module will take role of applying and modifying standards applied by the classification module and will

input standards to examine connection between each class. At this time, the basic infrastructure for supporting ubiquitous environment must have been structured and communication between each class will base on the wireless communication.

3.1 Medical Information Community System Information Classification

Standard classification of suggested medical community system is defined in the Table 1 to manage healthcare systems.

Table 1. Classifications for standard classification of the suggested medical community system

Classification	Define	Example
Task and management	Provide tasks required to execute medical activities	CRM
Diagnosis system	System that is involved with medical activities directly	PACS
Additional system	Systems providing additional support during diagnosis	OCS, POS
Sensor based system	Used sensor system	RFID Mobile
Communication method	Cable or wireless method	POC
Data system	Stores information	ERM

There are 6 classifications for standard classification of the suggested medical community system. System used to provide tasks required to execute medical activities are classified in to task and management system and systems that is involved with medical activities directly is defined as diagnosis system. Systems providing additional support during diagnosis, such as stock management and diagnosis delivery system, are defined as additional system and sensor driven systems such as RFID is defined as sensor based system. Communication method will be defined following cable or wireless method and lastly, system like ERM that stores information is defined as data system. Classified systems within the community system classification may alter depending on the situation. Specially, systems having high relativity in information processing will be treated as being the same Class, allowing easy construction of community.

3.2 Designing Suggested Medical Information Community System

Medical information community should be designed basing on virtual scenario. This is a countermeasure of the research trying to support various problems that may arise during the medical information processing through situational class bases. The scenario of this research will be creation of mobile device community between the medical information system within the hospital and the mobile devices of the patient when the patient visits the hospital.

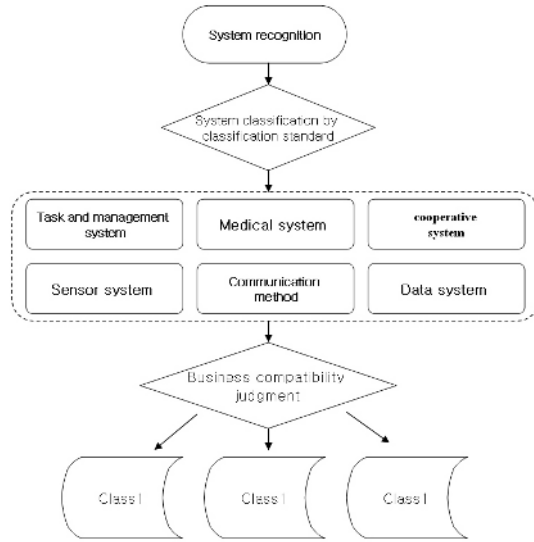


Fig. 3. Classification procedures

Following the scenario, the classification will under go procedures as shown in the Figure 3.

At this time, there is high tendency for the definition of work compatibility section to alter depending on the situational recognition. Therefore there is a need for the Community management to assign a standard by recognizing the situation. Specially, consideration for compatibility and data migration is necessary when a new system is introduced to the existing system.

Result of classification of each class following the introduction of scenario is as follows in Figure 4.

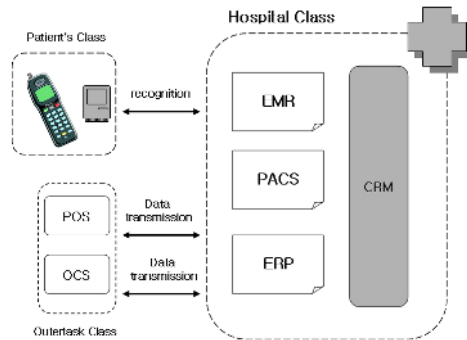


Fig. 4. Class structure for Medical Information system

In the classification of each class following scenario introduction, classes are classified in to Patient class, Hospital Class and Outer Task Class. With the consideration of portability and convenience, the Patient Class will consist of sensor based system and wireless communication method system based on the RFID mobile. The Hospital Class will be comprised of Electronic Medical Diagnosis Record (ERM), Medical Image Storing Information System (PACS) and the Enterprise Resource Management System (ERP), furthermore, the Outer Task Class will be comprised of Diagnosis Delivery System (OCS) and Stock Management System (POS).Cable and wireless communication between each class has been assumed to be possible following the result of medical information system class structures. Also, sensor information of Patient Class was assumed to provide required patient information to necessary systems within the Hospital Class.

3.3 Application of Class Based Medical Information Community System

By applying suggested medical information system to actual emergency medical situations, following structures as shown in <Figure 5> had been designed.

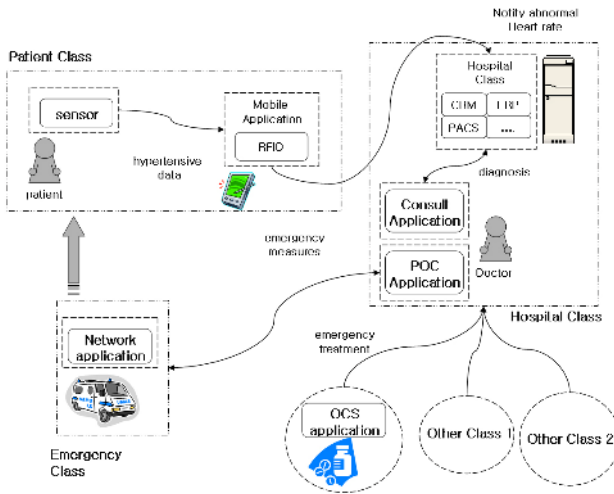


Fig. 5. Using of Medical information community system

Abnormal heart beat of patients with heart problems among Patient Class will be detected through sensor recognition and this information will be transmitted to the Hospital Class through mobile devices. At this time, sensor information will be analyzed, classified and transmitted to doctors then will trigger occurrence of emergency situation. After this, patient transfer will be ordered by communicating with ambulance while the emergency treatments will be executed through remote diagnosis system. Mutual interlock among systems plays an important role in complex situations like this. As a result of applying medical information community system suggested in this research, to a specific example, we were able to design class classification and structure where addition of class can be done easily. Furthermore,

community classified in to class based can be simply expanded, provide environment that can cope with limited services while providing much effective services to users utilizing medical information as user unit gets recognized as a class.

3.4 System Evaluation

Various kinds of community method are being researched to design effective systems within the ubiquitous environment. The research aimed to provide smooth mutual interlock among systems by grouping classified class based systems and focused on easy scalability and modification. The basic model of the research was; each class contains its own applications and these services mutually interlock in complex platforms. The community considered in this research is a basis for providing combined management of medical equipments and applications within environment where compatibility is causing problems during migration, and is suitable for designing medical information system within the ubiquitous environment.

4 Conclusion

The class based community system suggested to design medical information system, allows expansion of possible service targets by designing class units for the complex structured medical information, reduces unnecessary waste of resources such as repeated by interlocking systems and will provide convenient managements by classifying individual users into objects.

Lastly, further researches on community management module and security protection scheme for classes are required, to clearly define changes in classes following situational changes.

References

1. Mohan Kumar, et al., "PICO : A Middleware Framework for pervasive Computing." IEEE Persive Computing, July/September 2003.
2. Bin Wang, John Bodily, and Sandeep K. S. Gupta, "Supporting Persistent Social Group in Ubiquitous Computing Environments Using Context-Aware Ephemeral Group Service". Proc. 2nd IEEE International Conference on Pervasive Computing and Communications (PerCom), Orlando, FL, March 2004, pp.287-296.
3. Martin, T., Jovanov, E., and Raskovic, D., Issues in Wearable Computing for Medical Monitoring Applications: A Case Study of a Wearable ECG Monitoring Device. in Proceedings of ISWC 2000, (Atlanta, U.S.A., 2000).
4. Istepanian, R.S.H., Jovanov, E., and Zhang, Y.T., Guest Editorial Introduction to the Special Section on M-Health: Beyond Seamless Mobility and Global Wireless Health-Care Connectivity. in IEEE Transactions on Information Technology in Biomedicine, 8(4). December 2004. 405-414.
5. Jovanov, E., Milenkovic, A., Otto, C., and de Groen, P.C., A Wireless Body Area Network of Intelligent Motion Sensors for Computer Assisted Physical Rehabilitation. in Journal of NeuroEngineering and Rehabilitation, 2 (6). March 2005.

Intelligent Anonymous Secure E-Voting Scheme

Hee-Un Park and Dong-Myung Shin

Korea Information Security Agency, Seoul, Korea
{hupark, dmshin}@kisa.or.kr

Abstract. Various selection processes are present in the real world. Among them, some processes are open to everyone and others allow only members to validate them. These choice and selection types have been developed using cryptography protocol model. But, further study of definite anonymous secure selection protocols based on a cryptography is needed. In this paper, we propose the 'Intelligent Magic Sticker' scheme which is anonymous secure selection to get the anonymity of an originator, and which blinds the selection result from other entities on the open network. Because this approach protects a selection result of an originator using one-way anonymous authentication service, although the selection information is opened, the Magic Sticker scheme can be used in electronic voting, fair E-money and electronic bidding etc

Keyword: Anonymous Secure Selection, Intelligent Magic Sticker, Anonymous Authentication.

1 Introduction

A new media paradigm called 'Information Society' supports real life digital information services tremendously in the digital open network. In the background of these services, the various cryptography schemes are applied to it to get the safety and trusty. The cryptographic schemes are applied and adapted to many areas such as the digital signature service for the integrity, authentication, blinding service for the privacy, and anonymity of a user identifier etc.

Above services are based on a sending/receiving information entity and a message security. In a different way, a research for 'one-way anonymous authentication service' has been in progress to keep the anonymity of the information originator and authenticating the message based on the protocol for a specific group and information. Among them, electronic voting, electronic bidding, anonymous conference, and fair E-money can be representative examples. These services have following features:

- This service is based on specific group members.
- Figuring out an originator's real identity from his/her transmission information is impossible.
- The authentication to prevent forgery and falsification of information is possible.
- When the illegal information is made by the originator, group members can identify him/her.
- The third party can not confirm the selected information without the originator's permission.

Above special features are important means for offering safety to different sets of users with logically contradicting requirements. As you see, 'Anonymous Secure Selection Scheme' an originator processes secure selection in according to free will and guarantees the anonymity of his/her identity[1].

In the open network, some researches are proposed to solve the anonymous secure selection service only in parts, but integration mechanisms have not been studied yet. Therefore in this paper, we propose the 'Intelligent Magic Sticker' scheme which is anonymous secure selection to get the anonymity of an originator, and which blinds the selection result from other entities on the open network. Also we apply this scheme to the electronic voting system and certify the efficiency of the system.

2 Cryptographic Elements

In this chapter, we describe the Magic Sticker scheme that can be applied to anonymous secure selection service. This scheme is composed of three parts: TC(Trust Center), Confirmer, and Originator. Also, because information confirming is processed only by an originator, and his/her identity is anonymous from other parts, this scheme can be applied to one-way anonymous authentication services.

2.1 Intelligent Magic Sticker

Magic Sticker is a 2D holography film type sticker that includes more than two images and shows different images according to the degree of polarization angle θ . Just in time, an originator can choose a collected spot CS_k in specific image Img_k on Magic Sticker, and generates secure selection information D_i through selection information v_i , and enters D_i into this spot as degree of angle θ . In this section, we describe the Magic Sticker scheme scenario based on TC, Confirmer, and Originator.

Phase 1. TC sets the originator group $OG_i(i = \{1, 2, \dots, n\})$ that is a service group. Then it generates image set $Img_k(k = \{1, 2, \dots, m\})$. It makes duplicate images and grants the order numbers n_i to this duplicate images.

Phase 2. Confirmer sets image spot information CS_k which is used in generating the originator's selection result D_i . Because CS_k is a tool for an originator's decision making on selection processing, this is independently constructed with image set Img_k .

Phase 3. TC and Confirmer sends his/her Img_k and CS_{ki} to a fair distributor in secure way.

Phase 4. The distributor gives Img_k and CS_k to originators randomly using the confirming group membership process. Using this process, an originator can keep the anonymity and can be authenticated by the distributor.

Phase 5. The originator calculates a selection result D_i using the image information Img_k , collected spot information CS_{ki} , and the random value θ . When all processes are

finished, it sends the selection result and ordering number n_i to TC and Confirmer in a secure way.

Phase 6. TC and Confirmer exchange their secret information which is Img_k and CS_k .

Phase 7. The originator sends its θ to TC and Confirmer in a secure way. Using the received information D_i from Originator, TC and Confirmer can verify the selection information v_i . So, even with the processing entities of TC and Confirmer, they can not verify the selection information v_i without the Originator's authorization.

If a third party needs to confirm D_i , it must know above secret information. Although this secret information is disclosed, the proposed scheme provides anonymous secure selection, since this keeps the anonymity of the originator.

2.2 Secure Selecting Using Signcryption

In this section, we discuss how an originator operates secure selection using the distributed information and verifies the selection information. In addition, we show that this scheme can be applied to real world, through signcryption scheme introduce in the secure selection phase[2].

1) System parameters

- p, q : Large prime numbers, $p-1/q$
- g : Integer with order q modulo p chosen randomly from $[1, \dots, p-1]$
- H : One-way hash function
- $x_{gi} \in \mathbb{Z}_q^*$: The private key list for group signature
- $y_{gi} = g^{x_{gi}} \text{ mod } p$: The public key list for group certification

2) Protocol Schema

– Originator

P 1. He/She generates θ in this way.

$$\theta, x \in_R [1, 2, \dots, q-1]$$

P 2. After deciding a selection information v_i , he/she calculates D_i using a order number n_i , image information Img_k , collected spot CS_{ki} and his/her random number θ in which way.

$$D_i = (n_i \parallel Img_k * (v_i \oplus \theta) \oplus CS_{ki})$$

P 3. He/She generates k and k' to encrypt and sign D_i .

$$k = H(y_{TC}^x \text{ mod } p) \text{ (} y_{TC} \text{ is TC's public key)}$$

$$k_1 \parallel k_2 = k$$

$$k' = H(y_{Cer}^x \text{ mod } p) \text{ (} y_{Cer} \text{ is Confirmer's public key)}$$

$$k'_1 \parallel k'_2 = k'$$

P 4. Originator processes the encryption and signature as follows;

Note. In this phase we describe the transmitted information to Confirmer using k' .

$$c' = E_{k'_1}(D_i)$$

$$\begin{aligned} r' &= \text{KH}_{k_2'}(D_i) \\ R' &= g^{r'} \bmod p \\ s' &= x/(r + x_{gi}) \bmod q \end{aligned}$$

P 5. Originator sends (c, R, s) and (c', R', s') to TC and Confirmer.

$$\begin{aligned} (c, R, s) &\rightarrow \text{TC} \\ (c', R', s') &\rightarrow \text{Confirmer} \end{aligned}$$

– TC and Confirmer

P 6. When secure selection phase is finished, TC and Confirmer exchange his/her secret information using other's public key.

P 7. They decrypt and verify D_i using the received information from Originator.

Note. In this phase, we describe the Confirmer verification process.

$$\begin{aligned} k' &= H((y_{gi} * R')^{s' + x_{cer}} \bmod p) \\ k_1' \parallel k_2' &= k' \\ D_i' &= D_{k_1'}(c') \end{aligned}$$

When $R' = g^{\text{KH}_{k_2'}(D_i)} \bmod p$ then D_i is accepted.

– Originator

P 8. When the protocol is finished, it sends θ to TC and Confirmer secure way to verify v_i .

– TC and Confirmer

P 9. They verify v_i using the received information θ and exchanged values.

$$\begin{aligned} (\text{Img}_k * (v_i \oplus \theta) \oplus \text{CS}_{ki}) \oplus \text{CS}_{ki} &= \text{Img}_k * (v_i \oplus \theta) \\ (\text{Img}_k * (v_i \oplus \theta)) / \text{Img}_k &= (v_i \oplus \theta) \\ ((v_i \oplus \theta) \oplus \theta) &= v_i \end{aligned}$$

The proposed schema gives the anonymity of the originator, because this does not permit the confirmation of an originator's identity, but allows the confirmation only to membership using the group signature by TC and the confirmer. Also, since TC and the confirmer can not verify the originator's selection value without θ , we know that this schema provides the secure selection.

3 Intelligent E-Voting Scheme

In this chapter, we describe the electronic voting that is based on new one-way anonymous authentication services. On the open network, the efficiency of electronic voting is an interesting field on research.

In spite of its importance, when the voting is operated in the open network, the electronic voting scheme has some security requirements and if these requirements are not met, the voting scheme is not to be trusted.

Therefore, in this paper, we look into the requirements for electronic voting's security, and we propose a new electronic voting scheme that satisfies the all the

requirements. Especially we apply the Magic Sticker scheme for anonymous secure selection to the electronic voting system and analyze whether this satisfies all the requirements.

3.1 Requirement

1) General Requirements

The most important characteristic in the voting scheme is an ability to maintain voting security or anonymity. This means that the voting scheme must prevent the third party from knowing the relationship between the voting and the voter. Also to maintain the safety for electronic voting, more ability is required. This is described in the electronic voting's general requirement.

- Privacy : This feature ensures that an individual vote will be kept secret from any coalition of parties that does not include the voter himself/herself.
- Democracy : "One man one vote one time".
- Authentication : Only authorized voters can vote
- Fairness : No one should be able to determine the voting value through the other's results.
- Integrity : No one should be able to manipulate other's vote. Such an act should be detected.
- Universal Verifiability : Any party, including a passive observer, can convince himself/herself that the voting is fair, i.e., that the published final tally is computed fairly from the ballots that were correctly casted.

2) Specific Requirements

Since the electronic voting is operated on the open network, a step for confirming the voting result must be performed. Also since the voting result is sent to CTA and to a tally center after the voting process is over, the voting manager's illegal act and a vote buying action must be tolerated. Therefore, the electronic voting system must satisfy the following the specific requirements:

- Receipt Free : This feature enables voters to hide how they have voted even from a powerful adversary who is trying to coerce it[3][4][5][6].
- Robustness : This feature ensures that the system can recover from the faulty behavior of any coalition parties[7][8][9].

3.2 New Proposal

The proposed scheme applies the Magic Sticker scheme to the electronic voting system to satisfy general and specific requirements. This scheme disapproves the forgery and confirming of voting result from the third party and provides private voting that satisfies the anonymity.

1) System parameter

In this section, we describe system parameters used in the new proposed scheme.

- V_i , G_i , T_i , CTA : Voter, Voting booth, Tally center, and CTA(Central Tabulate Agency)

- $v_i, V_{\text{result}i}$: Voting value and result of the voter i
- $r_{\theta i}, r_{\text{salt}i}, r_{\text{seed}}$: Voter i 's random number θ , CTA's salt value, and Tally center's seed value
- ID_i : Voter i 's alias identity generated by CTA
- p, q : Large prime number, $p-1/q$
- g : Integer with order q modulo p randomly chosen from $[1, \dots, p-1]$
- $K_{S^*} \in_{RZ^*_q}$: $*$'s private key
- $K_{P^*} = g^{K_{S^*}} \bmod p$: $*$'s public key
- $EK_{P(S)^*}, DK_{P(S)^*}$: $*$'s public key or private key encryption and decryption
- H : One-way hash function
- $x_{gi} \in_{RZ^*_q}$: The private key list for group signature
- $y_{gi} = g^{x_{gi}} \bmod p$: The public key list for group certification

2) System protocol

Phase 1. Preliminary phase

Step 1.1 CTA

- (1) CTA writes out the voting list with confirming voting objects and broadcasts the information related to the voting to all voters.

Voting information $\rightarrow V_i$

- (2) It generates a salt value $r_{\text{salt}i}$ and an alias identity ID_i ($i = \{1, 2, \dots, n\}$), which are given to voter. Here, $r_{\text{salt}i}$ is stored in a voting token and a voter, and voting booth can view this information.
- (3) CTA signs ID_i , and concatenates ID_j ($j \subset i$) with a salt value $r_{\text{salt}j}$ which are Digital Voting Token(DVT) elements. It then sends a signed and encrypted DVT with a voting booth's public key to a voting booth G_i .

Note. the number of DVT is based on voter's number in a voting list.

$EKP_{G_i}(EKS_{CTA}(ID_j || r_{\text{salt}j})) \rightarrow G_i$

Step 1.2 Tally center T_i

T_i generates and signs a seed value r_{seed} and sends a signed and encrypted r_{seed} along with a voting booth's public key to G_i . At this time, only T_i takes the seed value.

$EKP_{G_i}(EKS_{T_i}(r_{\text{seed}})) \rightarrow G_i$

Phase 2. Voting phase

Step 2.1 Voter V_i

When the voting day, a voter V_i authenticates himself to CTA and joins a voting booth G_i on the Internet.

$EKS_{V_i}(\text{Voting list}) \rightarrow \text{CTA}$

Here, a voting list will be compared with a tally center's summary sheet, in the confirming phase.

Step 2.2 Voting booth G_i

He/She sends one of the DVTs to V_i randomly.

DVT $\rightarrow V_i$

Step 2.3 Voter V_i

- (1) V_i chooses two random numbers $r_{_0i}$ and x .
 $r_{_0i}, x \in_R [1, 2, \dots, q-1]$
- (2) He/She selects a voting value v_i .
- (3) He/She generates the voting results $V_{_resulti}$ using his/her alias identity ID_i , CTA's salt value $r_{_salti}$, T_i 's $r_{_seed}$, and $r_{_0i}$ in this way.

$$V_{_resulti} = ID_i \parallel (v_i \oplus r_{_seed}) \parallel (r_{_salti}(v_i \oplus r_{_0i}) \oplus r_{_seed}) \parallel \dots \parallel (v_k \oplus r_{_seed})$$
 Note. $V_{_n} = \{\text{selection element1}, \dots, \text{selection elementn}\}$, $v_i \in V_{_n}$, $V_{_s} \subset V_{_n}$, $v_j, v_k \in V_{_s}$
- (4) V_i calculates k and k' for group signcryption as follows.

$$k = H(K_{P_CTA}^x \text{ mod } p)$$

$$k_1 \parallel k_2 = k$$

$$k' = H(K_{P_Ti}^x \text{ mod } p)$$

$$k'_1 \parallel k'_2 = k'$$
- (5) He/She processes the encryption and signature.
 - For CTA :

$$c = E_{k_1}(V_{_resulti})$$

$$r = KH_{k_2}(V_{_resulti})$$

$$R = g^r \text{ mod } p$$

$$s = x/(r + x_{grpi}) \text{ mod } q$$
 - For T_i :

$$c' = Ek_1'(V_{_resulti})$$

$$r' = KH_{k_2'}(V_{_resulti})$$

$$R' = g^{r'} \text{ mod } p$$

$$s' = x/(r' + x_{grpi}) \text{ mod } q$$
- (6) V_i sends the generated information (c, R, s) and (c', R', s') , and send it to CTA and a tally center T_i .
 $(c \parallel R \parallel s) \rightarrow \text{CTA}$
 $(c' \parallel R' \parallel s') \rightarrow T_i$

Phase 3. Confirming phase

Step 3.1 CTA and Tally center T_i

- (1) When the voting time is over, CTA and T_i exchanger their secret information secure way.

$$EKP_{T_i}(EKS_{CTA}(ID_i \parallel r_{_salti})) \rightarrow T_i$$

$$EKP_{CTA}(EKS_{T_i}(r_{_seed})) \rightarrow \text{CTA}$$
- (2) They certify the received voting information from a voter V_i , and decrypts the voting result $V_{_resulti}$.
 - CTA :

$$k = H((y_{gi} * R)^{s * KS_{CTA}} \text{ mod } p)$$

$$k_1 \parallel k_2 = k$$

$$V_{_resulti} = D_{k_1}(c)$$
 When $R = g^{KH_{k_2'}(V_{_resulti})} \text{ mod } p$, then $V_{_resulti}$ is accepted.
 - T_i :

$$k' = H((y_{gi} * R')^{s' * KS_{T_i}} \text{ mod } p)$$

$$k'_1 \parallel k'_2 = k'$$

$$V_{_resulti} = D_{k'_1}(c')$$
 When $R' = g^{KH_{k_2'}(V_{_resulti})} \text{ mod } p$, then $V_{_resulti}$ is accepted.

Step 3.2 Voter V_i

A voter concatenates r_{-0i} , ID_i and a voting result to confirm, and sends this information to a tally center encrypting it with public key of CTA. At this time, a voting result to confirm is chosen in $V_{-resulti}$ by the voter V_i .

$$EKP_{CTA}(ID_i || r_{-salti}(v_i \oplus r_{-0i}) \oplus r_{-seed} || r_{-0i}) \rightarrow TCA$$

$$EKP_{Ti}(ID_i || r_{-salti}(v_i \oplus r_{-0i}) \oplus r_{-seed} || r_{-0i}) \rightarrow T_i$$

Step 3.3 CTA and Tally center T_i

(1) They verify v_i using the received θ value from the voter and exchange information between CTA with T_i .

$$(r_{-salti} * (v_i \oplus r_{-0i}) \oplus r_{-seed}) \oplus r_{-seed} = r_{-salti} * (v_i \oplus r_{-0i})$$

$$(r_{-salti} * (v_i \oplus r_{-0i})) / r_{-salti} = (v_i \oplus r_{-0i})$$

$$(v_i \oplus r_{-0i}) \oplus r_{-0i} = v_i$$

(2) Tally center T_i shows the recovered information to a voter V_i in which way.

$$ID_i || v_i$$

(3) After confirming the voting result, T_i opens the total voting results and compares number of voters with the number gotten from the voting list.

$$v_{1k}, \dots, v_{nk} \quad (n = \text{a number of voters}, k \in \{\text{voting results}\})$$

(4) CTA calculates H_{V_result} using one-way hash function as follows. It then compares its number of voters with the number gotten from the tally center's summary sheet.

$$H_{V_result} = H(v_{1k} \oplus \dots \oplus v_{nk})$$

Step 3.4 All voting entities

They verify the total voting result.

$$H(v_{1k} \oplus \dots \oplus v_{nk}) = H_{V_result}$$

Phase 4. Opening

If the voting results have no problem, CTA announces the election result to all members.

Status-based protocol was proposed for the safety of user location for any event in which the session is not terminated safely through the use of the flag. However it does not have a verification process for T therefore if T is altered, authentication is not possible even with transmission from associated tag. In addition if a malicious third party disrupts the last session of a tag whose flag value is 0, the next authentication protocol is processed with the flag value of 1. Here, ID of the database is renewed, but the tag is not able to renew the ID therefore, synchronization problems arise.

3.3 Investigating the Proposal

In the proposed scheme, only authenticated voters can obtain entrance into the voting booth, and get the voting token that is signed digitally by CTA. Because this scheme uses the Magic Sticker, the third party can not associate the voter to his/her voting result and ID_i . Also because CTA confirms the voter using voting list, one person for

one voting principle is assured. A third party can not vote for other voter and the conduct other illegal acts. After the voting is over, when the voting result is open to the public board, the other voters can not trace the voting value using the result. Because tally results are confirmed by only the voters, this scheme retains the receipt free and robustness in voting value and checks the conspiracy of the third party.

- Privacy : Hence this scheme uses the Magic Sticker that is based on one-way anonymous authentication, the third party can not confirm the voter from his/her alias identity ID_i and the voting result.
- Democracy and Authentication : When a voter enters into a voting booth, he/she signs the voting list digitally and will be authenticated by CTA. So that one voter can cast only one vote.
- Fairness : All voting results are opened after the voting process. So that, no one would not be able to determine the voting value through the other's results.
- Integrity : To take integrity, voting information is signed and encrypted with Signcryption scheme by a voter V_i .
- Universal Verifiability : After the voting process, every one confirms the voting result using the CTA and the tally center's voting value ' H_{V_result} '.
- Receipt Free: Since the proposed scheme uses Magic Sticker that satisfies anonymous secure selection, only the voter V_i can confirm the voting result in CTA and tally center. So, the third party can not verify the voting result.
- Robustness : Although all voting member including voters, voting booths, CTA and tally centers, may try illegal acts, robustness is kept because Magic Sticker supplies one-way anonymous authentication service to the proposed scheme.

Therefore, the proposed E-voting scheme using Magic Sticker supports the anonymity of the voter and secure selection.

4 Conclusion

In the information society, new network infra services are required, and various cryptography related schemes are studied to get the authentication and trusty. Among these services, one-way anonymous authentication service is focused on anonymity of the information originator in a group different from peer-to-peer services. Therefore when an anonymous secure selection scheme must be applied to this service, group oriented anonymous authentication can be used.

In this paper, we proposed the Magic Sticker scheme for anonymous secure selection and discussed the efficiency involved with the signcryption. At the same time, we applied the Magic Sticker scheme, using the signcryption, to the electronic voting service which has had the security problems. We showed that the safety and trusty are satisfied in the proposed scheme. Because the proposed scheme satisfies the Receipt free and Robustness required, it can prevent the voting entity's illegal acts, and carry out the anonymous secure selection.

References

1. K. Viswanathan, C. boyd and E. Dawson, "Secure Selection protocols," The International Conference on Information Security and Cryptology, ICISC '99, pp.117-131, 1999. 12.
2. Y. Zheng, "Signcryption and Its Application in Efficient Public Key Solutions," Proc. ISW'97, LNCS 1397, pp.291-312, 1998.
3. D. Chaum, "Elections with Unconditionally Secret Ballots and Disruptions Equivalent to Breaking RSA," Advances in Cryptology, Proceedings of EUROCRYPT '88, pp.177-181, 1988.
4. J. Cohen and M. Fischer, "A Robust and Verifiable Cryptographically Secure Election Scheme," Proceedings of the 26th Annual IEEE symposium on the Foundations of Computer Science, pp.372-382, 1985.
5. K. Koyama, "Secure Secret Voting System Using the RSA Public-Key Cryptosystem," IEICE Trans., Vol.J68-D, No.11, pp.19556-1965, 1985.
6. J. Benaloh, "Secret Sharing Homomorphism : Keeping shares of a Secret," Advances in Cryptology, Proceedings of Crypto '86, pp.251-260, 1986.
7. J. Benaloh, "Verifiable secret-ballot elections," Ph.D.thesis, Yale university, Technical report 561, 1987.
8. J. Benaloh and M. Yung, "Distributing the Power of a Government to Enhance the Privacy of Voters," Proceedings of the 5th ACM Symposium on the Principles in distributed Computing, pp.53-62, 1986.
9. J. Benaloh and D. Tuinstra, "Receipt Free Secret Ballot Elections," proceedings of the 26th ACM Symposium on the Theory of Computing, pp.544-553, 1994.
10. D. Chaum, "Untraceable Electronic Mail, Return Addresses, and Digital Pseudonyms," Communications of the ACM Vol.24, No.2, pp.84-88, 1981.
11. T. Asano, T. Matsumoto and H. Imai, "A Scheme for fair Electronic Secret Voting," technical Report, IEICE Japan, ISEC 90-35, pp.21-31, 1990.
12. C. Park, K. Itoh and K. Kurosawa, "Efficient anonymous channel and all/nothing election scheme," Proc. EUROCRYPT '93, Springer LnCS 765, pp.248-259, 1994.
13. K. Sako and J. Kilian, "Secure Voting Using Partially Compatible Homomorphisms," Advances in Cryptology, Proceedings of Crypto'94, pp.411-424, 1994.
14. K. Sako and J. Kilian, "Receipt Free Mix Type Voting Scheme-A Practical Solution to the Implementation of a Proceedings of EUROCRYPT'95, pp.393-403, 1995.
15. V. Niemi and A. Renvall, "How to prevent buying of votes in computer elections," ASIACrypto '94 pp.164-170, 1994.
16. D. Chaum, "Blind Signature for Untraceable Payments," Advances in Cryptology Proceedings of CRYPTO '82, pp.199-203.
17. K. Iversen, "A cryptographic scheme for computerized general elections," Proc. CRYPTO '91, Springer LNCS 576, pp. 405-409, 1992.
18. M. Akiyama, Y. Tanaka, T. Kikuchi and H. Uji, "Secret ballot Systems Using Cryptography," IEICE Trans., Vol.J67-A, No.12, pp. 1278-1285, 1984.
19. K. Ohta, "An Electrical Voting Scheme Using a Single administrator," 1998 Spring National Convention Record, IEICE, A-294, 1988.
20. Z. He and Z. Su, "A new practical secure e-voting scheme," IFIP/SEC '98 14th International Information Security Conference, 1998.
21. B. Schoenmakers, "A Simple publicly verifiable secret sharing and its application to electronic voting," LNCS 1666, Advances in Cryptology - CRYPTO '99, pp. 148-164, 1999.

22. H. U. Park and I. Y. Lee, "A Study on Preventing from Buying of Votes on Electronic elections," Proceedings of The 9th KIPS Spring Conference, vol 5, no 1, 1998. 4.
23. H. U. Park, H. G. Oh and I. Y. Lee, "A Study on The Protocol of Detection Illegal Act for Central Tabulating Agency," Proceedings of The 1th Korea Multimedia Society Spring Conference, vol 1, no 1, pp163-168, 1998. 6.

Actively Modifying Control Flow of Program for Efficient Anomaly Detection

Kohei Tatara¹, Toshihiro Tabata², and Kouichi Sakurai³

¹ Graduate School of Information Science
and Electrical Engineering, Kyushu University, Japan

² Graduate School of Natural Science and Technology, Okayama University, Japan

³ Faculty of Information Science and Electrical Engineering, Kyushu University,
Japan

tatara@itslab.csce.kyushu-u.ac.jp, tabata@cs.okayama-u.ac.jp,
sakurai@csce.kyushu-u.ac.jp

Abstract. In order to prevent the malicious use of the computers exploiting buffer overflow vulnerabilities, a corrective action by not only calling a programmer's attention but expansion of compiler or operating system is likely to be important. On the other hand, the introduction and employment of intrusion detection systems must be easy for people with the restricted knowledge of computers. In this paper, we propose an anomaly detection method by modifying actively some control flows of programs. Our method can efficiently detect anomaly program behavior and give no false positives.

1 Introduction

We can restrain programmers from writing source codes which includes some bugs, by calling their attentions. But, as long as they are likely to overlook them, any approaches available for the users count for much. Methods using compilers may accompany the recompilation of application programs. Other methods may require the users' knowledge of vulnerabilities.

Some approaches within the pale of the operating system conversion, can suppress the influence on the operating system to minimum. Against this background, the topic of this paper focuses on them. Typically, the role of detecting the illegal use of a computer is referred to as anomaly detection system. This is classified into the follows: misuse detection system, which finds some signatures of invasive act, or anomaly detection system, which catches on the anomalous behaviors derived from the normal ones. The main advantage of anomaly detection systems is that they may detect novel intrusions, as well as various kinds of known intrusions. However, the systems can be complicated and increase the overheads. Our method can be categorized into the anomaly detection, but have following characteristics.

1. Our method can detect the buffer overflow in real time and prevent the attacker from using the computer without authorization.

2. Our method can keep the overhead for anomaly detection to a minimum. Therefore, it is easy for users to introduce our method without degrading the system performance.
3. Our method does not require the recompilation of application programs. This provides relatively easy transition from the current system. Therefore, the users need little knowledge for installation and employment.

In order for intrusion detection systems to be accepted as practical tools, a capability to detect the illegal use of computers with high accuracy and with low overhead is essential. Therefore, it is very important to satisfy the first and second requirements. Moreover, users' knowledge of operating systems and application programs is often limited. Because it is desirable that the cost for introduction and employment is as low as possible, the third requirement is expected to encourage users.

2 Related Work and Motivation

We introduce some approaches addressing intrusion detection from the side of the operating system. They do not have to recompile each application program in introducing it into the system.

The Openwall [1] has functionality of not allowing processes to run any executable codes in the stack area. Currently, some CPUs support the so-called NX (Non-Execute) bit which prohibits execution of codes that is stored in certain memory pages to prevent the intrusions exploiting buffer overflow vulnerabilities as well as Openwall. However, these are not perfect solutions, because there exists another attack which cannot be detected with non-execution of stack [2].

The StackGuard [3] is able to detect the actual occurrence of buffer overflow by checking whether a certain value, called canary, inserted in stack is changed or not. However, it cannot decide whether local variables are correct or not. When an attacker overwrites a function pointer, it may still fail to detect executions of unauthorized codes.

Prasad et al. [4] proposed a method for detecting the stack overflow that uses rewriting execution code on IA-32 in detail. They clarified the coverage and made it to the measure of effectiveness. Their system can detect stack overflow by building the mechanism of RAD [5] into the execution code on Intel 32-bit Architecture (IA-32). Then, they supposed that their disassembler can catch prologue and epilogue of the function, and the existence of the area to insert the jmp instruction as a precondition [4]. We think therefore that it is appropriate that our method also puts same assumption as theirs. That is, we assume the existence of the one similar to the frame pointer or it in the function.

3 Our Proposal

3.1 Modification of the Control Flow

When loading a program into the memory area, our method modifies the control flow of it. Then, it inserts an additional flow, which provides the verification

process, called verification function, to check the legitimate use in the vicinity of each function call. The process of the modified function works as follows.

1. When the function is called, a stack frame is allocated for the return address. The return address in this stack frame is set to the next instruction. When the function call finishes, the process will be resumed at the return address set in the stack frame.
2. Next, the size of a pointer is taken from the value of the stack pointer. The address of the verification function is to be put in this reserved area. Then, the value of the frame pointer is put in the stack. The value of the frame pointer is newly replaced by the value of the stack pointer. After a stack frame is allocated for the local variables, the process of the original (unmodified) function will be carried out.
3. At the end of the function, the address of the verification function is overwritten by using the frame pointer in the reserved area. Therefore, in spite that the function is finished successfully or abnormally, the control of the process will be transferred to the verification function.

3.2 Adjustment for Function Call

If there exists the function call in the coverage of our method, it will be modified as follows. When calling another function, the return address that is supposed to be put in the stack is exclusive OR'ed with the random value p . The result is put in the stack as if it is the return address.

Next, the process of the function call is modified so that the control is transferred to the address specified by the operand of it, which is exclusive OR'ed by random value p . Also, the operand itself of the function call is correspondingly modified in advance. Especially, if the operand can be considered as one of the registers or a pointer to memory, we will try to find the instruction which provides the input to it. If a certain constant value is found in the instruction, it will also be exclusive OR'ed with the random value p . The value of p is chosen at every execution of the program. Here, we note that if the control is transferred to the raw address specified by the external input is called, we cannot locate such value. Such programs are not included in the coverage of our method. Because the p is not only secret (only known by the operating system) but also fresh (different at each execution), the operating system can only run the program normally.

3.3 Process of Verification Function

The verification function takes the random value p and the (exclusive OR'ed) return address as arguments. It can verify that the function was called from the valid origin. If p is not correct, the return address may probably be corrupted. In other word, we can decide that there exists the invasive action using the buffer overflow. In our method, the return address given as an argument will be

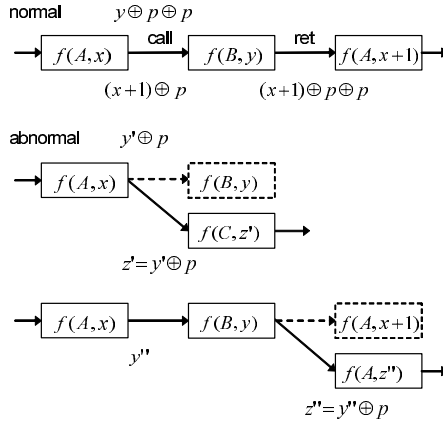


Fig. 1. State transition associated with function call

exclusive OR'ed with the random value p once again, and the control will be transferred to the address pointed by the result. For this reason, only when p is correct, the execution will be resumed correctly.

3.4 Procedure of Detecting Anomalous Behavior

We explain the procedure where our method detects the invasion act. Figure 1 shows that the function A calls another function B . The top of these flows represents normal execution, and the lower flows shows that the return address or the function pointer are illicitly replaced.

$f(A, x)$ indicates that CPU is running x -th instruction in the function A . In our method, when A calls B , both the operand of the function call and the return address are exclusive OR'ed with the random value p . If its operand (e.g., function pointer) is illicitly replaced by the pointer to the address of a shellcode or a shared library (z'), the control is transferred to the address $z' \oplus p$. This address may point to an illegal address, and cause an error. Also, it is assumed that an attacker can replace the return address with an address (z''). Regardless of buffer overflow, the verification function is certainly called. Therefore, when the process is restarted at $(A, z'' \oplus p)$. If the attacker does not know the random value p , he cannot succeed in any intrusion.

4 Implementation

4.1 Modification of Program

In this chapter, we touch on capability of the implementation in an existing system. Specifically, we use Linux operating system (Kernel 2.4.19) working on Intel 32-bit Architecture, and take GNU C Compiler (gcc-2.96). But, we believe that our method is applied for another platform.

```

_func1:
  pushl   %ebp
  movl   %esp, %ebp
  subl   $24, %esp
  movl   $_func2, -4(%ebp)
  movl   $LC1, 4(%esp)
  movl   $1, (%esp)
  movl   -4(%ebp), %eax
  call   *%eax
  leave
  ret
  ...
_func2:
  pushl   %ebp
  movl   %esp, %ebp
  subl   $8, %esp
  incl   8(%ebp)
  movl   12(%ebp), %eax
  movl   %eax, 4(%esp)
  movl   $LC0, (%esp)
  call   _printf
  leave
  ret

```

Fig. 2. Original code

Figure 2 shows the original (assembler) code before our method is applied. This code means that a function `func2` is called in a function `func1`. The modified (assembler) code is shown in Figure 3. We schematically explain the behavior when running it. Note that the code modification is performed at the binary level. Here, for ease of explanation, we show the assembler codes.

1. Firstly, when `func1` is called, a stack frame is reserved for the address of the verification function. At once, the frame pointer `%ebp` is put in the stack. the stack pointer `%esp` is used as a new `%ebp`. Then, 24-bytes are subtracted from the stack pointer to obtain a memory space for local variables.
2. The part, where there exists a call instruction (`call _func2`), is modified so that both the return address put in the stack and the operands of the call instruction are exclusive OR'ed with the random value p .
3. When the process of the original code is finished, the address of verification function is written in the reserved area by using the frame pointer `%ebp`. The verification function is then started.

4.2 Assuring the Consistency in the Vicinity of Modification

The most important thing to change the control flow is to minimize the effect on the original program. Also, it is important that we avoid using the variable parameters amenable to environment or external input. As shown in Figure 3, we rewrite the call instruction so that the return address is exclusive OR'ed with $p = 0x12345678$? before stepping into the next function `func2`. However, in the verification function, the result of it is exclusive OR'ed once again. Definitely, the consistency can be preserved in the vicinity of the application of our method.

The call instruction,

```

_func1:
  call  _trampoline11
  nop
  movl  $_trampoline20, -4(%ebp)
  movl  $LC1, 4(%esp)
  movl  $1, (%esp)
  movl  -4(%ebp), %eax
  jmp   _trampoline12
  nop
  nop
  ...

_trampoline10:
  xorl  $0x12345678, (%esp)
  jmp   _func1
_trampoline11:
  movl  (%esp), %eax
  pushl %ebp
  movl  %esp, %ebp
  subl  $24, %esp
  jmp   *%eax
_trampoline12:
  call  *%eax
  movl  $_ver, 4(%ebp)
  leave
  ret
_ver:
  xorl  $0x12345678, (%esp)
  ret
  ...

_func2:
  call  _trampoline21
  nop
  incl  12(%ebp)
  movl  16(%ebp), %eax
  movl  %eax, 4(%esp)
  movl  $LC0, (%esp)
  jmp   _trampoline22
  nop
  nop
  ...

_trampoline20:
  xorl  $0x12345678, (%esp)
  movl  $_func2, %eax
  xorl  $0x12345678, %eax
  ...
  xorl  $0x12345678, %eax
  jmp   *%eax
_trampoline21:
  movl  (%esp), %eax
  pushl %ebp
  movl  %esp, %ebp
  subl  $8, %esp
  jmp   *%eax
_trampoline22:
  call  _printf
  movl  $_ver, 4(%ebp)
  leave
  ret
  ...

```

Fig. 3. Modified code

```

movl  $_func2, -4(%ebp)
...
movl  -4(%ebp), %eax
call  *%eax

```

is adjusted as follows.

```

movl  $_trampoline20, -4(%ebp)
...
movl  -4(%ebp), %eax
jmp   _trampoline12
...
_trampoline12:
  call  *%eax
  ...
_trampoline20:
  xorl  $0x12345678, (%esp)
  movl  $_func2, %eax
  xorl  $0x12345678, %eax
  ...
  xorl  $0x12345678, %eax
  jmp  *%eax
  ...

```

By inserting the “trampoline” function, we do not have to consider any effects on former and latter instructions. Thus, as shown in Figure 2, 3, we can find that the size of `func1` and `func2` are kept during the modification. If the size was changed, we also had to update all the values of addresses in the latter instructions.

However, in the `func2`, the references to the arguments are modified as follows.

```
incl 8(%ebp) → incl 12(%ebp)
```

This is because we reserved the stack frame for the address of the verification function.

5 Security and Efficiency

5.1 Resistance for Intrusion

An attacker can hijack the control flow by overwriting the return address or the function pointer. But, the buffer available to the attacker is likely to be limited. Therefore, we assume that the attacker certainly uses functions or libraries supplied by the operating system. When she invades the system without using them, our method cannot work well. Namely, if she only tries to overwrite some variables, the system will condone her offence.

Modifying a return address of the function call prevents the attacker from replacing them with addresses of shellcode or shared libraries. On the other hand, the threat of replacement of function pointer cannot be eliminated in this scheme. We cannot decide whether an input to the function pointer is correct, because the pointer variable can be changed in the normal process. Our method is also resistant to this attack by modifying the operands of function call.

5.2 Possibility of Intrusion

The modification of a program code is carried out when loading it into the memory space. In order for the attacker to get the correct return address and modify it, she must guess the random value p . Possibility of succeeding in the attack is very small, when the attacker cannot guess it, because we assume that p is very large value (about 2^{32}).

5.3 Efficiency

Our method is applied to each function. We measured how much overhead is required to adapt our modification comparing the basic performance. Specifically, we saw the overhead per function call provided by executing the original code and modified code in our method (Figure 2, 3). Our machine has Pentium4 1GHz, 512MB RAM. The result of fifteen sets of one million executions shows that the (average) increase of execution time is 0.255137 seconds, and the relation between the increase and the number of execution times in Figure 4. In order to be accepted as practical art, our method should be applied to the

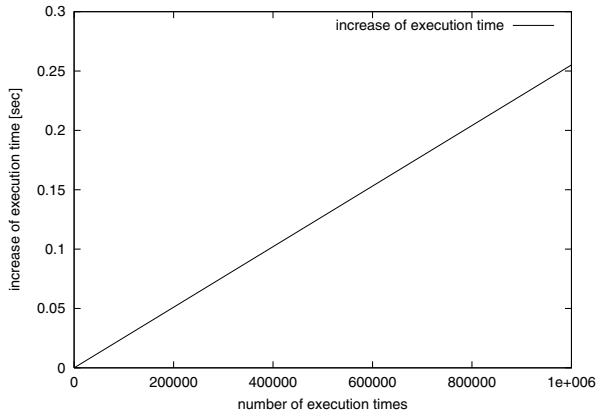


Fig. 4. Overhead per function call (x-axis: number of execution times, y-axis: increase of execution time [sec])

application programs widely used, and be evaluated in spite that this is negligible small value.

6 Conclusion

The case of abuse of computers using buffer overflow are continually reported and considered as serious problem. Currently, some CPUs support the so-called NX bit (Non-Execute) which prohibits the execution of code that is stored in certain memory pages to prevent the intrusion exploiting buffer overflow vulnerability. However, this is not a perfect solution, because there exists another attack which cannot be detected with non-execution of stack.

In this paper, we proposed an anomaly detecting method by modifying the control flow of the program. Our method has advantage of no false positives and reducing the overhead. It is also expected that it only takes the costs of the introduction and employment, and encourages users to install it. Future work is further evaluation of anomaly detection system.

References

1. Openwall Project, Linux kernel patch from the Openwall project, <http://www.openwall.com/linux/> (accessed 2004-01-20).
2. Linus Torvalds, <http://old.lwn.net/1998/0806/a/linus-noexec.html> (accessed 2004-02-13)
3. P. Wagle and C. Cowan. StackGuard: SimpleStack Smash Protection for GCC. In *Proceedings of the GCC Developers Summit*, pp. 243–255, May 2003.
4. M. Prasad and T. Chiueh. A Binary Rewriting Defense Against Stack-based Buffer Overflow Attacks. In *Proceedings of Usenix Annual Technical Conference*, Jun. 2003.
5. T. Chiueh and F. Hsu, RAD: A compile time solution for buffer overflow attacks. In *Proceedings of 21st IEEE International Conference on Distributed Computing Systems (ICDCS)*, Apr. 2001.

Intelligent Method for Building Security Countermeasures

Tai-hoon Kim¹ and Sun-myoung Hwang²

¹ Department of Information Electronics Engineering, Ewha Womans University, Korea
taihoonn@empal.com

² Department fo Computer Engineering, Daejun University, Korea
sunhwang@dragon.taejon.ac.kr

Abstract. It is critical to note that a weakness (or security hole) in any component of the IT systems may comprise whole systems. Applying of Security engineering and intelligent methods are needed to reduce security holes may be included in the software or systems. Therefore, more security-related intelligent researches are needed to reduce security weakness may be included in the systems. This paper proposes some intelligent methods for reducing the threat to the system by applying security engineering, and for building security countermeasure.

Keywords: Security countermeasure, security hole, security engineering.

1 Introduction

With the increasing reliance of society on information, the protection of that information and the systems contain that information is becoming important. In fact, many products, systems, and services are needed and used to protect information. The focus of security engineering has expanded from one primarily concerned with safeguarding classified government data to broader applications including financial transactions, contractual agreements, personal information, and the Internet. These trends have elevated the importance of security engineering [1].

Considerations for security are expressed well in some evaluation criteria such as ISO/IEC 21827, the Systems Security Engineering Capability Maturity Model (SSE-CMM), and ISO/IEC 15408, Common Criteria (CC) [2].

It is essential that not only the customer's requirements for software functionality should be satisfied but also the security requirements imposed on the software development should be effectively analyzed and implemented in contributing to the security objectives of customer's requirements. Unless suitable requirements are established at the start of the software development process, the resulting end product, however well engineered, may not meet the objectives of its anticipated consumers. The IT products like as firewall, IDS (Intrusion Detection System) and VPN (Virtual Private Network) are made to perform special functions related to security, and used to supply security characteristics. But the method using these products may be not the

perfect solution. Therefore, when making some kinds of software products, security-related requirements must be considered.

2 A Summary of Security Engineering

Security engineering is focused on the security requirements for implementing security in software or related systems. In fact, the scope of security engineering is very wide and encompasses:

- The security engineering activities for a secure software or a trusted system addressing the complete lifecycle of: concept definition, analysis of customer's requirements, high level design and low level design, development, integration, installation and generation, operation, maintenance and de-commissioning;
- Requirements for product developers, secure systems developers and integrators, organizations that develop software and provide computer security services and computer security engineering;
- Applies to all types and sizes of security engineering organizations from commercial to government and the academe.

The security engineering should not be practiced in isolation from other engineering disciplines. Maybe the security engineering promotes such integration, taking the view that security is pervasive across all engineering disciplines (e.g., systems, software and hardware) and defining components of the model to address such concerns.

The main interest of customers and suppliers may be not improvement of the development of security characteristics but performance and functionality. If developers consider some security-related aspects of software developed, maybe the price or fee of software more expensive. But if they think about that a security hole can compromise whole system, some cost-up will be appropriate.

3 Applying Security Engineering to Build Countermeasures

In general, threat agents' primary goals may fall into three categories: unauthorized access, unauthorized modification or destruction of important information, and denial of authorized access. Security countermeasures are implemented to prevent threat agents from successfully achieving these goals.

Security countermeasures should be considered with consideration of applicable threats and security solutions deployed to support appropriate security services and objectives. Subsequently, proposed security solutions may be evaluated to determine if residual vulnerabilities exist, and a managed approach to mitigating risks may be proposed.

Countermeasures must be considered and designed from the starting point of some IT system design or software development processes. The countermeasure or a group of countermeasures selected by designers or administrators may cover all the possibility of threats.

But a problem exists in this situation. How and who can guarantee that the countermeasure is believable?

Security engineering may be used to solve this problem. In fact, the processes for building of security countermeasures may not be fixed because the circumstances of each IT system may be different.

We propose a method for building security countermeasures as below.

3.1 Threats Identification

A ‘threat’ is an undesirable event, which may be characterized in terms of a threat agent (or attacker), a presumed attack method, a motivation of attack, an identification of the information or systems under attack, and so on.

Threat agents come from various backgrounds and have a wide range of financial resources at their disposal. Typically Threat agents are thought of as having malicious intent. However, in the context of system and information security and protection, it is also important to consider the threat posed by those without malicious intent. Threat agents may be Nation States, Hackers, Terrorists or Cyber terrorists, Organized Crime, Other Criminal Elements, International Press, Industrial Competitors, Disgruntled Employees, and so on.

Most attacks always aim at getting inside of information system, and individual motivations of attacks to “get inside” are many and varied. Persons who have malicious intent and wish to achieve commercial, military, or personal gain are known as hackers (or cracker). At the opposite end of the spectrum are persons who compromise the network accidentally. Hackers range from the inexperienced Script Kiddie to the highly technical expert.

3.2 Determination of System Security Level and Robustness Strategy

Robustness strategy should be applied to all components of a solution, both products and systems, to determine the robustness of configured systems and their component parts. It applies to commercial off-the-shelf (COTS), government off-the-shelf (GOTS), and hybrid solutions. The process is to be used by security requirements developers, decision makers, information systems security engineers, customers, and others involved in the solution life cycle. Clearly, if a solution component is modified, or threat levels or the value of information changes, risk must be reassessed with respect to the new configuration [3].

Various risk factors, such as the degree of damage that would be suffered if the security policy were violated, threat environment, and so on, will be used to guide determination of an appropriate strength and an associated level of assurance for each mechanism. Specifically, the value of the information to be protected and the perceived threat environment are used to obtain guidance on the recommended evaluation assurance level (EAL).

Furthermore, to decide systems security level, EAL is not a perfect one. So we should decide TL (Threat Level) and AL (Asset Level) to get more exact SL (Security Level). About the decision of SL, please recommend our report [4].

Table 1. Determination of Security Level by Threat Level and Asset Level

Asset Level	Threat Level					
	TL1	TL2	TL3	TL4	TL5	TL6
AL1	SL1	SL1	SL1	SL1	SL2	SL2
AL2	SL1	SL1	SL2	SL2	SL3	SL3
AL3	SL1	SL2	SL2	SL3	SL3	SL4
AL4	SL1	SL2	SL3	SL3	SL4	SL4

3.3 Determination of Components Level

To provide adequate information security countermeasures, selection of the desired (or sufficient) components level by considering particular situation is needed. An effective security solution will result only from the proper application of security engineering skills to specific operational and threat situations.

In this step, we should determine EAL (Evaluation Assurance Level) for products, STL (Security Technology Level) for technology, SSL (Security Staff Level) for operators and SOL (Security Operation Level) for systems operation environments.

Table 2. Determination of Components Level

Security Level	Components Level			
	Evaluation Assurance Level	Security Technology Level	Security Staff Level	Security Operation Level
SL1	EAL2 or more	STL1	SSL1	SOL1
SL2	EAL3 or more	STL2	SSL2	SOL2
SL3	EAL4 or more	STL2	SSL2	SOL2
SL4	EAL6 or more	STL3	SSL3	SOL3

3.4 Selection of Security Services

In general, primary security services are divided five areas: access control, confidentiality, integrity, availability, and non-repudiation. But in practice, none of these security services is isolated from or independent of the other services. Each service interacts with and depends on the others.

For example, access control is of limited value unless preceded by some type of authorization process. One cannot protect information and information systems from unauthorized entities if one cannot determine whether that entity one is communicating with is authorized.

In actual implementations, lines between the security services also are blurred by the use of mechanisms that support more than one service.

3.5 Determination of Assurance Level

The discussion of the need to view strength of mechanisms from an overall system security solution perspective is also relevant to level of assurance. While an underlying methodology is offered by a number of ways, a real solution (or security product) can only be deemed effective after a detailed review and analysis that consider the specific operational conditions and threat situations and the system context for the solution.

Assurance is the measure of confidence in the ability of the security features and architecture of an automated information system to appropriately mediate access and enforce the security policy. Evaluation is the traditional method ensures the confidence. Therefore, there are many evaluation methods and criteria exist. In these days, many evaluation criteria such as ITSEC are replaced by the Common Criteria.

The Common Criteria provide assurance through active investigation. Such investigation is an evaluation of the actual product or system to determine its actual security properties. The Common Criteria philosophy assumes that greater assurance results come from greater evaluation efforts in terms of scope, depth, and rigor.

4 Conclusions and Future Work

As mentioned earlier, security should be considered at the starting point of all the development processes. Making an additional remark, security should be implemented and applied to the IT system or software products by using the security engineering.

This paper proposes some methods for reducing the threat to the system by applying security engineering, and proposes a method for building security countermeasure. But this method we proposed can't cover all the cases. Therefore, more detailed research is needed. And the research for generalizing these processes may be proceeded, too.

References

1. ISO. ISO/IEC 21827 Information technology – Systems Security Engineering Capability Maturity Model (SSE-CMM)
2. ISO. ISO/IEC 15408-1:1999 Information technology - Security techniques - Evaluation criteria for IT security - Part 1: Introduction and general model
3. Tai-hoon Kim and Seung-youn Lee: Security Evaluation Targets for Enhancement of IT Systems Assurance, ICCSA 2005, LNCS 3481, 491-498

4. Tai-hoon Kim: Draft Domestic Standard-Information Systems Security Level Management, TTA, 2005
5. Tai-Hoon, Kim: Approaches and Methods of Security Engineering, ICCMSE 2004
6. Tai-Hoon Kim, Seung-youn Lee: Design Procedure of IT Systems Security Countermeasures. ICCSA (Computational Science and Its Applications) 2005: LNCS 3481, 468-473
7. Tai-Hoon Kim, Seung-youn Lee: A Relationship Between Products Evaluation and IT Systems Assurance. KES (Knowledge-Based Intelligent Information and Engineering Systems) 2005: LNCS 3681, 1125-1130.
8. Tai-Hoon Kim, Seung-youn Lee: Intelligent Method for Building Security Countermeasures by Applying Dr. T.H. Kim's Block Model. KES (Knowledge-Based Intelligent Information and Engineering Systems) 2005: LNCS 3682, 1069-1075
9. Tai-Hoon Kim, Chang-hwa Hong, Myoung-sub Kim: Towards New Areas of Security Engineering. RSFDGrC (Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing) 2005: LNCS 3642, 568-574
10. Tai-Hoon Kim, Haeng-Kon Kim: A Relationship between Security Engineering and Security Evaluation. ICCSA (Computational Science and Its Applications) 2004: LNCS 3046, 717-724

Intelligent Frameworks for Encoding XML Elements Using Mining Algorithm

Haeng-Kon Kim

Department of Computer Information & Communication Engineering, Catholic University
of Daegu, 712-702, South Korea
hangkon@cu.ac.kr

Abstract. Mining of association rules is to find associations among data items that appear together in some transactions or business activities. As of today, algorithms for association rule mining, as well as for other data mining tasks, are mostly applied to relational databases. As XML being adopted as the universal format for data storage and exchange, mining associations from XML data becomes an area of attention for researchers and developers. The challenge is that the semi-structured data format in XML is not directly suitable for traditional data mining algorithms and tools. In this paper we present an intelligent encoding method to encode XML tree-nodes. This method is used to store the XML data in 2-dimensional tables that can be easily accessed via indexing using knowledge. The hierarchical relationship in the original XML tree structure is embedded in the encoding. We applied this method in some applications, such as mining of association rules.

1 Introduction

XML is a language specifying semi-structured data. It is rapidly emerging as a new standard for data representation and exchange on the Web. It is also a set of the rules and guidelines describing semi-structured data in plain text rather than proprietary binary representations. The vast majority of the data reside in relational databases, the most widely used data storage and management format. Relational databases are well-defined and applications based on relational databases are in general very robust and efficient. Many algorithms and methodologies have been developed to apply to relational databases. The very basic feature of relational databases is that the data are stored in *tables*, or 2-dimensional form, in which rows represent data objects and columns specify the attributes of the objects. That is, the values of an object's attributes are "encapsulated" in a row (or a line for plain text data) that can be retrieved by column-access methods (or separated by some "delimiters" such as space, comma, tab, etc. for plain text data). This table-based data representation is very convenient for algorithms to access and process the data. For example, most data mining algorithms [9] work on relational databases. In particular, mining of association rules works on transactions, each of which consists of one or more items. A transaction is represented by one row of data values, each of which is an item of the transaction.

These algorithms all assume that the data items are arranged in this way, either in relational tables or in plain text.

Considering the task of mining association rules, the main difficulty is to find what constitute a transaction and how to find the items in the transactions that are the same. First, let's review the basic concept of association rule mining. Mining of association rules, also called market basket analysis in business, is to find interesting association or correlation among data items that often appear together in the given data set.

There are several algorithms for finding association rules, for example, the *Apriori Algorithm* [1,2]. These algorithms iteratively calculate frequent item sets of progressively increasing sizes, starting from singleton sets, that satisfy the minimum support and confidence levels. The iteration steps are quite straightforward, largely due to the simplicity of the way the input data are structured as a 2-dimensional table. If the data are given in XML format, however, the algorithms may be complicated in locating transactions and the item sets in the data. However, the levels in traditional multilevel data mining tasks are "predetermined" based on the schema defined on relational databases or data cubes, whereas the "levels" in XML data are not well-defined because they are semi structured and there may not be a schema defined on the XML documents. Even a schema exists, there may be missing elements and varying number of sub-elements of the same tag name, among other flexible features of XML.

In this paper, We present a solution to the problem of mining association rules on XML data by an encoding method that assigns a unique code to each different tag name and using 2-dimensional array (or vector of vectors) to store the data with the tag codes to index the collection. The code of an element embeds the codes of its children, so that the parent-child relationship can be easily determined by simple operations on the codes. With this coding scheme, we can also easily determine if two elements are the same because their children's codes are embedded in the parents' codes. This coding method is in fact equivalent to the recursive definition of structure equivalence in programming languages. And, we give a new method that organizes the XML data into 2-dimensional arrays for easy operations. The data mining method we use is based on the *Apriori Algorithm* – a basic algorithm for association rule mining. We use bottom-up algorithms to get all the rules for any level including cross-levels.

2 Related Work

A lot of work have been done on association rule mining since it was first introduced by Agrawal, Imielinski, and Swami [1]. Shortly after, Agrawal and Srikant, and Man- nila et al. published the Apriori algorithm, independently [3] and jointly [2]. Several variations of the Apriori algorithm were proposed, including use of hash table to improve the mining efficiency, partitioning the data to find candidate item sets, reducing the number of transactions [3], and dynamic itemset counting [5]. Various extensions of association rule mining were also studied [8] and is still an active research area.

Multilevel association rule mining finds rules at different abstraction levels of the concept hierarchy of the data items. Mining association rules at different levels

requires that the data items are categorized in a concept hierarchy before the algorithms can be applied. For example, the data should contain the information “HP 960 inkjet printer is an HP printer, which is a printer.” This information needs to be provided by the applications in question. Data in XML format are naturally form a hierarchy. However, it in general does not reflect the multilevel abstraction of the data items. The hierarchy in an XML file mostly represents the “has-a” relationship between a higher level item and its children rather than the “is-a” relationship as used in multilevel association rule mining.

Hence the multilevel mining methods do not directly apply to XML data, or at least not easily. Although applying data mining algorithms to XML data has not been extensively studied as of today, there are some research work reported. Some commercial products such as XMLMiner [7] are also available. Buchner et al. [6] outlined the XML mining and pointed out research issues and problems in the area. We do not use any specific query language to query the XML data for association rules; rather, we simply apply the Apriori algorithm to find all rules (of sizes from 2 to the maximum size allowed by the data) using user provided parameters (minimum support and minimum confidence).

3 Frameworks for Encoding of XML Elements

Because XML data can be abstracted to a tree structure by an XML parser, we use the tree terminologies (root, level, sub-tree, node, parent, child, sibling, leaf, path, etc.) throughout this paper. An XML element starts with a start-tag (e.g. <name>) and ends with an end-tag (e.g. </name>). Between the two tags is the “actual data” of the element, which may in turn contain sub-elements. For the purpose of association rule mining, we consider the first level elements as “transactions.” The sub-level elements are considered items in the transactions.

To keep track of the relationships between the nodes on different levels, we use several data structures to hold some information coming from the DTD and the DOM tree produced by an XML parser.

First, the nodes in the tree are encoded so that each node is assigned a unique number that embeds in it the location information of the node as well as the relationship between the node and its parent and siblings. Two “2-dimensional arrays” (vector of vectors) are used to hold the values of the nodes and information about the “transactions.” These 2-D arrays are indexed by tag names and elements’ values for fast access. First, let’s introduce the encoding of the nodes.

3.1 Intelligent Coding for XML Elements

Let an element E be a node in the DOM tree constructed from a given XML file. E and its siblings (all nodes having the same parent of E) are *ordered* based on the order in which the elements appear in the DTD file under the same parent tag. Nodes under different parents may have the same ordinal number but they are on different paths.

For example, the tags <TITLE>, <ARTIST>, <COUNTRY>, and <COMPANY> have ordinal numbers 1, 2, 3, and 4, respectively under their parent tag <CD>. The tags <firstname> and <lastname> are assigned ordinal numbers 1 and 2 under parent <ARTIST>.

Now, we can encode E with an integer C of n digits where n is the number of nesting levels of the XML as $C = d1,d2,d3...dn$ where di is the ordinal number of the node along the path from the root to E, or 0 if the level of E in the tree is less than i. That is, the code C is 0-filled on the right if E is not a leaf node. For example, there are three nesting levels in the example XML file and hence each node is encoded using a 3-digit number (3-integer number, in fact). The tag <CD> will be encoded as 100, <TITLE> 110, <ARTIST> 120, <firstname> 121, <lastname> 122, <COUNTRY> 130, <COMPANY> 140, etc. The tree nodes and their codes are shown in Figure 1. Tree nodes encoding of the example XML. This encoding method is also easy to apply for node insertion and deletion. Once the nodes are encoded, we can store the values of the elements in a 2-D array (lengths of the rows may differ, as vector of vectors), called the *Value Table*. We can also store information about the “transactions” (considering the first-level tag as a transaction) in a *Transaction Table*.

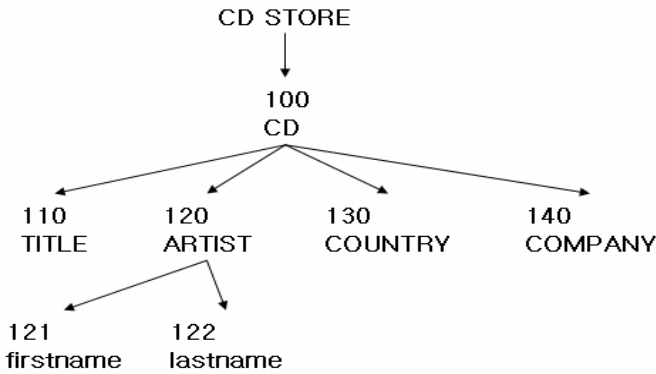


Fig. 1. Tree node encoding of the example XML

3.2 Value Table

Each XML element has a “value” as given in the XML file and its code as discussed above. Only the codes of the elements are used during the mining of association rules.

However, we need to show the values of the elements in the association rules when the rules are found. The *Value Table* serves the purpose of code-value mapping by storing the “values” of the elements in an array (again, can be considered a vector of vectors) with the elements’ encodings to index the rows. We get the structure of XML file from its DTD and make an index (i.e. encoding) for each tag name. The indexes are stored in a hash table. All occurrences of the same tag name are stored in a vector pointed to by the index.

Consider the previous example, the Value Table with the hash header will look like the following shown in Figure 1. Note that the “lengths” of the rows are different because there may be different number of values for each tag. That is the reason we have mentioned the “equivalence” between array and vector of vectors. Hash header Value Table Figure 1. Value Table with hash header. Here, each element (tag name) is encoded in a 3-digit number as discussed before and shown in the hash header. All distinct values of the same tag name are stored in the same row pointed to by the code of the tag (e.g. 100 for <TITLE>). If a node is a composite element (e.g. <CD>, or <ARTIST> with sub-elements <firstname> and <lastname>), the “value” of the element is the combination of the location of the “values” of all its children. For example, the third <CD> element is stored in column 3 of the 1st row, which stores for all CD’s. Its value is “.3.1.1.1,” meaning that it has 4 children (of codes 110, 120, 130, and 140), in columns 3, 1, 1, and 1, CD STORE respectively. Namely, “Highland,” “Sam Brown,” “UK,” and “A and M.” For the same token, the second <CD> element is stored in column 2 of the row for all CD’s. Its value is “.2.2.2.2” indicating that the values of its 4 children are all in column 2 of their respective rows. The second child (an <ARTIST> element) is itself a composite element with value “.1.1” indicating that it has two children in column 1 of the row for <firstname> (first child of <ARTIST>) and column 1 of the row for <lastname> (second child of <ARTIST>). This is shown in the shaded cells in Fig. 1.

3.3 Transaction Table

The Transaction Table holds the information of the transaction sets. Because XML is of a multilevel tree structure, it is a primary task to organize the transaction table in such a way that it will support easy retrieval of information at all levels related to association rules without storing redundant data. The transaction table is information-encoded instead of the traditional transaction table as classical association rule mining algorithms would apply to. Each row in the Transaction Table represents a “transaction,” although not in the traditional sense. Columns of the table are the tag names. A cell in the table is an encoded string, which is a concatenation of the position of the element in the hierarchy and its value index. That is, let T be the Transaction Table. $T[r][c]$ represents the item of transaction r with tag name c . The value stored in the cell is $A = T[r][c] = tv$ where t is the code of the tag and A represents the value of the item in the Value Table cell $V[t][v]$. The advantage of this representation is the easy cross-reference between V and T and the simplicity of decomposition of the digit sequence to get the indexes. In our implementation, we encode the information into bits, combine them together and convert to an integer just to reduce space. This requires fewer bits than using the object ID or barcode methods. The Transaction Table for the example in our previous discussion is shown in Table 1.

In the table, each cell is a 4-digit number, which is a concatenation of a 3-digit number and a 1-digit number. For example, $T[1][120]=1202$ is concatenation of 120 and 2, meaning that the code of the item’s tag name is 120 (which is <ARTIST>) and the value of the item is stored in $V[120][2]$ (which is .1.1 indicating a composite value with <firstname> and <lastname>). Using this organization of the Transaction Table

Table 1. Value table with hash header

Hash header		Value Table			
<CD>	100	→	.1.1.1.1.1	.2.2.2.2	.3.1.1.1
<TITLE>	110	→	Moonlight	Flying	Highland
<ARTIST>	120	→	Sam	.1.1	
<firstname>	121	→	Savage		
<lastname>	122	→	Rose		
<COUNTRY>	130	→	UK	EU	
<COMPANY>	140	→	A and M	Polydor	

Table 2. Transaction Table for the example XML

tid	Tag name index						
	100	110	120	121	122	130	140
0	1001	1101	1201			1301	1401
1	1002	1102	1202	1211	1221	1302	1402
2	1003	1103	1201			1301	1401

and Value Table, we have avoided storing redundant data (same value but with different tag names) and still kept the hierarchical relationships between the elements intact. Another problem is about missing data (DTD may define an element that can appear zero or more times). When processing the XML data, we should consider the different structures of the transactions including the missing and disordering of its sub-elements appearing in the XML file. The approach we use is to use an array to represent the pattern of a “right” element whose sub elements are all present and in the order given in the DTD. The actual XML data are checked against the pattern. If a sub-element is missing or not in the “right” order according to the DTD, the “digit” for the sub element in the encoding would be filled with a special symbol.. Hence, the code of an element is a sequence of integers separated by a dot. Each integer is of $\log_{10}(N)+1$ digits where N is the number of children of the node. Decomposing the sequence is just a matter of going through the sequence, only a bit more work than if they were digits.

3.4 Algorithms

Here are the algorithms for encoding of the XML elements and for building the Value Table and Transaction Table.

Encoding

```

get structure information from DTD;
get children from root into a collection of vectors;
for each vector in the collection
    for each element of this vector
        while the element has children
            get its children into a new vector;
    
```

```

add it to the collection;
N = number of children;
encoding each child (i.e. using  $\log_{10}(N)+1$ ),
and push it into a hash map;

```

Building Value Table and Transaction Table

```

parse XML document to get the DOM tree;
trans = 0;
for each sibling of the first child of root
  trans ++;
  col = V[trans].size;
  if this node is a leaf element
    V[trans][col++] = attribute name
  else
    value = concatenation of the "values" along
             the path from root;
    V[trans][col++] = value;
    T[trans][tag code] = concatenation of code of
    tag name and col of the value in V;

```

4 Mining Association Rules

We use the Apriori algorithm in this paper. It is the very basic and an influential algorithm for mining frequent itemsets for association rules dealing with the presence/absence of items. We shall briefly describe the basic idea of the Apriori algorithm below, and then discuss some modifications so that it becomes more flexible.

4.1 Apriori Algorithm

A set of items is called an *itemset*. An itemset of size k is called a *k-itemset*. An itemset that satisfies the minimum support level is called a *frequent k-itemset*, denoted L_k . The task of association rule mining is to find frequent k -itemsets of a given k . The Apriori algorithm starts with frequent 1-itemsets, L_1 , and then iteratively finds L_{i+1} from L_i until $i+1=k$. Each iteration step consists of a join operation and a prune operation. The join operation joins L_i with itself to produce a candidate set C_i , which is a superset of L_{i+1} and in general quite large. The prune operation deletes the members in C_i that are not frequent (i.e. those that do not meet the minimum support requirement) to produce L_{i+1} . This pruning significantly reduces the size of C_i and at the same time still guarantees that nothing useful is removed. This is possible because of the Apriori property that says all nonempty subsets of a frequent itemset must also be frequent.

Once the frequent k -itemsets are found, it is straightforward to generate association rules that also satisfy the minimum confidence level:

1. For each frequent itemset L , generate all nonempty subsets of L .
2. For each nonempty subset S of L , output the rule $S \Rightarrow L - S$ if the rule satisfies min-confidence.

It is clear that all association rules generated by this algorithm are of size $|L|$.

4.2 Modifications to the Algorithm

As mentioned above, the traditional Apriori algorithm is to find the k -frequent item sets for a given k . That is, all association rules resulted from the algorithm contains k items. However, if we want to find all association rules of sizes from 2 to k , we should have kept those frequent item sets in L_i that were pruned while calculating L_{i+1} but they did satisfy the minimum support in the previous iteration. These i -item sets can generate association rules of size i although they won't be in rules of size $i+1$. We added, therefore, an additional data structure (a Collection) to store the frequent i -item sets during the iteration so that smaller sized association rules will also be found, rather than just the longest possible rules. Another modification is to allow the user to select the elements he or she is interested in for finding associations. This may not be desirable as far as data mining is concerned (DM is supposed to find something unexpected), but sometimes the user does know for sure some items are not interesting. By selecting only certain items would cut the sizes of the item sets and result in much faster computation. This can be done by presenting to the user the list of items extracted from the DTD. For a data set given in the next section that has four levels in the hierarchy, it will find over one hundred rules (many of which are not at all interesting to the user), but by selecting three items, the number of rules found is reduced to only 11, and the time spent on the computation is cut by 2/3.

5 Summary and Future Work

In this paper we have discussed a method for mining association rules from XML data. Although both data mining and XML are by themselves known technologies if considered alone, the combination of the two is still a research area that is attracting more and more attentions.

The major difficult lies on the fact that the XML format is semi-structured and does not particularly suitable for the existing data mining algorithms. Several problems we have to deal with, including (a) items in "transactions" may be cross-levels and spread in different sub-trees, (b) there may be missing data as, and (c) duplicate tag names. An encoding scheme is used to assign each XML element a unique code that is used to index to the Value Table and the Transaction Table. The two tables, as "2-dimensional arrays," store the values of the elements and keep the hierarchical relationships among the elements. Using this internal data structure and indexing method, it is easy to retrieve XML data involved in "transactions." We used a modified Apriori algorithm to find frequent itemsets for mining association rules. The method was implemented and tested on some simple XML files and produced correct results. We are currently testing on large XML files to study the performance of the method. For future work, we plan to study ways of applying other data mining algorithms to XML data using the encoding approach and data structures. We will also be working on formulating DTD/Schema as an XML "standard" for data mining.

References

1. R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the ACM-SIGMOD Intl. Conf. on Management of Data (SIGMOD'93)*, 207-216, Washington, DC, July 1993.
2. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, "Fast discovery of association rules," in Fayyad, Piatetsky-Shapori, Smyth, and Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, 307-328, AAAI/MIT Press, 1996.
3. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proceedings of the Intl. Conf. on Very Large Databases (VLDB'94)*, 487- 499, Santiago, Chile, Sept. 1994.
4. Y. Aumann and Y. Lindell, "A statistical theory for quantitative association rules." *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, San Diego, CA, 261-270, Aug. 1999.
5. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket analysis," in *Proceedings of the Intl. Conf. on Very Large Databases (VLDB'97)*, 265-276, Tucson, AZ, May 1997.
6. A. G. Buchner, M. Baumgarten, M. D. Mulvenna, S.S. Anand, "Data mining and XML: current and future issues," *Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00)*, 127-131, Hong Kong, 2000.
7. Edmond, "XMLMiner, XMLRule and metarules white paper," Sciento Inc. April 2002.
8. H. Lu., L. Feng, and J. Han, "Beyond in transaction association analysis: mining multidimensional inter-transaction association rules," *ACM Transactions on Information Systems*, 423-454, Vol. 18, No. 4, October 2000.
9. J. Han, and M. Kamber, *Data Mining, Concepts and Techniques*, Morgan Kaufmann, CA, USA, 2001.
10. J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," in *Proceedings of the Intl. Conf. on Very Large Databases (VLDB'95)*, 420-431, Zürich, Switzerland, Sept. 1995.

Graphical Knowledge Template of CBD Meta-model

Haeng-Kon Kim

Department of Computer Information & Communication Engineering, Catholic University of Daegu, 712-702, South Korea
hangkon@cu.ac.kr

Abstract. There are several component reference models for component development. However, there is few integrated and generic reference model among reference models. That results in the problem of interoperability among component designs. In this paper, we propose an integrated component meta-model to support consistency and interoperability between component designs. Also we validate a proposed meta-model through graph theory. We expect that new meta-model will be added and extended because proposed meta-model is represented with UML's class diagram.

1 Introduction

Currently interests of component-based software development are being increased. For example, there are CBD96, RUP, Catalysis, Advisor, Fusion, and so on. Also, there are several CASE tools like as Rose, Together, and COOL series and technology platforms such as EJB, CCM, .NET and so on.[1]. These various methods, tools, and platforms have not a standard reference model, but a unique model. Each component reference model provides different notations and modeling elements for the same concept. A few reference models reflect characteristics of components fully on its meta-model. This raises the problems of inconsistency and low interoperability between components developed by different component reference model. Also, different reference models increase difficulties of communication between component designers and developers. In order to address the problems, we suggest a generic reference model, which is integrated and unified several reference models, as a forms of meta-model based on UML's class diagram [8].

The structure of this paper is as follows. Section 2 reviews existing component reference models as related approaches. Section 3 describes a generic and unified component reference model which integrates existing component reference models and suggests specification level and implementation level. Section 4 introduces graph grammar of component to verify proposed reference model, and validates proposed reference model based on proposed graph. Finally, concluding remarks and future works are described in Section 5.

2 Limitations of Existing Researches

2.1 SEI's Component Reference Model

SEI defines a component as a software implementation executable in physical and logical device. Therefore, a component implements one or more interfaces [2]. This

reflects that a component has a contract. Components developed independently depend on specific rules and different components can be interoperable with standard methods. Also, components can be executable dynamically in run time. Component-based system is a system developed based on independent component types executes specific roles in a system. Types of each component are represented with each interface.

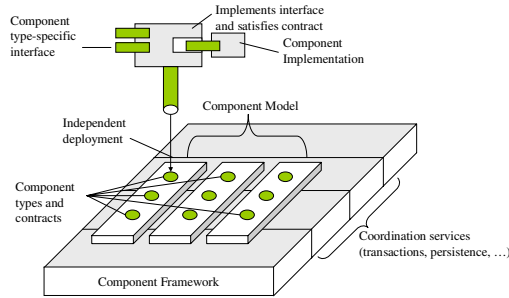


Fig. 1. Component Reference Model of SEI

A component model defines a set of specifications of component types, interfaces, and interaction pattern between component types. And component model is represented as a specification of standard and contract for component developer. Dependency of specific component model is a property that distinguishes one component from other components.

2.2 Perrone’s Component Reference Model

Perrone[3] defines a component as an unit which consists one or more classes and provides functions of classes with interfaces as depicted in Figure 2. In this research, a component is described as the concept of larger unit than the concept of existing class. Also, a component is described a unit which encapsulates separated problem domain.

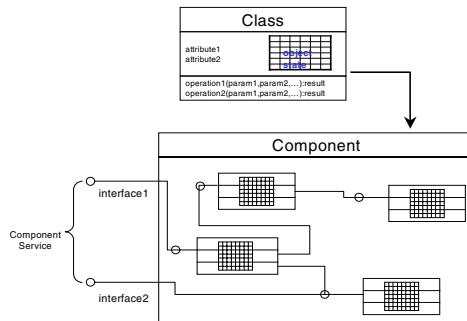


Fig. 2. Perrone’s Component Reference Model

Figure 2 represents basic elements contained in a component model. In this figure, Perrone defines that a component model is a component itself. Also, Perrone defines that a container is an environment in which a component is operated or worked. A container provides services that components require to send or receive messages in a standard way.

2.3 CORBA Component Reference Model

CORBA Component Model(CCM) separates a component into two-phases[6]. The one is basic component, and the other is extended component. Basic component provides a mechanism componentized CORBA object and can be mapped or integrated into EJB component. The extended component provides many functions than basic component. Both basic component and extended component are managed by component home.

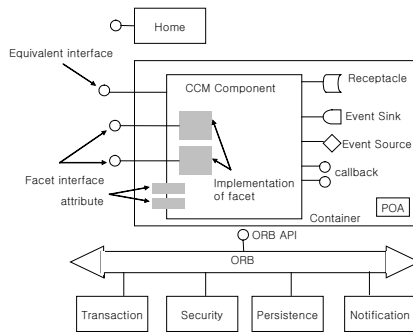


Fig. 3. CCM 's Reference Model

2.4 EJB's Component Reference Model

EJB component model is similar to CCM component model[7]. A component is executed in container and managed by home object. However, besides CCM component, EJB component is referenced by one component interface. Therefore, a bean does not have several interfaces. EJB component model is separated into local or remote interface and internal implementation logic.

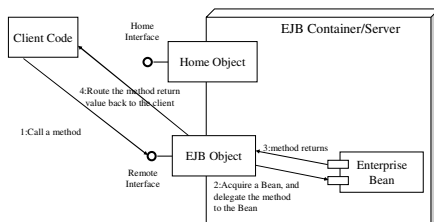


Fig. 4. EJB Reference Model

3 Generic Component Reference Model

In this section, we suggest a new component reference model based on component reference model of section 2. Figure 5 is a generic component reference model. Figure 5 describes both structural elements and dynamic elements of a component. Dynamically component workflows are occurred through calling operations of provide interface in a component depicted in Figure 5.

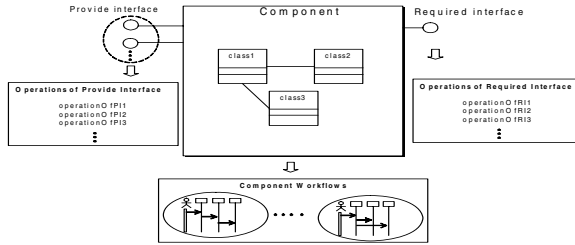


Fig. 5. Generic Component Reference Model

Figure 6 describes meta-model of static elements of a component with UML's class diagram[8].

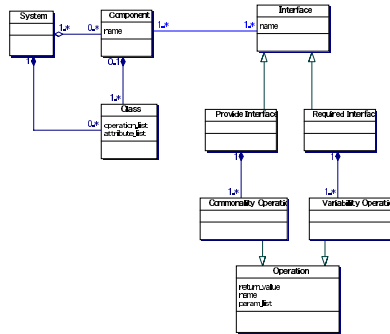


Fig. 6. Meta-Model of Static Elements in a Component

3.1 Meta-model of Specification Level

Specification level meta-model describes common information among CBD methodologies. Minimal meta-model of specification level represents component definition commonly contained in all of CBD methodologies.

A component consists of component declaration and component definition. Component declaration contains one or more interfaces containing one or more method declarations. Interface is classified into provided interface and required interface. Method declaration only contains method's signature, zero or more

pre-conditions and post-conditions. There are one or more classes in component definition. These classes implement interfaces declared in component declaration. Definition of class and attribute is based on UML’s definition. Also, component specification contains one or more interface specifications and components.

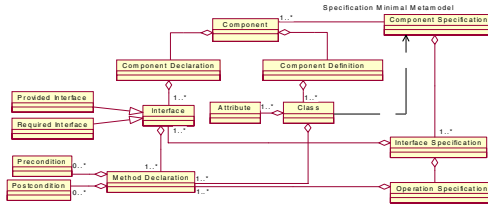


Fig. 7. Minimal Meta-Model of Specification Level

Maximal meta-model of specification level is depicted in Figure 8. Maximal meta-model represents information defined in CBD methodologies.

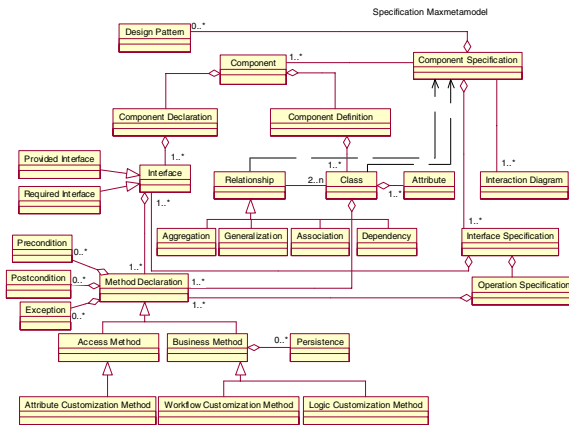


Fig. 8. Maximal Meta-model of Specification Level

In the specification maximal meta-model, method declaration is classified into access method and business method. Attribute customization method is a subclass of access method, because attribute customization method contains methods for variant attribute type as well as variant attribute values while access method contains get or set methods.

3.2 Meta-model of Implementation Level

Implementation level meta-model is used to implement components using specific component platforms such as EJB, CCM, and COM. Meta-model of implementation

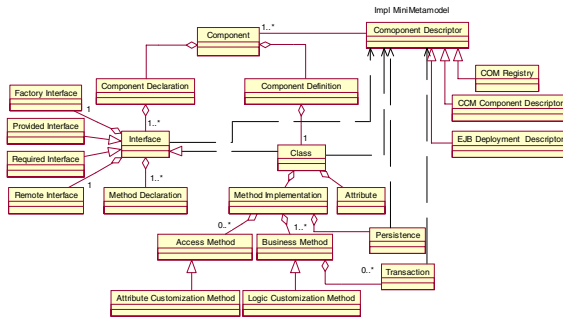


Fig. 9. Minimal Meta-Model of Implementation Level

is divided into two types; minimal meta-model(Figure 9) and maximal meta-model(Figure 10).

Minimal meta-model of implementation level represents additional information related with component implementation. For example, information related with transaction is reflected on business method. While component information is described in component specification in specification level meta-model, it is described in component descriptor in implementation level meta-model. Maximal meta-model of implementation level describes comprehensive of component platforms. Therefore, there is additional information needed in component implementation according to component platform.

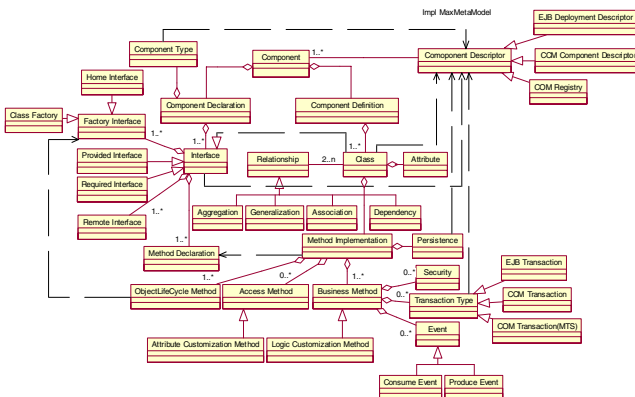


Fig. 10. Maximal Meta-Model of Implementation Level

4 Verification

In this section, we validate or verify whether proposed meta-model is well defined or not by using OMG's Meta Object Facility(MOF).

4.1 Verification of Meta-model Through MOF

MOF defines meta meta-model required composing, validating, and transforming expressible all of meta models including UML meta model.[10]. In order to verify whether component meta-model is well defined or not, we first should prove that proposed component meta-model is instance of MOF. And then, we should prove that proposed meta-model conforms to rules of MOF. The fact of component meta-model confirming to MOF can be proven by the relationship between MOF and component meta-model.

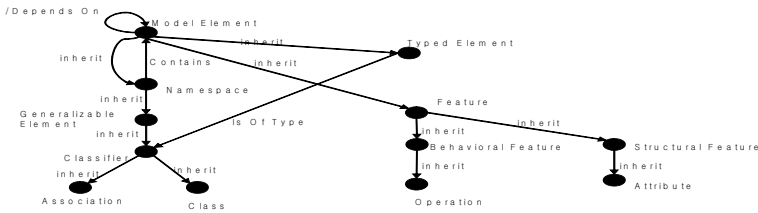


Fig. 11. BMOF Graph

MOF basic elements can be regarded as a graph including node and arc. A node represents an element of MOF, and arc means the relationship between elements. Therefore, MOF might be transformed into BMOF graph such like Figure 11. Finally, we also can transform minimal meta-model of specification level into a graph such like Figure 12.

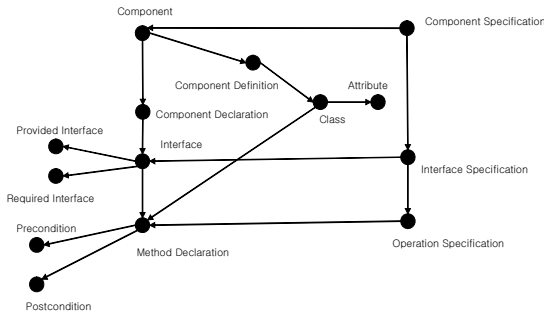


Fig. 12. Graph for Minimal Meta-Model of Specification Level

A graph can be described a pair of node and arc. Therefore, BMOF is represented as follows:

$$\mathbf{BMOF} = \langle N_b, A_b \rangle$$

$N_b = \{ \text{Model Element, Namespace, GeneralizableElement, Package, Classifier, Association, Class, Typed Element, Parameter, Structural Feature, Feature, Behavioral Feature, Operation, Attribute} \}$

$A_b = \{ \text{Depends On, Contains, ...} \}$

Minimal meta-model of specification level, **SMGraph**, is represented as follows:

SMGraph = $\langle N_s, A_s \rangle$

$N_s = \{ \text{Component, Component Specification, Component Declaration, Component Definition, Interface Specification, Operation Specification, Class, Attribute, Interface, Provided Interface, Required Interface, Precondition, Postcondition, Method Declaration} \}$

$A_s = \{ A_{\text{component_definition}}, \dots \}$

BMOF' becomes a basis to check the consistency of **SMGraph**. In order to have consistency, there should be nodes of **BMOF'** mapped into all nodes of **SMGraph**. Also, if there is an arc relating two nodes for any two nodes in **SMGraph**, there should be an arc for two nodes of **BMOF'**. Expression of those is as follows:

$F : \text{first}(\text{SMGraph}) \rightarrow \text{first}(\text{first}(\text{BMOF}'))$

$\forall n_i, n_j (n_i \neq n_j) \in \text{dom } F, \text{arc}(n_i, n_j) \in \text{second}(\text{SMGraph}) \cdot$

$\exists \text{arc}(F(n_i), F(n_j)) \in \text{rand } F$

In order to prove the consistency of **SMGraph**, the function from **SMGraph** into **BMOF'** should be defined. Function **F** is defined as follows:

$F(\text{component}) = \text{Classifier} \dots \dots \dots (1)$

$F(\text{Component Specification}) = \text{Package} \dots \dots \dots (2)$

$F(\text{Component Declaration}) = \text{Package} \dots \dots \dots (3)$

$F(\text{Component Definition}) = \text{Package} \dots \dots \dots (4)$

$F(\text{Interface}) = \text{Classifier} \dots \dots \dots (5)$

$F(\text{Provided Interface}) = \text{Classifier} \dots \dots \dots (6)$

$F(\text{Required Interface}) = \text{Classifier} \dots \dots \dots (7)$

$F(\text{Attribute}) = \text{Attribute} \dots \dots \dots (8)$

$F(\text{Class}) = \text{Class} \dots \dots \dots (9)$

$F(\text{Method Declaration}) = \text{Operation} \dots \dots \dots (10)$

$F(\text{Precondition}) = \text{Constraint} \dots \dots \dots (11)$

$F(\text{Postcondition}) = \text{Constraint} \dots \dots \dots (12)$

$F(\text{Interface Specification}) = \text{Package} \dots \dots \dots (13)$

$F(\text{Operation Specification}) = \text{Package} \dots \dots \dots (14)$

All arcs of **SMGraph** are mapped into all arcs of **BMOF'**. It means that all nodes of **BMOF'** are kinds of **Model Element**, and there are relationship of "Depends On" between **Model Element**.

Following mappings for maximal meta-model of specification level are added.

$F(\text{Design Pattern}) = \text{Classifier} \dots \dots \dots (16)$

$F(\text{Interaction Diagram}) = \text{Classifier} \dots \dots \dots (17)$

$F(\text{Exception}) = \text{Exception} \dots \dots \dots (18)$

$F(\text{Persistence}) = \text{Tag} \dots \dots \dots (19)$

$F(\text{Relationship}) = \text{Association} \dots \dots \dots (20)$

Following mappings for minimal meta-model of implementation level are added.

$F(\text{Component Descriptor}) = \text{Costraint} \dots \dots \dots (21)$

$F(\text{Transaction}) = \text{Tag} \dots \dots \dots (22)$

Maximal meta-model of implementation level includes following mappings.

F(Security)=Tag.....(23)

F(Transaction Type)=Tag.....(24)

F(Event)=Operation.....(25)

5 Conclusion Remarks

We define and propose meta-models for component with respect to generic view, specification view, and implementation view. Also, each meta-model of specification level and implementation level is divided into minimal meta-model and maximal meta-model. In order to verify the correctness and soundness of proposed meta-models, we use graph theory and MOF.

We expect that models of various methodologies and platforms will be integrated as well as new model elements are added or extended easily by applying proposed meta-model.

References

1. heineman, G. T., Council, W. T., Component-based Software Engineering, Addison Wesley, 2001.
2. Bachman, F., et. al., "Volume 2, Technical Concepts of Component-based Software Engineering", Carnegie Mellon Software Engineering Institute, 2000.
3. perrone, p., Building Java Enterprise Systems with J2EE, Sams Publishing, 2000.
4. Butler Group, "Catalysis: Enterprise Components with UML", at URL: <http://www.catalysis.org>, pp.2, 1999.
5. Desmon F. D'Souza and Alan Cameron Wills, Objects, Component and Frameworks with UML, Addison Wesley, 1999.
6. OMG, Final FTF Report of the Component December 2000, OMG Inc., 2001.
7. Roman, E., Mastering Enterprise JavaBeans, Jon Wiley and Sons, Inc., 2002.
8. UML Specification v1.4, OMG, Inc., September, 2001.
9. D. Harel, A. Naamad, "The STATEMATE semantics of Statecharts," ACM Transactions on Software Engineering and Methodology, Vol.5, No.4, pp.293-333, 1996.
10. Meta Object Facility Specification, OMG, URL: <http://www.omg.org>.
11. Akehurst, D.H, "Model Translation: A UML-based specification techniques and active implementation approach", PhD Thesis, University of Kent at Canterbury, 2000.

Two-Phase Identification Algorithm Based on Fuzzy Set and Voting for Intelligent Multi-sensor Data Fusion

Sukhoon Kang

Department of Computer Engineering, Daejeon University
96-3 Yongun-Dong, Dong-Gu, Daejeon, Korea 300-716
shkang@dju.ac.kr

Abstract. Multi-sensor data fusion techniques combine data from multiple sensors in order to get more accurate and efficient meaningful information through several intelligent process levels that may not be possible from a single sensor alone. One of the most important parts in the intelligent data fusion system is the identification fusion, and it can be categorized into physical models, parametric classification and cognitive-based models. In this paper, we present a novel identification fusion method by integrating two fusion approaches such as the parametric classification techniques and the cognitive-based models for achieving high intelligent decision support. We also have confirmed that the reliability and performance of two-phase identification algorithm never fall behind other fusion methods. We thus argue that our heuristics are required for effective decision making in real time for intelligent military situation assessment.

1 Introduction

Multi-sensor data fusion is an emerging intelligent technology applied to military applications such as automatic identification, analysis of battlefield situations and threat assessments. We can obtain more accurate and efficient meaningful information through multi-sensor data fusion that may not be possible from a single sensor alone. Especially, intelligent identification fusion is very important in order to estimate and correspond to battlefield situation as well as analyze threat assessments exactly. Most of identification fusion methods used in multi-sensor data fusion systems have a tendency of missing sensors' uncertain reports even if sometimes those reports help user to decide target's identity. The purpose of this paper is to provide an intelligent two-phase identification fusion method in which we integrate two different identification fusion approaches in order to consider even sensors' uncertain reports and improve processing time.

2 Related Work

In recent years, many theories and research projects about data fusion have been addressed by world-wide scientists and applications of data fusion span military

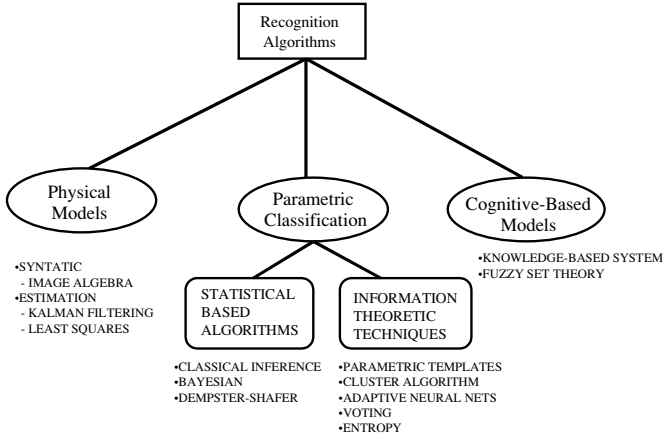


Fig. 1. Taxonomy of identification fusion algorithms

problems such as automatic identification of targets, analysis of battlefield situations and threat assessments [1]. Among the algorithms used in multi-target multi-sensor data fusion, identification fusion algorithm can be categorized as shown in Fig. 1.

Physical models attempt to model accurately observable or calculable data and estimate identity by matching predicted observations with actual data. Parametric classification seeks to make an identity declaration based on parametric data, without utilizing physical models. A direct mapping is made between parametric data and a declaration. We may further subdivide these into *statistically based techniques* and *information theoretic techniques*. Cognitive-based models seek to mimic the inference processes of human analysts in recognizing identity [2]. In this paper, we present the two-phase heuristic identification method in which we graft cognitive-based models and parametric classification techniques, that is, we apply fuzzy set theory to phase-I fusion in order to consider even sensors' uncertain reports and voting method to phase-II fusion in order to improve processing time.

3 Generation of Feature Membership Function for Intelligence

In this section, we shows the generation of *feature membership function* corresponding to each signal feature of emitting target referring to target database for intelligence.

The basic concept of fuzzy set [3] seeks to address problems in which imprecision is an inherent aspect of a reasoning process. The identification method which will be provided in this paper is composed of two-phase fusion, and we apply fuzzy set theory to the phase-I fusion.

ELINT stands for ELectronic INTelligence, and refers to intelligence-gathering by use of electronic sensors [4]. First, we assume that there are N number of ELINT sensors which detect emitting targets on the battlefield, and among the signal features

which are detected by each ELINT sensor. We consider only signal frequency, SP(Scan Period), PW(Pulse Width) and PRI(Pulse Repetition Interval), because those four features are the most important characteristics for intelligent identification fusion in order to identify each target. In order to recognize the identity value of the target, we suppose that there may exist four kinds of emitting targets (A, B, C and D) on the battlefield and there is a target database table as shown in Table 1.

Table 1. An example of target database table

Features \ Target	A	B	C	D
Avg. Frequency (MHz) $\sigma(\text{Frequency}) = 5 \text{ MHz}$	310	330	315	320
Avg. SP (Sec) $\sigma(\text{SP}) = 1 \text{ sec}$	5.7	4.5	5.5	6
Avg. PW (μsec) $\sigma(\text{PW}) = 1 \mu\text{sec}$	4.0	5.5	3.7	4.2
Avg. PRI (μsec) $\sigma(\text{PRI}) = 1 \mu\text{sec}$	4.5	6.0	5.0	5.5

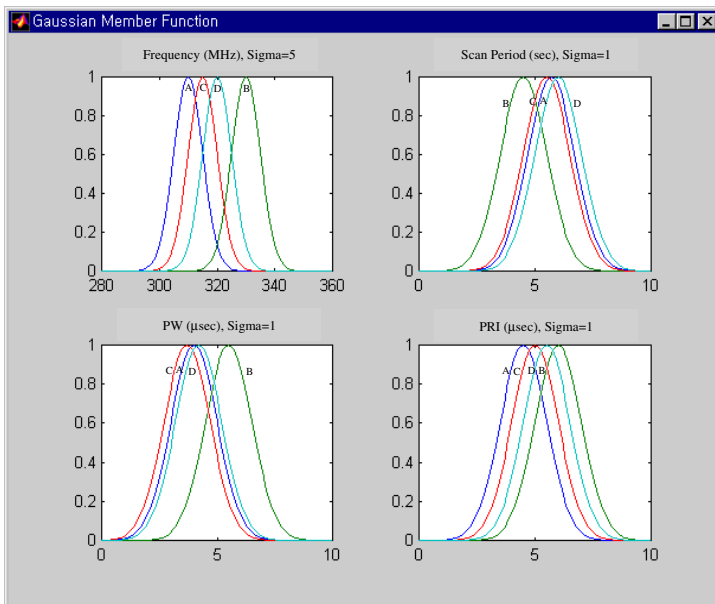


Fig. 2. Generation of feature membership functions of each target

σ represents a resolution of ELINT sensor for detecting each feature of emitting target. With those features of each target and σ in this table, we can create feature membership functions using Gaussian function as shown in Fig. 2.

4 Intelligent Two-Phase Identification Fusion Algorithm

The overall structure for intelligent processing is composed of two-phase identification fusion as shown in Fig. 3. First, when ELINT sensor detects a target on the battlefield, it declares the value of signal features such as frequency, SP, PW, PRI. With these sensor reports, phase-I fusion generates grade of each feature by applying to feature member functions, and run a Max-Min algorithm. In phase-II fusion, we decide the final identity of the target by applying to voting method using the results of phase-I fusion as inputs.

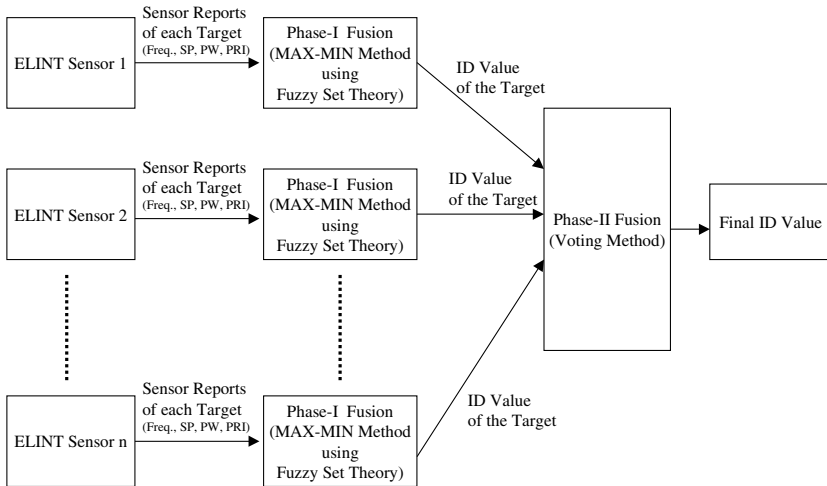


Fig. 3. Overall structure of two-phase identification

4.1 Phase-I Fusion

When an ELINT sensor detects emitting targets on the battlefield by acquiring features such that signal frequency, SP, PRI and PW, phase-I fusion algorithm declares a grade ($0 \leq \text{grade} \leq 1$) of each feature by applying to each feature membership function for all detected targets. That is, if we represent the order of each grade as the pair of [grade of frequency, grade of SP, grade of PW, grade of PRI], each value in the bracket has the value between 0 and 1. Thus, if there is N number of target information in the database table, then N pairs of grades are generated for a detected target. Then, Max-Min algorithm is applied, that is, we select the target which has the minimum grade from each pair. With the N number of minimum grades, we select the target which has the maximum grade as the result of phase-I fusion. The detail structure of phase-I fusion is shown in Fig. 4.

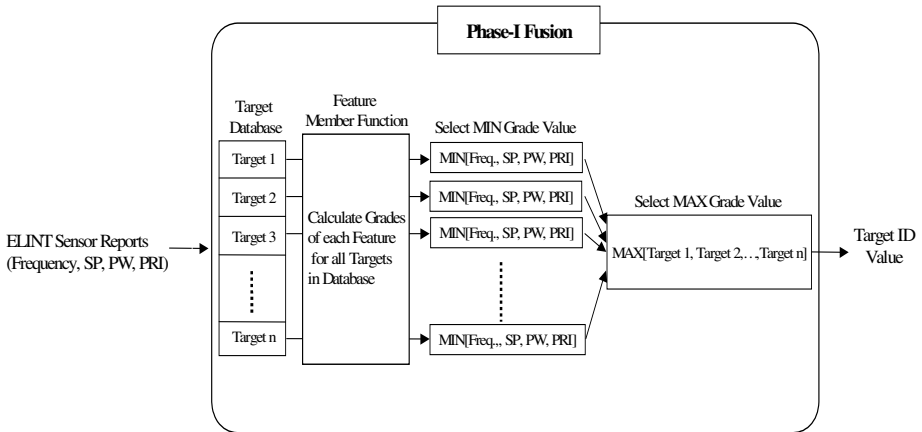


Fig. 4. Structure of phase-I fusion

For example, if an ELINT sensor has detected an emitting target acquiring following feature values:

Detected signal frequency value = 321 MHz

Detected SP value = 4.8 sec

Detected PW value = 5.2 μ sec

Detected PRI value = 6.7 μ sec.

Phase-I fusion algorithm generates a grade of frequency, SP, PW and PRI applying to each feature member function for four number of targets which exist in target database (see table1). Figure 5 shows the process of this step. For detected features, each pair of grades is generated such that target A, target B, target C and target D have a pair of grade [0.05, 0.58, 0.43, 0.1], [0.3, 0.96, 0.98, 0.8], [0.4, 0.72, 0.29, 0.23], and [0.87, 0.4, 0.6, 0.47]. Among those pairs of grade, if we select minimum value of grade from each pair, then we can get 0.05 for target A, 0.3 for target B, 0.23 for target C and 0.4 for target D. Finally, we choose the target D as the result of phase-I fusion by selecting the maximum value of grade.

With the N number of phase-I fusion results by applying this way to N number of ELINT sensors, we process phase-II fusion.

4.2 Phase-II Fusion

In phase-II fusion, we apply phase-I fusion results to voting methods in phase-II fusion in order to get final target identity value. Voting methods address the identification fusion problem by using a democratic process. The decision making from N sensors simply counted as votes with a majority or plurality decision rule. For example, for 10 sensors, if 6 sensors declare the target is type A whereas 4 sensors declare the target is type B, we decide that target is type B. Although conceptually simple, the reason we adopt voting method is due to the fact that the voting technique can be valuable, especially for real-time techniques, when accurate a priori statistics are not available, or may be attractive from an overall cost-benefit point of view [2].

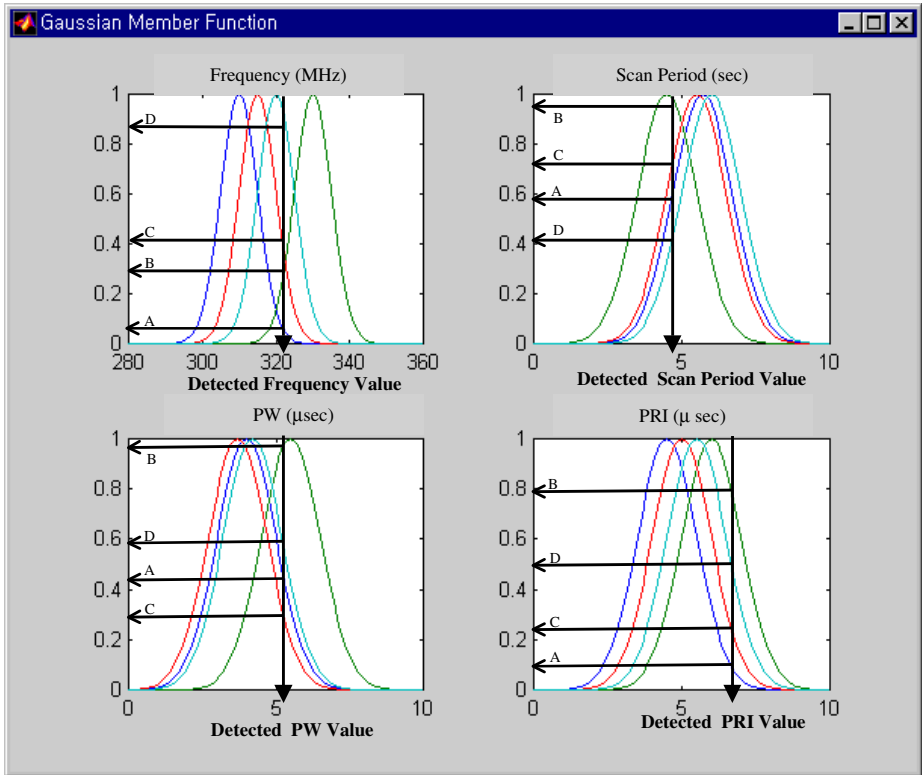


Fig. 5. Extraction of gades from each membership function

5 Performance Analysis

In order to analyze the performance of two-phase identification fusion method, we compared this method with Bayesian method and Dempster-Shafer method focusing on processing time and accuracy. The typical characteristics of these two methods are summarized as following [5]:

- Bayesian method:
 - (1) It is mathematically well established.
 - (2) Processing time is generally faster than Dempster-Shafer method.
 - (3) It is difficult to define prior likelihoods Requirement that competing hypothesis should be mutually exclusive.
- Dempster-Shafer method:
 - (1) It deals with all possible hypotheses combination.
 - (2) It Provides detail fusion information to user.
 - (3) Processing time is exponentially increased as hypothesis is getting increased.

For the empirical analysis, we created a scenario in which there are 10 emitting targets on the battlefield and we let the four ELINT sensors detects those ten targets. When each ELINT sensor detects targets, we iterated the process of the identification fusion 300 times in order to get much precise results. Fig. 6 illustrates the fusion results with a graph of accuracy versus time in seconds.

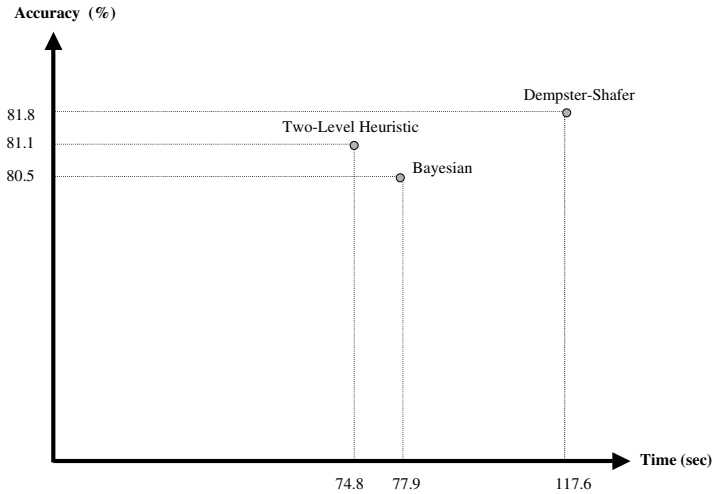


Fig. 6. Comparison of Bayesian, Dempster-Shafer and Two-phase identification method

The accuracy which is illustrated on the graph represents the average value of accuracy corresponding to 300 times iteration and time represents the total processing time corresponding to 300 times iteration. Additionally, we applied *Nearest Neighborhood Association method* for the identification fusion and performed comparisons under circumstance of Pentium-IV processor.

Regarding accuracy of neural network classifiers as a property of size of the data set, it seems that an increase in the size of the data set dose not directly result in improved accuracies. But, accuracy of neural network is improved when the quality of the data set is improved. On the other hand, accuracy of neural network is decreased when there is redundancy or outlier in the data set.

The numerical value itself of processing time is not important, because it totally depends on the computer system performance. However, we can infer from this analysis that two-phase heuristic identification fusion method took less processing time than any other two fusion methods, and moreover, the accuracy of this fusion method could be comparable with other two methods.

6 Conclusions

In this paper, we present an intelligent multi-sensor data fusion scheme based on two-phase identification algorithm by grafting two data fusion approaches such as the *parametric classification techniques* and the *cognitive-based models* through the

enough tests. We also confirmed that the performance of two-phase identification never falls behind other fusion methods. Based on proposed intelligent fusion method, if we utilize different kinds of sensors for the two-phase heuristic identification fusion method such as MTI sensor, SAR sensor, EO sensor, etc. and apply weighted voting method to the phase-II fusion, then the result of the identification algorithm would be much reliable.

References

1. David L. Hall, James Linas, An Introduction to Multisensor Data Fusion, Proceeding of the IEEE, Vol. 85, No. 1, 1997
2. Edward Waltz and James Linas, Multisensor Data Fusion, Artech House, Norwood, Massachusetts, 1990
3. Zadeh, L.A., Fuzzy Sets and Systems, Amsterdam, North-Holland Press, 1978
4. Richard G. Wiley, Electronic Intelligence, Artech House, Norwood, Massachusetts, 1982
5. David L. Hall, Mathematical Techniques in Multisensor Data Fusion, Artech House, Norwood, Massachusetts, 1992

*ui*H-PMAC Model Suitable for Ubi-Home Gateway in Ubiquitous Intelligent Environment

Jong Hyuk Park¹, Sangjin Lee¹, Byoung-Soo Koh², and Jae-Hyuk Jang³

¹ Center for Information Security Technologies, Korea University,
5-Ka, Anam-Dong, Sungbuk-Gu, Seoul, Korea
{hyuks00, sangjin}@korea.ac.kr

² DigiCAPs Co., Ltd., Jinjoo Bldg., Bangbae-Dong, Seocho-Gu, Seoul, Korea
bskoh@digicaps.com

³ Division of Computer Engineering, Daejon University, Dong-Gu, Daejon, Korea
good4u@zeus.dju.ac.kr

Abstract. In this paper, we propose a *ui*H-PMAC (*ubi-intelligent* Home - Privilege Management Access Control) model which is suitable to access control of ubiquitous intelligent environment. The model considers the temporary limit conditions as access control of home device or user location information considering the characteristics of members who could limit the use of resource. In addition, the model provides authentication policy which authenticates not only devices but also domains and access control policy which considers inheritance of role. Finally, the model can be applied various and scalable access control policies which are suitable to the characteristics of intelligent digital environment as considering the user location information as a temporary limit condition.

1 Introduction

Ubiquitous computing was proposed by Mark Weiser of Xerox Palo Alto Research Center in 1998. It is the concept that all objects and spaces are intellectualized and useful services are provided through the interactions among many computers which are not recognized to users through the unlimited connection in anytime and anywhere. For providing these ubiquitous services, automatic and intellectualized service, made-to-order service for user, user convenience or transparency and interoperability among heterogeneous devices should be needed [1-4].

Moreover, the solutions for network authentication, authority's permission for access of resource, access control and authentication problems about multimedia data policy and security are needed in security side [5]. One of important services in ubiquitous service is access control. As it is a mean to use and modify the resource and multimedia contents within each device, it is a technical method that only identified or authenticated user could access the information of home within permitted scope.

¹ This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

As these method, there are DAC (Discretionary Access Control) which is an early access control method or MAC (Mandatory Access Control), RBAC (Role Based Access Control) [6, 7] which is an extended access control method by role concept. T-RBAC (Task-Role Based Access Control) Model [8], which is suitable to enterprise environment. But, these models' shortcomings is not support the limitation of resource use in ubiquitous environment, access control function which considers location of user and access control function according to security policy between device and user. Although user are identical, different access rights could be issued according to the location that user intends to access to resource.

Furthermore, home devices should be controlled or restricted according to the characteristics of family members. For example, in case of a visitor, the use of service within home should be limited and in case of specific user, the access rights could be issued about correspondent device but the correspondent device could be restricted within specified time.

For solving these problems, GTRBAC (Generalized Temporal Role Based Access Control) [9] which could limit the use of resource by time (period) and GTRBAC which prevents the misuse of authority by applying PBDM(Permission Based Delegation Model) and provides flexible delegation policies, were proposed but these could not control the permission of rights.

In this paper, we propose extended RBAC delegation model (*uiH-PMAC*) that home gateway could easily control the inheritance of rights as a domain role and provide flexible delegation policies. In addition, the proposed model is extended to suit the elaborate and complex access control required from ubiquitous intelligent environment by considering temporary control conditions about the user location information.

The rest of this paper is organized as follows. In section 2, we describe the proposed model through the formal specification after explaining the characteristics of it. In section 3, we discuss the characteristic of proposed model through the comparison analysis with existing access control model and then we come to a conclusion in section 4.

2 *uiH-PMAC* Model

A technical part in ubiquitous intelligent environment has been more complicated and diversified according to the rapid development of information infrastructure and security technology. Therefore, the unified access control technology about many users and various devices on ubiquitous intelligent environment are needed. Specially, the rights management system and access control technology of users about broadcasting contents, devices and multimedia content services in ubiquitous intelligent environment are required such as follows: It should be support roles based on user location. In addition, it should be that one user supports the roles of various devices and the role of various users. Furthermore, it should be support role inheritance and multi-stage delegation.

Figure 1 shows the *uiH-PMAC* model architecture for home gateway in ubiquitous intelligent environment. A *uiH-PMAC* model limits the use of resource according to the user location information in basic RBAC model and considers the restriction of user in each domain.

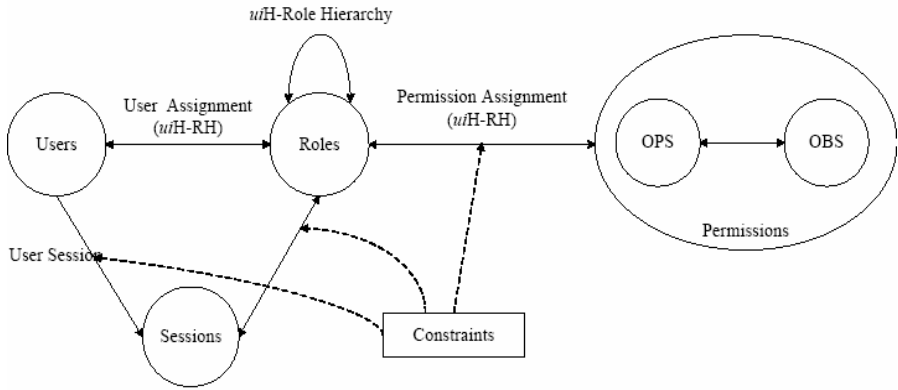


Fig. 1. uiH-PMAC Architecture

It assigns the essential role among sub-roles to user by uiH PMAC user-role allocation relationship and the rights which is allotted to other sub-roles is allocated to user by inheritance relationship of sub-role level when a user makes a essential role active.

- **Users:** Users who uses the each device or sensor network in ubiquitous intelligent environment.
- **Roles:** It means Security Policy DB and it composed of Domain Role, Users Role, Device Role and Privilege.
- **OPS:** Operations
- **OBS:** Objects
- **Permissions:** {Operations} * {Object}

2.1 uiH-RH: ubi-intelligent Home Role Hierarchy

Figure 2 shows ubi-intelligent Home Role Hierarchy (uiH-RH) and it satisfies the basic rights inheritance among each role. Home gateway receives the signal value of sensor installed within home network as an input and analyzes whether it can apply to correspondent role.

Devices and Users which are under the control of F-Domain (Full-Domain) permits the rights to all users and devices within home. Furthermore, P-Domain (Partial-Domain) is just possible to inherit the designated role in role analysis and security policy processes and a role is just assigned to one device in Pr-Domain (Private Domain).

2.2 Formal Specification of uiH-PMAC Model

In this subsection, we explain uiH-PMAC model for Ubi-Home Gateway. The notations in table 1 are used throughout this paper.

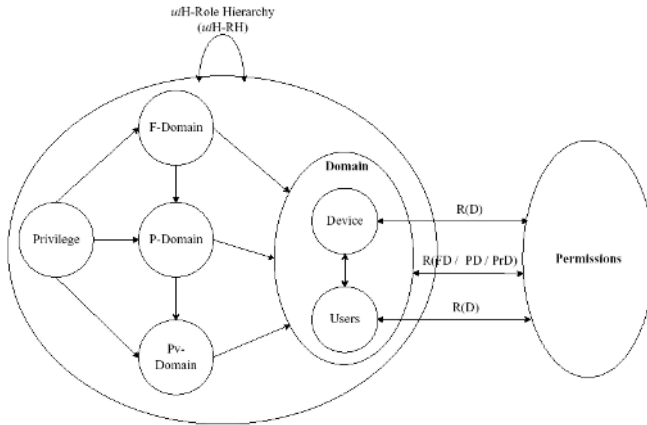


Fig. 2. Role Hierarchy in *uiH-PMAC* Model

Table 1. Notations appeared in formal specification of *uiH-PMAC* model

Notation	Description
U, R, P, S, L, D	User, Role, Privilege, Session, Location, Device
DO	Domain which includes device and user.
OP, OB	Operation, Object
SC	Security Class, security level about each element
$R(U) / R(D)$	User Role / Device Role
$R(FD) / R(PD)$	Full-Domain Role / Partial-Domain Role
$R(PrD)$	Private-Domain Role
$P(U) / P(D)$	User Privilege / Device Privilege
$P(FD) / P(PD)$	Full-Domain Privilege / Partial-Domain Privilege
$P(PrD)$	Private-Domain Privilege
PL	Privilege of User or Device

The formal specification of the proposed model is as follows.

- $U=\{u_i|i=1, \dots, n\}$: A user exactly corresponds to a role
- $R=\{r_i|i=1, \dots, n\}$: A role exactly corresponds to operation.
- $OP=\{op_i|i=1, \dots, n\}$: It is the all acts about objects supported from system and it is described through confidentiality and definition of permission relationship.
- $SC=\{sc_i|i=1, 2, 3, 4\}$ $sc_1 \subset sc_2, sc_2 \subset sc_3, sc_3 \subset sc_4$: It is the relationship about process action among roles, operation, the privilege of low class is added to upper class.
- $U \rightarrow R$: It issues the relationship between user and role.
- $R \rightarrow SC$: It issues the security class to role.
- $OP \rightarrow SC$: It issues the security class to operation.

- $OB \rightarrow SC$: It issues the security class to object.
- $R(U)$: Users * Roles

2.3 uiH-PMAC Model Scenario

Figure 3 shows Block-diagram among objects at uiH-PMAC model

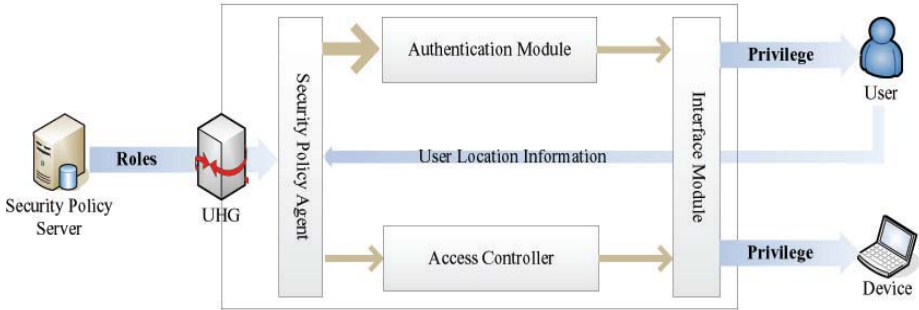


Fig. 3. Block-diagram among objects at uiH-PMAC model

- Security Policy Server: It manages DB and sends the Role information according to the request of Ubi-Home Gateway.
- Security Policy Agent: It sends the information of user or device to SPS (Security Policy Server).
- Access Controller: It handles and controls the rights information of user or device.
- Authentication Module: It authenticates a user or device.
- Interface Module: It provides the interface between Ubi-Home Gateway and User, Ubi-Home Gateway and device.

Table 2. Notations appeared in a scenario of the uiH-PMAC model

Notation	Description
User	User in Home Network
UHG	Ubi-Home Gateway
USN_device	USN Device
Home_device	Device in Home Network
Device(Client)	Client Device
user_cert	User Certificate
User_Info	User Information
UserID	User ID
User_Location_Info	User Location Information
Device_Info	Device Information
Device_Cert	Device Certificate

Home Gateway is controlled by Security Policy in the proposed model and it handles the access control of client, user, device and domain. Also, the user information is transmitted to Security Policy Agent by USN. Security Policy Server handles the Privilege control of multimedia contents and Privilege control of device of user as it lets Access Control know the information of correspondent user.

A User requests user authentication to manage permission of Ubi-Home Gateway and Ubi-Home Gateway transmits the result message after confirming the user authentication information.

$$1) \text{ User} \rightarrow \text{UHG}_{\text{user_cert}}, \text{UHG} \rightarrow \text{User}_{\text{OK}}$$

Ubi-Home Gateway receives ID and information of user through sensor network module of user. This information is utilized as metadata information when a user uses a other device and the access rights of correspondent device is issued.

$$2) \text{ User} \rightarrow \text{USN_device}_{(\text{UserID} \cap \text{User_Info})}$$

$$3) \text{ USN_device} \rightarrow \text{UHG}_{(\text{UserID} \cap \text{User_Location_Info})}$$

USN Device transmits the location and authentication information of user from Ubi-Home Gateway and it collects the location information of correspondent user based on message received from Ubi-Home Gateway.

$$4) \text{ USN_device}_{(\text{UserID} \cap \text{User_Info})} \rightarrow \text{Device}$$

User receives authentication information from device to confirm the authentication and rights information of device..

$$5) \text{ UHG} \rightarrow \text{Device}_{\text{Info_Req}}$$

$$6) \text{ Device} \rightarrow \text{UHG}_{(\text{Device_Cert} \cap \text{Device_Info})}$$

Ubi-Home Gateway authenticates a device and issues the usage rights of correspondent content after it confirms whether a correspondent user has the usage rights to use a device or not.

$$7) \text{ UHG} \rightarrow \text{Device}_{\text{Cert_Auth}}$$

$$8) \text{ Device} \rightarrow \text{HomeDevice}_{(\text{Device_Cert} \cap \text{User_Content})}$$

3 Comparison Between the Proposed Model and Existing Model

In this paper, we propose model where delegation between device and user for access control of ubiquitous intelligent environment (UIE). In addition, its model provide access control for resource considering location information of user.

In this section, we compare the proposed model with the existing model [6, 7, 9, 10] such as seen at table 3.

Table 3. Comparison between proposed model and existing model

Characteristics	RBAC[6,7]	TRBAC[10]	GTRBAC[9]	uiH-PMAC
Consideration of role and rights (UIE)	○	X	○	○
Consideration of user location information (UIE)	X	X	X	○
Inheritance of role (UIE)	X	X	○	○

- **Consideration of user location information (UIE):** The consideration of user location information means that the Privilege of correspondent domain is issued according to the location of user. In the existing model, the consideration of user location does not exist. The proposed model considers the user location information of ubiquitous intelligent environment as it flexibly applies the user location information using USN such as the scenario of the subsection 3.3.
- **Inheritance of role (UIE):** The inheritance of role means the role inheritance of user or device within domain. The consideration about domain does not exist in the existing model. The proposed model considers the role inheritance of user or device in domain environment as it classifies and applies the domain by three kinds of domain.
- **Delegation according to the user location information (UIE):** Delegation according to the location information of user means that power is delegated according to user location. The consideration for the user location information and the delegation policy according to the location information do not exist in the existing model. The proposed model the power delegation in cast of moving to other domain as it considers not only a user or device but also the user location or the Privilege of user such as the scenario of the subsection 3.3.

4 Conclusion

In this paper, we proposed a new access control model considering the user location information and domain concept in ubiquitous intelligent environment that were not supported in the existing RBAC. In addition, the model was considered the temporary limit conditions as access control of home device or user location information considering the characteristics of members who could limit the use of resource. Furthermore, the model can be provided authentication policy which authenticates not only devices but also domains and access control policy which considers role inheritance.

In the future, we will need to study the model which considers security policy. Finally, we will study model is able to control various contents and devices in ubiquitous intelligent environment.

References

1. Mark Weiser: The Computer for the 21st Century, Scientific American (1991), 94-104
2. Jong Hyuk Park, Heung-Soo Park, Sangjin Lee, Jun-Choi: Intelligent Multimedia Service System based on context awareness in Smart Home, KES 2005, Springer-LNAI, Vol. 3681 (2005), 1146 -1152

3. M. Satya: IEEE Pervasive Computing Magazine, <http://www.computer.org/pervasive>
4. Jong Hyuk Park, Jun-Choi, Sangjin Lee, Hye-Ung Park, and Deok-Gyu Lee: User-oriented Multimedia Service using Smart Sensor Agent Module in the Intelligent Home, CIS 2005, Springer-LNAI, Vol. 3801, (2005) 313 – 320
5. Frank Stajano: Security for Ubiquitous Computing, Wiley (2002)
6. David F. Ferraiolo and D. Richard Kuhn: Role-based access controls, 15th NIST-NCSC National Computer Security Conference, Baltimore (1992), 554-563
7. Gregory Tassej, Michael P. Gallaher, Alan C. O'Connor, Brian Kropp: The Economic Impact of Role-Based Access Control, NIST Planning Report 02-1 (2002)
8. Sejong Oh and Seog Park: Task-Role Based Access Control (T-RBAC): An Improved Access Control Model for Enterprise Environment, DEXA 2000
9. James B.D. Joshi, Elisa Bertino, Usman Latif, and Arif Ghafoor: A Generalized Temporal Role-Based Access Control Model, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 1 (2005)
10. Bertino, E., Bonatti, P. A. and Ferrari E. TRBAC: A Temporal Role-Based Access Control Model, ACM Transaction on Information and System Security (TISSEC), Vol. 4, No. 3 (2001), 191-233

A One-Time Password Authentication Scheme for Secure Remote Access in Intelligent Home Networks

Ilsun You

Department of Information Science, Korean Bible University,
205 Sanggye-7 Dong, Nowon-ku, Seoul, 139-791, South Korea
isyou@bible.ac.kr

Abstract. One of the important services that intelligent home networks provide is to remotely control home appliances in home network. However, the remote control service causes intelligent home networks to have various security threats. Thus, for the service, intelligent home networks should provide strong security services, especially user authentication. In this paper, we provide a public key based one-time password authentication scheme for secure remote access in intelligent home networks. To provide 2-factor strong authentication conveniently and cost effectively, we adopt and enhance YEH-SHEN-HWANG's authentication scheme. Since our scheme uses a server-side public key to address the vulnerabilities of YEH-SHEN-HWANG's scheme, it can securely perform user authentication, server authentication and session key distribution without any pre-shared secret, while defending against server compromise.

1 Introduction

Nowadays, much interest has risen in intelligent home networks [1]. One of the important services that intelligent home networks provide is to remotely control home appliances in home network. This service enables residential users to remotely access and control home appliances such as TVs, lights, washing machines, and refrigerators using their handheld devices. For example, from their office, they can turn on or turn off their gas range using their cellular phone.

However, in spite of such convenience, the remote control service causes intelligent home networks to have various security threats such as masquerade, denial of service attacks and so forth. Furthermore, wireless links, whereby handheld devices are connected to intelligent home networks, are particularly vulnerable to passive eavesdropping, active replay attacks, and other active attacks. Thus, for the service, intelligent home networks should provide strong security services, especially user authentication.

In this paper, we provide a public key based one-time password authentication scheme for secure remote access in intelligent home networks. To provide 2-factor strong authentication conveniently and cost effectively, we adopt and enhance YEH-SHEN-HWANG's authentication scheme that is a one-time password authentication

scheme using smart cards [2]. Especially, we use a server-side public key to address the vulnerabilities of YEH-SHEN-HWANG's scheme. Thus, our scheme can securely perform user authentication, server authentication and session key distribution without any pre-shared secret, while defending against server compromise. Since handheld devices can be used instead of smart cards in intelligent home networks, additional authentication devices are not required for the deployment of our scheme.

The rest of the paper is organized as follows. Section 2 describes related works and section 3 reviews YEH-SHEN-HWANG's scheme, and describes their weaknesses. In section 4, we propose a public key based one-time password authentication scheme using smart cards. Section 5 analyzes the proposed scheme. Finally, section 6 draws some conclusions.

2 Related Work

In this section, we describe the *S/KEY* authentication scheme, its variants and YEH-SHEN-HWANG's scheme on which our scheme is based.

The *S/KEY* one-time password scheme is designed to counter replay attacks or eavesdropping attacks [3,4]. Although the *S/KEY* scheme thus protects against passive attacks based on replaying captured reusable passwords, it is vulnerable to server spoofing attacks, preplay attacks and off-line dictionary attacks [2,5]. Several researches have been conducted to solve these drawbacks of the *S/KEY* scheme. Michell and Chen proposed two possible solutions to resist against server spoofing attacks and preplay attacks [5]. Yen and Liao proposed a scheme that uses a shared tamper resistant cryptographic token, including the SEED, to prevent off-line dictionary attacks [6]. Recently, YEH, SHEN and HWANG proposed a one-time password authentication scheme which enhances the *S/KEY* scheme to resist against the above attacks [2]. However, since the scheme uses SEED as a pre-shared secret and the user's weak pass-phrase, exposure of SEED causes the scheme to retain the flaws of the *S/KEY* scheme [7]. Furthermore, it is vulnerable to some attacks such as stolen-verifier attacks, denial of service attacks and denning-sacco attacks [7-10]. Obviously, this scheme with the drawbacks cannot satisfy high-level security that the remote control service requires. In [11], the *S/KEY* based authentication scheme using a server-side public key was proposed. But, the scheme is vulnerable to denial of service attacks, while providing no way to verify the server's public key.

3 YEH-SHEN-HWANG's Scheme

YEH-SHEN-HWANG's scheme is composed of registration stage, login stage and authentication stage [2]. The scheme uses SEED as a pre-shared secret to resist against server spoofing attacks, preplay attacks and off-line dictionary attacks. Also, it uses smart cards to securely preserve a pre-shared secret SEED and simplify the user login process. Moreover, it provides a session key to enable confidential communication over the network.

3.1 Notation

- K is the user secret, C is the client and S is the server
- id is the user's identity and IP_X denotes the IP address of X
- SC_X denotes X 's smart card and
- SCN_X denotes the serial number of X 's smart card
- N is a permitted number of login times and D is a random number
- $H()$ denotes a collision-resistant hash function
- $H^X(m)$ means that the message m is hashed X times
- KR_X denotes the private key of X and KU_X denotes the public key of X
- PKC_X denotes X.509 public key certificate of X
- $P_X(m)$ means that the message m is encrypted with the public key of X
- $P_X^{-1}(m)$ means that the message m is encrypted with the private key of X
- p_t means t th one-time password
- \oplus means Exclusive-OR operation and $|$ means concatenation

3.2 Registration Stage

- (1) $C \rightarrow S: id$
- (2) $C \leftarrow S: N, SEED \oplus D, H(D)$
- (3) $C \rightarrow S: p_0 \oplus D$, where $p_0 = H^N(K \oplus SEED)$

It is assumed that the server initially issues a smart card containing a pre-shared secret $SEED$, a large random number it generates, to the client.

3.3 Login and Authentication Stages

- (1) $C \rightarrow S: id$
- (2) $C \leftarrow S: CN, SEED \oplus D, H(D) \oplus p_{t-1}$, where $CN = N - t$
- (3) $C \rightarrow S: p_t \oplus D$, where $p_t = H^{CN}(K \oplus SEED)$

After receiving the message of step (3), the server XORs the received $p_t \oplus D$ with D to obtain the fresh one-time password p_t . If the hash value of p_t is equal to p_{t-1} stored in the server, then the client is authenticated. Finally, the server updates the last password with p_t and the counter value with CN in its database. D , randomly generated by the server, can be used as a session key to enable confidential communication between the server and the client.

3.4 Weaknesses of YEH-SHEN-HWANG's Scheme

Stolen-Verifier Attack: If an attacker steals a user's $SEED$, he/she can use the pre-shared secret as a verifier to mount off-line dictionary attacks, while trying to extract the user's secret K from p_t . Thus, exposure of $SEED$ causes YEH-SHEN-HWANG's scheme to be still vulnerable to off-line dictionary attacks and preplay attacks. Such vulnerability does not allow the scheme to achieve the strength of the S/KEY scheme that no secret information need be stored on the server and defend against server compromise.

Denning-Sacco Attack: YEH-SHEN-HWANG's scheme is vulnerable to the Denning-Sacco attack based on a compromised session key D [8,9]. The compromised session key D can be used to extract SEED, which allows an attacker to mount preplay attacks or off-line dictionary attacks.

Denial of Service Attack: During the registration stage, an attacker can replace N of step (2) with \underline{N} or $p_0 \oplus D$ of step (3) with an equal-sized random number r . The values, \underline{N} and r , cause the server to extract the wrong initial password p_0' , resulting in desynchronization between the server and the client.

Inconveniences: YEH-SHEN-HWANG's scheme causes the following inconveniences to the user. First, as the scheme provides no method for the server to securely distribute SEED to the user, the user should request the administrator to directly generate SEED and store it in the smart card in the case of reinitializing the user's login information. Second, because of using the value derived from the pass-phrase as the user secret K , the user should input both the Personal Identification Number (PIN) of the smart card and the pass-phrase to authenticate himself/herself.

4 Proposed Scheme

In this section, we use a server-side public key to improve the vulnerabilities of YEH-SHEN-HWANG's scheme. The stolen-verifier attack, the Denning-Sacco attack and the inconvenience of reinitializing the user's login information result from a pre-shared secret SEED and the user's weak pass-phrase. Therefore, it is desirable to strengthen the user secret K rather than SEED. For that, our scheme uses the user secret randomly generated and stored in the user's smart card. Also, public key, instead of a pre-shared secret SEED, is used to prevent server spoofing attacks and preplay attacks. Our scheme is composed of registration stage, login stage and authentication stage.

4.1 Preliminary

It is assumed that the server initially issues a smart card containing the server's X.509 public key certificate PKC_S to the user. If PKC_S is expired or compromised, it can be updated through an out-of-band method. In addition to PKC_S , the smart card contains a randomly generated large number D and $P_S^{-1}(H(id \oplus D))$ used to authenticate the user in the registration stage. Those values are used once in the first registration of the user. To reinitialize the user's login information, the client uses the large random number D shared with the server in the login stage. Before starting the registration stage, the client randomly generates a large number K and stores it in the smart card. Unlike existing S/KEY schemes, the large number K is used as the user secret instead of the value derived from the user's pass-phrase. Thus, our scheme can protect against off-line dictionary attacks, while defending against server compromise.

As shown in Fig. 1, a smart card can be in any one of several statuses: issued, initialized, registered and ready.

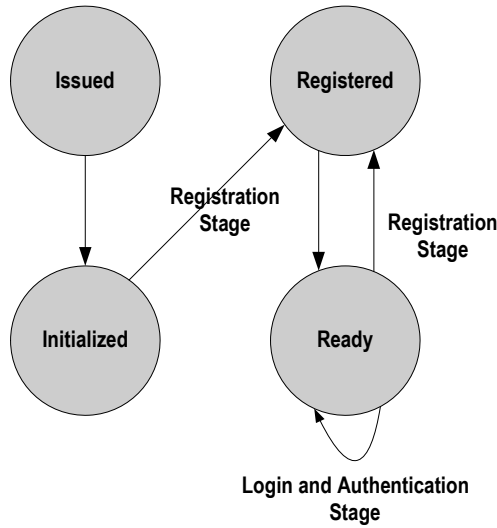


Fig. 1. Smart Card Status

Issued: In this status, the smart card is issued and then passed to its user. The smart card SC_{id} and the user information stored in the server database $User_{id}$ are as follows.

- $SC_{id} = \{ SCN_{id}, id, D, P_S^{-1}(H(id \oplus D)), PKC_S \}$
- $User_{id} = \{ id, SCN_{id} \}$

Initialized: In this status, the user randomly generates his secret key K and sets Personal Identification Number (PIN) through the smart card utility.

- $SC_{id} = \{ SCN_{id}, id, K, D, P_S^{-1}(H(id \oplus D)), PKC_S \}$
- $User_{id} = \{ id, SCN_{id} \}$

Registered: In this status, p_0 , N and SEED are initialized in both the smart card and the server database.

- $SC_{id} = \{ SCN_{id}, id, K, N, SEED, p_0, PKC_S \}$
- $User_{id} = \{ id, SCN_{id}, N, SEED, p_0 \}$

Ready: This status allows the user with the smart card to be authenticated by the server.

- $SC_{id} = \{ SCN_{id}, id, K, CN, SEED, p_t, PKC_S \}$
- $User_{id} = \{ id, SCN_{id}, CN, SEED, p_t \}$

4.2 Registration Stage

- (1) $C \rightarrow S: id, H(SCN_{id} \oplus IP_C), P_S(D), P_S^{-1}(H(id \oplus D))$
- (2) $C \leftarrow S: N, SEED, H(N \oplus SEED \oplus D)$
- (3) $C \rightarrow S: p_0 \oplus D, H(p_0 \parallel D)$

For the first registration, the client uses the D stored in the smart card to authenticate the user. To prevent denial of service attacks, the server firstly verifies the received

$H(SC_{id} \oplus IP_C)$ before the expensive public key operations for $P_S(D)$ and $P_S^{-1}(H(id \oplus D))$. Later, to reinitialize the user's login information such as SEED, N and p_0 , the client should pass both the login stage and the authentication stage to be mentioned below, and use the large number D shared with the server in the login stage. In this case, the client goes through step (2) - (3), omitting step (1).

4.3 Login and Authentication Stages

- (1) $C \rightarrow S: id, H(IP_C \oplus SCN_{id} \oplus p_{t-1}), P_S(D)$
- (2) $C \leftarrow S: CN, SEED, H(CN \oplus SEED \oplus D)$
- (3) $C \rightarrow S: p_t \oplus H(D)$

For the t th login, the client generates a large random number D, and uses the server's public key KU_S to encrypt D. To prevent denial of service attacks, the server firstly verifies $H(IP_C \oplus SC_{id} \oplus p_{t-1})$ before the expensive public key operation for $P_S(D)$. Upon receiving the message of step (3), the server compares the hash value of the received p_t with p_{t-1} stored in its database. If they are equal, then the user is authenticated. In this case, the server updates the last password with p_t and the counter value with CN in its database, and sends the accept message to the client. Later, the randomly generated D can be used as a session key to enable confidential communication between the server and the client. Furthermore, it is used to reinitialize the user's login information such as SEED, N and p_0 as mentioned in 3.2.

5 Analysis of the Proposed Scheme

In comparison to YEH-SHEN-HWANG's scheme, our scheme has an additional cost, two public key operations, which may result in performance degrade. However, with such an additional cost, the proposed scheme improves the drawbacks of YEH-SHEN-HWANG's scheme. In this section, our scheme's security is analyzed.

Server Spoofing Attack: The client can ensure that only the server with the private key which corresponds to PKC_S stored in the smart card can obtain D and make the message for step (2) of the registration or login stage. Also, the client randomly generates D in each session. Thus, attackers without the server's private key cannot mount server spoofing attacks.

Preplay Attack: Before receiving the fresh one-time password, the server should prove the possession of the private key as mentioned above. Because attackers cannot impersonate the server to the client without the server's private key, it is difficult for attackers to cheat the client for the fresh one-time password that is guaranteed to be valid at some time in the future.

Off-Line Dictionary Attack: As our scheme uses the user secret randomly generated and stored in the user's smart card, it can resist against off-line dictionary attacks. Additionally, the user is required to input just the PIN of the smart card.

Denial of Service Attack: Our scheme uses $H(SCN_{id} \oplus IP_C)$ or $H(IP_C \oplus SCN_{id} \oplus p_{t-1})$ to check the client's request before the computationally expensive public key operation. Also, as the server computes $H(N \oplus SEED \oplus D)$ in step (2) of the registration

stage, the client can ensure that N and SEED are not changed. Thus, attackers cannot change N or SEED to desynchronize the initial password p_0 between the server and the client. In addition, due to $H(p_0 \parallel D)$, attackers cannot replace p_0 with a wrong initial password p_0' during step (3) of the registration stage.

Stolen-Verifier and Denning-Sacco Attacks: By using the large random number stored in the user's smart card as the user secret, our scheme can prevent off-line dictionary attacks. Therefore, even if attackers steal the user's login information such as CN, SEED and p_t from the server, they cannot use the stolen information as a verifier to mount off-line dictionary attacks. Furthermore, attackers cannot mount Denning-Sacco attacks based on a compromised session key D . Because of not being vulnerable to the above attacks, our scheme can truly achieve the strength of the S/KEY scheme that no secret information need be stored on the server.

Man-in-the-middle and Active Attacks: Since the client can verify the server's public key through PKC_s , our scheme prevents man-in-the-middle-attacks. Also, in our scheme, the randomly generated number D can be used as a session key to enable confidential communication between the server and the client. Our scheme can thus defeat active attacks.

6 Conclusions

In this paper, we have proposed a public key based one-time password authentication scheme for secure remote access in intelligent home networks. To provide 2-factor strong authentication conveniently and cost effectively, we have adopted and enhanced YEH-SHEN-HWANG's authentication scheme. Since our scheme, unlike YEH-SHEN-HWANG's scheme, uses a server-side public key, it can authenticate the server and distribute a session key without any pre-shared secret. Furthermore, the user can remotely and securely reinitialize the user's login information such as N , p_0 and SEED. Also, to prevent off-line dictionary attacks, our scheme uses the randomly generated user secret that is stored in the user's smart card instead of the one derived from the user's pass-phrase. Therefore, our scheme is not vulnerable to stolen-verifier attacks and Denning-Sacco attacks. In addition, our scheme can provides a session key to enable confidential communication between the server and the client. It can thus defeat active attacks. From the viewpoint of the user, our scheme simplifies the registration or login process by using smart cards. The user just inputs the PIN of the smart card to authenticate himself/herself.

Most importantly, our scheme can truly achieve the strength of the S/KEY scheme that no secret information need be stored on the server.

References

1. H. Sun, "Home Networking," Mitsubishi Electric Research Laboratories, 2004, <http://www.merl.com/projects/hmnt/>
2. T.C. Yeh, H.Y. Shen and J.J. Hwang, "A Secure One-Time Password Authentication Scheme Using Smart Cards," IEICE IEICE Transaction on Communication, vol.E85-B, no.11, pp.2515-2518, Nov. 2002.

3. N. Haller, C. Metz, P. Nesser and M. Straw, "A one-time password system," RFC 2289, Feb. 1998.
4. N. Haller, "The S/KEY one-time password," RFC 1760, Feb. 1995.
5. C.J. Mitchell and L. Chen, "Comments on the S/KEY user authentication scheme," ACM Operating Systems Review, vol.30, no.4, pp.12-16, Oct. 1996.
6. S.M. Yen and K.H. Liao, "Shared Authentication Token Secure against Replay and Weak Key Attacks," Information Processing Letters, vol.62, pp.77-80, 1997.
7. I. You and K. Cho, "Comments on YEH-SHEN-HWANG's One-Time Password Authentication Scheme," IEICE Transaction on Communication, vol E88-B, no.2 pp.751-753, Feb. 2005.
8. D. Denning and G. Sacco, "Timestamps in Key Distribution Systems," Communications of the ACM, vol.24, no.8, pp.533-536, Aug. 1981.
9. S. Kim, B. Kim, S. Park and S. Yen, "Comments on Password-Based Private Key Download Protocol of NDSS'99," Electronics Letters, vol.35, no.22, pp.1937-1938, 1999.
10. W.C. Ku, C.M. Chen and H.L. Lee, "Cryptoanalysis of a Variant of Peyravian-Zunic's Password Authentication Scheme," IEICE Transaction on Communication, vol.E86-B, no.5, pp.1682-1684, May. 2003.
11. I. You and K. Cho, "A S/KEY Based Secure Authentication Protocol Using Public Key Cryptography," The KIPS Transactions: Part C, Vol. 10-C, No.6, 2003.

An Intelligent and Efficient Traitor Tracing for Ubiquitous Environments*

Deok-Gyu Lee, Seo Il Kang, and Im-Yeong Lee

Multimedia-Hall, No. 607, #646, Eupnae-ri, Shinchang-myun,
Asan-siChoongchungnam-do, Korea
{hbrhcdb, kop98, imylee}@sch.ac.kr
<http://sec-cse.sch.ac.kr/index.html>

Abstract. Broadcast encryption has been applied to transmit digital information such as multimedia, software and paid TV programs on the open networks. One of key factors in the broadcast encryption is that only previously authorized users can access the digital information. If the broadcast message is sent, first of all, the privileged users will decode the session key by using his or her personal key, which the user got previously. The user will get the digital information through this session key. As shown above, the user will obtain messages or session keys using the keys transmitted from a broadcaster, which process requires effective ways for the broadcaster to generate and distribute keys. In addition, when a user wants to withdraw or sign up, an effective process to renew a key is required. It is also necessary to chase and check users' malicious activities or attacking others. This paper presents a method called Traitor Tracing to solve all these problems. Traitor tracing can check attackers and trace them. It also utilizes a proactive scheme for each user to have effective and intelligence renewal cycle to generate keys.

1 Introduction

With the development of digital technology, various kinds of digital content have been developed. The digital information can be easily copied which has created a lot of damages so far. For example, such digital information stored in CDs or diskettes can be easily copied using CD writers or diskette drivers. The digital information can also be easily downloaded and shared on the Internet. Under these circumstances, broadcast encryption has been applied to transmit digital information such as multimedia, software and paid TV on the open networks. The public key method, one of ways to provide a key, has two kinds of keys; an encoding key of a group to encode a session key and a lot of decoding keys to decode the session key. Therefore, a server can encode the session key and each user can decode the session key using different keys.

* This research was supported by MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

One important thing for the broadcast encryption is that only previously privileged user can get the digital information. When a broadcast message is delivered, a privileged user can get the digital information using his or her personal key which the user got previously. The most important one for the broadcast encryption is the process to generate, distribute and renew keys. However, even for the broadcast encryption, illegal users can not be traced if the users use keys maliciously. The method to trace illegal users is to trace them with an hidden simple digital information. In this paper, users who are involved in or related to piracy will be called as traitors, and a technique to trace such traitors as traitor tracing. In this paper, when to trace traitors, proactive techniques are applied to a key which is provided to the users to protect illegal activities beforehand. When a user is involved in illegal activities, an effective way to trace the traitors is presented. This means a new way to trace traitors. In the suggestion method, the broadcast message is composed into block right and block cipher. When a personal key is provided after registration. The key can be transformed effectively from the attackers. In addition, it is designed to sort traitors from privileged users more effectively. After summarizing the broadcast encryption and traitor tracing, this paper will explain each step of suggestion method. Through the suggestion method, we will review the suggestion process and make a conclusion.

2 Overview of Broadcast Encryption and Traitor Tracing

2.1 Overview of Broadcast Encryption

Broadcast encryption is one of protocols to communicate between one sender and a lot of receivers. A sender who has one encoded key can broadcast the encoded message. Only privileged receivers can get a message with a key to decode the message. Broadcast encryption can be applied for many kinds of scenarios where content providers send out many kinds of encoded information and only the privileged users can decode such information. One example can be Pay-TV which has provided many kinds of broadcast encryption technique. The first suggested method for broadcast encryption presented by Cho and two other colleagues is composed into three steps.

- Initiation by a content provider: A content provider will generate necessary information for all users, which step is called as initial information.
- User information initiation: This is a step for a user to register his or her personal information to the content provider. After this step, the user's information will be stored as his or her personal key. The content provider will renew the user's initial information after the user initiation step.
- Session Transmission: Content data will be encoded as session keys. These session keys will be transmitted by divided into small parts called session. Each session will be transmitted as an encoded form into different session part. Privileged users will get session keys to get real data by decoding part of session keys using his or her personal keys.

2.2 Overview of Traitor Tracing

Traitor tracing has the purpose to find traitors who copy and distribute digital information illegally. This technology hides unknown simple digital information into the

digital information to trace traitors. This paper calls a traitor who is related to piracy such as copying and distributing digital information illegally. The technique to trace traitors is called as traitor tracing. Through the tracing of traitor, we can prohibit the piracy such as illegal copying of the information. Such activities will enable smooth flow of digital information.

How to trace traitors will be explained simply in the following part.

- A basic method to trace traitors: A basic way to trace traitors is to transmit encoded digital information to each user. A privileged user will have a decoder to decode the encoded digital information into the original digital information. After making a pirate decoder, a traitor will distribute the encoded digital information to the users who are registered to the traitor. Traitor tracing is to find these kinds of pirate decoders.

However, there are problems to find these kinds of pirate decoders. The stated method can not be applied in finding the pirate decoders in the following case. Such cases include when the pirate decoder is not used but an privileged decoder is used to decode the original digital information which is redistributed by the traitors. Or when the decoding key is open on the Internet, the method will not work.

In these cases, special information should be inserted in the digital information to trace. When digital information with a hidden code is transmitted illegally, we can find out who violates through the hidden information

- Safety of the attached information: It is possible for the digital information itself to change due to the attached information which will be used to find traitors. Only when original digital information does not change even with the additional information, the traitor will not find out the hidden information. Also, it should not be possible to correct or delete the added information by changing the attached digital information.
- Traitor Tracing: It is possible to mis-recognize a privileged user as an illegal user when you trace the traitors with the hidden digital information which would cause a lot of problems. We need to use such technique to find all of traitors but not the privileged users.

Watermarking can be used for this purpose. Digital watermarking is a technology to insert specific information into the digital information. The added information can be extracted whenever necessary without destroying the original digital information.

2.3 Proactive Secret Sharing

Proactive Secret Sharing is a new notion suggested to undertake safely the secrete sharing process for the mobile adversary which is more powerful adversary model than the existing one. In the mobile adversary model, the attacker invades the protocol periodically for each cycle. Therefore, in this protocol, during a specific cycle, if the attacker in the current cycle is not considered, more powerful attack can be done using the collected protocol from the previous cycle.

Proactive secret sharing is a notion suggested to process safe and secret sharing under the circumstances with mobile attackers. In 1995, Jarecki has suggested a solution for proactive secret sharing through periodic sharing renewal. It processes under the status to distribute sharing after secret sharing with random secret information. This is a safe process than the existing sharing by adding new sharing renewal value.

Simply summarizing the sharing renewal process, it is to make the constant of each participant's polynomial into zero in the secret sharing protocol of Pederson which has no dealers.

When a polynomial used for the secret sharing before the sharing renewal is $f_{old}(x)$, the new sharing from the share renewal result is the same as the result distributed from the polynomial $f_{new}(x)$.

$$f_{new}(x) = f_{old}(x) + \sum_{i \in QUAL} f_i(x)$$

The sum of coefficient for the newly added $f_i(x)$ can be found after finding out each new coefficient of polynomials from all participants who participated in the protocol. The attacker having the information of $f_{old}(x)$ cannot find out the information for the new polynomial $f_{new}(x)$. Therefore, the attacker cannot find out the information for the new sharing.

Jarecki's share renewal protocol will have the calculated value by all participants who acted as a dealer under the secret sharing protocol verified by Pederson. Therefore, as stated above, each participant should perform all of $O(kn)$ times of modular exponential operation. Generally, k is the value of $O(n)$, which should perform $O(n^2)$ of modular exponential operation finally. There would be a great amount of calculation. If the number of participants increases, it will consume a great deal of time to perform the share renewal protocol.

3 Proposed Scheme

The suggestion method is composed of a content provider and n users. Each user will receive one's personal keys. The personal key is required to decode the session key included in the block right. Three personal keys will be transmitted to each user. Through the synchronizing process between the users and the content provider, the broadcasted message will be decoded by each personal keys.

Let assume that all of three users are A, B, and C. The keys transmitted to A are $i-3$, $i-2$, and $i-1$. For B, they are $i-1$, i , and $i+1$ and for C, $i+1$, $i+2$, and $i+3$. There are two purposes to distribute triple keys for each user. First, the key of each user will be used interchangeably so that the attacker cannot use the same key with the users. Secondly, when the user uses the keys illegally, the triple keys will be shifted overall.

When the keys are used illegally, the keys will form an intersection set which will enable to trace the traitors correctly. The steps and assumptions of the suggestion method are summarized in the next part.

- Initial Stage: This is a step to assume the system variables. A content provider will assume the number of participants, and generate public keys and personal keys. The number of keys will be three times more than the number of users.
- Registration Stage: This is a protocol between a content provider and the users who want to receive digital information. The user will get a couple of keys from the content provider.
- Broadcast message encoding stage: This is a stage to generate encoded broadcast message. The user will synchronize a key to use with the content providers. If the transmitted broadcast message is used illegally, the decoded key included in the illegal decoder will be one of the personal keys distributed to each user from the content provider.
- Decoding Stage: Each user will get the digital information from the broadcast message through the decoding process using his or her personal key. At this point, the user can not distinguish his or her own keys. The user can simply decode using his or her keys.
- Key renewal Stage: The content provider will broadcast the related information with the key change using the value of new polynomial. Each user will change his or her personal key using the transmitted value. At this point, the user will perform activities with his or her triple keys, which will be processed automatically.
- Tracing traitors and withdrawal Stage: When a content provider finds pirate decoder, the provider will continue the key renewal process by entering the broadcast message into the decoder to find traitors. The pirate decoder cannot distinguish whether it is to renew the key or to trace traitors. Also, among the triple keys of illegal users, that user who has two keys will be definitely traitors.

3.1 Basic Ground

This paragraph will introduce polynomial interpolation, and DDHP(Decision Diffie-Hellman Problem) which are the basis in the suggestion method.

3.1.1 Polynomial Interpolation

If we assume $f(x) = \sum_{i=0}^z a_i x^i$ as a z dimensional polynomial, each user will be given the sharing information of $(x_i, f(x_i))$. From the Lagrange Interpolation, users of

$0, 1, \dots, z$ can calculate the constant a_0 of the polynomial.
$$\sum_{i=0}^z \left(f(x_i) \bullet \prod_{0 \leq j \neq i \leq z} \frac{x_j}{x_j - x_i} \right)$$

where $\lambda_i = \prod_{0 \leq j \neq i \leq z} \frac{x_j}{x_j - x_i}, 0 \leq i \leq z$ is a Lagrange coefficient. Therefore, when

$(x_0, g^{rf(x_0)}), (x_1, g^{rf(x_1)}), \dots, (x_z, g^{rf(x_z)})$ are given, $g^{ra_0} = \prod_{i=0}^z (g^{rf(x_i)})^{\lambda_i}$ can be calculated for random r .

3.1.2 Decision Diffie-Hellman Problem

Let's assume G as a group having a large decimal order q to consider the two probability distribution ensemble R and D .

$R = (g_1, g_2, u_1, u_2) \in G^4$, where g_1 and g_2 are generators. $D = (g_1, g_2, u_1, u_2)$ where g_1 and g_2 are generators of G_q . As for $r \in Z_q$, $u_1 = g_1^r$ and $u_2 = g_2^r$. It is a problem to distinguish DDHP probability distribution ensemble R and D . This is a difficult probability algorithm un-distinguishable by calculation from the distribution $|\Pr[A(R_n)=1] - \Pr[A(D_n)=1]| \geq 1/n^c$ for all of the probability polynomial time algorithm A and quite large number c .

3.2 Characteristics of Proposed Scheme

The suggested method has the following characteristics. First of all, the content provider will insert renewal factors to renew easily the user's keys to set up the system algorithm and related variables, and to keep safely the key of users from illegal users. The users will register to the content providers later. At this point, the allocated personal keys will be transmitted.

When the content provider wants to broadcast a message, only the users who has the encoded encryption for broadcasting data and privileged personal keys can acquire the session keys to constitute a block right. The user can decode using his or her personal keys to acquire messages.

At the stage to transmit and to decode the encoded message, it is required to synchronize messages to use keys between users and servers. When a content provider discovers a traitor and wants to find out a privileged user, and the privileged users decode the broadcasted data after combining his or her personal keys into un-privileged keys, the content provider can find out the used personal keys by checking the relationship between the input and output of the messages.

We assume that the maximum illegal usage by users is k , extracted critical value from traitors is z (because the number of personal keys of traitors is three, the user can be found out exactly.) and the group with the largest decimal order q is G_q .

3.2.1 The Initial Stage and the Process of Encoding and Decoding the Broadcasted Message

Step 1. The content provider selects random number (β) after predicting users $(i = 1, \dots, 2n + 2)$. The random number at this point will be used as a renewal factors to renew users' keys.

Step 2. The content provider selects the polynomial $f(x) = \sum_{i=0}^z \beta_i \alpha_i x^i$ having coefficient z on the Z_q space. The content provider will publish the public keys after making the polynomials into the secrete keys just as the following.

$$\langle g, g^{\beta_0 \alpha_0}, g^{f(1)}, \dots, g^{f(z)} \rangle \text{ Published.}$$

Step 3. When the content provider registers users, he or she will transmit personal keys $((i, f(i-1)), (i, f(i)), (i, f(i+1)))$ to the users. The user will try to verify the correctness of his or her keys. At this point, the content provider will induce to verify with the first key.

$$g^{\beta_0 \alpha_0} = \prod_{t=0}^z g^{f(x_t) \lambda_t}, x_0 = 1, x_2 = 2, \dots, x_{z-1} = z, x_z = i$$

When this equation is verified, the user will acquire personal keys.

Step 4. The content provider calculates the block right after selecting the unused information and the random number $r \in Z_q$.

$$(j_1, f(j_1)), (j_2, f(j_2)), \dots, (j_z, f(j_z)), C = \langle sg^{r\beta_0 \alpha_0}, g^{r(j_1, g^{f(j_1)})}, \dots, g^{r(j_z, g^{f(j_z)})} \rangle$$

After encoding the message into session keys, C and encoded message are transmitted.

Step 5. The process to acquire session keys from the C and encoded message from content provider follows below.

$$s = sg^{r\beta_0 \alpha_0} / \left[(g^r)^{f(i) \lambda_z} \cdot \prod_{t=0}^{z-1} (g^{f(x_t)}) \right], x_0 = j_1, x_2 = j_2, \dots, x_{z-1} = j_z, x_z = j_i$$

3.2.2 The Key Renewal Stage

Step 1. A user j requests his or her withdrawal to the content provider.

Step 2. The content provider deletes the renewal factor of a user j from the renewal factor β to renew the existing users' personal key.

Step 3. The content provider will transmit to the user the renewed personal keys after deleting the renewal factors of the withdrawn users. Because three pairs of keys are transmitted for each user, a user j can not use the enabling block to utilize the unused sharing information after fixing the sharing information of the three $B = (\beta_{i-1}, \beta_i, \beta_{i+1})$ related to the renewal.

$$(\beta_{i-1}, g^{r\beta_{i-1}}), \dots, (\beta_{i+1}, g^{r\beta_{i+1}}) \Rightarrow (j_1, g^{r\beta_{i-1}}), \dots, (j_{z-3}, g^{r\beta_{i+1}})$$

3.2.3 The Conspirator Tracing Stage

The conspirator tracing will apply the suggested method by WGT in the key renewal process. WGT suggested two kinds of conspirator tracing method. We will explain this method with the suggested method.

Step 1. The content provider will constitute the user's set $\{c_1, c_2, \dots, c_m\}$, $(m \leq k)$ assuming as a traitor.

Step 2. To trace a traitor, the information will be added like the following block right.

$$\langle sg^{r\beta_0 \alpha_0}, g^{r(c_1, g^{f(c_1)})}, \dots, g^{r(c_m, g^{f(c_m)})} \rangle$$

The other method is a way only for the pirate users to decode the block right. The content provider will select a new polynomial $h(x)$ which does not match with $f(x)$ with other solutions, while $\{(c_1, f(c_1)), \dots, (c_m, f(c_m))\}$ are part of solutions.

After a traitor is discovered, the content provider can make illegal personal keys not to decode the broadcasted data with the keys of traitors. If we assume $\{c_1, c_2, \dots, c_m\}$, ($m \leq k$) as a discovered conspiracy by the content provider, the sharing information can be extracted while not changing the personal keys of other users. The content provider will fix the first m of sharing information as block right with $(c_1, g^{rf(c_1)}), \dots, (c_m, g^{rf(c_m)})$. The other $z - m$ unused sharing information of $(j_1, g^{rf(j_1)}), \dots, (j_m, g^{rf(j_m)})$ will constitute the block right.

4 Analysis of Suggested Scheme

This paper suggests to provide triple of keys to the users to trace users with effective methods for key generation and encoding of broadcast encryption than the existing method. The safety of the suggested method is based on the problems of discrete logarithm. Compared to the existing method, this suggested method provides increased effectiveness for the participation of users, key renewal, and the withdrawal of users. This chapter will review the suggestion method.

4.1 Key Renewal

The existing KPS(Key Predistribution Scheme) will transmit message by encoding with generated and shared keys. After confirming the transmitted message, the key will be newly generated after one session. As for the key, after the attack, the key will not be renewed but will be generated again.

However, in the suggested methods, the existing users can renew and use the keys after signing up or withdrawing. The key renewal will insert β factor, the key renewal factor at the initial key generation stage. When a user withdraws one's membership his or her own will or by force, the server will provide after deleting the β factor, the key renewal information of the withdrawn user. The user will finish the key renewal process with a simple calculation.

4.2 Recalculation of the Initial Estimation Error

As for the suggested method, the server should set up and manage the system. If the server manages mobile users, the estimation of the users should be correctly measured. Therefore, if there is any error for the initial estimation, recalculation or additional calculation should be performed. However, for the existing method, there was not any ways to calculate the incorrect estimation about the users.

This paper suggests the estimation of calculation of the users to set up a system through the calculation with r . Also, as for the randomized number r , the number will be generated on the space Z_q and a problem will be solved to make more users than the previously estimated users.

4.3 A Safe Scheme Against the Selected Encoded Message Attack

It is very good character to design a safe encoded system against the adaptive chosen encoded message attack. In this chapter, by transforming the suggested method of Boneh and Franklin, we suggest a safe tracing scheme against adaptive chosen encoding message attack from traitors.

The content provider will select $a, b, x_1, x_2, y_1, y_2 \in Z_q$ and a polynomial $f(x) = \sum_{i=0}^z \beta_i \alpha_i x^i$ having z on the space Z_q . The secrete key of a content provider is $\langle f(x), a, b \rangle$. The public key $\langle g, g^{\beta_0 \alpha_0}, g^{f(1)}, \dots, g^{f(x)}, g^{r\beta_a}, g^{r\beta_b}, c, d, H \rangle$ will be provided for all users. The content provider will provide personal keys to the users when the content provider registers a user. The user will confirm the correctness of the received value following the stated method in the previous chapter.

The content provider will select randomly the unused sharing information of z to calculate the block right based on the information. The user will acquire the session key using his or her personal keys from the broadcasted block right. After verification, we can confirm the safety from the selected encoded attack.

5 Conclusion

Broadcast encryption has been applied for various kinds of digital information such as multimedia, software and paid TV programs on the open internet. The important one in the broadcast encryption is that only the privileged users can get the digital information. When the broadcast message is transmitted, the privileged users can get the digital information using the previously received personal keys. As stated above, the user can utilize the transmitted key from the broadcaster to acquire message and session keys. In this stage, we need a process to generate and share keys.

In addition, effective key renewal will be required to withdraw or sign up. Only the privileged users can receive the information for the broadcasted message. The privileged users can obtain the session keys using previously transmitted personal keys. This paper applied a proactive method to provide keys to users to protect users' illegal activities beforehand when tracing traitors. Based on this scheme, when a user commit an illegal activity, the traitor can be traced effectively.

This means a new way to trace traitor. The broadcasted message is composed of block right, block renewal, and block code. In addition, because the user will receive a personal key during registration, it is designed more effectively to detect traitors from the original users to trace traitors after transforming the keys effectively from the attackers.

References

1. Amos Fiat and Moni Naor, "Broadcast Encryption", *Crypto'93*, pp. 480-491, 1993
2. A. Narayana, "Practical Pay TV Schemes", *to appear in the Proceedings of ACISP03*, July, 2003

3. C. Blundo, Luiz A. Frota Mattos and D.R. Stinson, "Generalized Beimel-Chor schemes for Broadcast Encryption and Interactive Key Distribution", *Theoretical Computer Science*, vol. 200, pp. 313-334, 1998.
4. Carlo Blundo, Luiz A. Frota Mattos and Douglas R. Stinson, "Trade-offs Between Communication and Storage in Unconditionally Secure Schemes for Broadcast Encryption and Interactive Key Distribution", *In Advances in Cryptology - Crypro '96, Lecture Notes in Computer Science 1109*, pp. 387-400.
5. Carlo Blundo and A. Cresti, "Space Requirements for Broadcast Encryption", *EUROCRYPT 94, LNCS 950*, pp. 287-298, 1994
6. Donald Beaver and Nicol So, "Global, Unpredictable Bit Generation Without Broadcast", *EUROCRYPT 93*, volume 765 of Lecture Notes in Computer Science, pp. 424-434. Springer-Verlag, 1994, 23-27 May 1993.
7. Dong Hun Lee, Hyun Jung Kim and Jong In Lim, "Efficient Public-Key Traitor Tracing in Provably Secure Broadcast Encryption with Unlimited Revocation Capability", *KoreaCrypto 02'*, 2003
8. D. Boneh and M. Franklin, "AN Efficient Public Key Traitor Tracing Scheme", *CRYPTO 99, LNCS 1666*, pp. 338-353, 1999
9. Dani Halevy and Adi Shamir, "The LSD Broadcast Encryption Scheme", *Crypto '02, Lecture Notes in Computer Science*, vol. 2442, pp. 47-60, 2002.
10. Ignacio Gracia, Sebastia Martin and Carles Padro, "Improving the Trade-off Between Storage and Communication in Broadcast Encryption Schemes", 2001
11. Juan A. Garay, Jessica Staddon and Avishai Wool, "Long-Lived Broadcast Encryption", *In Crypto 2000, volume 1880 of Springer Lecture Notes in Computer Science*, pages 333--352, 2000.
12. Michel Abdalla, Yucal Shavitt, and Avishai Wool, "Towards Marking Broadcast Encryption Practical", *IEEE/ACM Transactions on Networking*, 8(4):443--454, August 2000.
13. Yevgeniy Dodis and Nelly Fazio, "Public Key Broadcast Encryption for Stateless Receivers", *ACM Workshop on Digital Rights Management*, 2002
14. R. Ostrovsky and M. youg, "How to withstand mobile virus attacks", *in Proc. 10th ACM symp. on principles of Distributed Computation*. pp. 51-61, 1991
15. H.K.A. Herzberg, S. Jarecki and M. Yung, "Proactive Secret Sharing or: How to Cope With Perpetual Leakage", *Crypto95, LNCS*. 1995

e-Business Agent Oriented Component Based Development for Business Intelligence

Ho-Jun Shin and Bo-Yeon Shim

CSPI, Inc. #201, Hyunsan B/D, 108-7, Yangjae-Dong, Seocho-Ku, Seoul, 137-891,
Rep. of Korea
{hjshin, byshim}@cspi.co.kr

Abstract. Agent technology becomes more and more importance in Business Intelligence domain. The concepts and technology have been brought to a stage where they are useable in real applications, and there is a growing understanding of how to apply them to practical problems. Component methodologies have proved to be successful in increasing speed to market of software development projects, lowering the development cost and providing better quality.

In this paper, we propose systemical development process using component and UML(Unified Modeling Language) technology to analysis, design and develop e-business agent for business intelligence. The ebA-CBD(e-business Agent-Component Based Development) process is an attempt to consider all of the best features of existing AOSE(Agent Oriented Software Engineering) methodologies while grounding agent-oriented concepts in the same underlying semantic framework used by UML, the standard modeling language for Object Oriented Software Engineering. Finally we describe how these concepts may assist in increasing the efficiency and reusability in business intelligence application and e-business agent development in business intelligence environment.

Keywords: e-Business Agent, Business Intelligence, ebA-CBD Reference Architecture, ebA-Spec, Component Based Development.

1 Introduction

Recently the software lifecycle is getting shorter and web service for paradigm of next generation information technology is more focused on while e-business model has developed very rapidly. Therefore, the development of software is more functional, various, stable software, the key of business intelligence domain. According to these requirements, not only the component having exchangeable module that performs independent business and function in software system but also the utilization of agent in e-business domain become more notice. It is important to produce the agent service based on component technology that is change to replacement and portability toward developing the software having high productability[1][2].

In this paper, we propose ebA-CBD process for business intelligence. This proposed process applies ebA model notation and role model, goal model, architecture model and interaction model in the view of agent. Simultaneously, these 4 models

considered as ebA model. The ebA Model can define agent characteristics and the relations among agents. At the same time, component is possibly constructed on ebA specification. In addition the process is presented though of e business agent models the case study of component information search agent.

2 Related Works

2.1 Agent Concept Model

An agent is an atomic autonomous entity that is capable of performing some useful function. The functional capability is captured as the agent's services. A service is the knowledge level analogue of an object's operation. The quality of autonomy means that an agent's actions are not solely dictated by external events or interactions, but also by its own motivation. We capture this motivation in an attribute named purpose. The purpose will, for example, influence whether an agent agrees to a request to perform a service and also the way it provides the service. Software Agent and Human Agent are specialization of agent[3].

Figure 1 gives an informal agent-centric overview of how these concepts are inter-related. The role concept allows the part played by an agent to be separated logically from the identity of the agent itself. The distinction between role and agent is analogous to that between interface and class: a role describes the external characteristics of an agent in a particular context. An agent may be capable of playing several roles, and multiple agents may be able to play the same role. Roles can also be used as indirect references to agents. This is useful in defining re-usable patterns. Resource is used to represent non-autonomous entities such as databases or external programs used by agents. Standard object-oriented concepts are adequate for modeling resources[4].

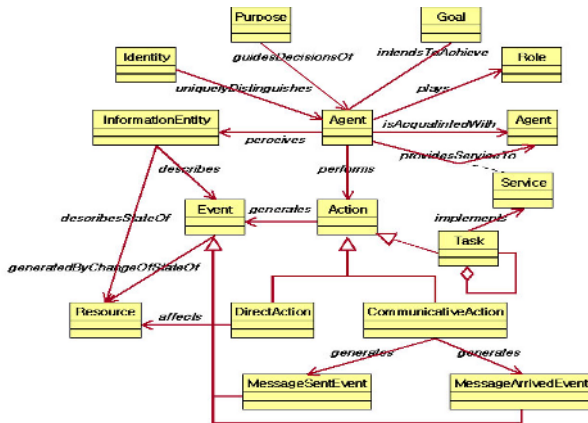


Fig. 1. Agent Concept Model

2.2 Business Intelligence

Despite the technology being available since the late seventies, it is only in the last decade that business intelligence has moved into mainstream business conscience, helped in part by the increasingly competitive business environment.

The focus in BI has traditionally been on analysis of cleaned historical data in a data warehouse that is updated periodically from operational sources e.g. daily, weekly or monthly. The data is largely static and updating is usually scheduled for low use periods. The data warehouse is quarantined from the operational databases.

The increased use of the web and interactive business has driven the need for web site analysis, including click-stream analysis and real time response, which has led to increased interest in real-time business intelligence. Given that traditional BI has usually relied on batch updating, at best daily, "real-time" has tended to be considered as anything better than that. The Data Warehousing Institute notes "it is more practical to focus on "right time" than "real time"" The terminology is further confused by the concept of Business Activity Monitoring (BAM) which refers to the automated monitoring of business-related activity affecting an enterprise. Some vendors refer to BAM as Real-Time BI while others separate the issues. BAM applications monitor and report on activities in the current operational cycle[5][6].

3 ebA-CBD Process

As we suggested ebA-CBD reference architecture in previous research[7], component development process based architecture is a set of activities and associated results, which lead to the production of a component as shown in figure 2. These may involve the development of component from ebA specification by using UML model. Here, our main concern is the specification workflow.

In addition, we consider systemical development process using AUML and ebA model technology to analyze, design, and develop e-business agent. The domain analysis specification, design model, implemented component, which are produced though the process, are stored in the repository[8].

3.1 ebA Requirements Identification Phase

The requirement of agent should be first identified in desired business system. The primary property of agent is able to analyze after that the description for specific agent platform and the sorts of essential properties should be understood. At the same time, it is very important to consider weather the requirement, which is already defined, is corresponding to agent type in reference architecture and what business concept is focused on. For the e-business domain analysis, UML approach is used. The Diagrams used in problem domain analysis are use case diagram. Use case diagram is a diagram that shows a set of use cases and actors and their relationships. It supports the behavior of a system by modeling static aspects of a system.

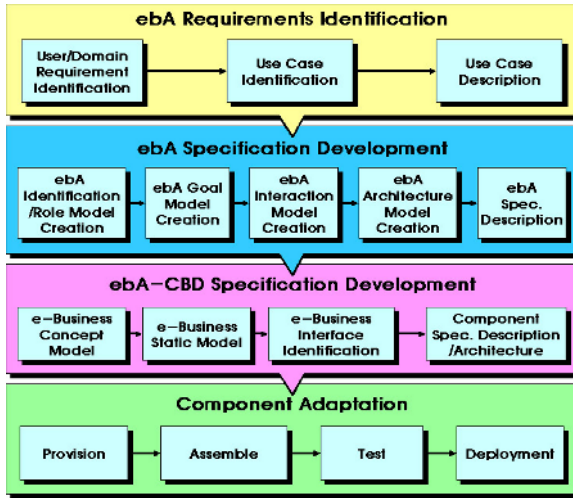


Fig. 2. ebA-CBD process

Also, domain analysis is presented on entire domain concept and scenario using activity diagram. Requirement analysis is defined through use case diagram, and use case description.

3.2 ebA Specification Development

Agent specification based on user’s requirement creates ebA Specification and 4 models. These products that are acquired though the specification phase, becomes the main resource to identify and development new components. As mentioned, in order to provide a further degree of expressiveness, ebA Model extends the UML meta-model by adding a number of elements to it. This section describes the notation that ebA Model represents graphically the instances of these new meta-elements in the diagrams.

Figures 3 provide a summary of the symbols representing the ebA Model concepts and relations respectively. It is notice that symbols are associated to elements natively included in the UML meta-model, which doesn’t mention here.

The usages of relationships are as follows:

- Implication : This relation links one or more elements that have an attribute of type state to a single element that has an attribute of type state.
- Assignment : This relation links an element of type AutonomousEntity to an element that has an attribute of type AutonomousEntity. The semantics are such that the assignment from one AutonomousEntity to another following the direction of the arrow.
- Data flow : This relation links a DataProsumer to an InformationEntity that is produced or consumed. This is the same relation as the ObjectFlow relation defined in UML and therefore the same symbol is used here.

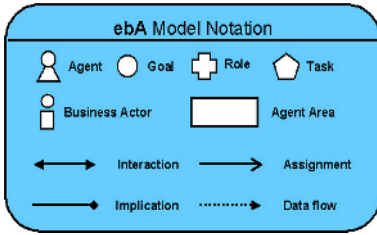


Fig. 3. ebA Model Notation

E-Business Agent Agent Type	System-Level Agent(00)			General Business Activity Agent(10)				Personal Agent (20)	System Level Agent (30)	Security Agent (40)
	Information Agent (01)	Transaction Agent (02)	Workflow Agent (03)	Marketing Agent (11)	Legal Agent (12)	Financial Agent (13)	Negotiation/Coordinating Agent (14)			
Software Agent (SWA)										
Autonomous Agent (AIA)										
Interactive Agent (IA)										
Adaptive Agent (ADA)										
Mobile Agent (MA)										
Coordinative Agent (CA)										
Intelligent Agent (ITA)										
Wrapper Agent (WA)										
Middle Agent (MA)										
Interface Agent (IFA)										
Information Agent (IA)										
Smart Agent (SA)										
Hybrid Agent (HA)										
Heterogeneous Agent (HGA)										

Fig. 4. ebA-CBD reference metrics

3.2.1 ebA Identification and Role Model Creation

This focuses on the individual Agents and Roles. For each agent/role it uses scheme supported by diagrams to its characteristics such as what goals it is responsible for, what events it needs to sense, what resources it controls, what tasks it informs how to perform, 'behavior rules', etc.

Agent identification uses ebA-CBD reference metrics as figure 4. Agent naming is referenced from use case description and role model is created using ebA model notation. Also, ebA-CBD reference metrics used to give a classification code.

3.2.2 ebA Goal Model Creation

This shows Goals, Tasks, States and the dependencies among them. Goals and Tasks are both associated with States, so that they can be linked by logical dependencies to form graphs that show e.g. that achieving a set of sub-goals implies that a higher level Goal is achieved, and how Tasks can be performed to achieve Goals. Graphs showing temporal dependencies can also be drawn, and we have found UML Activity Diagram notation useful here.

3.2.3 ebA Interaction Model Creation

This model highlights which, why and when agents/roles need to communicate leaving all the details about how the communication takes place to the design process. The interaction model is typically refined through several iterations as long as new interactions are discovered. It can be conveniently expressed by means of a number of interaction diagrams. This model is interaction centric and shows the initiator, the responders, the motivator of an interaction plus other optional information such as the trigger condition and the information achieved and supplied by each participant.

3.2.4 ebA Architecture Model Creation

This model shows agents relationship to negotiate and coordinate in agent area. It considers the business actor and domain concept. An agent area is where software agents meet and interact in the target architecture. The agent areas can be distributed on different hosts, and facilitate means for efficient inter-agent communication.

There are two main types of agent area. One is the area where the agents advertise their capabilities, communicate with other agents. The other is the user-client where the user interacts with agents.

3.2.5 ebA Specification Description

The ebA specification description is based previous models as role model, goal model, interaction model and architecture model. ebA specification is shown functional and non-functional elements–agent name, e-business type, general agent type, identification code, access information, Produce information, related agent, information model and operation model. It presents how to make specification in Example of ebA-Specification. The functional elements are described to use class diagram and sequence diagram.

3.3 ebA-CBD Specification Development

We have attempted summarize the process tasks into the four stages: e-business concept model, e-business static model, e-business interface identification and component spec description. The specification development takes as its input from requirements a use case model, ebA models and an ebA-spec. It also uses information about existing software assets, such as legacy systems, packages, and databases, and technical constraints, such as use of particular architectures or tools. It generates a set of component specifications and component architecture. The component specifications include the interface specifications they support or depend on, and the component architecture shows how the components interact with each other.

The identified information based on component users and performance must be provided in specification form for integration. Also, this information can be provided and acquired by producer, consumer and agent in interoperating system. The information of component design and development, and also functional and non-functional information must be provided by producer, and agent must provide the commercial information with this. This information is the important standard for choice and the ground for reuse to acquire the component. Figure 5 shows this information.

Item	Description
Category	Component family of Business domain
Component Diagram	Relationship between component
Component Name	Identified component name
Classification Code	Classification code of component based on ABCD Architecture
Short Description	Describe about component function, motive, constraint, etc.
Glossary	Describe concept of glossary related component specification
Component Context Diagram	Main function of Component
Component Interaction Diagram	Relationship between component
Component Sequence Diagram	Operational sequence of component
Component Diagram	Represent of required and provide interface
Component State Diagram	Represent of operation change
Interface Description	Pre/Post condition, input/output result
Usage Scenario	Scenario for component usage
Quality Attribute	Non-functional(Quality) attribute

Fig. 5. Component specification

3.4 Component Adaptation

These outputs are used in the provisioning phase to determine what components to build or buy, in the assembly phase as an input to test scripts.

The provisioning phase ensures that the necessary components are made available, either by building them from scratch, buying them from a third party, or reusing, integrating, mining, or otherwise modifying an existing component or other software. The provisioning phase also includes unit testing the component prior to assembly.

The assembly phase takes all the components and puts them together with existing software assets and a suitable user interface to form an application that meets the business need. The application is passed to the test phase for system and user acceptance testing.

4 Example of ebA-Specification

In this thesis, we focused on ebA specification so, case study only mentioned ebA requirements identification and specification development. The example is component information search agent. The agent finds the information of the component to be registered newly.

The requirement identification simply presented by use case diagram as figure 6. The actors are user and agent that is agent family. Also, the ebA candidate identify based on use case diagram and using ebA-CBD reference metrics. Figure 6 presented identified ebA as user, search and collection agent.

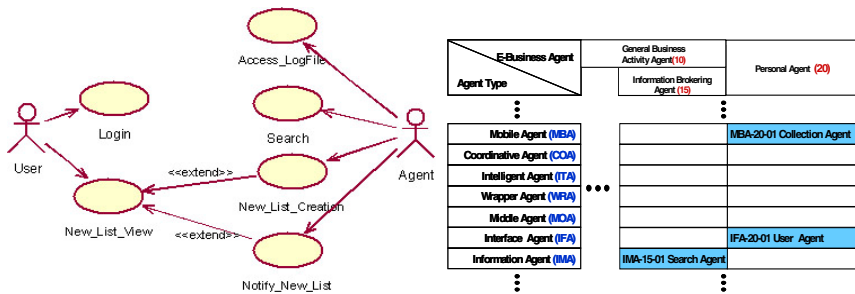


Fig. 6. Use Case Diagram and ebA-CBD Metrics of Component Information Search Agent

Figure 7 represents overall role of component information search agent. The roles describe the external characteristics of identified agent. Also, the goal model in Figure 7 shows the main goal of the component information search agent. The CreateNewList goal is achieved when lower goal successfully completed.

The following figure 8 shows as an example the interaction model describing the InformationRequest interaction between the Component Information Gather and Component Information Assistant roles. The architecture model represents overall system structure of Component Information Search Agent. Search Agent finds the information of the component to be registered newly, Collection Agent periodically update and gather to component information list and User Agent provide to alert service and manage log file.

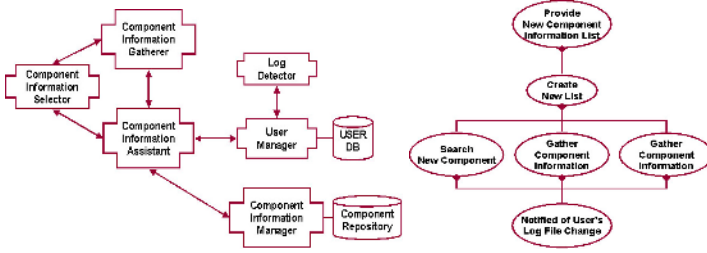


Fig. 7. Role Model and Goal Model of Component Information Search Agent

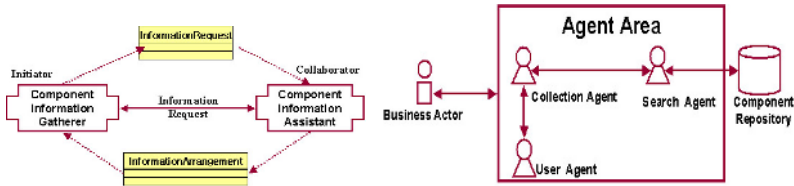


Fig. 8. Interaction Model and Architecture Model of Component Information Search Agent

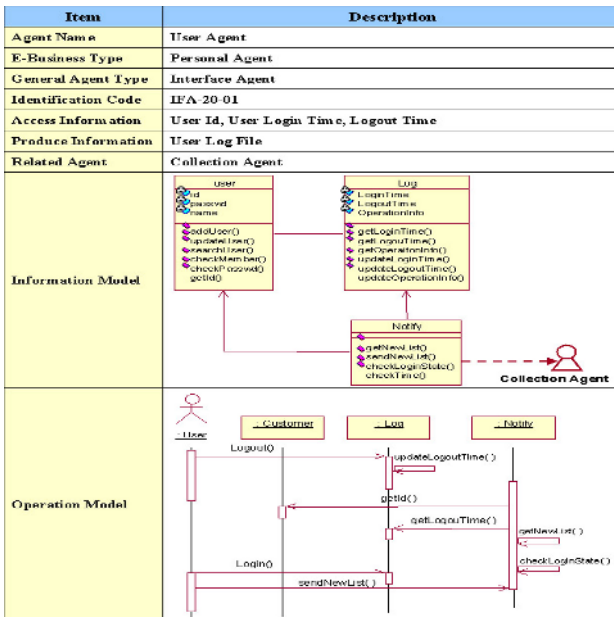


Fig. 9. ebA-Spec. of User Agent

Figure 9 present User Agent specifications consist of ebA resource information and two models. The resource information is basic elements such as name, e-business type and general agent type according to ebA-CBD reference architecture and classification code. Information and operation model provide inter/external structure and inter-operation of agents. These referred ebA-CBD specification development phase.

5 Conclusion

In this paper, each characteristic in the view of agent and component are defined and ebA-CBD process is proposed to develop e-business agent for business intelligence. Defining ebA specification of whole entire agent has e-business agent information more systemical and intuitional then also includes more information. Introducing the systemical process and ebA-CBD reference model of e-business agent can provide the efficiency by the component easily. Also, the specification can be the guideline to choose desired component and be reused as based more for component creation.

For the further works, the definition and detail methods should be required based on ebA specification for component development and assemble. Furthermore, the comparison and verification are needed though the cases study of implementation.

References

1. Martin L. Griss and Gilda Pour, "Accelerating Development with Agent Components", IEEE Computer, pp. 37-43, May. 2001.
2. Hideki Hara, Shigeru Fujita and Kenji Sugawara, "Reusable Software Components based on an Agent Model", Proceedings of the 7th International Conference on Parallel and Distributed Systems Workshops, 2000.
3. OMG Agent PSIG, "Agent Technology Green Paper", <http://www.objs.com/agent/>, 2000.
4. EURESOMP, "MESSAGE: Methodology for Engineering Systems of Software Agents", EURESCOMP Project P907 Publication, 2000.
5. Williams, S. and Williams, N., "The Business Value of Business Intelligence", Business Intelligence Journal, Vol. 8, No. 4, 2004.
6. Andreas Seufert and Josef Schiefer, "Enhanced Business Intelligence - Supporting Business Processes with Real-Time Business Analytics", Proceedings of the 16th International Workshop on Database and Expert Systems Applications, pp. 919-925, Aug. 2005.
7. Ho-Jun Shin, Haeng-Kon Kim, "CBD Reference Architecture through E-Business Agent Classification", In Proceedings 2nd International Conference on Computer and Information Science, Aug. 2002, pp. 653-658.
8. H.K. Kim, "Component Repository and Configuration Management System", ETRI Final Research Report, 2000.
9. Golfarelli, S. Rizzi and I. Cella, "Beyond data warehousing: What's next in business intelligence?", Proceedings of the 7th International Workshop on Data Warehousing and OLAP, ACM Press, 2004.
10. Martin L. Griss, "Agent-Mediated E-Commerce Agents, Components, Services, Workflow, UML, Java, XML and Games...", In Proceedings the Technology of Object-Oriented Languages and System, Keynote Presentation, 2000.

New Design of PMU for Real-Time Security Monitoring and Control of Wide Area Intelligent System

Hak-Man Kim¹, Jin-Hong Jeon², Myong-Chul Shin^{3,*}, and Tae-Kyoo Oh⁴

¹ Korea Electrotechnology Research Institute
Fusion Technology Research Lab.
Uiwang-city, Gyeonggi, S. Korea
hmkim@keri.re.kr

² Korea Electrotechnology Research Institute
Electric Power Research Lab.
Changwon-city, Gyeongnam, S. Korea
jhjeon@keri.re.kr

³ Sungkyunkwan University
Dept. of Information & Communication Eng.
Suwon-city, Gyeonggi-do, S. Korea
mcsin@skku.edu

⁴ Korea Electrotechnology Research Institute
Uiwang-city, Gyeonggi-do, S. Korea
tkoh@keri.re.kr

Abstract. Security monitoring and control are very important for stable operation in wide area intelligent system. Electrical power grid is one of a wide area system and is progressing to intelligent system. For real-time security monitoring and control of electrical power grid, measurement of control variables such as voltage, current, real power, reactive power, power factor and system frequency must be synchronized because sub-control sites are distributed widely. Global Positioning System (GPS) is used for synchronization. In this paper, we introduce the new design concept of Phasor Measurement Unit (PMU) for security monitoring and control of electrical power grid, which is one of the wide area intelligent systems. PMU is applied d-q transformation and Phase Locked Loop (PLL) to improve the robustness of measurement. PMU is tested on three-phase 380V distribution line with resistor load.

1 Introduction

In recent decades, electrical power grid has become more and more intelligent system. While intelligent system leads to greater efficiency and reliability, it also brings new sources of vulnerability through the increasing complexity and external threats. Due to these environments, security monitoring and control are very important for stable control in electrical power grid.

* Corresponding author.

For accurate real-time security monitoring and control of electrical power grid, control variables such as voltage, current, real power, reactive power, power factor and system frequency should be measured by synchronization. Global Positioning System (GPS) is designed primarily for navigational purposes, but it furnishes a common-access timing pulse, which is accurate to within 1 microsecond at any location on earth. The system uses transmissions from a constellation of satellites in non-stationary orbits at about 10,000 miles above the earth's surface. For accurate acquisition of the timing pulse, only one of the satellites need be visible to the antenna. The antenna is small and can be easily mounted on the roof of a substation control house. The experience with the availability and dependability of the GPS satellite transmissions has been exceptionally good [1]. Phase measurement units using synchronization signals from the GPS satellite system have been developing. In general, Zero-crossing and DFT are used to extract phasor information [2, 3].

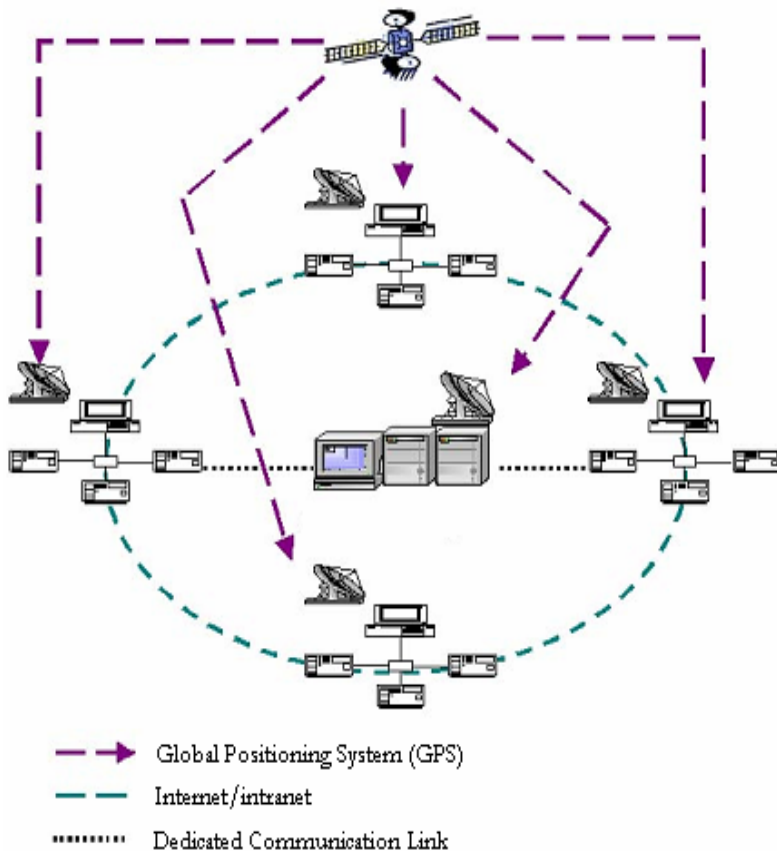


Fig. 1. Schematic structure for wide area monitoring and control in wide area system

In this paper, we suggest new concept of Phasor Measurement Unit (PMU) which is applied d-q transformation and Phase Locked Loop (PLL) to improve the robustness of measurement. The suggested PMU is tested on three-phase 380V distribution line with resistor load in laboratory and by simulation.

2 New PMU design

Fig 1 shows a schematic structure for wide area monitoring and control in electrical power grid. Since electrical power grid is wide area system, sub-control sites are distributed widely. Without accurate time reference, each measurement has trivial time difference. For accurate monitoring and control, phasor measurement using synchronization signals from the GPS satellite system is needed.

PMU is designed by IEEE Std. 1344-1995(IEEE Standard for Synchrophasors for Power System). PMU specifications are summarized at Table 1. To obtain measured data from wide area, PMU is received 1pps reference synchronized signal from GPS.

Table 1. PMU specification

Items	Specification
Analog input channels	12
Analog signal conditioning	2 nd low pass filter
A/D converter	1.25MHz, 12Bit
DSP	TMS320C32-60
RAM	128k*8bit
GPT	Motorola UT model
Delay time [4,5]	1μsec

Fig. 2 shows interrupt timing diagram for synchronized data sampling. Using interrupt timing diagram, total timing errors are calculated as follows.

- GPS receiver error: 160 nsec
- 1pps interrupt operation time: 495 nsec

Maximum and minimum interrupt delay is as follows.

- Maximum interrupt delay: $160 + 495 = 655$ nsec
- Minimum interrupt delay: $160 - 495 = 335$ nsec

Upper results are satisfied delay time of Table 1.

In general, Zero-crossing and DFT are used to extract phasor information [2, 3]. In this study, dq-transformation and Phase Locked Loop (PLL) are used to extract phasor information from measured data. Fig. 3 shows phasor measurement algorithm by dq-transformation and PLL.

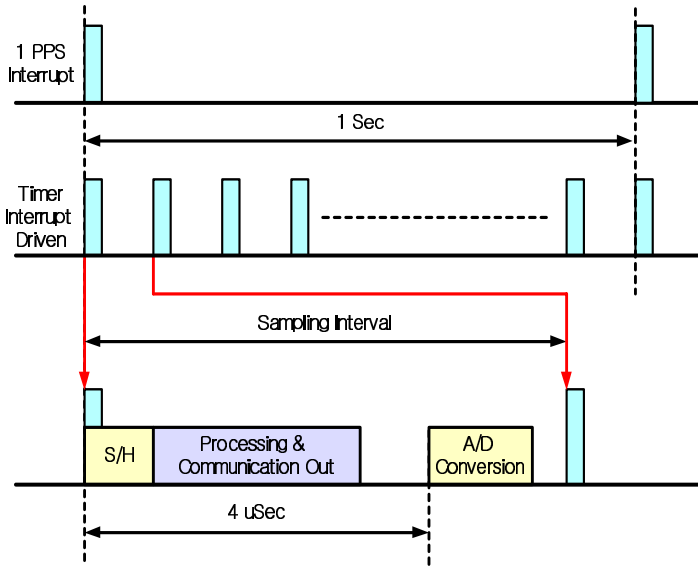


Fig. 2. Interrupt timing diagram for synchronized data sampling

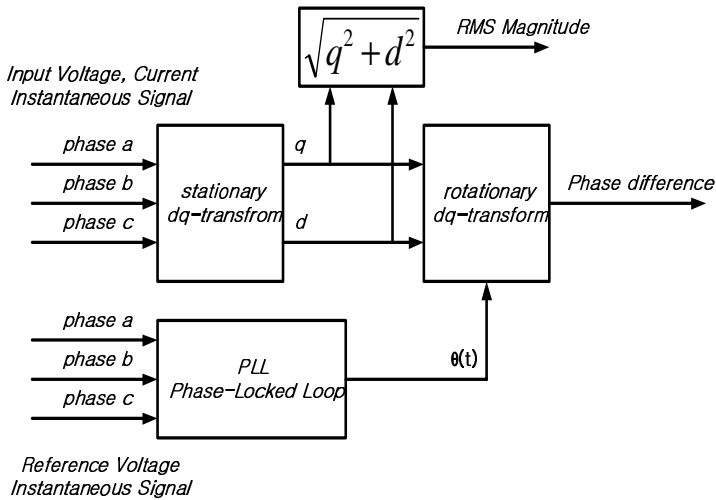


Fig. 3. Phasor measurement algorithm by dq-transformation and PLL

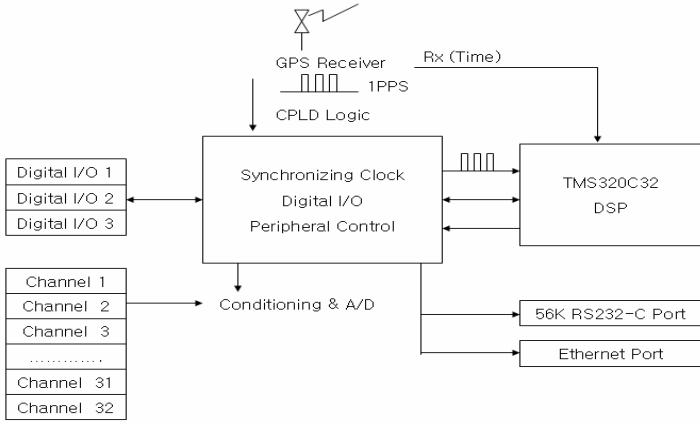


Fig. 4. PMU configuration

PMU is composed by four parts such as DSP, analog circuit, EPLD circuit and GPS module. Fig. 4 shows PMU configuration.

3 Performance Test

Performance of PMU is tested on three-phase 380V distribution line resistor load with by simulation. Test object is to measure input phasor according to frequency change, harmonics and noises. Fig. 5 and Fig. 6 show phasor measurement results of input voltage and input current, respectively.

Also, the developed prototype PMU is tested on three-phase 380V distribution line with resistor load in laboratory. Fig. 7 and Fig. 8 show phasor measurement results of input voltage and input current, respectively.

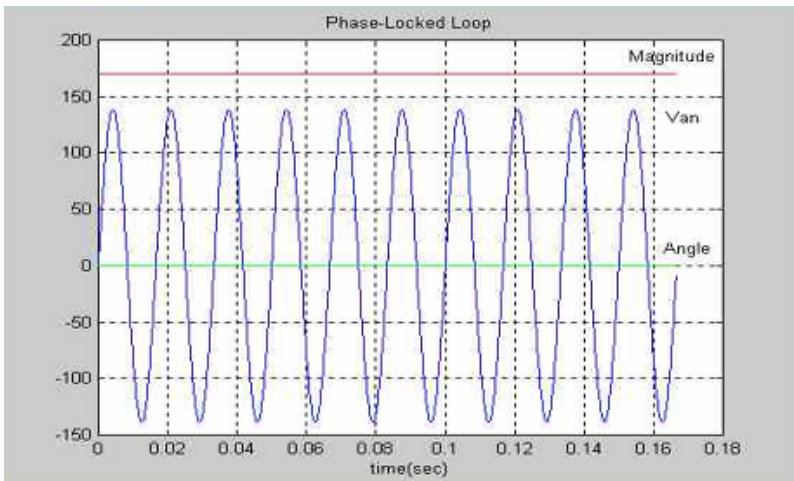


Fig. 5. Phasor measurement result of input voltage

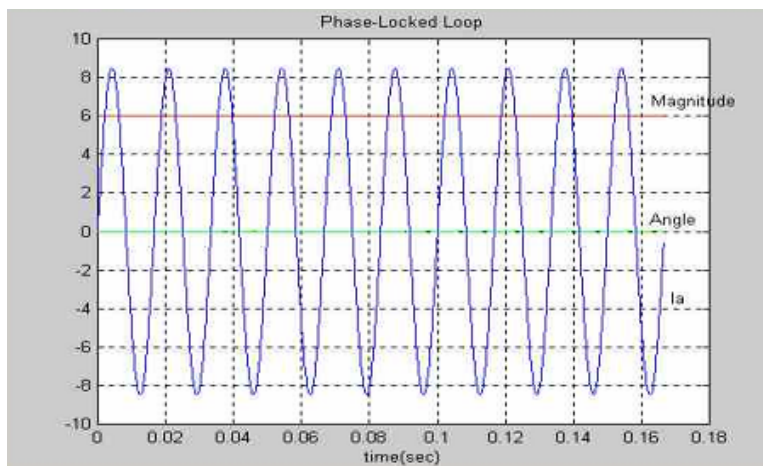


Fig. 6. Phasor measurement result of input

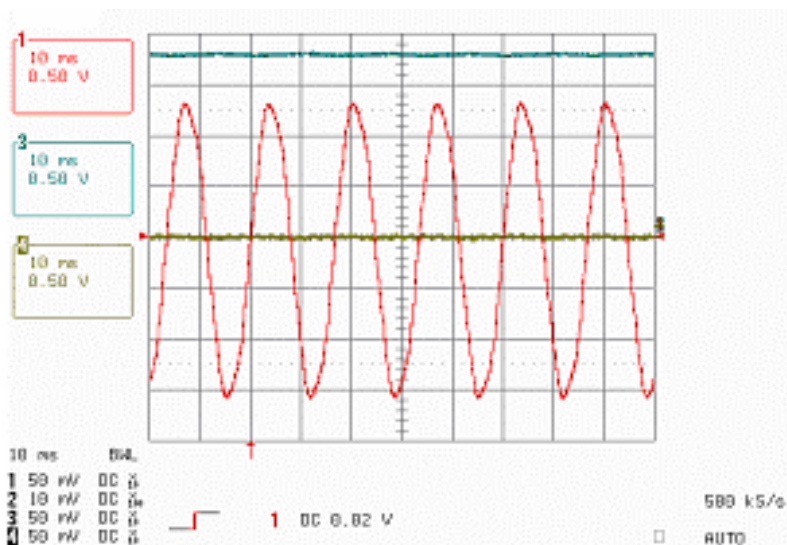


Fig. 7. Phasor measurement result of input voltage (5ms/div, 100/div, 100°/div)

Test results of prototype PMU is almost same. The trivial error is related to noise and harmonics of laboratory. Also, the results show correctness and robustness to measure voltage and current inputs with frequency change, harmonics and noises.

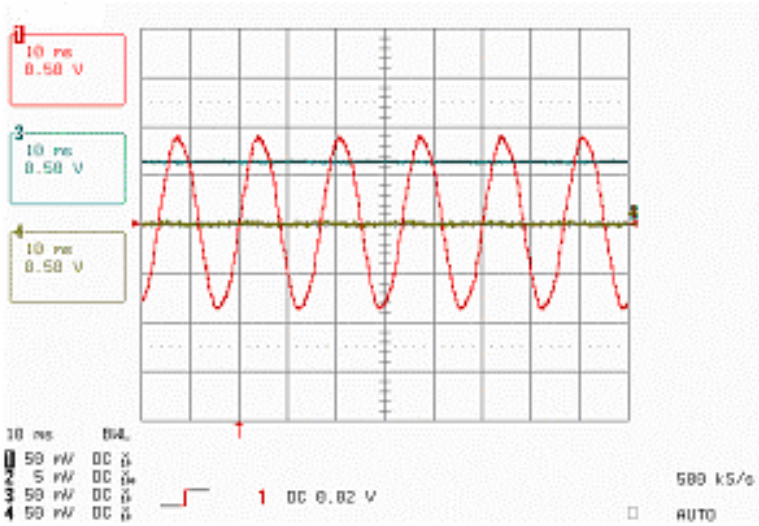


Fig. 8. Phasor measurement result of input current (5ms/div, 2.25/div, 100°/div)

4 Conclusion

We suggested new design of PMU using synchronization signals from the GPS satellite system for monitoring security and control of electrical power grid. PMU uses d-q transformation to extract phasor information. The suggested PMU is tested on three-phase 380V distribution line with resistor load in laboratory and by simulation. Test results of prototype PMU is almost same as simulation results. The trivial error is related to noise and harmonics of laboratory. Also, The test results showed that PMU could measure voltage and current inputs with frequency change, harmonics and noises with correctness and robustness because of applying d-q transformation and PLL.

References

1. A.G. Phadke, “Synchronization Phasor Measurements”, IEEE Computer Applications in Power, April (1993) 10-15
2. A.G. Phadke et al, “A New Measurement Technique for Tracking Voltage Phasor and Rate of Change of Frequency”, IEEE (1993)
3. R. Jay Murphy and R.O. Burnett, Jr. “Phasor measurement hardware and Application”, 48th Annual Georgia Tech Protective Relaying Conference, Georgia institute of Technology Atlanta, May (1994)
4. IEEE Power Engineering Society: IEEE Standard for Synchrophasors for Power Systems, the Institute of Electrical and Electronics Engineering Inc. (1996)
5. IEEE Power Engineering Society: IEEE Standard Common Format for Transient Data Exchange (COMTRADE) for Power Systems, the Institute of Electrical and Electronics Engineering Inc. (1991)

Security Intelligence: Web Contents Security System for Semantic Web*

Nam-deok Cho¹, Eun-ser Lee², and Hyun-gun Park³

¹ Chung-Ang University, 221, Huksuk-Dong, Dongjak-Gu, Seoul, Korea
ndcho@softcamp.co.kr

² Soong-Sil University, 511 Sangdo-dong, Dongjak-gu, Seoul 156-743, South Korea
eslee1@ssu.ac.kr

³ Soong-Sil University Computer Institute, 511 Sangdo-dong, Dongjak-gu, Seoul 156-743,
South Korea
gatepark007@hanmail.net

Abstract. WWW (World Wide Web) has incurred the problem that users are not necessarily provided with the information they want to receive, at a time when the amount of information is explosively increasing. Therefore, Tim Berners-Lee proposed the Semantic Web as the new web paradigm to solve this problem. But there is always a security problem in Semantic Web such as WWW and the study about this is insufficient. Therefore, the authors of this paper propose that the Security Intelligence system should be used to present semantic information using ontology and prevents that users flow out the information. This system is an ACM(Access Control Matrix) based access control model basically, and It is a system that prevent information leakage by user's deliberation and by user's mistake. It can be also used for WWW.

1 Introduction

The World Wide Web, which was introduced by Tim Berners-Lee in 1989, has brought an Internet revolution in the late 20th century because of convenience at use. But It has incurred the problem that users are not necessarily provided with the information they want to receive, at a time when the amount of information is explosively increasing. Accordingly, W3C (World Wide Web Consortium) proposes the Semantic Web[1]. The Semantic Web is defined to give clear meaning and definition to information exposed in the web, and involves extension of the web so as to enable people to work in cooperation with computers[2][3]. To show semantic information that is well defined for users according to this paradigm a study of a browser system using ontology needs to be achieved[4][5][6]. But, security problem is weighed in this Semantic Web such as World Wide Web. "exposing data hiding in documents, servers and databases "Eric Miller, Semantic Web activity lead for the W3C, said to several hundred conference participants[7] and expressed gravity of security problem. This paper propose that the Security Intelligence system should be used to present

* This work was supported by the Soongsil University Research Fund.

semantic information using ontology and prevents that users flow out the information. Security function that proposed in this paper is not a function that users can do freely about the information that provided them. If users pass the user certification successfully, Authority of users is controlled. For example, even if a user can see semantic information because of having read authority, the user can not print, if he has not print authority. This function prevents information leakage by a mistake and by user's deliberation. It is method that prevents information leakage through client program function including web browser. Because user that pass the certification successfully can flow out the information through the browser's function such as print, source view, screen capture etc, this method is needed.

Chapter 2 discusses the Semantic Web, Web Security etc., and chapter 3 describes the design and implementation of the Security Intelligence. Chapter 4 shows the results of the Security Intelligence along with its estimates, while finally chapter 5 concludes the paper.

2 Base Study

2.1 Semantic Web

Semantic Web that Tim Berners-Lee defines is as following.

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [3] Ultimate purpose of Semantic Web develop the standard and technology that help computer can understand better information in web, and support semantic search, data integration, navigation and automation of task etc. if describes in detail, it enables following work

- When search information, it bring more correct information.
- Integrates and compares information of different alloplasm source.
- Correlates meaningful and descriptive information about certain resources.
- attaches detail information on the web for automation of web service.

If see in the point that improve current web, the hierarchical structure of Semantic Web seems Fig.1

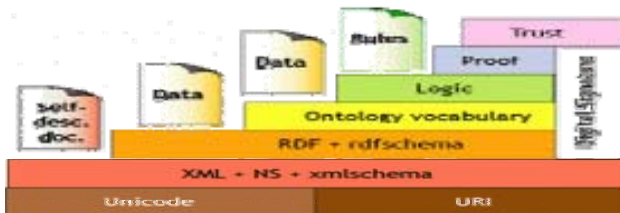


Fig. 1. Hierarchical structure of Semantic Web

Most low floor is consisted of URI(Uniform Resource Identifier) and unicode that is addressing method to access the resource in web protocol. Next floor is XML (eXtensible Markup Language) and namespace that can define concept as modularity. RDF (Resource Description Framework) and RDF schema to describe resource are located in next floor. Ontology exists in next floor, and technology element for lawyer, logic, proof etc. is located in the above hierarchy.

2.2 Web Security Technique

2.2.1 Channel-Based Method

There is method that applies an encryption technology about TCP connection to be delivered with hypertext transfer protocol message existing between hypertext transfer protocol class and TCP class. In this case, It provides equal encryption service about all hypertext transfer protocol messages of uniformity TCP connection, and this is known as channel base (Channel-based) method and SSL (Secure Socket Layer) is representative protocol. SSL by channel base method developed in Netscape owes in Netscape's supply and it was settled by factual web security standard[8]. IETF developed TLS(Transport Layer Security) Protocol 1.0 using SSL3.0[9]. SSL (Fig 2) can be applied to other applications such as Telnet, FTP etc as well as hypertext transfer protocol for web, because it is protocol that exist between Internet application and TCP/IP communication protocol class. Although this point makes it web security standard, weak point that is not offering function such as digital signature in Internet electronic commerce is subject to supplement.

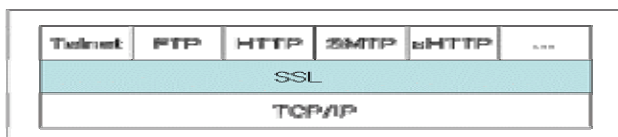


Fig. 2. Relation with SSL and other protocol

2.2.2 Content-Based Method

In higher level of hypertext transfer protocol, to embody security function is content-based web security method in connection with existing encryption system that stability such as Kerberos or PGP is recognized.

This method installs en/decryption program excluding server or browser. This method embodies web security function as do not require modification entirely to existent web system. Outside program that achieve en/decryption function separately, was interlinked to server and client and this outside program is used in encryption of hypertext transfer protocol request and response message(Fig 3). Message transmits on each outside program just before it leaves and transfers browser with server and encrypted message transmits to server and browser again and transfers through the Internet. The best advantage of this method is that do not require any modification to existent web system. Also, encryption module can be used for security function

support to several application programs in addition to web therefore there is advantage that may not install various encryption program on a system. Because it is located on web application outside, there is shortcoming that execution time is delayed by accomplishing unnecessary processing procedure[10].



Fig. 3. Example of outside program

2.4 Related Work

2.4.1 MagPie

MagPie is plug-in program of Internet Explorer. It is a Semantic Web browsing system that shows semantic information using ontology that can speak as most important element of Semantic Web. Basically, It is Active X control program and when it is consented with word that is suitable in domain and word that exist in lexicon, show semantic information of word that user wants[4][5].

2.4.2 SecuIntanet

There is security system for enterprise's intranet system. most enterprise intranet systems process user information for security and access authentication purposes. However, this information is often captured by unauthorized users who may edit, modify, delete or otherwise corrupt this data. Therefore, a method is needed to prevent unauthorized or erroneous access and modification of data through the intranet, and SecuIntranet is a suitable system about it. It is an ACM based web security access control (Web Security Access Control) system basically[12].

3 Design and Implementation

This system performs in a basic client/server environment. Fig. 4 shows the overall structure of the system.

3.1 Summary

First, user is downed Active X control program in main page that provide web service. This program is the program that control various functions for actuality security as program that is run when user connects relevant Web page. Then, user performs login and downloads authority that is allocated beforehand to server. And user requests semantic information that oneself wants to server. This time, if user has a read authority, Server supplies semantic information that user wants by html form through Ontology Agent. And, the information is downloaded to user encrypting, and user can

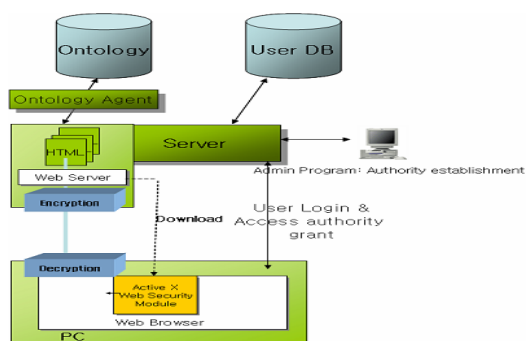


Fig. 4. System Configuration

show the information doing decryption in Active X control program that downed beforehand in user PC. When the information is inspected, web page with semantic information receives various function limitations. For example, user can not print information of the page if the user has no print Authority. This function also achieves at Active X control program that downed beforehand.

3.2 Active X Web Security Module

This system performs security function in client. Module that performs this security function is supplied by Active X control form, and when user connects to server, user is downloaded. The function of this module is divided into greatly two. First, the web page is included semantic information that is offered in server by encrypting state, it performs function that does decryption this. The decryption key is received when user performs login. Second, it performs various function controls of browser. It prevents the function, print, source view, screen capture etc. Control list is showed in 3.6. User receives limitation in case of not prevent unconditionally but has no authority.

3.3 User Authentication

User performs login through the web. User is given access authority about semantic information from server by authenticated ID. User authentication for this is required following point.

- User DB should be constructed.
- User information and En/Decryption Key must be in User DB.
- The access authority should be established beforehand in user and this is given when succeed login.

3.4 Ontology Agent

Ontology Agent accepts information that user wants as keyword and it extracts semantic information that is correct in the keyword in Ontology that already constructed and performs function that makes it by file of html form. Ontology in this paper limits by climate field. And server encrypts this file and sends to User PC. This time, if user

has read authority for the semantic information, one can show the information if not, one can't show.

3.5 En/Decryption

As preceding section referred, html file that made by Ontology Agent is downloaded being encrypted. User can see this file doing decryption by Active X Web Security Module that is downloaded in client. In this system, en/decryption module can do file whole to doing en/decryption but it can also offer part encryption function to reduce load. Fig. 5 is proto-type of encryption API of this system. Text strings in the web pages are encoded by providing an input value to the source string and returning the result as shown in Fig. 5. To encode the whole web page, the file pathname is given as the input value to the source string as shown.

```

Function FileEncrypt()
{
    var bRe = DSSLATL.RequestFileEncrypt(source_string, Authority, dest_string);
}

// [In] source_string : data for encryption
// [In] Authority : User Authority Information
// [Out] dest_string : result data for encryption
    
```

Fig. 5. Encryption Algorithm

This ensures the security of the web page when it is downloaded to the client's PC.

3.6 Authority Control

Authorized users can get permission or limit the permission assigned during server authentication. Fig. 6 shows control list.

Read	Read the Web Page	<input type="checkbox"/> Disable <input checked="" type="checkbox"/> Enable
Print	Print the Web Page	<input type="checkbox"/> Disable <input checked="" type="checkbox"/> Enable
Source Viewing	Source Viewing of the Web Page	<input type="checkbox"/> Disable <input checked="" type="checkbox"/> Enable
Capture	Capture the Web Page	<input type="checkbox"/> Disable <input checked="" type="checkbox"/> Enable

Fig. 6. List for Access Control

Above control function is performed in client and this system can prevent information leakage by user's mistake as well as information leakage by user's deliberation by performing relevant function. For example, read, print, source view and various function of web browser etc.

3.7 Administrator Program

This system supplies administrator program that manages user's authority. Security administrator can change user Authority about semantic information in administrator program. Fig. 7 shows Authority establishment dialog of administrator program.

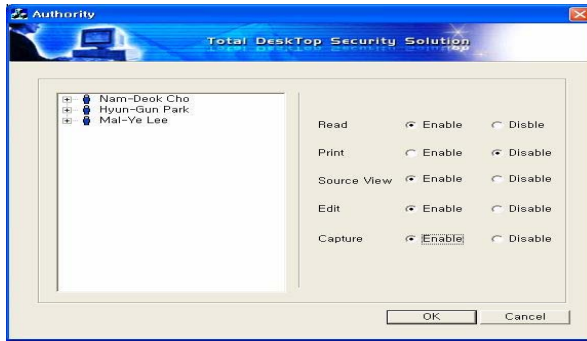


Fig. 7. Administrator Program

4 Result and Estimation

4.1 Result

Design and implementation for Security Intelligence system discussed in the previous chapter. In this chapter the results and estimations of the authors' system will be discussed. Fig. 8 is example Web page of this system. When user clicks each words, this system shows semantic information that use Ontology.

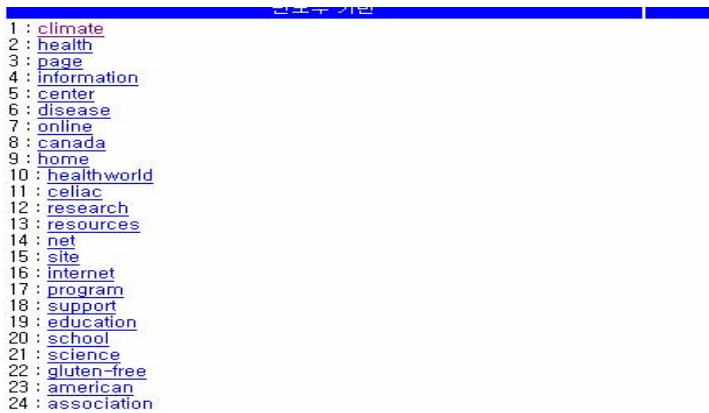


Fig. 8. Example Web page

Fig. 9 shows semantic information when clicked "Climate" of example web page. This semantic information is information that is defined beforehand using Ontology.

Fig. 9 is screen when user has read authority. If user has no read authority, the information is not seen. And fig. 10 shows screen opened by notepad.

This means that relevant page has encrypted, and when a user inspects in web browser, it is decrypted by real-time. Also, even if a user can show the semantic

information because one has read authority, in this system a user can not perform relevant function (print, source view, screen capture) if one has no relevant authority. Fig. 11 shows that a user who has no print authority tries to print through the print menu of web browser.

There is no examination associated with this course. Instead you are asked to do what all scientists do: keep a record of what you end of each chapter you will find a special section on assessment activities (as distinct from activities embedded in the text). through the course you need to keep your Assessment Activities as Star Office word processor documents. We realise that for reasons you may not be able to undertake all the Assessment Activity tasks so you are only required to submit 9 out of the 12 task required to submit these Assessment Activity tasks electronically at the end of the course. See the guide to "Using the Elect System" for instructions on how to do this.

1.2 Weather and Climate

"Climate is what you expect, weather is what you get"

Edward Lorenz

This section introduces you to the differences between what we mean by the terms "weather" and "climate". Make a start by trying the following

Question:

The organisers of a charity afternoon garden party wish to take out insurance against it raining on the day of the party. Who the following groups be concerned with weather or climate?

- (a) The organisers
- (b) The statisticians who calculate the insurance premium

Fig. 9. Semantic information about "Climate" that use Ontology

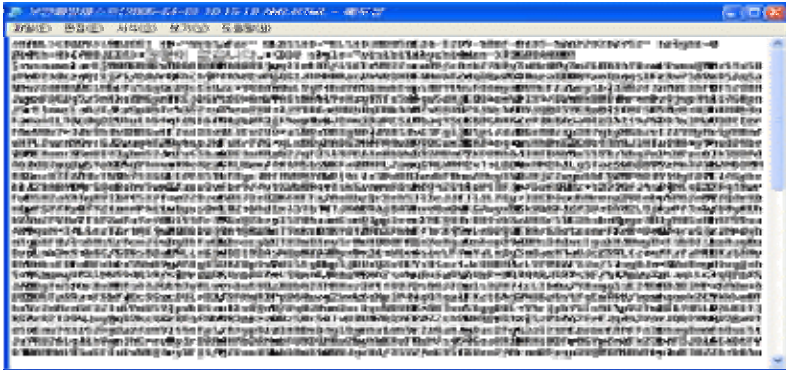


Fig. 10. Encrypted semantic information

If summarize above function, when semantic information that user requires is transmitted from server to client, it is encrypting state, and the information is decrypted and inspected in client, and function of web browser is limited because of security although the information is inspected.

4.2 Estimation

Security Intelligence that is proposed in this paper is a system that control so that user do not flow out the information as well as may show wanting semantic information using Ontology by applying security to Semantic Web that is web technology next generation.

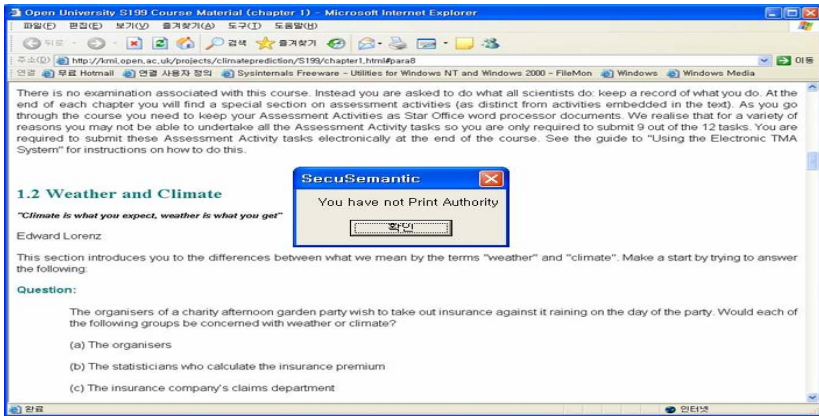


Fig. 11. The screen that a user has no print authority

In this system, a user can inspect web page including semantic information, if he has read authority after user's authentication is succeeded, and even if the web page is inspected, if he has no more authority about relevant web page, user can not flow out the information. That is, if there is no source view authority or print authority, screen capture authority, a user can not flow out the information even if he uses special program and print screen key etc in client PC. This function is available because it is achieved in client.

Therefore, this system can be called suitable system to secure about sensitive semantic information that is constructed using Ontology. Fig. 12 shows a chart providing comparisons of other relevant systems.

System	MagPie	SecuIntranet	this system
Semantic Information Extract	○	×	○
Web Page Encryption	×	○	○
Access Control	×	○	○
Generality	○	×	○

Fig. 12. Chart comparing other systems

5 Conclusion

Semantic Web is web technology next generation invented by Tim Berners-Lee who proposes World Wide Web. This Semantic Web by using Ontology can show more correct information that users wanted, but security about the information becomes problem.

This paper proposes system that presents semantic information using ontology and prevents that users flow out the information. When semantic information that user requires is transmitted from server to client, it is encrypting state, and the information

is decrypted and inspected in client, and function of web browser is limited because of security.

Hereafter, by way of research tasks, Resources in web can be file, image, animation etc. as well as web page and the study of security about its information is defined in Ontology is also needed.

References

- [1] <http://www.w3.org/2001/sw/>
- [2] Tim Berners-Lee, Semantic Web, W3C, <http://www.w3.org/2000/Talks/1206-xml2k-tbl/>, 2000.
- [3] Tim Berners-Lee, James Hendler, and Ora Lassila, Scientific American article: The Semantic Web, Scientific American issue, May 17, 2001.
- [4] Dzbor, M., Domingue, J., and Motta, E.: Magpie: Towards a Semantic Web Browser. Proc. of the 2nd Intl. Semantic Web Conf. (ISWC). 2003. Florida, USA
- [5] Domingue, J., Dzbor, M. and Motta, E.: Magpie: Supporting Browsing and Navigation on the Semantic Web. Proc. of Intl. Conf. on Intelligent User Interfaces (IUI). 2004. Portugal.
- [6] Quan, Dennis and Karger, David R : How to Make a Semantic Web Browser. In Proceedings International WWW Conference, New York, 2004. USA.
- [7] http://news.com.com/2100-1032_3-5605922.html
- [8] A.Freier,P.Karlton and P.Kocher, "The SSL Protocol Version 3.0", <http://www.netscape.com/eng/ssl3/3-spec.ps>, 1996.3.
- [9] T.Dierks, et, al., "The TLS Protocol 1.0" RFC2246, 1999.1.
- [10] J. Weeks, etc, "CCI-Based WebSecurity : A Design Using PGP", WWW Journal 95, 1995.
- [11] B.Rescorla, et, al, The Secure HyperText Transfer Protocol, RFC 2660, 1999.8.
- [12] ND Cho, Hg Park, "Design and Implementation of ACM-based Web Security Access Control System for Intranet Security", Korea Information Processing Society Journal, 12-5, pp643-648 2005.10

Performance Analysis of Location Estimation Algorithm Using an Intelligent Coordination Scheme in RTLS

Seung-Hee Jeong¹, Hyun-Jae Lee¹, Joon-Sung Lee², and Chang-Heon Oh¹

¹ School of Information Technology, Korea University of Technology and Education,
Byeoncheon-myeon, Cheonan-si, Chungcheongnam-do, (330-708 Korea)

{maju9797, present7, choh}@kut.ac.kr

² SAMSUNG ELECTRONICS

junsung72.lee@samsung.com

Abstract. In this paper, we proposed a high precision location estimation algorithm using an intelligent coordination scheme in 2.45GHz RTLS and analyzed an average estimation error distance at 2D coordinates searching-area (300m×300m and 250m×250m). An average error distance was reduced as the number of available reader and received sub-blink increased. Proposed location estimation algorithm satisfied the RTLS specification requirements, 3m radius accuracy when the number of available reader is greater than 3 and the received sub-blink number exceeds 2 times in 250m×250m. Also, an average error distance was saturated within 0.5m~2.5m in case of 250m×250m when the number of available reader is greater than 4 regardless of the number of sub-blink. Also we confirmed that the 3m radius accuracy is 78 percent, and the 2.348m radius accuracy is 72 percent in 300m×300m searching-area when the number of available reader and received sub-blink was 8 and 4 times, respectively. In that case, we confirmed that 3m radius accuracy is 99 percent, and the 2.348m radius accuracy is 97 percent in 250m×250m searching-area.

1 Introduction

The RFID (Radio Frequency IDentification) technologies attach electronic tag at thing serve to various services which can be remote processing and information exchange between management and thing, etc. The RFID can also recognize physical location and condition for user[1]. Therefore, recent growth of interest in pervasive computing and location aware system provides a strong motivation to develop the techniques for estimating the location of devices in both outdoor and indoor environments[2],[3].

Thus, we propose high precision location estimation algorithm with an intelligent coordination scheme, and analyzed performance of this algorithm in 2.45 GHz band RTLS. The organization of this paper is as follows. In section 2, we introduce the overview of RTLS. In section 3, we introduce enhanced TDOA adopting an

intelligent coordination scheme for high precision location estimation. Simulation results are analyzed in section 4. Some conclusions are presented in section 5.

2 The Overview of RTLS (Real Time Location System)

We are known 125kHz, 135kHz, 13.56MHz, 433MHz, 860~960MHz, 2.45GHz, 5.8GHz as available RFID frequency. Among this, RTLS is real time location estimation with RFID technology in 2.45GHz. The system utilizes RTLS tags that autonomously generate a direct-sequence spread spectrum radio frequency. These tags shall transmit at a power level that can facilitate reception at ranges of at least 300 meters open field separation between the tag and reader. Each reader shall be capable of receiving and processing data from a minimum of 120 tags per second. RTLS enables the user to locate, track and manage assets within the infrastructure of the system on a real-time basis. The nominal location data provided by the RTLS shall be within a 3 meter radius of the actual location of the transmitting RTLS tags.

The structure of tag signal designed sub-blink and blink. 1 blink is consisting of sub-blink as 1 to 8. Figure 1 shows the elements of RTLS infrastructure constitute the RTLS transmitters (Tags), RTLS server (Reader), and RTLS application program interface[4].

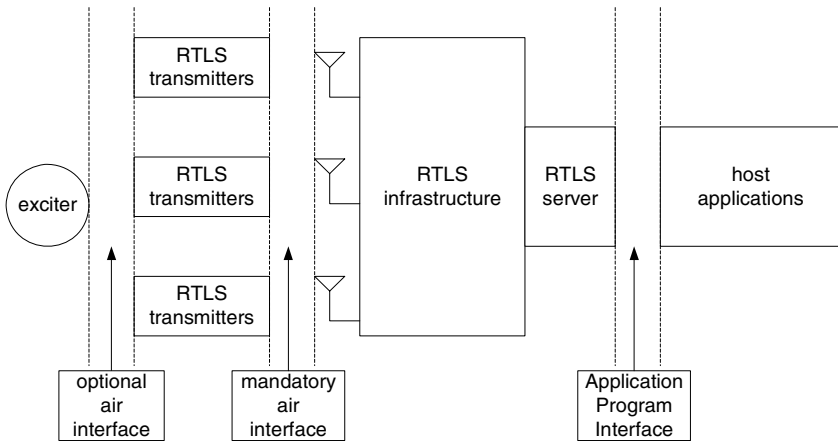


Fig. 1. Elements of RTLS infrastructure

3 Enhanced TDOA with an Intelligent Coordination Scheme

Techniques for estimating the location have been many approaches to get accurate position of things or object. TOA (Time Of Arrival), TDOA (Time Of Difference Arrival), and ROA (Received signal strength Of Arrival) are location aware methods which calculate the relative distance between reader and a tag. TOA uses the time

information of received signals from the tag to calculate distance. Because, this method requires highly accurate time synchronization among all of tags and readers[2]. ROA technique uses the power of the received signal at the receiver. Furthermore, ROA method needs considerable efforts to obtain an empirical radio propagation model beforehand. However, In case of TDOA, synchronized readers receive signals from a tag and calculate time differences between times on which each reader received signals from the tag, which uses two signals with different propagation speeds such as RF and acoustic signals[5]. Thus, TDOA technique is the most suitable method for a RTLS (real time location system) specification without synchronization among all of tags and readers and without accurate GPS time.

We used the time information which received signal to available reader given by equation (1). Where X and Y are unknown actual tag position, X_i and Y_i are available reader position; C is speed of light, respectively. $R(i, j)$ is time difference from tag to available each readers. Afterward, this time difference has include measurement error time, which is uniformly distributed random error time within 32.76nsec[6],[7].

$$R(i, j) = \frac{\sqrt{(X - X_i)^2 + (Y - Y_i)^2}}{C} - \frac{\sqrt{(X - X_j)^2 + (Y - Y_j)^2}}{C}. \tag{1}$$

Figure 2 shows the intersection position of hyperbolic curve using t_1, t_2, t_3 . The RTLS system have data rate of 59.7Kbit per the 1 second. Transmission time is about 32.76nsec per 1 pseudo number code when 511 code length. Therefore, we generate received time from the tag to the reader within 32.76nsec, randomly.

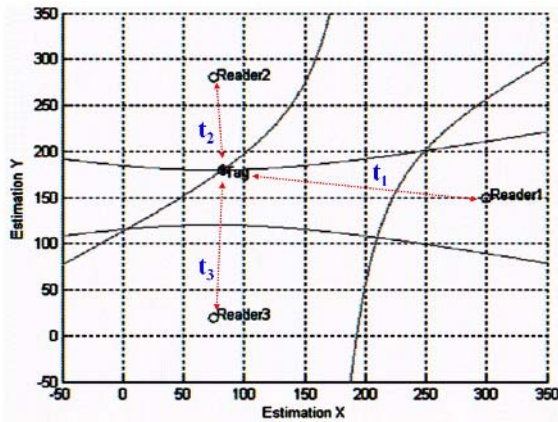


Fig. 2. Locating a tag through trilateration

We propose location estimation algorithm with intelligent coordination scheme for high precision accuracy. The algorithm of this paper executes each step as follows:

Step 1: Get arrival time from each reader

RTLS server has received the arrival time(t_1, t_2, t_3) of a tag signal from each reader. Even though transmit time is same at tag, available reader are placed at different

positions. Therefore, arrival time(t_1, t_2, t_3) is normally different. Then calculates time differences between each arrival time and stores the time difference at database. (Ref. Elements of RTLS infrastructure in figure 1)

Step 2: Calculate time difference between each arrival time

Proposed algorithm calculates intersection position of hyperbolic curve using measured t_1, t_2, t_3 . And then set virtual tag to valid intersection position. (Ref. Locating a tag through trilateration in figure 2, and equation (1))

Step 3: Virtual tag's location estimation

Virtual tag repeats step 1 and produces the intersection position among hyperbolic curve as 2 to 4 with 3 readers for intelligent coordination scheme. (Normally, we have two virtual tag positions. We assume that $t_1=t_1', t_2=t_2'$ and $t_3=t_3'$.)

Step 4: Location decision with intelligent coordination scheme

The location estimation algorithm with intelligent coordination scheme compares time difference information(t_1', t_2', t_3' and t_1'', t_2'', t_3'') of virtual tags with measured time difference information(t_1, t_2, t_3) in database of RTLS. And then select the estimation position of the unknown tag. At that time, we use several decision and exception rules. For example, when the virtual tag's position has an imagination root, eliminate the estimated location. So, we can estimate location of tag in real time and improves accuracy of estimation using above four steps. Also, we can manage each tag efficiently as using database information and algorithm with intelligent coordination scheme.

4 Simulation and Results

In this paper, we used the MATLAB as simulator, and assumed that each tag locates 3m interval to the edge. Each reader is locating within 0m~50m of searching-area, and has equal distance of between readers. We performed simulation of two cases which are 300m×300m, 250m×250m searching-area when the number of reader is fixed. An average error distance was reduced according to receiving number of sub-blink increased. Figure 3 shows the result of an average error distance in 300m×300m searching-area. However, in case of same receiving number of sub-blink, an average error distance is little increase as the number of reader increased. Figure 4 shows the result of average error distance according to the number of reader in the 250m×250m searching-area. In this case, we confirm that an average error distance was reduced when the number of reader and sub-blink are increased.

Also, confirm that the 3m radius within accuracy of RTLS specification is satisfied, if the number of sub-blink receives more than 2 when available reader is 3 to 8 from the result of simulation. Figure 5 and figure 6 shows the result of average error distance when the number of sub-blink fixed in 300m×300m, 250m×250m, respectively. If the number of sub-blink exceeds 7 when 3 available readers and the number of sub-blink exceed 2 when available reader exceed 4, which satisfied for the

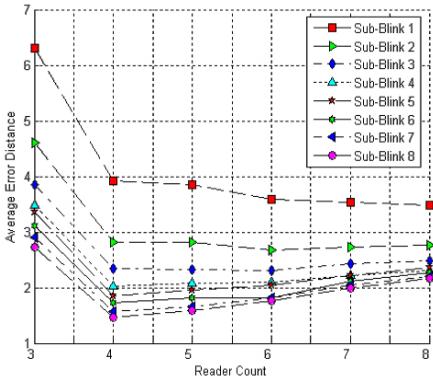


Fig. 3. Average error distance of each sub-blink according to the available reader (searching-area: 300m×300m)

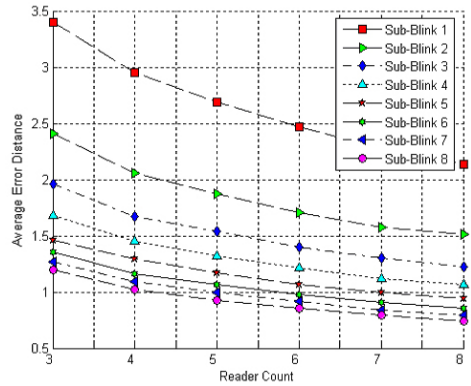


Fig. 4. Average error distance of each sub-blink according to the available reader (searching-area: 250m×250m)

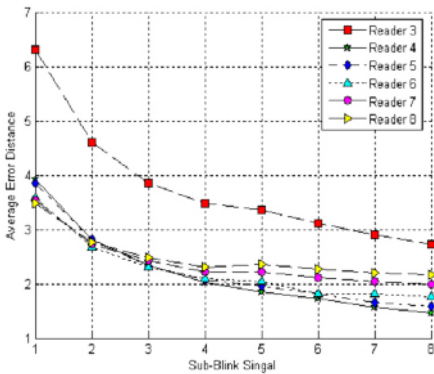


Fig. 5. Average error distance according to the number of sub-blink (searching-area: 300m×300m)

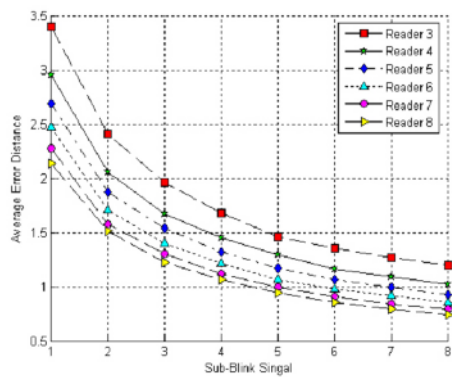


Fig. 6. Average error distance according to the number of sub-blink (searching-area: 250m×250m)

3m radius within accuracy of RTLS specification. Also, an average error distance is saturation within 0.5m~2.5m in 250m×250m when available readers exceed 4.

Figure 7 - figure 10 shows that distribute of error distance and percentage when the number of reader and sub-blink is 8 and 4, respectively. We confirmed that the 3m radius within accuracy is 78 percent, and the 2.348m radius within accuracy is 72 percent in 300m×300m searching-area when the number of available reader and received sub-blink was 8 and 4, respectively. In that case, we also confirmed that 3m radius within accuracy is 99 percent, and the 2.348m radius within accuracy is 97 percent in 250m×250m searching-area.

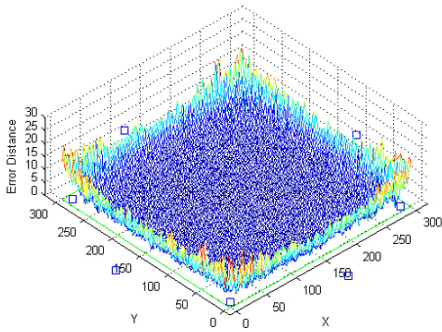


Fig. 7. Distribution of error distance according to searching-area (300m×300m)

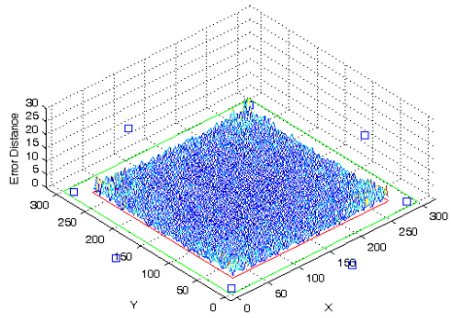


Fig. 8. Distribution of error distance according to searching-area (250m×250m)

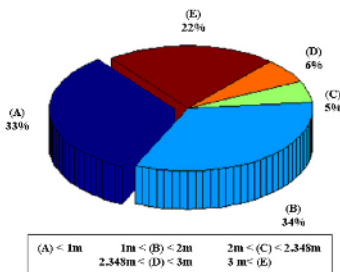


Fig. 9. Percentage of error distance according to searching-area (300m×300m)

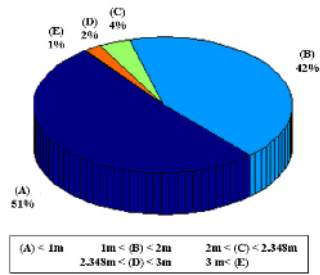


Fig. 10. Percentage of error distance according to searching-area (250m×250m)

5 Conclusions

In this paper, we proposed a high precision location estimation algorithm using an intelligent coordination scheme in 2.45GHz RTLS and analyzed an average estimation error distance at 2D coordinates searching-area (300m×300m and 250m×250m).

An average error distance was reduced as the number of available reader and received sub-blink increased. Proposed location estimation algorithm satisfied the RTLS specification requirements, 3m radius accuracy when the number of available reader is greater than 3 and the received sub-blink number exceeds 2 times in 250m×250m. Also, an average error distance was saturated within 0.5m~2.5m in case of 250m×250m when the number of available reader is greater than 4 regardless of the number of sub-blink. Also we confirmed that the 3m radius accuracy is 78 percent, and the 2.348m radius accuracy is 72 percent in 300m×300m searching-area when the number of available reader and received sub-blink was 8 and 4 times,

respectively. In that case, we confirmed that 3m radius accuracy is 99 percent, and the 2.348m radius accuracy is 97 percent in 250m×250m searching-area.

From the results, we verify that the proposed location estimation algorithm using an intelligent coordination scheme has more estimation accuracy and reasonable efficiency compare with the conventional RTLS.

References

1. T. G. Kanter: Attaching context-aware services to moving locations. *IEEE Internet Computing*, vol. 7, Iss. 2, Mar.-Apr. 2003.
2. Jochen schiller and Agnes voisard: *Location- Based Services*, Morgan Kaufmann, 2004.
3. Y. K. Lee, E. H. Kwon and J. S. Lim: Self Location Estimation Scheme Using ROA in Wireless. *EUC Workshops 2005, LNCS 3823*, pp.1169-1177, Dec 2005.
4. ISO/IEC JTC 1/SC 31/WG 5: Information technology automatic identification and data capture techniques — Real Time Locating Systems (RTLS) — Part 2: 2.4 GHz air interface. Feb. 2005.
5. Xingfa Shen, Zhi Wang, Peng Jiang, Ruizhong Lin, and Youxian Sun : Connectivity and RSSI Based Localization Scheme for Wireless Sensor Networks. *ICIC 2005, Part II, LNCS 3645*, pp. 578–587, 2005.
6. Y. T. Chan and K. C. Ho: A Simple and Efficient Estimator for Hyperbolic Location. *IEEE Transaction on signal processing*, vol. 42, no. 8, Aug. 1994.
7. B.T. Fang: Simple solutions for hyperbolic and related position fixes. in *IEEE Transactions on Aerospace and Electronic Systems*, vol. 26, Iss. 5, Sep. 1990.

Security Requirements for Ubiquitous Software Development Site

Tai-hoon Kim

Department of Information Electronics Engineering, Ewha Womans University, Korea
taihoonn@empal.com

Abstract. A PP (protection profile) defines an implementation-independent set of security requirements for a category of Target of Evaluations. Consumers or owners can therefore construct or cite a PP to express their security needs without reference to any specific IT products. Generally, PP contains security assurance requirements about the security of development environment for IT product or system and PP can applied to ubiquitous software development site. This paper proposes some security environments for ubiquitous software development site by analyzing the ALC_DVS.1 of the ISO/IEC 15408 and Base Practices (BPs) of the ISO/IEC 21827.

1 Ubiquitous Environment

Sensor Network is the information management that detects the condition of an item (the temperature, humidity, degree of contamination, a crevice, etc.) by attaching an RFID tag to it and sends the information to the network. Ultimately, all the objects are endowed with computing and communication abilities regardless of time, places, where, net-works, devices, and services. In order to realize USN, various RFID tags that offer sensor information should be developed as well as sensing functions so as to build networks among them.

USN is composed of a tag (a sensor) attached to an antenna which is installed on a reader. This reader is connected to information networks while the tag and reader communicate through an electric wave. It is powered by a built-in energy source or radio waves received. When a reader transmits radio waves through a tag, the tag uses the electric waves as an energy source. In turn, the activated tag sends its information to the reader. According to the energy source, tags is divided into active and passive types. A passive tag gets energy from the electric waves that a reader sends while an active tag has its own battery. The reader transmits collected information to Savant server and then ONS locates PML where the information exists so that it can obtain specific information of an item [1].

This paper proposes some assumptions about the ubiquitous software development site to describe security requirements of PP for ubiquitous software development environment.

2 Overview of Related Works

2.1 Radio Frequency Identification

RFID is, the Main Technology of USN. The frequency range of an RFID tag is usually 125 KHz and 13.56 MHz, which are used for a traffic card or admission control within short distances. As 900MHz and 2.4GHz are used, the available distance becomes longer and prices, inexpensive. As a result, the technology has been introduced to various areas such distribution, logistics, the environment, transportation, and so on. Besides, more sensing abilities are added and its use is extended to medical/security/national defense areas.

An RFID reader can read 100 tags per a second by adopting tag anti-collision algorithms but the technology aims at reading several hundreds. Since the use of mixed frequency range (13.56MHz, 900MHz, and 2.4GHz) is expected, a multi-frequency/multi-protocol reader should be developed. Currently, the sensing distance and accuracy of an RFID reader appears to be limited by the function of an antenna as well as surrounding circumstances. In order to enhance the sensor function, therefore, arrangement of 2~4 antennas is used. Beam-forming antennas, which control beams by responding to the environments, will be applied in the future. Research should be done on tag anti-collision also. There must be identifier code system that gives a tag identification numbers according to regions, for the technology will employ electric tags using identification schemes to identify an object. Related technologies are EPC (96bit) system developed by EAN and UCC, which leads international distribution standardization, and u-ID (126bit) system by Japan. Meanwhile, IPv6 of 128bit is being promoted for the Internet address system, indicating that providers need to scale up global efforts to establish code system standardization. The U.S. government is seeking to utilize RFID technology to reduce costs and enhance services. According to the GAO report in U.S., 13 of 24 federal bodies are using the RFID or planning to apply the technology [2].

2.2 Common Criteria

The multipart standard ISO/IEC 15408 defines criteria, which for historical and continuity purposes are referred to herein as the Common Criteria (CC), to be used as the basis for evaluation of security properties of IT products and systems. By establishing such a common criteria base, the results of an IT security evaluation will be meaningful to a wider audience.

The CC is presented as a set of distinct but related parts as identified below.

Part 1, Introduction and general model, is the introduction to the CC. It defines general concepts and principles of IT security evaluation and presents a general model of evaluation.

Part 2, Security functional requirements, establishes a set of functional components as a standard way of expressing the functional requirements for TOEs (Target of

Evaluations). Part 2 catalogues the set of functional components, families, and classes.

Part 3, Security assurance requirements, establishes a set of assurance components as a standard way of expressing the assurance requirements for TOEs. Part 3 catalogues the set of assurance components, families and classes.

2.3 Protection Profile

A PP defines an implementation-independent set of IT security requirements for a category of Target of Evaluations (TOEs). Such TOEs are intended to meet common consumer needs for IT security. Consumers can therefore construct or cite a PP to express their IT security needs without reference to any specific TOE.

The purpose of a PP is to state a security problem rigorously for a given collection of systems or products (known as the TOE) and to specify security requirements to address that problem without dictating how these requirements will be implemented. For this reason, a PP is said to provide an implementation-independent security description.

2.4 ALC_DVS

ALC_DVS.1 component consists of one developer action element, two evidence elements, and two evaluator action elements.

Contents and presentation of evidence elements of ALC_DVS.1 component are described as like following (Requirements for content and presentation of evidence are identified by appending the letter 'C' to the element number):

ALC_DVS.1.1C The development security documentation shall describe all the physical, procedural, personnel, and other security measures that are necessary to protect the confidentiality and integrity of the TOE design and implementation in its development environment.

ALC_DVS.1.2C The development security documentation shall provide evidence that these security measures are followed during the development and maintenance of the TOE.

2.4 SSE-CMM

Modern statistical process control suggests that higher quality products can be produced more cost-effectively by emphasizing the quality of the processes that produce them, and the maturity of the organizational practices inherent in those processes.

More efficient processes are warranted, given the increasing cost and time required for the development of secure systems and trusted products. The operation and maintenance of secure systems relies on the processes that link the people and technologies. These interdependencies can be managed more cost effectively by emphasizing the quality of the processes being used, and the maturity of the organizational practices inherent in the processes.

The SSE-CMM model is a standard metric for security engineering practices covering:

- The entire life cycle, including development, operation, maintenance, and de-commissioning activities
- The whole organization, including management, organizational, and engineering activities
- Concurrent interactions with other disciplines, such as system, software, hardware, human factors, and test engineering; system management, operation, and maintenance
- Interactions with other organizations, including acquisition, system management, certification, accreditation, and evaluation.

3 Performance Analyses

3.1 Comparison in Process Area

The SSE-CMM has two dimensions, “domain” and “capability.” The domain dimension is perhaps the easier of the two dimensions to understand. This dimension simply consists of all the practices that collectively define security engineering. These practices are called Base Practices (BPs).

The base practices have been organized into Process Areas (PAs) in a way that meets a broad spectrum of security engineering organizations. There are many ways to divide the security engineering domain into PAs. One might try to model the real world, creating process areas that match security engineering services. Other strategies attempt to identify conceptual areas that form fundamental security engineering building blocks. The SSE-CMM compromises between these competing goals in the current set of process areas.

Each process area has a set of goals that represent the expected state of an organization that is successfully performing the PA. An organization that performs the BPs of the PA should also achieve its goals.

There are eleven PAs related to security in the SSE-CMM, and we found next three PAs which have compliance with ALC_DVS.1 component:

- A01 Administer Security Controls
- PA08 Monitor Security Posture
- PA09 Provide Security Input

3.2 Comparison in Base Practice

All of the BPs in each PA mentioned earlier need not have compliance with the evidence elements of ALC_DVS.1. But if any BP included in the PA is excluded or failed when the evaluation is preceded, the PA itself is concluded as fail.

Evidence element ALC_DVS.1.1C requires that the development security documentation shall describe all the physical, procedural, personnel, and other security measures that are necessary to protect the confidentiality and integrity of the TOE design and implementation in its development environment. But ALC_DVS.1.1C dose not describe what are the physical, procedural, personnel, and other security

measures. Evidence element ALC_DVS.1.2C requires that the development security documentation shall provide evidence that the security measures described in ALC_DVS.1.1C are followed during the development and maintenance of the TOE.

Some BPs contains example work products, and work products are all the documents, reports, files, data, etc., generated in the course of performing any process. Rather than list individual work products for each process area, the SSE-CMM lists Example Work Products (EWPs) of a particular base practice, to elaborate further the intended scope of a BP. These lists are illustrative only and reflect a range of organizational and product contexts. As though they are not to be construed as mandatory work products, we can analysis the compliance between ALC_DVS.1 component and BPs by comparing evidence elements with these work products. We categorized these example work products as eight parts:

1. Physical measures related to the security of development site and system.
2. Procedural measures related to the access to development site and system.
3. Procedural measures related to the configuration management and maintenance of development site and system.
4. Procedural measures (contain personnel measures) related to the selection, control, assignment and replacement of developers.
5. Procedural measures (contain personnel measures) related to the qualification, consciousness, training of developers.
6. Procedural measures related to the configuration management of the development work products.
7. Procedural measures related to the product development and incident response in the development environment.
8. Other security measures considered as need for security of development environment.

Categorized eight parts above we suggested are based on the contents of evidence requirement ALC_DVS.1.1C, and contains all types' measures mentioned in ALC_DVS.1.1C. But the eight parts we suggested may contain the possibility to be divided to more parts.

We can classify work products included in BPs according to eight parts category mentioned above. Next table 1 describes the result.

From the table above, we can verify that some BPs of SSE-CMM may meet the requirements of ALC_DVS.1.1C by comparing the contents of evidence element with work products.

The requirements described in ALC_DVS.1.2C can be satisfied by records express the development processes, and some BPs can meet the requirements of evidence element ALC_DVS.1.2C. We researched all BPs and selected the BPs which can satisfy the requirements of evidence element ALC_DVS.1.2C. We list BPs related to ALC_DVS.1.2C as like:

- BP.08.01: Analyze event records to determine the cause of an event, how it proceeded, and likely future events.
- BP.08.02 Monitor changes in threats, vulnerabilities, impacts, risks, and the environment
- BP.08.03 Identify security relevant incidents

- BP.08.04 Monitor the performance and functional effectiveness of security safeguards
- BP.08.05 Review the security posture of the system to identify necessary changes
- BP.08.06 Manage the response to security relevant incidents
- And all BPs included in PA01

Table 1. Categorization of work products

Number of category	Work Products	Related BP
1	control implementation	BP.01.02
	sensitive media lists	BP.01.04
2	control implementation	BP.01.02
	control disposal	BP.01.02
	sensitive media lists	BP.01.04
	sanitization, downgrading, & disposal	BP.01.04
	architecture recommendation	BP.09.05
	implementation recommendation	BP.09.05
	security architecture recommendation	BP.09.05
	users manual	BP.09.06
3	records of all software updates	BP.01.02
	system security configuration	BP.01.02
	system security configuration changes	BP.01.02
	records of all confirmed software updates	BP.01.02
	security changes to requirements	BP.01.02
	security changes to design documentation	BP.01.02
	control implementation	BP.01.02
	security reviews	BP.01.02
	control disposal	BP.01.02
	maintenance and administrative logs	BP.01.04
	periodic maintenance and administrative reviews	BP.01.04
	administration and maintenance failure	BP.01.04
	administration and maintenance exception	BP.01.04
	sensitive media lists	BP.01.04
	sanitization, downgrading, and disposal	BP.01.04
	architecture recommendations	BP.09.05
	implementation recommendations	BP.09.05
	security architecture recommendations	BP.09.05
administrators manual	BP.09.06	
4	an organizational security structure chart	BP.01.01
	documented security roles	BP.01.01
	documented security accountabilities	BP.01.01
	documented security authorizations	BP.01.01
5	sanitization, downgrading, and disposal	BP.01.04
	user review of security training material	BP.01.03
	logs of all awareness, training and education	BP.01.03

	undertaken, and the results of that training	
	periodic reassessments of the user community level of knowledge, awareness and training with regard to security	BP.01.03
	records of training, awareness and educational material	BP.01.03
6	documented security responsibilities	BP.01.01
	records of all distribution problems	BP.01.02
	periodic summaries of trusted software distribution	BP.01.02
	sensitive information lists	BP.01.04
	sanitization, downgrading, and disposal	BP.01.04
7	periodic reassessments of the user community level of knowledge, awareness and training with regard to security	BP.01.03
	design recommendations	BP.09.05
	design standards, philosophies, principles	BP.09.05
	coding standards	BP.09.05
8	philosophy of protection	BP.09.05
	security profile	BP.09.06
	system configuration instructions	BP.09.06

Therefore, if the PA01, PA08 and PA09 are performed exactly, it is possible ALC_DVS.1 component is satisfied. But one more consideration is needed to meet the requirements completely.

4 Assumptions

Now, we can describe some assumptions like as:

- Adequate communications exist between the component developers and between the component developers and the IT system developers.
- The development site will be managed in a manner that allows it to appropriately address changes in the IT System.
- The security auditor has access to all the IT System data it needs to perform its functions.
- The threat of malicious attacks aimed at entering to site is considered low.
- There will be one or more competent individuals assigned to manage the environments and the security of the site.
- Administrators are non-hostile, appropriately trained and follow all administrator guidance.
- There will be no general-purpose computing or storage repository capabilities (e.g., compilers, editors, or user applications) not used for developing in the site.
- Anybody cannot gain access to recourses protected by the security countermeasures without passing through the access control mechanisms.
- Physical security will be provided within the domain for the value of the IT assets.
- The security environment is appropriately scalable to provide support to the site.

5 Conclusions

In general, threat agents' primary goals may fall into three categories: unauthorized access, unauthorized modification or destruction of important information, and denial of authorized access. Security countermeasures are implemented to prevent threat agents from successfully achieving these goals.

This paper proposes some assumptions about the development site to describe security environments of PP for software development site.

In these days, some security countermeasures are used to protect development site. But the security countermeasures should be considered with consideration of applicable threats and security solutions deployed to support appropriate security services and objectives. Maybe this is one of our future works.

References

1. Ubiquitous IT Korea forum : <http://www.ukoreaforum.or.kr/ukorea/index.php>
2. Seok-soo Kim, Gilcheol Park, Kyungsuk Lee and Sunho Kim: Ubiquitous Military Supplies Model based on Sensor Network. International Journal of Multimedia and Ubiquitous Engineering 2006: Vol.2, No.1, SERSC
3. Tai-Hoon Kim, Seung-youn Lee: A Relationship Between Products Evaluation and IT Systems Assurance. KES (Knowledge-Based Intelligent Information and Engineering Systems) 2005: LNCS 3681, 1125-1130.
4. Tai-Hoon Kim, Seung-youn Lee: Intelligent Method for Building Security Countermeasures by Applying Dr. T.H. Kim's Block Model. KES (Knowledge-Based Intelligent Information and Engineering Systems) 2005: LNCS 3682, 1069-1075

Paraconsistent Artificial Neural Network: Applicability in Computer Analysis of Speech Productions

Jair Minoro Abe^{1,3}, João Carlos Almeida Prado², and Kazumi Nakamatsu⁴

¹ Information Technology Dept., ICET – Paulista University – Brazil

² Faculty of Philosophy, Letters and Human Sciences – University of São Paulo – Brazil

³ Institute For Advanced Studies – University of São Paulo – Brazil

⁴ School of Human Science and Environment/H.S.E. – University of Hyogo – Japan

jairabe@uol.com.br, joaocarlos@autobyte.com.br,
nakamatu@shse.u-hyogo.ac.jp

Abstract. In this work we sketch how Paraconsistent Artificial Neural Network – PANN – can be useful in speech signals recognition by using phonic traces signals. The PANN is built based on Paraconsistent Annotated Logic $E\tau$ and it allows us to manipulate uncertain, inconsistent and paracomplete information without trivialization.

1 Introduction

Many pattern recognition applications use statistical models with a large number of parameters, although the amount of available training data is often insufficient for robust parameter estimation. In order to overcome these aspects, a common technique to reduce the effect of data sparseness is the divide-and-conquer approach, which decomposes a problem into a number of smaller subproblems, each of which can be handled by a more specialized and potentially more robust model. This principle can be applied to a variety of problems in speech and language processing: the general procedure is to adopt a feature-based representation for the objects to be modeled (such as phones or words), learn statistical models describing the features of the object rather than the object itself, and recombine these partial probability estimates. Although this enables a more efficient use of data, other interesting techniques have been employed for the task. One of the most successful theories is the so-called artificial neural networks - ANN.

ANN are computational paradigms based on mathematical models that unlike traditional computing have a structure and operation that resembles that of the mammal brain. ANN or neural networks for short, are also called *connectionist systems*, *parallel distributed systems* or *adaptive systems*, because they are composed by a series of interconnected processing elements that operate in parallel. Neural networks lack centralized control in the classical sense, since all the interconnected processing elements change or “adapt” simultaneously with the flow of information and adaptive rules.

One of the original aims of ANN was to understand and shape the functional characteristics and computational properties of the brain when it performs cognitive

processes such as sensorial perception, concept categorization, concept association and learning. However, today a great deal of effort is focused on the development of neural networks for applications such as pattern recognition and classification, data compression and optimization. Most of ANN known is based on classical logic or extensions of it.

In this paper we are concerned in applying a particular ANN, namely the paraconsistent artificial neural network – PANN, introduced in [4] which is based on paraconsistent annotated logic $E\tau$ [1] to speech signals recognition by using phonic traces signals. The PANN is capable of manipulating concepts like uncertainty, inconsistency and paracompleteness in its interior.

2 Background

Paraconsistent Artificial Neural Networks – PANN is a new artificial neural network introduced in [4]. Its basis leans on paraconsistent annotated logic $E\tau$ [1]. Let us present it briefly.

The atomic formulas of the logic $E\tau$ is of the type $p_{(\mu, \lambda)}$, where $(\mu, \lambda) \in [0, 1]^2$ and $[0, 1]$ is the real unitary interval (p denotes a propositional variable). $p_{(\mu, \lambda)}$ can be intuitively read: “It is assumed that p ’s favorable evidence is μ and contrary evidence is λ .” Thus,

- $p_{(1.0, 0.0)}$ can be read as a true proposition.
- $p_{(0.0, 1.0)}$ can be read as a false proposition.
- $p_{(1.0, 1.0)}$ can be read as an inconsistent proposition.
- $p_{(0.0, 0.0)}$ can be read as a paracomplete (unknown) proposition.
- $p_{(0.5, 0.5)}$ can be read as an indefinite proposition.

We introduce the following concepts (with $0 \leq \mu, \lambda \leq 1$): Uncertainty Degree: $G_{un}(\mu, \lambda) = \mu + \lambda - 1$; Certainty Degree: $G_{cc}(\mu, \lambda) = \mu - \lambda$; An order relation is defined on $[0, 1]^2$: $(\mu_1, \lambda_1) \leq (\mu_2, \lambda_2) \Leftrightarrow \mu_1 \leq \mu_2$ and $\lambda_1 \leq \lambda_2$, constituting a lattice that will be symbolized by τ .

With the uncertainty and certainty degrees we can get the following 12 output states: *extreme states* that are, False, True, Inconsistent and Paracomplete, and *non-extreme states*.

Table 1. Extreme and Non-extreme states

Extreme States	Symbol	Non-extreme states	Symbol
True	V	Quasi-true tending to Inconsistent	$QV \rightarrow T$
False	F	Quasi-true tending to Paracomplete	$QV \rightarrow \perp$
Inconsistent	T	Quasi-false tending to Inconsistent	$QF \rightarrow T$
Paracomplete	\perp	Quasi-false tending to Paracomplete	$QF \rightarrow \perp$
		Quasi-inconsistent tending to True	$QT \rightarrow V$
		Quasi-inconsistent tending to False	$QT \rightarrow F$
		Quasi-paracomplete tending to True	$Q\perp \rightarrow V$
		Quasi-paracomplete tending to False	$Q\perp \rightarrow F$

Some additional control values are:

- V_{cic} = maximum value of uncertainty control = Ft_{ct}
- V_{cve} = maximum value of certainty control = Ft_{ce}
- V_{cpa} = minimum value of uncertainty control = $-Ft_{ct}$
- V_{cfa} = minimum value of certainty control = $-Ft_{ce}$

For the discussion in the present paper we have used: $Ft_{ct} = Ft_{ce} = 1/2$.

All states are represented in the lattice τ of the next figure.

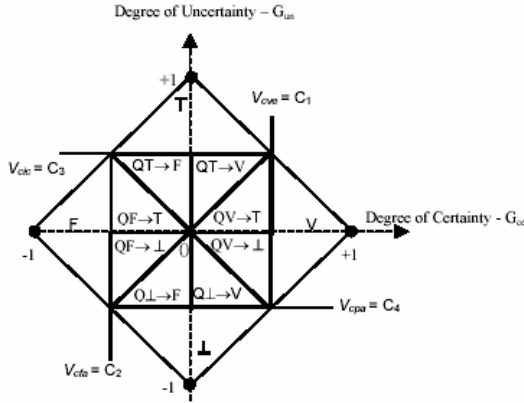


Fig. 1. Output lattice τ 3

3 The Main Artificial Neural Cells

In the PANN the main aim is to know how to determine the certainty degree concerning a proposition, if it is False or True. Therefore, for this, we take into account only the certainty degree G_{ce} . The uncertainty degree G_{un} indicates the ‘measure’ of the inconsistency or para-completeness. If the certainty degree is low or the uncertainty degree is high, it generates an indefinision.

The resulting certainty degree G_{ce} is obtained as follows:

If: $V_{cfa} \leq G_{un} \leq V_{cve}$ or $V_{cic} \leq G_{un} \leq V_{cpa}$
 $\Rightarrow G_{ce} = \text{Indefinition}$

For: $V_{cpa} \leq G_{un} \leq V_{cic}$

If: $G_{un} \leq V_{cfa} \Rightarrow G_{ce} = \text{False}$ with degree G_{un}

$V_{cic} \leq G_{un} \Rightarrow G_{ce} = \text{True}$ with degree G_{un}

A Paraconsistent Artificial Neural Cell – PANC – is called *basic* PANC when given a pair (μ, λ) is used as input and resulting as output: G_{un} = resulting

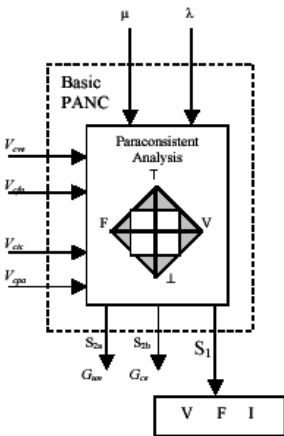


Fig. 2. Basic cell

uncertainty degree, G_{ce} = resulting certainty degree, and X = constant of Indefinition.

Using the concepts of *basic* PANC we can obtain the family of PANC considered in this work, as described in Table 2 below:

Table 2. Paraconsistent Artificial Neural Cells

PANC	Inputs	Calculations	Output
Analytic connection – PANNac	$\mu, \lambda, Ft_{ct}, Ft_{ce}$	$\lambda_c = 1 - \lambda, G_{un}, G_{ce}, \mu_r = (G_{ce} + 1)/2$	If $ G_{ce} > Ft_{ce}$ then $S_1 = \mu_r$ and $S_2 = 0$; If $ G_{un} > Ft_{ct}$ and $ G_{un} > G_{ce} $ then $S_1 = \mu_r$ and $S_2 = G_{un} $, if not $S_1 = 1/2$ and $S_2 = 0$
Maximization – PANNmax	μ, λ	none	If $\mu > \lambda$, then $S_1 = \mu$, if not $S_1 = \lambda$
Minimization – PANNmin	μ, λ	none	If $\mu < \lambda$, then $S_1 = \mu$, if not $S_1 = \lambda$
Complementation – PANNco	μ, Ft_{ct}	$\mu_c = 1 - \mu$	$\mu_r = (\mu_c - \mu + 1)/2$; if $\mu_r \geq Ft_{ct}$, if not $\mu_r = 1/2$
Decision – PANNde	μ, λ, Ft_{de}	$Vl_r = (1 - t_{de})/2, Vl_v = (1 + t_d)/2, \mu_r = (\mu - \lambda + 1)/2$	If $\mu_r \geq Vl_v$, then $S_1 = 1$ (V), if $\mu_r \leq Vl_r$, then $S_2 = 0$ (F), if not, constant to be determined by the application

3 Using PANN in Speech Production Recognition

Through a microphone hardwired to a computer, a sonorous signal can be caught and transformed to a vector (finite sequence of natural numbers x_i) through a digital sampling. This vector characterizes a sonorous pattern and it is registered by the PANN. So, new signals are compared, allowing their recognition or not. For the sake of completeness, we show some basic aspects of how PANN operates. Let us take three vectors: $V_1 = (2, 1, 2, 7, 2)$; $V_2 = (2, 1, 3, 6, 2)$; $V_3 = (2, 1, 1, 5, 2)$. The favorable evidence is calculated as follows: given a pair of vectors, we take ‘1’ for equal elements and ‘0’ to the different elements, and we figure out its percentage.

Comparing V_2 with V_1 : $1 + 1 + 0 + 0 + 1 = 3$; in percentage: $(3/5)*100 = 60\%$

Comparing V_3 with V_1 : $1 + 1 + 0 + 0 + 1 = 3$; in percentage: $(3/5)*100 = 60\%$

The contrary evidence is the weighted addition of the differences between the different elements, in module:

Comparing V_2 with $V_1 = 0 + 0 + 1/8 + 1/8 + 0 = (2/8)/5 = 5\%$

Comparing V_3 with $V_1 = 0 + 0 + 1/8 + 2/8 + 0 = (3/8)/5 = 7.5\%$

Therefore, we can say that V_2 is ‘closer’ to V_1 than V_3 . We use a PANN to recognize this technical system.

We can improve this technical method by adding more capabilities to the PANN, like ‘proximity’ concept and ‘recognizing’ level.

Also, the PANN has the capability of adjusting its own recognizing factor through the recognizing factor internal to the Neural Cell, that can be propagated to higher neural levels. Thus, the PANN can improve its capability in each recognizing speech.

Another important functionality aspect of the PANN is the processing velocity, so we can work in real time producing and identifying speech.

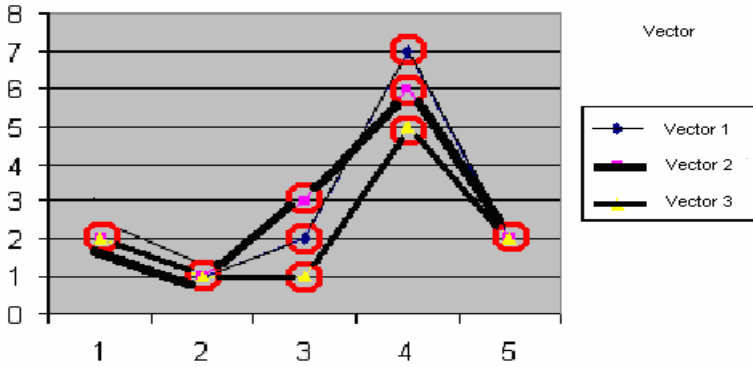


Fig. 3. Vector's representation

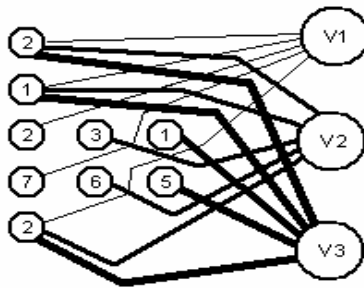


Fig. 4. PANN and layers

4 Practical Results

To test the theory presented here, we develop one computer system with the capability of capturing and converting a speech signal as a vector. After this, we analyze the percentage recognizing results shown below. With these studies we point out some of the most important features of PANN: firstly, the PANN recognition becomes 'better' in every new recognition step, so it is a consequence of discarding contradicting signals and recognizing them by proximity, without trivializing the results. Finally, the performance and efficiency of the PANN is enough to recognize in real time, any speech signal.

Now we show how the PANN was efficient in formants recognition. The tests were made in Portuguese and 3 pairs of syllables were chosen 'FA-VA', 'PA-BA', 'CA-GA' presenting one articulation and differences in sonority (see table 2). The speaker is an adult, masculine sex, 42 years old, Brazilian, from São Paulo city.

Table 3 shows the recognizing capability. The recognizing percent in the first column is 100% because the PANN is empty and the syllables are just being learned. The process of recognition is made in the computer system as follows: in the step 2 the speaker says, for instance, the syllable 'FA'. Then the PANN gives an output with

the calculations (favorable/contrary evidences, G_{cc} , G_{un}) and asks to the speaker (operator) if the data is acceptable or not. If the answer is ‘Yes’, the PANN keep the parameters for the next recognition. If the answer is ‘Not’, the PANN recalculate the parameters in order to criticize the next recognition, till such data becomes belongs to False state (fig. 1), preparing for the next step to repeat the process (in this way, improves the recognition). This is performed by the neural cell PNACde (see table 3):

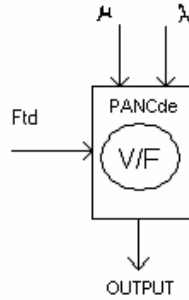


Fig. 5. PANCde

Table 3. Syllable recognition

Syllable	Steps										Average
	1	2	3	4	5	6	7	8	9	10	
FA	100%	87%	88%	91%	90%	92%	95%	94%	94%	95%	91,78%
VA	100%	82%	85%	87%	88%	90%	94%	92%	96%	95%	89,89%
PA	100%	83%	86%	90%	88%	95%	89%	94%	95%	95%	90,56%
BA	100%	85%	82%	87%	90%	89%	92%	95%	95%	97%	90,22%
CA	100%	82%	87%	89%	88%	92%	90%	94%	92%	95%	89,89%
GA	100%	84%	88%	92%	90%	89%	95%	95%	95%	92%	91,11%

Table 4. Recognition of pairs of syllables with one different speech articulation

Pairs	Steps										Average
	1	2	3	4	5	6	7	8	9	10	
FA-VA	70%	67%	72%	59%	65%	71%	64%	69%	66%	63%	66,60%
PA-BA	51%	59%	49%	53%	48%	52%	46%	47%	52%	48%	50,50%
CA-GA	62%	59%	61%	62%	58%	59%	60%	49%	63%	57%	59,00%

These adjustments are made automatically by the PANN except only the learning steps, which is made the intervention of the operator to feed the Yes/No data. More details are to found in [4]. Thus, from the second column on, PANN is recognizing and learning adjusts simultaneously, as well as adapting such improvements; this is the reason that the recognizing factor increases. In the example, we can see that after the sixth speech step, the PANN is able to recognize efficiently every signal with recognizing factor higher than 88%. Every signal lower than this factor can be considered as unrecognized.

Table 4 shows the recognition factor percentage when PANN analyzes a syllable with one different speech articulation. As we can see, when the PANN was learned the 'FA' syllable (10 times) and it is asked the 'VA' syllable for recognizing, the recognizing factor is never higher than 72%. For the remaining pairs of syllables, this factor showed lower.

6 Conclusions

The variation of the analyzed values are interpreted by the PANN and adjusted automatically by the system. Due the PANN structural construction, the network is able to identify small variations between the pairs of syllables chosen. One central reason is its capability of proximity recognition and discarding contradictory data without trivialization. In the examples above, we can define as recognized if the factor is higher than 88%, and non-recognized, if the factor is lower than 72%. The difference of 16% (between recognition and non-recognition) is enough to avoid mistakes in the interpretation of the results. Thus, PANN shows itself as a superior system in being capable to manipulate the factors described showing high accuracy in data analysis.

The results presented in this paper show that PANN can be a very efficient structure for speech analysis. Of course, new concepts are necessary for a more complete study of speech production, but this is in course. We hope to say more in forthcoming papers.

References

- [1] J. M. Abe, "Fundamentos da Lógica Anotada" (Foundations of Annotated Logics), in Portuguese, Ph. D. Thesis, University of São Paulo, São Paulo, 1992.
- [2] J. I. Da Silva Filho & J. M. Abe, Para-Analyzer and Inconsistencies in Control Systems, Proceedings of the IASTED *International Conference on Artificial Intelligence and Soft Computing* (ASC'99), August 9-12, Honolulu, Hawaii, USA, 78-85, 1999.
- [3] J. I. Da Silva Filho & J. M. Abe, Paraconsistent analyzer module, *International Journal of Computing Anticipatory Systems*, vol. 9, ISSN 1373-5411, ISBN 2-9600262-1-7, 346-352, 2001.
- [4] J. I. Da Silva Filho & J. M. Abe, *Fundamentos das Redes Neurais Paraconsistentes – Destacando Aplicações em Neurocomputação*, in Portuguese, Editôra Arte & Ciência, ISBN 85-7473-045-9, 247 pp., 2001.
- [5] A. P. Dempster, Generalization of Bayesian inference, *Journal of the Royal Statistical Society*, Series B-30, 205-247, 1968.
- [6] R. Hecht-Nielsen, *Neurocomputing*. New York, Addison Wesley Pub. Co., 1990.
- [7] T. Kohonen, *Self-Organization and Associative Memory*. Springer-Verlag, 1984.
- [8] B. Kosko, *Neural Networks for signal processing*. USA, New Jersey, Prentice-Hall, 1992
- [9] R. Sylvan & J. M. Abe, On general annotated logics, with an introduction to full accounting logics, *Bulletin of Symbolic Logic*, 2, 118-119, 1996.
- [10] L. Fausett, *Fundamentals of Neural Networks Architectures, Algorithms and Applications*, Prentice-Hall, Englewood Cliffs, 1994.
- [11] M.J. Russell & J.A. Bilmes, Introduction to the Special Issue on New Computational Paradigms for Acoustic Modeling in Speech Recognition, *Computer Speech and Language*, 17, 107-112, 2003.

Intelligent Paraconsistent Logic Controller and Autonomous Mobile Robot Emmy II

Jair M. Abe^{1,3}, Cláudio R. Torres^{2,6}, Germano L. Torres⁶, Kazumi Nakamatsu⁵, and Michiro Kondo⁴

¹ Institute For Advanced Studies – University of São Paulo - Brazil

² Universidade Metodista de São Paulo, São Paulo, Brazil

³ Information Technology Dept., ICET – Paulista University – Brazil

⁴ School of Information Environment - Tokyo Denki University - Japan

⁵ School of Human Science and Environment/H.S.E. - University of Hyogo - Japan

⁶ GAIA – Grupo de Aplicações de Inteligência Artificial, UNIFEI – Federal University of Itajubá – Brazil

jairabe@uol.co.br, c.r.t@uol.com.br, germano@iee.efei.br,
nakamatu@shse.u-hyogo.ac.jp, kondo@sie.dendai.ac.jp

Abstract. In this work we present a logic controller based on Paraconsistent Annotated Logic named Paracontrol, which can be applied to resolve conflicts and to deal with contradictions and/or paracompleteness, by implementing a decision-making in the presence of uncertainties. Such controller was implemented in a real autonomous mobile robot Emmy II.

1 Introduction

The logical controller system proposed in this paper – Paracontrol - is an improvement of the logical controller proposed in some previous works [5], [4] and implemented in a real autonomous mobile robot Emmy [6]. Both controllers are based on a new class of non-logical system: the Paraconsistent Annotated Logic $\mathcal{E}\tau$ [1]. The atomic formulae of the logic $\mathcal{E}\tau$ is of the type $p_{(\mu, \lambda)}$, where $(\mu, \lambda) \in [0, 1]^2$ and $[0, 1]$ is the real unitary interval with the usual order relation and p denotes a propositional variable. There is an order relation defined on $[0, 1]^2$: $(\mu_1, \lambda_1) \leq (\mu_2, \lambda_2) \Leftrightarrow \mu_1 \leq \mu_2$ and $\lambda_1 \leq \lambda_2$. Such ordered system constitutes a lattice that will be symbolized by τ . $p_{(\mu, \lambda)}$ can be intuitively read: “It is believed that p ’s belief degree (or favorable evidence) is μ and disbelief degree (or contrary evidence) is λ .” So, we have some interesting examples:

- $p_{(1.0, 0.0)}$ can be read as a true proposition.
- $p_{(0.0, 1.0)}$ can be read as a false proposition.
- $p_{(1.0, 1.0)}$ can be read as an inconsistent proposition.
- $p_{(0.0, 0.0)}$ can be read as a paracomplete (unknown) proposition.
- $p_{(0.5, 0.5)}$ can be read as an indefinite proposition.

Note, the concept of paracompleteness is the “dual” of the concept of inconsistency.

The Paracontrol is the electric-electronic materialization of the Para-analyzer algorithm [5], which is basically an electronic circuit, which treats logical signals in a context of logic $\mathcal{E}\tau$. Such circuit compares logical values and determines domains of a state lattice corresponding to output value. Favorable evidence and contrary evidence degrees are represented by voltage. Certainty and contradiction degrees are determined by analogues of operational amplifiers. The Paracontrol comprises both analog and digital systems and it can be externally adjusted by applying positive and negative voltages. The Paracontrol was tested in real-life experiments with an autonomous mobile robot Emmy, whose favorable/contrary evidences coincide with the values of ultrasonic sensors and distances are represented by continuous values of voltage.

The Emmy can act in an adequate way in “special” situations, such as when she faces contradictory information: one sensor may detect an obstacle ahead (for instance, a wall) while the other detects no obstacle (for example, it may be in direction to an opened door). In such situation, the Emmy stops and turns 45° to the more free direction. Next, if in a new measure, there is no inconsistency, the robot may take another decision, for example, to go ahead. This work presents some improvements regarding the controller described. The proposed system, in which we keep the same name, Paracontrol, uses 6 (six) logical states. The main characteristics are:

1. Velocity control: the Paracontrol allows the robot more softly movements, for instance, to brake “in a smoothly way”. Also, when it faces with a contradictory situation, the robot turns (right or left) “in a smoothly way”.
2. The new controller allows also backward motion. In some situations the robot may move backward or turns with a fixed wheel and the other spinning around backward. There aren’t these kinds of movements in the original Emmy robot.
3. The combination of the above characteristics and others presented in the original prototype makes the new one, a robot with more “sophisticated” movements. It represents an important step on the development of autonomous mobile robots.

The autonomous mobile robot built with the new Paracontrol was named Emmy II.

2 The Autonomous Mobile Robot Emmy II

The platform used to assemble the Emmy II robot measures approximately 23cm height and 25cm of diameter (circular format). The main components of Emmy II are a microcontroller of 8051 family, two ultrasonic sensors, and two DC motors. Figure 1 shows the Emmy II basic structure. The ultrasonic sensors are responsible in verifying whether there is any obstacle in front of the robot. The signals generated by sensors are sent to the microcontroller. These signals are used to determine the favorable evidence degree μ and the contrary evidence degree λ regarding the proposition “There is no obstacle in front of the robot”. Then the Paracontrol, recorded in the internal memory of the microcontroller uses in order to determine the robot movements. Also, the microcontroller is responsible to apply power to the DC motors (Figures 2, 3, & 4).

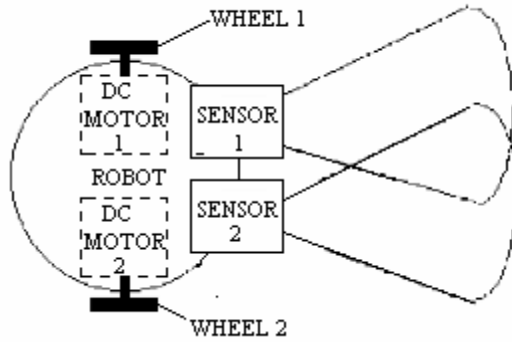


Fig. 1. Basic structure of Emmy II

The figure 2 shows the simplified block diagram of Emmy II.

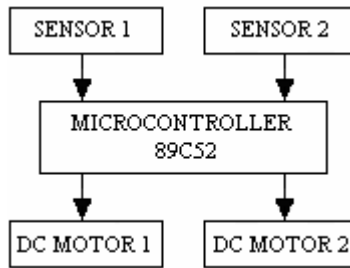


Fig. 2. Block diagram

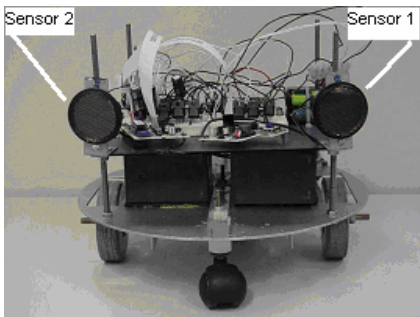


Fig. 3. Frontal vision of Emmy II

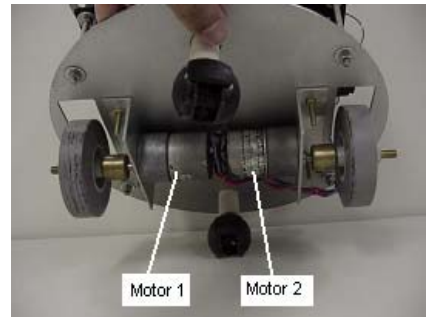


Fig. 4. Inferior vision of Emmy II

3 Emmy II Robot Program

The most important component is the 89C52 microcontroller, so it is responsible to determine the distances between sensor and the obstacles, and to calculate the favorable and contrary evidence degrees, to perform the Para-analyzer algorithm, and to generate the signals to the DC motors.

The possible movements are the following:

- Robot goes ahead. DC motors 1 and 2 are supplied for spinning around forward.
- Robot goes back. DC motors 1 and 2 are supplied for spinning around backward.
- Robot turns right. Just DC motor 1 is supplied for spinning around forward.
- Robot turns left. Just DC motor 2 is supplied for spinning around forward.
- Robot turns right. Just DC motor 2 is supplied for spinning around backward.
- Robot turns left. Just DC motor 1 is supplied for spinning around backward.

The signal generated by the sensor 1 is considered the favorable evidence degree and the signal generated by the sensor 2 is considered the contrary evidence degree. Thus, when there is an obstacle near the sensor 1, the favorable evidence degree is low and when there is an obstacle far from the sensor 1, the favorable evidence degree is high. Otherwise, when there is an obstacle near the sensor 2, the contrary evidence degree is high and when there is an obstacle far from the sensor 2, the contrary evidence degree is low. The decision-making of the robot of what movement to do is based on favorable/contrary evidences and the control system proposed, according to the state lattice corresponding to output value as shown in figure 5.

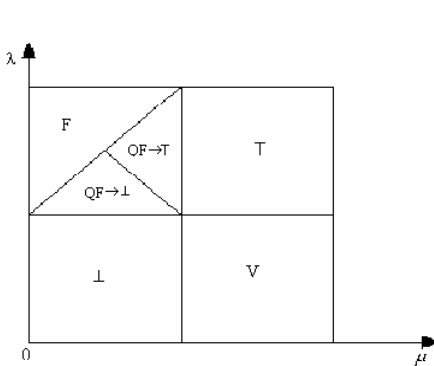


Fig. 5. Logical output lattice of Emmy II

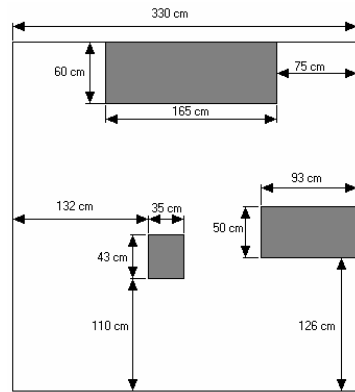


Fig. 6. Test environment of Emmy II

The verification of the favorable evidence degree and the contrary evidence degree, as well as the decision taken and the DC motors supply are made sequentially. Such sequence of actions is almost imperceptible when the robot is moving.

The decision for each logic state is the following:

Table 1. Logical states and action

Symbol	State	Action
V	True	Robot goes ahead
F	False	Robot goes back
\perp	Paracomplete	Robot turns right
T	Inconsistent	Robot turns left
$QF \rightarrow \perp$	Quasi-false tending to paracomplete	Robot turns right
$QF \rightarrow T$	Quasi-true tending to inconsistent	Robot turns left

The justification for each decision is the following:

When the logical state is true (V), it means that the front of the robot is free. So, the robot may go ahead.

In the inconsistency (T), favorable/contrary evidences μ and λ are high (i.e., belong to state T). It means that the sensor 1 is far from an obstacle and the sensor 2 is near an obstacle, so the left side is freer than the right side. Then, the action should be to turn left by supplying only DC motor 2 for spinning around forward and keeping DC motor 1 stopped.

When the paracompleteness (\perp) is detected, μ and λ are low. It means that the sensor 1 is near an obstacle and the sensor 2 is far from an obstacle, so the right side is freer than the left side. Then, the action should be to turn right by supplying only DC motor 1 for spinning around forward and keeping DC motor 2 stopped.

In the false state (F) there are obstacles very near in front of the robot. Therefore the robot should go back.

In the $QF \rightarrow T$ state, the front of the robot is obstructed but the obstacle is not so near as in the false state and the left side is a little bit freer than the right side. So in this case, the robot should turn left by supplying only the DC motor 1 for spinning around backward and keeping DC motor 2 stopped.

In the $QF \rightarrow \perp$ state, the front of the robot is obstructed but the obstacle is not so near as in the false state and the right side is a little bit freer than the left side. So, in this case the robot should turn right by supplying only the DC motor 2 for spinning around backward and keeping DC motor 1 stopped.

4 Tests and Conclusions

In order to verify Emmy II robot performance, we've performed 4 tests. Basically, counting how many collisions there were while the robot moved in an environment similar to the one showed in figure 6. The ultrasonic sensor used by Emmy II robot can't detect obstacles closer than 7,5 cm. The sensors transmit sonar pulses and wait for their echo in order to determine the distance between the sensors and the obstacles; nevertheless, sometimes the echo don't return; it can reflect to another direction. These are the main causes to the robot collisions. By adding more suitable sensors and adapting the Paracontrol we can expect better results

The time duration and results for each test were the following:

Table 2. Time period and number of collisions

Test	Time Period	Number of Collisions
1	3 minutes and 50 seconds	13
2	3 minutes and 10 seconds	7
3	3 minutes and 30 seconds	10
4	2 minutes and 45 seconds	10

The ultrasonic sensor used by Emmy II robot can't detect obstacles closer than 7,5 cm. The sensors transmit sonar pulses and wait for their echo in order to determine the distance between the sensors and the obstacles; nevertheless, sometimes the echo don't return; it can reflect to another direction. These are the main causes to the robot collisions. By adding more suitable sensors and adapting the Paracontrol we can expect better results.

The robot collision causes are the following:

Test 1: Collisions: 13.

Collisions caused by echo reflection: 4.

Collisions caused by too near obstacles: 9.

Test 2: Collisions: 7.

Collisions caused by echo reflection: 2.

Collisions caused by too near obstacles: 5.

Test 3: Collisions: 10.

Collisions caused by echo reflection: 5.

Collisions caused by too near obstacles: 5.

Test 4: Collisions: 10.

Collisions caused by echo reflection: 4.

Collisions caused by too near obstacles: 6.

There is another possibility of collision which is when the robot is going back. As there is no sensors behind the robot, it may collide. All these questions will be studied in forthcoming works.

References

1. J.M. Abe, "Fundamentos da Lógica Anotada" (Foundations of Annotated Logics), in Portuguese, Ph. D. Thesis, Universidade de São Paulo, São Paulo, 1992.
2. C.R. Torres, *Sistema Inteligente Paraconsistente para Controle de Robôs Móveis Autônomos*, MSc. Dissertation, Universidade Federal de Itajubá - UNIFEI, Itajubá, 2004.
3. J.M. Abe, Some Aspects of Paraconsistent Systems and Applications, *Logique et Analyse*, 157(1997), 83-96.
4. J.M. Abe & J.I. da Silva Filho, Manipulating Conflicts and Uncertainties in Robotics, *Multiple-Valued Logic and Soft Computing*, V.9, ISSN 1542-3980, 147-169, 2003.

5. J.I. da Silva Filho - *Métodos de Aplicações da Lógica Paraconsistente Anotada de Anotação com Dois Valores LPA2v com Construção de Algoritmo e Implementação de Circuitos Eletrônicos*, in Portuguese, Ph. D. Thesis, Universidade de São Paulo, São Paulo, 1999.
6. J.I. da Silva Filho & J.M. Abe – “Emmy: a paraconsistent autonomous mobile robot”, in Logic, Artificial Intelligence, and Robotics, *Proc. 2nd Congress of Logic Applied to Technology – LAPTEC’2001*, Edts. J.M. Abe & J.I. Da Silva Filho, Frontiers in Artificial Intelligence and Its Applications, IOS Press, Amsterdam, Ohmsha, Tokyo, Vol. 71, ISBN 1 58603 206 2 (IOS Press), 4 274 90476 8 C3000 (Ohmsha), ISSN 0922-6389, 53-61, 287p., 2001.

EVALPSN Based Intelligent Drivers' Model

Kazumi Nakamatsu¹, Michiro Kondo², and Jair M. Abe³

¹ University of Hyogo, HIMEJI, Japan

`nakamatu@shse.u-hyogo.ac.jp`

² Tokyo Denki University, INZAI, CHIBA, Japan

`kondo@sie.dendai.ac.jp`

³ University of Sao Paulo, Paulista University, SAO PAULO, Brazil

`jairabe@uol.com.br`

Abstract. We introduce an intelligent drivers' model for traffic simulation in a small area including some intersections, which models drivers' decision making based on defeasible deontic reasoning and can deal with minute speed change of cars in the simulation system. The intelligent model is computed based on defeasible deontic reasoning by a paraconsistent annotated logic program EVALPSN.

Keywords: EVALPSN(Extended Vector Annotated Logic Program with Strong Negation), defeasible deontic reasoning, paraconsistent annotated logic program, intelligent traffic simulation.

1 Introduction

EVALPSN(Extended Vector Annotated Logic Program) that can deal with defeasible deontic reasoning [7] was introduced by K.Nakamatsu et al.[1,2] and applied to various kinds of control such as traffic signal control [6] and robot action control [4,5,3]. In order to evaluate the traffic signal control [6], we made a traffic simulation system based on the cellular automaton method that simulates each car movement around a few intersections. Basically, in the cellular automaton method, roads are divided into many cells and each cell is supposed to have one car, and car movement is simulated based on a simple cell transition rule such that "if the next cell is vacant, the car has to move into the next cell". Therefore, it does not seem that the usual cellular automaton method can simulate each car movement minutely, even though it has many other advantages such as it does not cost long time for traffic simulation.

In this paper, we introduce an intelligent model to model drivers' speed control for their cars by EVALPSN defeasible deontic reasoning, which can be used for simulating each car movement minutely in the same framework of the cellular automaton traffic simulation method.

Generally, driving actions of human being such as putting brake to slow down the car can be regarded as the result of defeasible deontic reasoning to resolve conflicts. For example, if you are driving a car, you may catch conflicting informations "there is enough distance from your car to the precedent car for

speeding up your car” and “I am driving the car at the speed limit”. The first information derives permission for the action “speed up” and the second one derives forbiddance from it. Then the forbiddance defeats the permission and you may not speed up your car. On the other hand, if you catch the information “I am driving the car at much less than the speed limit” as the second one, then this information derives permission for speeding up your car and you may speed up your car. Therefore, as shown in the example, human being decision making for action control can be done by defeasible deontic reasoning with some rules such as traffic rules We formalize such a defeasible deontic model for car speed control action in the paraconsistent logic program EVALPSN and introduce a traffic simulation system based on the drivers' model.

This paper is organized as follows : first, we review EVALPSN very briefly and introduce defeasible deontic reasoning for the drivers' model in EVALPSN ; next, we describe some sample drivers' rules to control car speed and how those rules are translated into EVALPSN cluases ; and introduce the traffic simulation system based on the EVALPSN drivers' model. We omit the details of EVALPSN in this paper.

2 EVALPSN

In EVALPSN, each literal has a truth value called *extended vector annotation* explicitly, and it is formulated in a form $p : [(i, j), \mu]$, where (i, j) belongs to a complete lattice \mathcal{T}_v and μ does to a complete lattice \mathcal{T}_d . The first component (i, j) indicates the positive information degree i and the negative one j to support the literal p , and the second one μ is an index that represents deontic notion or paraconsistency. The intuitive meaning of each member in the lattice \mathcal{T}_d is ; \perp (unknown), α (fact), β (obligation), γ (non-obligation), $*_1$ (both fact and obligation), $*_2$ (both obligation and non-obligation), $*_3$ (both fact and non-obligation) and \top (paraconsistency). The complete lattice \mathcal{T}_e of extended vector annotations is defined as the product $\mathcal{T}_v \times \mathcal{T}_d$ described in **Fig.1**.

$$\mathcal{T}_v = \{ (x, y) \mid 0 \leq x \leq n, 0 \leq y \leq n, x, y \text{ and } n \text{ are non-negative integers} \},$$

$$\mathcal{T}_d = \{ \perp, \alpha, \beta, \gamma, *_1, *_2, *_3, \top \}.$$

3 Defeasible Deontic Drivers' Model

Suppose that a man is driving a car. Then, how does the car driver decide the next action for controlling car speed such as braking or acceleration ? It is easily supposed that, for example, if the traffic light in front of the car is red, the driver has to slow down the car, or if there is enough distance from the driver's car to the precedent car, the driver may speed up the car. If we model such drivers' car speed control, we should consider conflicting informations such as “traffic light is red” and “enough distance to speed up”, and its conflict resolving. It

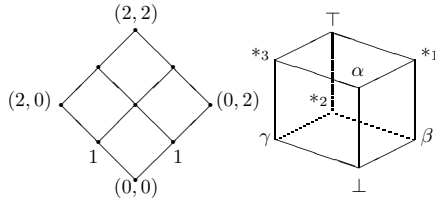


Fig. 1. Lattice $T_v(n = 2)$ and Lattice T_d

also should be considered that car drivers reason car speed control based on not only detected physical information such as the current car speed but also traffic rules such as “keep driving at less than speed limit”. For example, if a driver is driving a car over the speed limit of the road, the driver would slow down the car even if there is no car ahead of the car, then, it is supposed that there exists strong forbiddance from driving over the speed limit, and eventually it may turn into obligation to slow down the car. On the other hand, if a driver is driving a car at very slow speed, the driver would speed up the car even if the traffic light far ahead of the car is red, then, it is also supposed that there exist both strong permission and weak forbiddance to speed up the car, then only the permission is obtained by defeasible deontic reasoning, and eventually it may turn into obligation to speed up the car. Therefore, we can easily model such drivers’ decision making for car speed control by EVALPSN defeasible deontic reasoning as described in the above example. In this section, we introduce the EVALPSN drivers’ model that can derive the three car speed control actions, “slow down”, “speed up”, or “keep the current speed” in EVALPSN programming. We define the drivers’ model as follows.

3.1 Framework for EVALPSN Drivers’ Model

1. Forbiddance or permission for the car speed control action, “speed up” are derived based on the traffic rules,
 - it is obligatory to obey traffic signal,
 - it is obligatory to keep the speed limit, etc.,
 and the following detected information,
 - the object car speed,
 - the precedent car speed,
 - the distance between the precedent and objective cars,
 - the distance to the intersection or the curve ahead of the objective car ;
2. obligation for one of the three car speed control actions, “speed up”, “slow down” and “continue the current speed” is derived by defeasible deontic reasoning in EVALPSN programming ;
3. basically, a similar method to the cellular automaton method is used as the traffic simulation method.

3.2 Annotated Literals in Drivers' Model

In the EVALPSN drivers' model, the following annotated literals are used to represent various information,

$mv(t)$ represents one of the three car actions, "speed up", "slow down", or "keep the current speed" at the time t ; this predicate has the complete lattice \mathcal{T}_v of vector annotations,

$$\mathcal{T}_v = \{ \begin{array}{ll} (0, 0) \text{ "no information",} & (1, 0) \text{ "weak speed up",} \\ (0, 1) \text{ "weak slow down",} & (2, 0) \text{ "strong speed up",} \\ (0, 2) \text{ "strong slow down",} & \dots, (2, 2) \end{array} \},$$

for example, if we have the EVALP clause $mv(t) : [(0, 1), \beta]$, it represents the weak forbiddance from the action "speed up" at the time t , on the other hand, if it has the annotation $[(2, 0), \gamma]$, it represents the strong permission for the action "slow down", etc. ;

$v_o(t)$ represents the speed of the objective car at the time t , then we suppose the complete lattice of vector annotations for representing the objective car speed,

$$\mathcal{T}_v = \{(i, j) | i, j \in \{0, 1, 2, 3, 4, 5\}\},$$

we may have the following informal interpretation, if we have the EVALP clause $v_o(t) : [(2, 0), \alpha]$, it represents that the car is moving forward at the speed of over 20km/h at the time t , on the other hand, if we have the EVALP clause $v_o(t) : [(0, 1), \alpha]$, it represents that the car is moving backward at the speed of over 10km/h at the time t , etc. ;

$v_n(t)$ represents the speed of the precedent car at the time t ; the complete lattice structure and informal interpretation of the vector annotations are the same as the case of the predicate $v_o(t)$;

$v_o(s, t)$ represents the speed of the oncoming car at the time t , the vector annotations are as well as the predicate $v_o(t)$;

$d_p(t)$ represents the distance between the precedent and the objective cars at the time t ; the complete lattice of vector annotations for representing the distance,

$$\mathcal{T}_v = \{(i, j) | i, j \in \{0, 1, 2, \dots, n\}\},$$

if we have the EVALP clause $d_p(t) : [(2, 0), \alpha]$, it represents that the distance is more than 2 cells at the time t , moreover, if we have the EVALP clause $d_p(t) : [(5, 0), \beta]$, it represents that the distance has to be more than 5 cells at the time t , on the other hand, if we have the EVALP clause $d_p(t) : [(0, 3), \beta]$, it represents that the distance must not be more than 3 cells at the time t , etc. ;

$d_c(t)$ represents the distance from the objective car to the curve in front of the car at the time t ; the complete lattice structure and informal interpretation of the vector annotations are the same as the case of the predicate $d_p(t)$;

$d_f(t)$ represents the distance between the oncoming and the cars at the time t , the vector annotations are as well as the predicate $d_p(t)$;

$go(t)$ represents the direction where the objective car turns to at the time t ; this predicate has the complete lattice of vector annotations,

$$\mathcal{T}_v = \{ \begin{array}{ll} (0,0) \text{ "no information",} & (1,0) \text{ "right turn",} \\ (0,1) \text{ "left turn",} & (2,0) \text{ "right turn",} \\ (0,2) \text{ "left turn",} & \dots, (2,2) \end{array} \},$$

if we have the EVALP clause $go(t) : [(2,0), \alpha]$, it represents that the car turns to the right at the time t , if we have the EVALP clause $go(t) : [(0,2), \beta]$, it represents that the car must not turn to the right, that is to say, must turn to the right, etc..

3.3 Inference Rules in Drivers' Model

We also have some inference rules to derive the next car control action in the EVALPSN drivers' model and introduce the basic three inference rules, **Traffic Signal Rule**, **Straight Road Rule** and **Curve and Turn Rule**. We suppose that there is a cross intersection with a traffic light in front of the objective car in the following rules.

Traffic Signal Rule. if the traffic light indicates

- **red**, it is considered as there is an obstacle on the stop line before the traffic light, that is to say, there is strong forbiddance from entering into the intersection ;
- **yellow**, it is considered as the same as the red light rule except that if the distance between the car and the stop line is less than 2 cells, it is weakly permitted for entering into the intersection ;
- **green**, it has no forbiddance from going into the intersection except that if the car turning at the intersection, it is described in **Curve and Turn Rule**.

Straight Road Rule. if the road is straight, the objective car behavior is decided by

- distance between the precedent car and the objective car ;
- each speed of the precedent car and the objective car ;
- obeying the traffic rule, speed limit of roads and traffic signal, etc..

Curve and Turn Rule. if the objective car is headed to the curve or going to turn at the intersection, forbiddance to speed up the car is derived.

Basic idea of the EVALPSN Drivers' model based simulation is as follows : as the first step, forbiddance or permission for one of the car actions, "speed up" or "slow down", are derived by EVALPSN defeasible deontic reasoning ; as the next step, if the forbiddance for a car action is derived, the objective car has to do the opposite action ; if the permission for a car action is derived, the objective car has to do the action ; if neither forbiddance nor permission is derived, the objective car does not have to do any action, that is to say, it has to keep the current speed. We show an example for the EVALPSN drivers' model.

3.4 Example of EVALPSN Drivers' Model

suppose that the objective car is moving at the speed of 1, then we have the following EVALP clauses to reason the next action of the objective car according to the current information.

Case 1. If the distance between the precedent car and the objective car is longer than 2 cells, we have permission to accelerate the car at the time t . This rule is translated into the EVALP clause,

$$v_o(t):[(1, 0), \alpha] \wedge d_p(t):[(2, 0), \alpha] \rightarrow mv(t):[(0, 1), \gamma]. \quad (1)$$

Case 2. If the precedent car stopped at the next cell and the objective car is moving at the speed of 1, we have strong forbiddance from speed up at the time t , which means strong obligation to slow down. This rule is translated into the EVALP clause,

$$v_o(t):[(1, 0), \alpha] \wedge v_n(t):[(0, 0), \alpha] \wedge d_p(t):[(0, 0), \alpha] \rightarrow mv(t):[(0, 2), \beta]. \quad (2)$$

Case 3. If the precedent car is faster than the objective car whose speed is 1, we have permission to accelerate the objective car at the time t . This rule is translated into the EVALP clause,

$$v_o(t):[(1, 0), \alpha] \wedge v_n(t):[(2, 0), \alpha] \rightarrow mv(t):[(0, 1), \gamma]. \quad (3)$$

Case 4. If the car is moving at the speed of 3 and the distance between the car and the curve is 2 cells at the time t . This rule is translated into :

$$v_o(t):[(3, 0), \alpha] \wedge d_c(t):[(2, 0), \alpha] \wedge go(t):[(2, 0), \alpha] \rightarrow mv(t):[(0, 1), \beta] \quad (4)$$

If both the permission $mv(t):[(0, 1), \gamma]$ and the forbiddance $mv(t):[(0, 2), \beta]$ from speed up are derived, we have obligation to slow down the objective car at the next step by defeasible deontic reasoning, since the forbiddance is stronger than the permission.

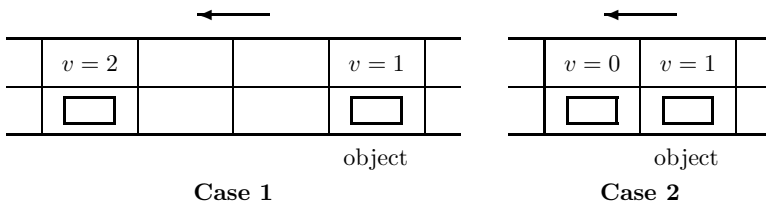


Fig. 2. Cell States in the **Case 1** and **2**

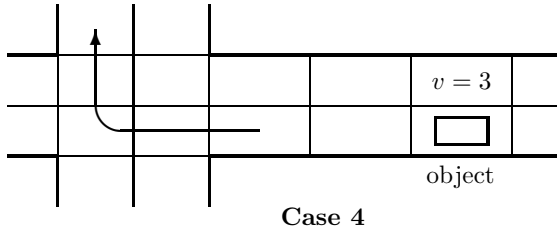


Fig. 3. Cell States in the Case 4

3.5 Traffic Signal Simulation System

The **Fig. 4** shows the drivers' model based traffic signal simulation around a typical cross intersection with traffic lights. In the figure, each square box with an integer 0 to 4 indicates a car, and the integer indicates its speed at that time. For example, if the integer 2 is attached to a car, it indicates that the car is moving at the speed of 20km/h. In the traffic signal simulation, one of the three car control actions, "speed up", "slow down" or "keep the current speed" is reasoned for each car in the simulation area based on EVALPSN drivers' model. Moreover, the simulation system can simulate the traffic signal control based on

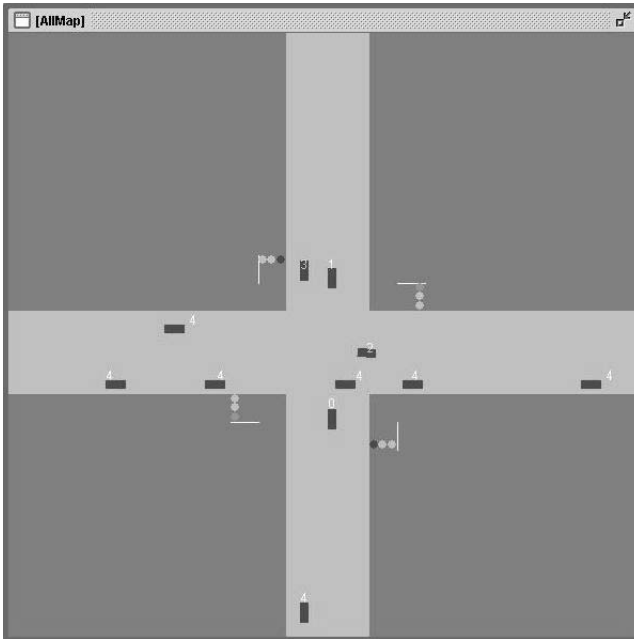


Fig. 4. Traffic Simulation at Intersection

EVALPSN defeasible deontic reasoning [6] in which the length of each traffic light (red, yellow, green, etc.) is controlled by EVALPSN programming.

4 Conclusion

In this paper, we have introduced a drivers' model based on EVALPSN as an application of the defeasible deontic model and its simulation system. We also have a similar problem in simulation of railway train operation. Actually, we have already developed train operators' model as another application of the defeasible deontic model and applied the defeasible deontic model to railway train simulation, which is used for simulating train speed precisely and can also be used for delay estimation or recovery of train schedules. We could not construct the precise railway operation simulator without the train operators' model based on EVALPSN defeasible deontic reasoning. However, the defeasible deontic drivers' model simulation costs much time to compute each car speed. Therefore, it is necessary to consider simulation time reduction when the drivers' model is implemented.

References

1. Nakamatsu,K., Abe,J.M., and Suzuki,A., A Defeasible Deontic Reasoning System Based on Annotated Logic Programming, *Proc. the Fourth International Conference on Computing Anticipatory Systems*, AIP Conference Proceedings **573**, pp.609–620, AIP, 2001.
2. Nakamatsu,K., Abe,J.M., and Suzuki,A., Annotated Semantics for Defeasible Deontic Reasoning, *Proc. the Second International Conference on Rough Sets and Current Trends in Computing*, LNAI **2005**, pp.432–440, Springer-Verlag, 2001.
3. Nakamatsu,K., Abe,J.M., and Suzuki,A., Defeasible Deontic Robot Control Based on Extended Vector Annotated Logic Programming, *Proc. the Fifth International Conference on Computing Anticipatory Systems*, AIP Conference Proceedings **627**, pp.490–500, AIP, 2002.
4. Nakamatsu,K., Mita,Y., and Shibata,T., “An Intelligent Action Control System Based on Extended Vector Annotated Logic Program and its Hardware Implementation”, *Intelligent Automation and Soft Computing*, **12**, TSI Press, 2006 (to appear).
5. Nakamatsu,K., Mita,Y., Shibata,T., and Abe,J.M., Defeasible Deontic Action Control Based on Paraconsistent Logic Program and its Hardware Implementation, *Proc. 3rd International Conference on Computational Intelligence for Modelling Control and Automation* (CD-ROM), 2003.
6. Nakamatsu,K., Seno,T., Abe,J.M., and Suzuki,A., “Intelligent Real-time Traffic Signal Control Based on a Paraconsistent Logic Program EVALP”, *Proc. the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, LNCS **2639**, pp.719–723, Springer-Verlag, 2003.
7. Nute,D.(ed.) *Defeasible Deontic Reasoning*, Synthese Library, **263**, Kluwer Academic Publishers, 1997.

The Study of the Robust Learning Algorithm for Neural Networks

Shigenobu Yamawaki

Department of Electric and Electronic Engineering,
School of Science and Engineering
Kinki University, Osaka, 577-8502, Japan
yamawaki@ele.kindai.ac.jp

Abstract. In this paper, we propose the robust learning algorithm for neural networks. The suggested algorithm is obtaining the expanded Kalman filter in the Krein space. We show that this algorithm can be applied to identify the nonlinear system in the presence of the observed noise and system noise.

1 Introduction

We know that neural networks are applied to identify the nonlinear system [1] and pattern recognition [2]. The error back propagation method applied to these conveniently is useful learning law of neural networks [3]. Furthermore, the error back propagation method using the least-squares method for their parameter estimation based on the error back propagation method is proposed [4]. These methods do not provide robustness, since these learning methods minimize fundamentally square error. On the other hand, in linear system theory, robust estimate methods are continually studied based on the Kalman filter algorithm [5], [6]. This report proposes the learning method that guaranteed robustness to the neural network based on the Kalman filter algorithm. The learning method of the neural network that repeats the following three procedures is suggested. The neural network is linearized to describe the input-output relation in the Markov process. The Markov parameters are presumed applying the expanded Kalman filter in the Klein space. The parameters of the neural network are determined by minimum realization of the estimated Markov parameter. The verification of robustness of this method is applied to identify the nonlinear system in presence of the noise.

2 The Robust Learning Algorithm of the Neural Network

In this paper, we consider the robust learning algorithm of the neural network (NN) as described as follows :

$$\left. \begin{aligned} x_N(t+1) &= A_N o_N(t) + B_N u(t) + \theta_N, \\ o_N(t) &= f(x_N(t)), \\ f(x_N(t)) &= [f_1(x_{N1}(t)) f_2(x_{N2}(t)) \cdots f_n(x_{Nn}(t))]^T, \\ f_i(x) &= \lambda \left\{ \frac{2}{1 - \exp(-x/q_s)} - 1 \right\} \\ y_N(t) &= C_N o_N(t) + v(t+1) \end{aligned} \right\} \quad (1)$$

where $x_N(t)$, $o_N(t)$ and $u(t)$ are n -dimensional states, the same dimensional output of the hidden layer and q -dimensional input of the NN at the step t . θ_N is the threshold value of the NN at the step t . The weights parameters A_N , B_N and C_N are appropriately size coefficient matrices of each layer of the NN. The sigmoid function $f_i(x)$ is the amplitude λ and q_s slope. The variable is p -dimensional expanded output of the NN. $v(t + 1)$ is a noise.

The learning law suggested consists of the following three steps.

1. The neural network is approximated to linear system.
2. Impulse response functions are estimated on Krein space.
3. The parameters of the neural network are determined from the minimum realization method.

The suggested learning law acquires the robustness by applying the Kalman filter algorithm on Krein space in the 2nd step. To overcome this, expanding the sigmoid function $f(x)$ into a Taylor series around the threshold and neglecting terms higher than the first order for linearization, we can obtain the following linear model of the neural network.

$$o_N(t + 1) = f(x_N(t + 1))_{x_N(t+1)=\theta_N} + A_L o_N(t) + B_L u(t) \tag{2}$$

$$y_L(t + 1) = C_N o_N(t + 1) + v(t + 1) \tag{3}$$

where the A_x , A_L and B_L and are expressed by

$$A_x = \frac{\partial}{\partial x_N^T} f_N(x_N(t))_{x_N(t)=\theta_N}, \quad A_L = A_x A_N, \quad B_L = A_x B_N.$$

The response of the neural network is shown with the impulse response functions as follows.

$$y_L(t + 1) = g_0 + g_1 u(0) + \dots + g_n u(n - 1) + \dots + v(t + 1) \tag{4}$$

Furthermore, the matrix expression of the linear model (4) is (5).

$$y_L(t + 1) = Z^T(t + 1)\Theta(t) + v(t + 1) \tag{5}$$

where $Z^T(t + 1) = [I_p u^T(t) \otimes I_p \dots u^T(t + 1 - k) \otimes I_p]$, $\Theta(t) = cs[g_0 g_1 \dots g_k]$ and $g_i = C_N A_L^{t+1-i} B_L$. $cs(\bullet)$ means the column string of the (\bullet) . We can derive the Kalman filter algorithm to estimate the impulse response function of a linear model that achieves the following performance function.

$$\sup_{\Theta(0), v(t)} \frac{\sum_{t=0}^n \|e_f(t)\|^2}{\left\| \Theta(0) - \hat{\Theta}(0) \right\|_{\Sigma_0^{-1}}^2 + \sum_{t=0}^n \|v(t)\|^2} < \gamma_f \tag{6}$$

where $e_f(t) = \hat{z}(t|t) - z(t)$ and $z(t)$ is the linear combination $L(t)$ and $\Theta(t)$, namely $z(t) = L(t)\Theta(t)$. $\hat{z}(i|i)$ is estimation of $z(t)$ using $\{y(t)\}_{k=0}^i$. $\gamma_f (> 0)$ is a given scalar.

Consequently, by applying the Kalman filter to the linear model in the Krein space, we can drive the robust learning algorithm for the neural network as

$$\hat{\Theta}_{t+1} = \hat{\Theta}_t + K_{t+1}(y(t+1) - y_{NL}(t+1)), \tag{7}$$

$$K_{t+1} = \hat{P}_{t+1|t} Z_{t+1} (Z_{t+1}^T \hat{P}_{t+1|t} Z_{t+1} + I)^{-1} \tag{8}$$

$$\begin{aligned} \hat{P}_{t+1|t} &= \hat{P}_{t|t-1} - \hat{P}_{t|t-1} [Z_t \ Z_t] \\ &\quad \times R_{et}^{-1} \begin{bmatrix} Z_t^T \\ Z_t^T \end{bmatrix} \hat{P}_{t|t-1} \end{aligned} \tag{9}$$

$$R_{et} = R_t + \begin{bmatrix} Z_t^T \\ Z_t^T \end{bmatrix} \hat{P}_{t|t-1} [Z_t \ Z_t] \tag{10}$$

$$R_t = \begin{bmatrix} I & 0 \\ 0 & -\gamma_f^2 I \end{bmatrix} \tag{11}$$

where $L(t) = Z^T(t) = Z_t^T$ and $y_{NL}(t+1)$ the is the output of (12) and (13). The parameters of the neural network are obtained from the impulse response function of the linear model using the minimum realization method of linear theory.

$$x_N(t+1) = \hat{A}_N o_N(t) + \hat{B}_N u(t) + \hat{\theta}_N \tag{12}$$

$$y_N(t+1) = \hat{C}_N o_N(t+1) \tag{13}$$

3 Examples

We apply to the identification problem of a nonlinear system in presence of the system noise and the observation noise to verify the suggested learning law; where $w^T(t) = [w_1(t)w_2(t)]$ and $v^T(t) = [v_1(t)v_2(t)]$ are given by the Gaussian white noise of the average zero, variance σ_v, σ_w , respectively. In the estimation, the number of data was taken to be 500. For $\gamma_f = 1.5$, the estimation result of the NN with $n = 6$ is shown in Fig. 1.

$$\left. \begin{aligned} x(t+1) &= \begin{bmatrix} 0.3 & 0.4 \\ -0.4 & 0.2 \end{bmatrix} x(t) + u_1(t) \begin{bmatrix} 0.0 & 0.2 \\ 0.3 & 0.0 \end{bmatrix} x(t) \\ &\quad + u_2(t) \begin{bmatrix} 0.0 & 0.4 \\ 0.0 & -0.2 \end{bmatrix} x(t) + \begin{bmatrix} 1.0 & 0.0 \\ 0.2 & 1.0 \end{bmatrix} x(t) + w(t) \\ y(t) &= \begin{bmatrix} 1.0 & -0.3 \\ 0.4 & 1.0 \end{bmatrix} x(t) + v(t) \end{aligned} \right\} \tag{14}$$

The simulation result of BP is also shown in Fig.1. It is clear from Fig.1 that the estimated accuracy of this method is obtained comparably BP method with the learning rate 0.01.

Then, for this method and BP, Table1 illustrates the evaluation result of the robustness over each of noise using (15).

$$\frac{|\text{AIC}_{\text{Noise}} - \text{AIC}_{\text{NoiseFree}}|}{|\text{AIC}_{\text{NoiseFree}}|} \tag{15}$$

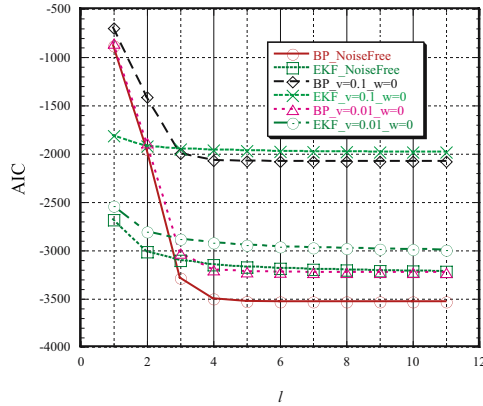


Fig. 1. The estimate result of each noise for this method and BP

Table 1. Robustness over each noise for this method and BP

	BP [%]	EKF [%]
$\sigma_v = 0.1, w(t) = 0$	41.1	38.3
$\sigma_v = 0.01, w(t) = 0$	8.7	6.9
$v(t) = 0, \sigma_w = 0.1$	45.8	33.3
$v(t) = 0, \sigma_w = 0.5$	86.4	86.5
$\sigma_v = 0.1, \sigma_w = 0.1$	60.6	55.0
$\sigma_v = 0.01, \sigma_w = 0.1$	53.8	51.2
$\sigma_v = 0.1, \sigma_w = 0.5$	87.5	88.7
$\sigma_v = 0.01, \sigma_w = 0.5$	84.1	83.0

In the example, we can see that this method is able to improve about 10% robustness rather than BP of a noise smaller than $\sigma_w = 0.5$. For the larger noise than $\sigma_w = 0.5$, it is shown that BP has more robustness than this method.

4 Conclusion

The robust algorithm of a neural network is suggested by obtaining the expanded Kalman filter in the Krein space. Namely, the proposed algorithm is deriving the linear model of a neural network, estimating impulse response functions using an expansion Kalman filter algorithm, and calculating the parameters of a neural network at the end. In case the system noise is small, it is shown that this is able to be obtained the estimate accuracy more robust about 10% than the classical BP method.

References

- [1] S. Chen, S. A. Billings and P. M. Grant : Non-linear system identification using neural networks; INT. J. CONTROL, Vol. 51, No. 6, 1191/1214, (1990)
- [2] C. M. Bishop; Neural Networks for Pattern Recognition; Oxford, U.K. Clarendon
- [3] R. J. Williams and D. Zipser: A Learning Algorithm for Continually Running Fully Recurrent Neural Networks; Neural Computation 1, 270/280, (1989)
- [4] S. Yamawaki: A study of Learning Algorithm for Expanded Neural Networks; Proc. KES 2002, 358/363, (2002)
- [5] R. E. Kalman: A new approach to linear filtering and prediction problem; J. Basic Eng., Vol.82, 35/45, (1960)
- [6] B. Hassibi, A. H. Sayed and T. Kailath: Linear Estimation in Krein Spaces Part I & Part II; IEEE Tran. A.C Vol. 41, No. 1, 18/33 & 34/49, (1996)

Logic Determined by Boolean Algebras with Conjugate

Michiro Kondo¹, Kazumi Nakamatsu², and Jair Minoro Abe³

¹ School of Information Environment, Tokyo Denki University, Japan
kondo@sie.dendai.ac.jp

² School of Human Science and Environment, University of Hyogo, Japan
nakamatu@shse.u-hyogo.ac.jp

³ Information Technology Dept. ICET - Paulista University, Sao Paulo, Brazil
Institute for Advanced Studies - University of Sao Paulo, Brazil
jairabe@uol.com.br *

Abstract. We give an axiomatic system of a logic characterized by the class of Boolean algebras with conjugate, which has a close connection with the theory of rough sets, and prove that the logic is decidable.

1 Introduction

In [1], J.Järvinen and J.Kortelainen considered properties of lower (upper) approximation operators in rough set theory by use of the algebras with conjugate pair of maps. Let B be a Boolean algebra. A pair (f, g) of maps $f, g : B \rightarrow B$ is called *conjugate* ([2]) if, for all $x, y \in B$, the following condition is satisfied:

$$x \wedge f(y) = 0 \iff y \wedge g(x) = 0$$

Moreover if a pair (f, f) is conjugate, then f is called *self-conjugate*. If a Boolean algebra has a pair of conjugate maps, then we say it simply the Boolean algebra with conjugate. By \mathbf{B} we mean the class of all Boolean algebras with conjugate.

To apply their algebraic results to the theory of data analysis, data mining and so on, we need to consider those results in the frame work of the theory of logic. In this short paper we consider the logic determined by the Boolean algebras with conjugate and show that the logic named by K_t^* is characterized by \mathbf{B} , that is, for the class Φ of all formulas of K_t^* ,

$\vdash_{K_t^*} A$ if and only if $\xi(A) = 1$ for every function $\xi : \Phi \rightarrow B$ on any $B \in \mathbf{B}$.

2 Tense Logic K_t^*

We define a certain kind of tense logic named K_t^* here. The logic is obtained from the minimal tense logic K_t by removing the axioms $(sym) : A \rightarrow GPA$, $A \rightarrow HFA$ and $(cl) : GA \rightarrow GGA$, $HA \rightarrow HHA$.

* This work was partially supported by Grant-Aid for Scientific Research (No. 15500016), Japan Society for the Promotion of Science.

Let Φ_0 be a countable set $\{p_0, p_1, p_2, \dots\}$ of propositional variables and $\wedge, \vee, \rightarrow, \neg, G, H$ be logical symbols. A formula of K_t^* is defined as follows:

- (1) Every propositional variable is a formula;
- (2) If A and B are formulas, then so are $A \wedge B, A \vee B, A \rightarrow B, \neg A, GA, HA$.

Let Φ be the set of all formulas of K_t^* . A logical system K_t^* has the following axioms and rules of inference ([3]):

- (1) $A \rightarrow (B \rightarrow A)$
- (2) $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$
- (3) $(\neg A \rightarrow \neg B) \rightarrow (B \rightarrow A)$
- (4) $G(A \rightarrow B) \rightarrow (GA \rightarrow GB), H(A \rightarrow B) \rightarrow (HA \rightarrow HB)$
- (MP) Deduce B from A and $A \rightarrow B$;
- (Nec) Deduce GA and HA from A .

We list typical axioms which characterize some properties of conjugate:

- (ext) : $GA \rightarrow A, HA \rightarrow A$
 (sym) : $A \rightarrow G\neg H\neg A, A \rightarrow H\neg G\neg A$
 (cl) : $GA \rightarrow GGA, HA \rightarrow HHA$

A well-known tense logic K_t is an axiomatic extension of K_t^* , which has extra axioms (sym) and (cl), that is,

$$K_t = K_t^* + (\text{sym}) + (\text{cl})$$

A formula A is called *provable* when there is a finite sequence $A_1, A_2, \dots, A_n (= A)$ ($n \geq 1$) of formulas such that, for every i ($1 \leq i \leq n$),

- (1) A_i is an axiom;
- (2) A_i is deduced from A_j, A_k ($j, k < i$) by (MP);
- (3) A_i is done from A_j ($j < i$) by (Nec).

By $\vdash_{K_t^*} A$ (or simply $\vdash A$), we mean that A is provable in K_t^* .

A relational structure (W, R) is called a *Kripke frame*, where W is a non-empty set and R is a binary relation on it. A map $v : \Phi_0 \rightarrow \mathcal{P}(W)$ is called a *valuation* and it can be extended uniquely to the set Φ of all formulas:

- (1) $v(A \wedge B) = v(A) \cap v(B)$
- (2) $v(A \vee B) = v(A) \cup v(B)$
- (3) $v(A \rightarrow B) = v(A)^c \cup v(B)$
- (4) $v(\neg A) = v(A)^c$
- (5) $v(GA) = \{x \in W \mid \forall y((x, y) \in R \implies y \in v(A))\}$
- (6) $v(HA) = \{x \in W \mid \forall y((y, x) \in R \implies y \in v(A))\}$

We denote the extended valuation by the same symbol v .

Since, for all formulas A and B

$$\vdash_{K_t^*} A \wedge \neg A \rightarrow B \wedge \neg B, \vdash_{K_t^*} A \vee \neg A \rightarrow B \vee \neg B,$$

we define symbols \perp and \top respectively by

$$\perp \equiv A \wedge \neg A, \quad \top \equiv A \vee \neg A.$$

Then for every formula $A \in \mathcal{F}$, we have

$$\vdash_{K_t^*} \perp \rightarrow A, \quad \vdash_{K_t^*} A \rightarrow \top.$$

A structure $\mathcal{M} = (W, R, v)$ is called a *Kripke model*, where (W, R) is a Kripke frame and v is a valuation on it. Let $\mathcal{M} = (W, R, v)$ be a Kripke model. For $x \in W$, a formula A is said to be *true* at x on the Kripke model \mathcal{M} if $x \in v(A)$, and denoted by $\mathcal{M} \models_x A$. If $v(A) = W$, that is, A is true at every $x \in W$ on the Kripke model \mathcal{M} , then A is called *true* on \mathcal{M} and denoted by $\mathcal{M} \models A$. Moreover A is called *valid* if A is true on every Kripke model \mathcal{M} and denoted by $\models A$.

It is easy to show the next result ([3]):

Theorem 1. (*Completeness Theorem*) *For every formula A , we have*

$$\vdash_{K_t^*} A \iff A : \text{valid}$$

We can get the next result by use of *filtration* method ([3]):

Theorem 2. *For every formula A , we have*

$$\vdash_{K_t^*} A \iff A : \text{true on any finite Kripke model } \mathcal{M}.$$

3 Boolean Algebra with a Conjugate Pair

Let $\mathcal{B} = (B, \wedge, \vee, ', 0, 1)$ be a *Boolean algebra*. A pair (φ, ψ) of maps $\varphi, \psi : B \rightarrow B$ is called *conjugate* if, for all $x, y \in B$,

$$x \wedge \varphi(y) = 0 \iff y \wedge \psi(x) = 0.$$

We define some properties about a map $\varphi : B \rightarrow B$ as follows:

$$\begin{aligned} \varphi : \text{extensive} &\iff x \leq \varphi(x) \quad (\forall x \in B) \\ \varphi : \text{symmetric} &\iff x \leq \varphi(y) \text{ implies } y \leq \varphi(x) \quad (\forall x, y \in B) \\ \varphi : \text{closed} &\iff y \leq \varphi(x) \text{ implies } \varphi(y) \leq \varphi(x) \quad (\forall x, y \in B) \end{aligned}$$

For a conjugate pair (φ, ψ) we have ([1])

$$\begin{aligned} \varphi : \text{extensive} &\iff \psi : \text{extensive} \\ \varphi : \text{symmetric} &\iff \varphi : \text{self-conjugate} \\ \varphi : \text{closed} &\iff \psi : \text{closed} \end{aligned}$$

We introduce two operators $\varphi^\partial, \psi^\partial$ for the sake of simplicity

$$\varphi^\partial(x) = (\varphi(x'))', \quad \psi^\partial(x) = (\psi(x'))' \quad (x \in B).$$

Then the conjugate pair (φ, ψ) can be represented by

$$\varphi(x) \leq y \iff x \leq \psi^\partial(y) \quad (x, y \in B).$$

That is, φ is a left adjoint of ψ^∂ . It is obvious from definition that

Proposition 1. *For every $x \in B$ we have*

$$\begin{aligned} \varphi : \text{extensive} &\iff \varphi^\partial(x) \leq x \\ \varphi : \text{symmetric} &\iff x \leq \varphi^\partial(\varphi(x)) \\ \varphi : \text{closed} &\iff \varphi^\partial(x) \leq \varphi^\partial(\varphi^\partial(x)) \end{aligned}$$

Let \mathbf{B} be a Boolean algebra with conjugate and $\xi : \Phi_0 \rightarrow B$ be a map. The function ξ can be extended to the set Φ of all formulas as follows:

- (1) $\xi(A \wedge B) = \xi(A) \wedge \xi(B)$
- (2) $\xi(A \vee B) = \xi(A) \vee \xi(B)$
- (3) $\xi(A \rightarrow B) = (\xi(A))' \vee \xi(B)$
- (4) $\xi(\neg A) = (\xi(A))'$
- (5) $\xi(GA) = (\varphi((\xi(A))'))' = \varphi^\partial(\xi(A))$
- (6) $\xi(HA) = (\psi(\xi(A)'))' = \psi^\partial(\xi(A))$

The following is easy to prove.

Lemma 1. *For every formula A , we have*

$$\vdash_{K_t^*} A \implies \xi(A) = 1 \text{ for all } \xi : \Phi \rightarrow B$$

We can show the converse direction of the above. At first we define a relation \equiv on the set Φ of formulas of $K_t^* : \text{For } A, B \in \Phi,$

$$A \equiv B \iff \vdash_{K_t^*} A \rightarrow B \text{ and } \vdash_{K_t^*} B \rightarrow A$$

As to the relation \equiv we have

Lemma 2. *\equiv is a congruence on Φ , that is, it is an equivalence relation and satisfies the compatible property : If $A \equiv B$ and $C \equiv D$, then*

$$\begin{aligned} A \wedge C &\equiv B \wedge D, \quad A \vee C \equiv B \vee D, \\ A \rightarrow C &\equiv B \rightarrow D, \\ \neg A &\equiv \neg B, \\ GA &\equiv GB, \quad HA \equiv HB \end{aligned}$$

Proof. We only prove that if $A \equiv B$ then $GA \equiv GB$. It follows from assumption that $\vdash A \rightarrow B$. From (Nec) we get

$$\vdash G(A \rightarrow B).$$

On the other hand, since $\vdash G(A \rightarrow B) \rightarrow (GA \rightarrow GB)$, we have from (MP)

$$\vdash GA \rightarrow GB.$$

Similarly, by $\vdash B \rightarrow A$, we get

$$\vdash GB \rightarrow GA.$$

This means that

$$GA \equiv GB.$$

Since \equiv is the congruence, we can define operations on Φ/\equiv : For $A, B \in \Phi$, we define

$$\begin{aligned} [A] \sqcap [B] &= [A \wedge B], \\ [A] \sqcup [B] &= [A \vee B], \\ [A]^* &= [\neg A], \\ \varphi([A]) &= [\neg G \neg A] = [FA], \\ \psi([A]) &= [\neg H \neg A] = [PA], \\ \mathbf{0} &= [\perp], \quad \mathbf{1} = [\top]. \end{aligned}$$

Lemma 3. $(\Phi/\equiv, \sqcap, \sqcup, *, \mathbf{0}, \mathbf{1})$ is a Boolean algebra with (φ, ψ) as a conjugate pair.

Proof. We show that (φ, ψ) is the conjugate pair. Let $[A], [B] \in \Phi/\equiv$. We have to prove

$$[A] \sqcap \varphi([B]) = \mathbf{0} \iff [B] \sqcap \psi([A]) = \mathbf{0},$$

that is,

$$[A \wedge FB] = \mathbf{0} \iff [B \wedge PA] = \mathbf{0}.$$

Suppose that $[A \wedge FB] = \mathbf{0}$. Since $\vdash A \wedge FB \rightarrow \perp$, we have $\vdash FB \rightarrow \neg A$. From (Nec) we get $\vdash HFB \rightarrow H\neg A$. Since $\vdash B \rightarrow HFB$, we also have $\vdash B \rightarrow H\neg A$. Thus we obtain $\vdash \neg(B \wedge PA)$, that is, $[B \wedge PA] = \mathbf{0}$. The converse is similar.

Lemma 4. For any formula $A \in \Phi$,

$$\vdash_{K_t^*} A \iff [A] = \mathbf{1} \text{ in } \Phi/\equiv$$

From the above, we can prove the next theorem.

Theorem 3. Let $A \in \Phi$. Then we have

$\vdash_{K_t^*} A$ if and only if $\xi(A) = 1$ for every function $\xi : \Phi \rightarrow B$ on any Boolean algebra B with conjugate.

Proof. To show the "only if part", we assume that $\not\vdash_{K_t^*} A$. Since Φ/\equiv is the Boolean algebra with conjugate, if we take a map

$$\xi : \Phi \rightarrow \Phi/\equiv, \quad \xi(A) = [A],$$

then on Φ/\equiv we get

$$\xi(A) \neq \mathbf{1}$$

by $\not\vdash_{K_t^*} A$.

We can characterize some logics by Boolean algebras with conjugate.

Theorem 4. Logical systems $K_t^* + (ext)$, $K_t^* + (sym)$, $K_t^* + (cl)$ are characterized respectively by the Boolean algebras with extensive, symmetric, closed conjugate, that is, for any formula $A \in \Phi$

- (1) $\vdash_{K_t^*+(ext)} A \iff \xi(A) = 1$ for every map $\xi : \Phi \rightarrow B$ on any Boolean algebra B with extensive conjugate;
- (2) $\vdash_{K_t^*+(sym)} A \iff \xi(A) = 1$ for every map $\xi : \Phi \rightarrow B$ on any Boolean algebra B with symmetric conjugate;
- (3) $\vdash_{K_t^*+(cl)} A \iff \xi(A) = 1$ for every map $\xi : \Phi \rightarrow B$ on any Boolean algebra B with closed conjugate.

Proof. We only show that $\xi(A) = 1$ for the typical axiom A in each case. Suppose that $\xi(A) = x \in B$.

(1) For an extensive conjugate (φ, ψ) , we have to prove that $\xi(GA \rightarrow A) = 1$. Since

$$\begin{aligned} \xi(GA \rightarrow A) = 1 &\iff \xi(GA) \leq \xi(A) \\ &\iff (\varphi(x'))' \leq x \\ &\iff x' \leq \varphi(x') \end{aligned}$$

and φ is extensive, we have $\xi(GA \rightarrow A) = 1$.

(2) Let (φ, ψ) be a symmetric conjugate. Since $\varphi = \psi$ by assumption, we have

$$\begin{aligned} \xi(A \rightarrow GPA) = 1 &\iff \xi(A) \leq \xi(GPA) \\ &\iff x \leq \varphi^\partial(\psi(x)) \\ &\iff x \leq \psi^\partial(\psi(x)) \\ &\iff \varphi(x) \leq \psi(x) \\ &\iff \varphi(x) \leq \varphi(x). \end{aligned}$$

Thus, $\xi(A \rightarrow GPA) = 1$.

(3) Suppose that (φ, ψ) is a closed conjugate. It follows from the assumption that $\varphi^\partial(x) \leq \varphi^\partial(\varphi^\partial(x))$ ($x \in B$) and hence that

$$\begin{aligned} \xi(GA \rightarrow GGA) = 1 &\iff \xi(GA) \leq \xi(GGA) \\ &\iff \varphi^\partial(x) \leq \varphi^\partial(\varphi^\partial(x)). \end{aligned}$$

This means that $\xi(A \rightarrow GPA) = 1$.

4 Decidability

It is well-known that the minimal tense logic K_t can be characterized by the class of *finite* Kripke models. Similarly we can show that K_t^* is characterized by the class \mathbf{B}^* of *finite* Boolean algebras with conjugate.

Suppose that $\not\vdash_{K_t^*} A$. There is a finite Kripke model $\mathcal{M}^* = (W, R, v)$ such that $x \notin v(A)$ for some $x \in W$, that is, $v(A) \neq W$. We construct a finite Boolean algebra B^* with conjugate from the finite Kripke model \mathcal{M}^* as follows:

$$B^* = \mathcal{P}(W)$$

$\varphi, \psi : B \rightarrow B$ are defined respectively by

$$\begin{aligned} \varphi(X) &= \{x \in B \mid R(x) \cap X \neq \emptyset\} \\ \psi(X) &= \{x \in B \mid R^{-1}(x) \cap X \neq \emptyset\}, \end{aligned}$$

where $R(x), R^{-1}(x)$ are defined by

$$R(x) = \{y \in B \mid (x, y) \in R\}, \quad R^{-1}(x) = \{y \in B \mid (y, x) \in R\}$$

We can prove the fundamental result.

Lemma 5. *B^* is a finite Boolean algebra with a conjugate pair $\varphi, \psi : B^* \rightarrow B^*$.*

Proof. It is sufficient to prove that $\varphi, \psi : B^* \rightarrow B^*$ are conjugate. That is, we have to prove that for $X, Y \subseteq W$ (i.e., $X, Y \in B^*$),

$$X \cap \varphi(Y) = \emptyset \iff Y \cap \psi(X) = \emptyset.$$

Suppose that $Y \cap \psi(X) \neq \emptyset$. Since $y \in \psi(X)$ for some $y \in Y$, it follows from definition of $\psi(X)$ that

$$\exists x \in X \text{ s.t. } (x, y) \in R.$$

We also have $(x, y) \in R$ and $y \in Y$. This implies that

$$R(x) \cap Y \neq \emptyset$$

and $x \in \varphi(Y)$. The fact that $x \in X$ means

$$x \in X \cap \varphi(Y), \text{ that is, } X \cap \varphi(Y) \neq \emptyset.$$

The converse can be proved similarly. Thus B^* is the finite Boolean algebra with the conjugate pair $\varphi, \psi : B^* \rightarrow B^*$.

Moreover if we take $\xi^* : \Phi \rightarrow B^*$ as

$$\xi^*(A) = v(A),$$

then we have $\xi^*(A) \neq 1$ from $v(A) \neq W$. This means that $\not\vdash_{K_t^*} A$ implies $\xi^*(A) \neq 1$ for some finite Boolean algebra with conjugate and $\xi^* : B^* \rightarrow B^*$. It is obvious the converse statement. We thus obtain the next result.

Theorem 5. *The logic K_t^* can be characterized by the finite Boolean algebras with conjugate.*

We can show the following similarly.

Theorem 6. *The logics $K_t^* + (ext)$, $K_t^* + (sym)$, $K_t^* + (cl)$ are characterized by the class of all finite Boolean algebras with extensive, symmetric, closed conjugate pair, respectively.*

Thus we can conclude that our logical systems $K_t^* + (ext), + (sym), + (cl)$ are decidable, that is, we can determine whether a given formula is provable or not by finite steps.

References

1. J.Järvinen and J.Kortelainen, A unifying study between modal-like operators, topologies, and fuzzy sets, TUCS Technical report, 642 (2004)
2. B.Jónsson and A.Tarski, Boolean algebras with operators. Part 1., American Journal of Mathematics, vol.73 (1951), 891-939
3. Goldblatt, R., Logics of time and computation, CSLI Lecture Notes No.7 (1987)
4. Lemmon, E.J., New foundation for Lewis modal systems, Journal of Symbolic Logic, vol.22 (1957), 176-186

An Intelligent Technique Based on Petri Nets for Diagnosability Enhancement of Discrete Event Systems

YuanLin Wen¹, MuDer Jeng^{1,*}, LiDer Jeng², and Fan Pei-Shu³

¹ Department of Electrical Engineering,
National Taiwan Ocean University, Keelung, 202, Taiwan
{D88530004, Jeng}@mail.ntou.edu.tw

² Department of Electrical Engineering,
Chung-Yuan Christian University, Chung-Li, 320, Taiwan
Lider@cycu.edu.tw

³ College of Mechanica and Electrical Engineering,
National Taipei University of Technology, Taipei, 106, Taiwan
Fanpishu@yahoo.com.tw

Abstract. This paper presents an intelligent systematic methodology for enhancing diagnosability of discrete event systems by adding sensors. The methodology consists of the following interactive steps. First, Petri nets are used to model the target system. Then, an algorithm of polynomial complexity is adopted to analyze a sufficient condition of diagnosability of the modeled system. Here, diagnosability is defined in the context of the discrete event systems theory, which was first introduced by Sampath [3]. If the system is found to be possibly non-diagnosable, T-components of the Petri net model are computed to find a location in the system for adding a sensor. The objective is to distinguish multiple T-components with the same observable event sequences. The diagnosability-checking algorithm is used again to see if the system with the newly added sensor is diagnosable. The process is repeated until either the system is diagnosable or diagnosability of the system cannot be enhanced.

1 Introduction

Discrete event system such as semiconductor manufacturing machines are very complex and expensive, and therefore their reliable operations are significantly important for reducing the production costs. When a fault is diagnosed, a recovery procedure is taken to bring the machine back to the normal operational status.

Diagnosability of discrete event systems was first defined formally by Sampath et al. [3]. The merit of their approach is that a model-based, systematic methodology is proposed to detect a specific type of failure in finite delay after its occurrence. In addition, they present a necessary and sufficient condition for checking diagnosability. One of the challenges of diagnosis is that not all events are observable. It is difficult to differentiate between a nominal and a faulty trace within a bounded delay from the time of the occurrence of a fault. Thus, the computational cost for checking diagnosability using the methods of Sampath and her follower [3] is quite high. On the other hand, some propose methods that take advantage of the conciseness of Petri

* Corresponding author.

net models [5] to solve the diagnosis problem [2]. In [4], an algorithm of polynomial complexity in the number of Petri net nodes is proposed for computing a sufficient condition of diagnosability.

When a system is non-diagnosable, it may be possible that an original undetectable fault is discovered by adding sensors. The concept is related to multi-sensor data fusion techniques [6] that combine data from multiple sensors to achieve improved accuracies and more specific inferences than could be achieved by the use of a single sensor alone. The goal of this paper is to enhance diagnosability of discrete event systems based on the data fusion concept and the Petri net modeling paradigm. Petri nets are a well-known intelligent technique for Discrete Event Systems. We adopt our previously proposed algorithm [4] of polynomial complexity in the number of net nodes for checking a sufficient condition of diagnosability of the target system. If the system is found to be possibly non-diagnosable, T-components of the Petri net model are computed to find a location in the system for adding a sensor. The objective is to distinguish multiple T-components with the same observable event sequences. The diagnosability-checking algorithm is used again to see if the system with the newly added sensor is diagnosable. The process is repeated until either the system is diagnosable or diagnosability of the system cannot be enhanced.

The remaining paper is organized as follows: Section 2 provides background concepts related to the discussion of this paper. Section 3 presents our diagnosability-checking algorithm and present the proposed diagnosability enhancement methodology. In Section 4, we demonstrate our approach using an example. Concluding remarks and future directions are given in Section 5.

2 Basic Definitions and Properties

This section briefly summarizes the diagnosability concepts as follows: The system of interest is modeled as an FSM, $G = (X, \Sigma, \delta, x_o)$, where X is the set of states, Σ is set of events, δ is the partial transition function, and x_o is the initial state. The event set $\Sigma = \Sigma_o \cup \Sigma_{uo}$ is composed of observable events Σ_o , and unobservable events Σ_{uo} . For the nontrivial diagnosis problem, failures are unobservable events $\Sigma_f \subseteq \Sigma_{uo}$. There can be several types of failure, with the possibility that several failures belong to the same type of failure: Σ_{f_i} . Collectively, we use the partition Π_f on Σ_f to represent such categorization of failure types.

The objective of the diagnosis problem is to identify the occurrence and type of, if any, failure events, based on the observable traces generated by the system. In doing so, the detection of the failure needs be done within finite steps of observation after the occurrence of the failure.

While not all of the systems allow eventual diagnosis of an occurred failure, some systems do. In this regard, the notion of diagnosability is defined [2] as follows: for a given system G with the associated language L , i.e., the set of all event traces, if s is a trace in L ending with a F_i -type failure, and V is a sufficient long (at least n_i events longer) trace obtained by extending s in L , then every trace ω in L that is

observation equivalent to ν , i.e., $M(\omega) = M(\nu)$, should contain in it a F_i -type failure. In other words, a system is called diagnosable, if all of the failure types defined, no matter how the system trajectory gets into a failure event, and no matter what the following system continuations can be, we can always infer retrospectively in finite steps n and declare that a failure of the particular type has occurred.

The following example informally illustrates the concept of non-diagnosable net.

Example 1. Fig. 1 shows an FSM model G of some system where *state 1* is the initial state. The event ξ_f is a failure and thus unobservable while all other events α, β, γ and δ are normal and observable. Fig.1 has three cycles $6 \xrightarrow{\beta} 7 \xrightarrow{\gamma} 8 \xrightarrow{\delta} 6$, $3 \xrightarrow{\beta} 4 \xrightarrow{\gamma} 5 \xrightarrow{\delta} 3$ and $9 \xrightarrow{\beta} 10 \xrightarrow{\gamma} 11 \xrightarrow{\delta} 9$ that correspond to the same observable mapping $\beta \gamma \delta$, where an observable mapping maps the unobservable labels to nulls and the observable labels to themselves. However, the former cycle is reached from a failure while the latter is not. Thus, the system is not diagnosable.

The disadvantage of the FSM approach is the computational inefficiency, since the number of states of a system is in general exponential to the size of the system.

In the following section, we will present our previous approach for checking diagnosability, which is based on net structures without enumerating the states.

3 A Polynomial Algorithm for Checking Diagnosability

We consider a system modeled by a Petri net, $G = (P, T, I, O, M_o)$. The set of places P is partitioned into the set of observable places P_o and the set of unobservable P_{uo} , i.e., $P = P_o \cup P_{uo}$ and $P_o \cap P_{uo} = \emptyset$. The set of transitions T is partitioned into the set of observable transitions T_o and the set of unobservable transition T_{uo} , i.e., $T = T_o \cup T_{uo}$ and $T_o \cap T_{uo} = \emptyset$. Labels are associated with places and transitions denoting conditions and events, respectively. Failures are labels associated to unobservable transitions. We call a label associated with an unobservable place or transition an unobservable label.

In a Petri net, a T-invariant may represent a firing sequence that takes the model from a marking back to that marking where a T-invariant is defined by the following equation:

$$A^T \bullet X = 0.$$

The incidence matrix $A = (A(t_j, p_i))$ is defined as follows:

$$A(t_j, p_i) = \begin{cases} 1 & \text{if } p_i \in O(t_j), p_i \notin I(t_j) \\ -1 & \text{if } p_i \in I(t_j), p_i \notin O(t_j) \\ 0 & \text{otherwise} \end{cases}$$

A T-invariant shows the possibility of regeneration of a certain marking of the model. A T-invariant X is said to be minimal if there is no other invariant X_l such that $X_l(t) \leq X(t)$ for all t . Thus, an elementary cycle of the state space of a Petri net, i.e., its reachability graph, may correspond to a minimal T-invariant of the net structure [7].

Since places are associated to labels, we will consider structural objects called T-components, which are related to minimal T-invariants, as follows: Let x be a minimal T-invariant of G . The subnet generated by $\bullet X \cup X \bullet$ is a T-component, denoted as C_T of G .

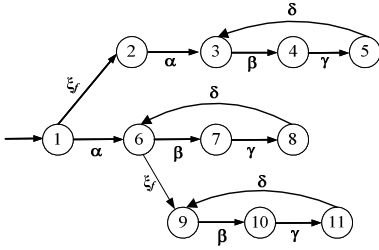


Fig. 1. An FSM model G of a system

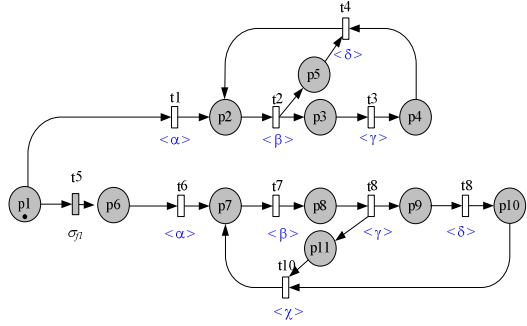


Fig. 2. A Petri Net model

Example 2. Fig. 2 shows a Petri net model G where event f_1 is a failure, which is unobservable. All places are unobservable and transition associated labels are shown in the parentheses following transition names. In the model, there are only two minimal T-invariants $\{t2, t3, t4\}$ and $\{t7, t8, t9, t10\}$, which generate two T-components $\{t2, p2, p3, t3, p4, t4, p5\}$ and $\{p7, t7, p8, t8, p9, t9, p10, t10, p11\}$ respectively.

From the above discussion, we obtain a sufficient condition for a system to be diagnosable as follows:

Property 1. The system G modeled by a Petri net is diagnosable if there do not exist two T-components with the same observable labels such that unobservable labels mapped into nulls.

Proof: If there are no two T-components with the same observable labels, then there are no two cycles in G with the same observable labels such that one cycle is reached from a failure event while the other is not reached from a failure event. Q.E.D.

Since T-invariants can be solved efficiently using linear programming [4], Property 1 can be transformed into the following property, which shows our diagnosability-checking algorithm:

Property 2. (diagnosability-checking algorithm) A system modeled by a Petri net G is diagnosable if the following linear programming problem (LPP) has no solution:

$$\begin{aligned}
 &LPP: \\
 &Z = \max \quad 0(X_1 + X_2) \\
 &Subject \ to \\
 &A^T X_1 = A^T X_2 = 0 \\
 &X_1 \neq X_2 \\
 &X_1, X_2 > 0 \\
 &\xi(C_T(X_1)) = \xi(C_T(X_2))
 \end{aligned} \tag{1}$$

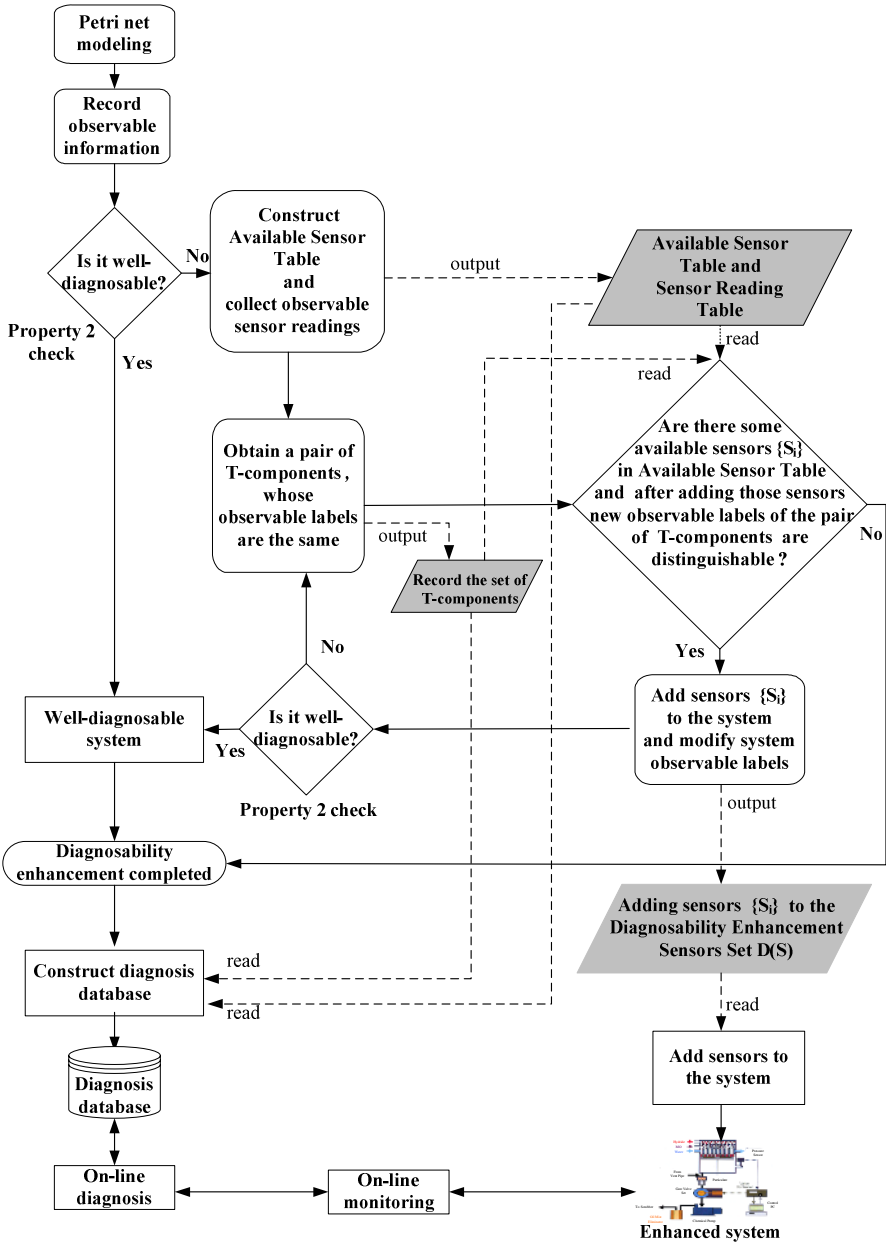


Fig. 3. Diagnosability Enhancement procedure

where A^T is the transpose of the incidence matrix A , X_1 and X_2 are any two T-invariants, and $\xi(C_T(X_1))$ and $\xi(C_T(X_2))$ are the observable label mapping (linear function) of the T-components $C_T(X_1)$ and $C_T(X_2)$ respectively.

Example 3. In Fig.2, since all places are unobservable, from the only two T-components $C_T(X_1) = \{t2, p2, p3, t3, p4, t4, p5\}$, $C_T(X_2) = \{p7, t7, p8, t8, p9, t9, p10, t10, p11\}$, we obtain their observable mapping $\xi(C_T(X_1)) = \{\beta, \gamma, \delta\}$, $\xi(C_T(X_2)) = \{\beta, \gamma, \delta, \chi\}$. Thus, Property 2 holds, which means that the system is diagnosable.

Definition 1. A systems G is called well-diagnosable if Property 2 holds. In other words, well-diagnosable nets can be checked with polynomial complexity.

The above sufficient condition for diagnosability is useful for designing a well-diagnosable system. Fig. 3 shows the proposed diagnosability enhancement procedure.

4 Example

We use an example to illustrate the above procedure. In the following, we assume fault events are independent on one another such that each sensor reading that depends on a single fault event.

Fig. 4 shows a Petri net model without sensor. Event σ_{f1} is the failure event, which is unobservable, and $\alpha, \beta, \gamma, \delta$ are normal and observable events. Table 2 shows its Available Sensors Table. In the model, there are two minimal T-invariants $\{t2, t3, t4\}$ and $\{t7, t8, t9\}$, which generate two T-components $C_T(X_1) = \{p2, t2, p3, t3, p4, t4, p5\}$ and $C_T(X_2) = \{p7, t7, p8, t8, p9, t9, p10\}$ respectively. The observable labels are $\xi(C_T(X_1)) = \xi(C_T(X_2)) = \{\beta, \gamma, \delta\}$, where $C_T(X_1)$ is not reachable from the failure event σ_{f1} but $C_T(X_2)$ are reachable from the failure event σ_{f1} . Since $\xi(C_T(X_1)) = \xi(C_T(X_2))$, the system is not well- diagnosable, i.e., the system is possibly non-diagnosable.

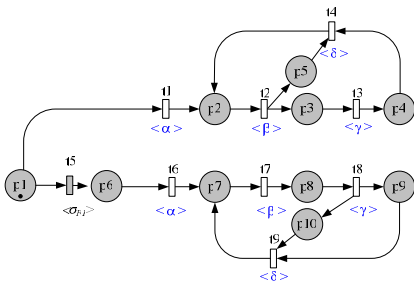


Fig. 4. Observable label mapping without sensor

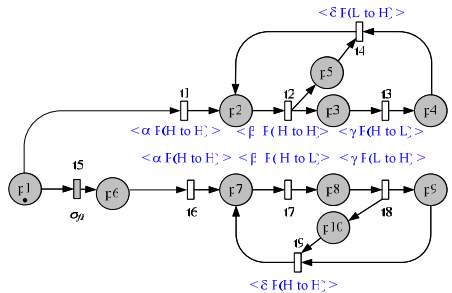


Fig. 5. New observable labels after the pressure sensor added

Table 1. Observable labels of T-components

X_i	Set of T-invariants X_i	Observable labels without sensor $\xi(C_T(X_i))$
X_1	{t2,t3,t4}	$\langle \beta, \gamma, \delta \rangle$
X_2	{t7,t8,t9}	$\langle \beta, \gamma, \delta \rangle$

Next we analyze to see if sensors are available for use, This requires adequately understanding the system functions and its components, and can only be done manually. Support that after the manual analysis, we have available sensors of Fig. 4 shown in Table 2.

Table 2. Available Sensors Table

No.	Failures	Available Sensors	Location
1	σ_{f1}	Pressure (S_p)	Location l_1 in the system(e.g. output of a pump)
2	σ_{f1}	Temperature(S_t)	Location l_2 in the System(e.g. process chamber)

Then, we try to add all available sensors to the system of Fig.4, and collect all those observable sensor readings in Table 3, where P(H to H) means that the reading of the pressure sensor changes from High to High.

For the pair of T-components with the same observable label as listed in Table 1, where $\xi(C_T(X_1)) = \xi(C_T(X_2)) = \langle \beta, \gamma, \delta \rangle$, we obtain that after adding the pressure sensor S_p , the new observable labels of the pair of T-components are distinguishable. Therefore, S_p is added to the Diagnosability Enhancement Sensors Set $D(S)$. The new observable labels after adding the pressure sensor S_p are shown in Fig. 5. In this example, adding the temperature sensor is useless since the pair of T-components still have the same observable label ($\langle \beta, T(H \text{ to } L), \gamma, T(L \text{ to } H), \delta, T(H \text{ to } H) \rangle$).

Table 3. Observable sensor readings of Fig.4

Transition	Event	Pressure Sensor	Temperature sensor
t1	$\langle \alpha \rangle$	P(H to H)	T(H to H)
t2	$\langle \beta \rangle$	P(H to L)	T(H to L)
t3	$\langle \gamma \rangle$	P(L to H)	T(L to H)
t4	$\langle \delta \rangle$	P(H to H)	T(H to H)
t6	$\langle \alpha \rangle$	P(H to H)	T(H to H)
t7	$\langle \beta \rangle$	P(H to H)	T(H to L)
t8	$\langle \gamma \rangle$	P(H to H)	T(L to H)
t9	$\langle \delta \rangle$	P(H to H)	T(H to H)

Next, we check the well-diagnosability property of the new system. Since $\xi(C_T(X_1)) = (\langle \beta, P(H \text{ to } H) \rangle, \langle \lambda, P(H \text{ to } H) \rangle, \langle \delta, P(H \text{ to } H) \rangle)$ and $\xi(C_T(X_2)) = (\langle \beta, P(H \text{ to } L) \rangle, \langle \lambda, P(L \text{ to } H) \rangle, \langle \delta, P(H \text{ to } H) \rangle)$, the new system with S_p added is a well-diagnosable system with respect to failure f_1 .

The Diagnosability Enhancement Sensors Set $D(S)$ has one sensor, pressure sensor S_p . Thus, we just need to add S_p sensors to the system to enhance its diagnosability. We can construct the diagnosis database shown in Table 4, i.e., when we on-line monitor the operation of the system, if the No.1 observable label of Table 4 is discovered, then we know that the f_1 fault event has happened.

Table 4. Fault diagnosis database of Fig. 4

No.	Observable label with pressure sensor $\xi(C_T(X_i))$	Fault event
1	$(\langle \beta, P(H \text{ to } L) \rangle, \langle \lambda, P(L \text{ to } H) \rangle, \langle \delta, P(H \text{ to } H) \rangle)$	f_1
2	Others	Normal

6 Conclusion

In this paper we have applied our previously proposed diagnosability-checking algorithm to designing well-diagnosable discrete event systems. The methodology is an iterative and systematic for enhancing system diagnosability by adding sensors. An algorithm of polynomial complexity is adopted to analyze a sufficient condition of diagnosability of the target system modeled by Petri nets. If the system is found to be possibly non-diagnosable, T-components of the Petri net model are computed to find a location in the system for adding a sensor. The objective is to distinguish multiple T-components with the same observable event sequences. The diagnosability-checking algorithm is used again to see if the system with the newly added sensor is diagnosable. The process is repeated until either the system is diagnosable or diagnosability of the system cannot be enhanced. To show the applicability of the proposed methodology, an example is given to illustrate our approach.

References

1. Jiang, S. Kumar, R. and Garcia, H. E.: Optimal Sensor Selection for discrete-Event Systems with Partial Observation. IEEE Trans. on Automatic Control, Vol. 48(3) 369-381
2. Jiang S., Huang Z., Chandra V., and Kumar R.: A Polynomial Algorithm for Testing Diagnosability of Discrete-Event Systems. IEEE Trans. on Automatic Control, Vol. 46(8) 1318-1321
3. Sampath M., Lafortune S., Sinnamohideen K., and Teneketzis D.: Diagnosability of Discrete-Event Systems. IEEE Trans. on Automatic Control, Vol. 40(9) 1555-1557
4. Wen Y. L., Jeng, M. D, Huang, Y. S.: Diagnosability of Semiconductor Manufacturing Equipment. Material Science Forum Vol. 505 1135-1140

5. Peterson J.L., Petri Net Theory and the Modeling of Systems. Prentice-Hall Englewood Cliffs, N.J. (1981)
6. Hall D. L., Llinas J.: An introduction to multisensor data fusion. Proceedings of the IEEE Vol. 85(1) 6-23
7. Desel J. and Esparza J.: Free Choice Petri Nets. Cambridge University Press (1995)

Fuzzy Logic Based Mobility Management for 4G Heterogeneous Networks

Jin-Long Wang and Chen-Wen Chen

Department of Information and Telecommunications Engineering
Ming Chuan University, Taipei 11120, Taiwan
j1wang@mcu.edu.tw

Abstract. Next-generation wireless networks will provide information transmission in Heterogeneous Wireless Networks. It not only offers a variety of wireless access services for integrating different wireless networks, but also takes into account the high bit rate, the QoS management, and the friendly mobility. In this article, the scheme of choosing the suitable network with the best service for a mobile terminal in a heterogeneous wireless network is studied. A new scheme based on fuzzy logic is proposed to employ the important traffic criteria, including bandwidth, dropping rate, blocking rate, signal, and velocity. Finally, the simulation is used to investigate the performance of proposed schemes. The simulation results show that the proposed schemes have better performance than conventional schemes.

Keywords: Fuzzy logic, mobility management, heterogeneous networks, 4G.

1 Introduction

The shift to third-generation in the radio access networks is presently ongoing. The third-generation mobile system provides data rates up to 384kb/s for wide-area coverage and up to 2Mb/s for local-area coverage. The worldwide introduction of WCDMA took place in 2001 and 2002, starting in Japan and continuing in Europe [1]. In the U.S., several 3G alternatives will be available including H2, Bluetooth, EDGE, WCDMA, and cdma2000 system. The EDGE, which has data rates upping to 60 kbps, is improved spectrum efficiency from GPRS. The Bluetooth with the low power and low-cost system offers a short distance transmission. Bluetooth can be used in among mobile terminals, such as mobile phones, PDA, laptop computer, and so on [2].

The next generation of wireless communication systems will be based on heterogeneous concepts and technologies, and some systems are extended from WCDMA, EDGE, cdma2000, Bluetooth, WLAN, and etc. These systems support “anytime, anywhere with anybody/anything” communication. A key component of these evolving systems is the multiplicity of access technologies as well as a diversity of terminals that allow users on the move to enjoy seamless high-quality wireless services irrespective of geographical location, speed of movement, and time of day [3].

Driven by the “anytime, anywhere” concept, new requirements for flexible network access have made their way into the telecommunication community. The present communication paradigm expects a user to be able to access its services independent of its location in a completely transparent way [4]. The mobile terminal should be able to select the best access technology among ad hoc, personal area network, WLAN, and cellular networks. This homogeneous, high-speed, secure, multi-service, and multiple-operator network is being developed in a context commonly referred to as fourth-generation (4G) networks or, sometimes, as beyond third generation—B3G [5].

In 4G technology, the important issues are how to access several different mobile and wireless networks, and how to select one of networks to be handed off. The proposed scheme is based on fuzzy logic to decide the target of handoff. There are many trade-offs between computing in the fuzzy and crisp domains [6]. Because the arrival time of the calls may vary significantly, their call duration times are vague and uncertain. The advantages of using linguistic variables and fuzzy functions to improve software reliability and reduce computation requirement has become well understood. Due to this nature, fuzzy logic seems to be the best way to approach the problem. We adopt the number of available channels, signal power, blocking rate, dropping rate, and velocity as the input variables for fuzzy sets and define a set of membership functions. The fuzzy logic decision process consists of three module: (1)Fuzzification, (2)Rule Evaluation, (3)Defuzzification.

The simulation results reveal that the proposed scheme yields better performance as compared to others. The fuzzy logic has higher utilization, lower blocking rate, and lower dropping rate.

The subsequent sections of this thesis are organized as following. In Section 2, the system model, 4G wireless networks, mobility management, mobile terminal how to roam in different system, and some handoff decision schemes are introduced. Then the Fuzzy Multiple Objective Handoff Decision Scheme will be proposed in Section 3. The experiment results are shown in Section 4. At last in section 5, we have a conclusion.

2 System Models and Related Works

This section will introduce the system model, the architectures of wireless LAN and CDMA2000 network. When the mobile terminal is handed off to the other cell, some management problem produced will be introduced. Finally, several existing handoff schemes will be illustrated.

2.1 System Model

A CDMA2000/WLAN is the cell overlapping system, where a CDMA2000 network overlaps with a set of WLANs. It is assumed that every cell has 64 channels both in these two kinds of networks.

Each CDMA2000 network cell, CD_i , overlaps with n WLAN cells, CC_j . For instance, in Figure 1, each CDMA2000 network CD_i covers 7 WLAN cells. Cell CD_4

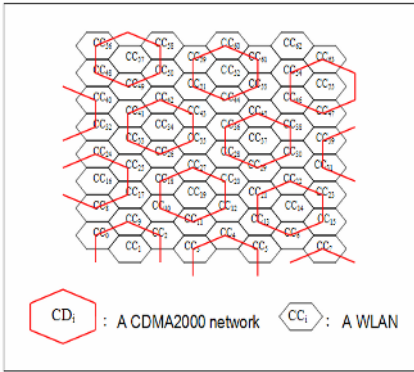


Fig. 1. WLANs and CDMA2000 network

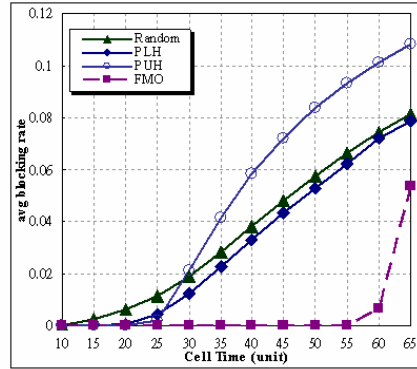


Fig. 2. Average Blocking Rate

in CDMA2000 network covers 7 WLANs, including CC₁₀, CC₁₁, CC₁₂, CC₁₈, CC₁₉, CC₂₀, CC₂₇. The MT may connect both with CDMA2000 network and WLAN. Therefore, if an MT should be handed off, it performs the handoff process, which determines and chooses the suitable cell of CDMA2000 or WLAN.

2.2 Wireless LAN

The WLANs standardized in the IEEE 802.11 standards. The infrastructure mode network architecture resembles the wide-area cellular networks and is of most interest to wireless Internet service providers (WISPs). When mobile terminal (MT) accesses the WLAN, MT must register to access point (AP). The packet transmissions between the AP and the MN can be optionally protected using a symmetric key-based RC4-based encryption called Wired Equivalency Privacy (WEP).

2.3 CDMA2000 Network

CDMA2000 networks consist of multiple base stations (BSs), each of which connects to a radio network controller (RNC). The data traffic of multiple RNCs will aggregate in packet data serving node (PDSN). The PDSN is also an interface between a RAN and a packet-switched network. The PDSN terminates a Point-to-Point Protocol (PPP) connection and maintains the session state for each MN in its serving area. PPP header and payload compression can be negotiated between the PDSN and the MN. The RNC manages several concurrent Radio Link Protocol (RLP) layer2 sessions with mobile nodes (MNs) and performs per-link bandwidth management functions [6]. While an MN moves from one RNC to the other RNC, the RLP will be re-established with the now RNC.

2.4 Handoff Decision Schemes

There are a lot of handoff decision schemes, such as conventional power level based handoff, user population based handoff, bandwidth based handoff, and fuzzy logic based handoff. Random Based Handoff, Power Level Based Handoff (PLH), Power

and User Population Based Handoff, and Fuzzy Logic Based Model are introduced in this section[5].

In Random Based Handoff scheme, the mobile terminal is handed off to the random neighboring cells. It does not consider any parameter which may influence the quality of network.

In the Power Level Based Handoff scheme, when an MT should perform the handoff, the power levels received from the candidate wireless networks are compared and the network with highest power above the predefined threshold is selected for handoff. Since the loading status of the selected wireless network is not took into account, this scheme results in non-uniform utilization of system resource, poor blocking rate, and high dropping rate.

In the Power and User Population Based Handoff scheme, the MT is handed off to the wireless network with the largest weight value. The weight value is computed in signal power and number of users. If the signal power is larger or unused channel is more, the weight value would be higher. In this scheme, the MT may choose the wireless network with larger power signal or low channel utilization.

Fuzzy logic has emerged as one of the most active fields for researching in the applications of simulating human decision-making. The Fuzzy Logic Model is efficient to be used to consider a lot of factors and is able to provide an effective handoff decision (Figure 3). The considered factors are resource management, mobility, power, propagation environments, service, and so on. There are three stages in the fuzzy logic based decision, including the fuzzification procedures, the weighting of the criteria are performed, and the final decision making [5].

3 Proposed Schemes

In this section, a new proposed scheme, called the Fuzzy Multiple Objective Handoff Decision Scheme, for handoff decision is presented.

3.1 Fuzzy Multiple Objective Handoff Decision Scheme

Handoff Decision is used to determine a cell for the handoff service. There are two steps in this scheme. In the first step, the weight of the criteria is determined and the fuzzy ranking procedure is executed. Then, the ranking procedure compares the performance of the different cells according to the specific handoff decision criterion. It is assumed that $C = \{ C_1, C_2, \dots, C_q \}$ is a set of q handoff decision criteria.

3.2 Criteria Weighting

The relationships among these decision criteria can be represented by a $q \times q$ matrix, as shown in equation (1), where the definitions of cv_{ij} are shown in equation (2) and g_{ij} is the important degree provided by wireless carrier administrators. The eigenvector can be derived from the matrix in equation (1), denoted as FW in equation (3). For normalization, each element of FW will be multiplied by a value α , as shown in equation (4). The obtained eigenvector can be used in the decision process.

$$B = \begin{bmatrix} CV_{11} & CV_{12} & CV_{13} & \cdots & CV_{1q} \\ CV_{21} & CV_{22} & CV_{23} & \cdots & CV_{2q} \\ CV_{31} & CV_{32} & CV_{33} & \cdots & CV_{3q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ CV_{q1} & CV_{q2} & CV_{q3} & \cdots & CV_{qq} \end{bmatrix} \tag{1}$$

$$\begin{cases} CV_{ij} = 1 & i = j \\ CV_{ij} = g_{ij} & i < j \\ CV_{ji} = 1/CV_{ij} & i > j \end{cases} \tag{2}$$

$$B \Rightarrow FW = [w_1 \ w_2 \ w_3 \ \cdots \ w_q]^T \tag{3}$$

$$F \Rightarrow \alpha FW = [\alpha w_1 \ \alpha w_2 \ \alpha w_3 \ \cdots \ \alpha w_q]^T \tag{4}$$

CV_{ij} indicates the important relationships between i and j criteria, where i, j are from 1 to q criterion; w_i indicates the weight of criterion, where i is from 1 to q criterion; α indicates that is a normalize value, w_i indicates the weight of criterion, where i is from 1 to q criterion.

3.3 Criteria Values

Five criteria values are evaluated during handoff process, in which c_1 represents degree of membership in the High bandwidth (HBW), c_2 represents degree of membership in the Low Blocking rate (LBK), c_3 represents degree of membership in the Low Dropping rate (LDR), c_4 represents the degree of membership in the High Velocity (HV) or degree of membership in the Low Velocity (LV) depending on whether the candidate cell belongs to CDMA2000 or not, and c_5 represents degree of membership in the High Signal (HS). For instance, if a mobile station is at CC_9 and ready to perform the handoff, the neighboring cells with CDMA 2000 network cells and WLAN cells are considered to be the candidates of the handoff. One of the candidates is CD_4 , of which the usability bandwidth rate is 0.25, the blocking rate is 0.6, the dropping rate is 0.3, the signal is 60, and the velocity is 70. Through the membership function, the fuzzy values can be obtained, such as $HBW(0.25)=0.75$, $LBK(0.6)=0.6$, $LDR(0.3)=0.3$, $HS(60)=0.6$, and $HV(70)=0.7$, as shown in Table 1.

Table 1. An illustration of Fuzzy Multiple Objective Handoff

Criteria	Input Value	Membership Function	Output Value
Bandwidth	0.25	HBW	$HBW(0.25) = 0.75$
Blocking rate	0.6	LBK	$LBK(0.6) = 0.6$
Dropping rate	0.3	LDR	$LDR(0.3) = 0.3$
Signal	60	HS	$HS(60) = 0.6$
Velocity	70	If “the candidate cell belongs CDMA2000 network” then “HV” else “LV”	If “the candidate cell belongs CDMA2000 network” then $HV(70) = 0.7$ else $LV(70) = 0.3$

3.4 Decision Value

After the weighting value and five criteria values are obtained, the decision value can be calculated in the following steps. First, the initial fuzzy decision (IFD) value is obtained according to the equation (5), in which five criteria are considered. Then, based on the equation (6), the final fuzzy decision (FFD) value can be derived.

$$IFD(P) = MFOV_1^{w_1} \cap MFOV_2^{w_2} \cap MFOV_3^{w_3} \cap MFOV_4^{w_4} \cap MFOV_5^{w_5} \quad (5)$$

$$FFD = \alpha IFD(Y) + \beta IFD(PY) - \gamma IFD(PN) - \lambda IFD(N) \quad (6)$$

$MFOV_t$ indicates fuzzy values, while $w_1, w_2, w_3, w_4,$ and w_5 indicate the weight values for five criteria. $\alpha, \beta, \gamma,$ and λ are weight value for different IFD function.

4 Simulation Results

This wireless network simulation is based on SMPL simulation system and built by C language. In this simulation, we have 16 CDMA2000 network cells, and 64 WLAN cells. Every CDMA2000 network cell overlays 7 WLAN cells and every cell has 64 channels. It is assumed that new call arrivals follow Poisson process with arrival rate 1, and handoff time follow exponential process. The call service time was from 10 to 70 unit times. The call service time was a Poisson process.

The proposed Fuzzy Multiple Objective Handoff Decision (FMO) scheme is compared with the Random Based Handoff scheme (Random), the Power Level Based Handoff (PLH) scheme, and the Power and User Population Based Handoff (PUH) scheme. There are three quality factors are used to estimate the schemes, which are blocking rate, dropping rate, and utilization.

4.1 Average Blocking Rate

The average blocking rate is defined as the total number of blocking calls in WLAN and CDMA2000 networks dividing total new calls in WLAN and CDMA2000.

It is expected to have the average blocking rate lower than 0.1. In Figure 2, the Random scheme, the PLH scheme, and the PUH scheme have higher blocking, while the FMO Scheme has lower blocking. Especially, the service time smaller than 50 units, their blocking rates approach to 0.

Since the Random scheme does not consider any factor, its blocking rate is higher. The PLH scheme considers with the signal element, but it just thinks about one factor, so their blocking rate is higher than the fuzzy scheme. In the PUH scheme, two factors, signal and bandwidth, are considered. If the signal is strong, but the bandwidth is few, the call will not be accepted and the blocking rate will increase. The FMO scheme is better than other schemes due to that it considers ten factors and also the criteria with different weighting. Thus the blocking rate of FMO scheme is lower.

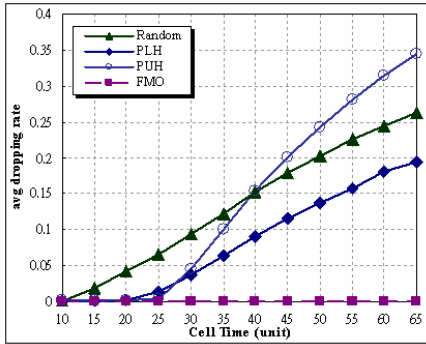


Fig. 3. Average Dropping Rate

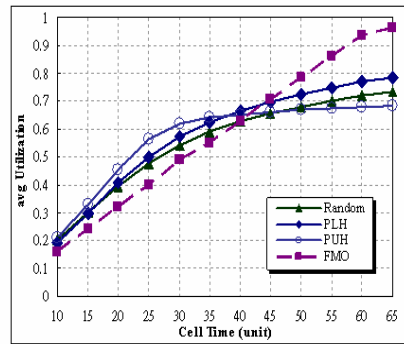


Fig. 4. Average Utilization

4.2 Average Dropping Rate

The average dropping rate means total number of dropping calls in WLAN and CDMA2000 networks divided by total handoff calls in WLAN and CDMA2000.

In Figure 3, the PUH scheme has the highest average dropping rate. This is because the handoff calls are more than the new calls, its dropping rate becomes the highest. Parts of dropping rates from the Random scheme and the PLH scheme are higher than 0.1. This is due to that they can not effectively deal properly with the resources in cells. On the other hand, the FMO Scheme is superior to the Random scheme, the PLH scheme, and the PUH scheme. Since they consider the five more influence factors, and make the resource distributive effectively, the dropping call is lower than the other schemes.

4.3 Average Utilization

The average utilization indicates the average of WLAN’s utilization and CDMA2000 network’s utilization.

It is expected to have the average utilization higher than 0.8. In Figure 4, the Random scheme, the PLH scheme, and the PUH scheme have the average utilization lower than 0.8. This indicates that the resources are not used effectively. The FMO Scheme has the higher average utilization than other schemes. When the service time is more than 50 units, the utilization is more than 0.8. Specially, when their service time is higher than 60 units, their average utilization is higher than 0.9. Thus, the proposed FMO scheme is able to reach the high utilization.

5 Conclusions

This paper has presented handoff schemes based on fuzzy logic for the heterogeneous wireless networks, including WLANs and CDMA2000 networks. The proposed handoff scheme, Fuzzy Multiple Objective Handoff Decision Scheme, improves the handoff decision for the overlap of coverage in the 4G heterogeneous wireless networks. It offers the advantages of lower blocking rate, lower dropping rate, and higher utilization in comparison with current schemes.

References

1. Frodigh, M., Parkvall, S., Roobol, C., Johansson, P., and Larsson, P.:Future-Generation Wireless Networks. IEEE Personal Communication, Volume: 8 (2001) 10-17.
2. Ali, S.I., Radha, H.:Hierarchical handoff schemes over wireless lan/nvan networks for multimedia applications. Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on , Volume: 2 (2003) 545-548.
3. Buddhikot, M.M., Chandranmenon, G., Han, S., Lee,Y.-W., Miller, S., Salgarelli, L.: Design and implementation of a WLAN/cdma2000 interworking architecture. Communications Magazine, IEEE , Volume: 41 , Issue: 11 (2003) 90-100.
4. Varshney, U., Jain, R.: Issues in Emerging 4G Wireless Networks. IEEE Computer, Volume: 34 (2001) 94-96.
5. Chan, P. M. L., Sheriff, R. E., Hu, Y. F., Conforto, P., Tocci, C.: Mobility Management Incorporating Fuzzy Logic for a Heterogeneous IP Environment. Communications Magazine, IEEE, Volume: 39, Issue: 12 (2001) 42-51.
6. Wang, J.-L.:Handover Management Based on Fuzzy Logic Decision for LEO Satellite Networks. Intelligent Automation and Soft Computing, Vol. 11, No. 2 (2005) 69-82.

On-Line Association Rules Mining with Dynamic Support

Hsuan-Shih Lee

Department of Shipping and Transportation Management
Department of Computer Science
National Taiwan Ocean University
Keelung 202, Taiwan
Republic of China

Abstract. In this paper, we use maximal itemsets to represent itemsets in a database. We show that the set of supreme covers, which are the maximal itemsets whose proper subsets are not maximal itemsets, induces an equivalence relation on the set itemsets. Based on maximal itemsets, we propose a large itemset generation algorithm with dynamic support, which runs in time $O(M'2^N + M'\log M)$, where N is the maximum number of items in a maximal itemset, M' is the number of the maximal itemsets with minimum support greater than the required support, and M is the number of the maximal itemsets.

1 Introduction

One of the research topics in data mining that attracts many researchers' attention is mining association rules from transaction database [1,2,3,4,5]. Databases are usually dynamic. How to utilize previously mined patterns for latter mining process has become important research topics. Incremental data mining methods are introduced to deal with such problems [6,7,8,9,11].

Mining association rules of a database can be decomposed into two stages, which are identifying large itemsets and generating associating rules from large itemsets. The later can be accomplished easily once large itemsets are identified. Therefore, the main focus of mining association rules is to identify large itemsets efficiently. One of the most known algorithms to generate large itemsets is the Apriori algorithm [3]. In this paper, we use maximal itemsets to keep track of the occurrences of itemsets. We show that the set of maximal itemsets which contain no other maximal itemsets induces an equivalence relation on the set of itemsets that occur in a database. Based on maximal itemsets, we propose a faster algorithm to generate large itemsets with dynamic support that runs in time of $O(M'2^N + M'\log M)$, where N is the maximum number of items in a maximal itemset, M' is the number of the maximal itemsets with minimum support greater than the required support, and M is the number of the maximal itemsets.

2 Maximal Itemset and Its Properties

Definition 1. Given a set of items $I = \{i_1, i_2, \dots, i_n\}$ and a database $B = \{t_1, t_2, \dots, t_n\}$ where B is a multiset and $t_i \subseteq I$, for itemset $X \subseteq I$, the number of the occurrences of X , denoted as $O_B(X)$, is defined to be the number of transactions in database B that contain X .

Definition 2. Given $X \subseteq I$, X is a maximal itemset in B if and only if $X = I$ or $O_B(X) > O_B(Y)$ for all $Y \subseteq I$ and $Y \supset X$. That is, an itemset is a maximal itemset if its number of occurrences is greater than occurrence of all its proper supersets or it has no proper superset.

A maximal itemset is maximal in the sense that the occurrences of the itemset would decrease if we add more items into the itemset. For example, in the sample database of Table 1, $\{Beer\}$ is maximal since the number of occurrences would decrease if other items are added into it.

Table 1. Sample database

Transaction	Items
t1	Bread, Jelly, PeanutButter
t2	Bread, PeanutButter
t3	Bread, Milk, PeanutButter
t4	Beer, Bread
t5	Beer, Milk

Lemma 1. [10] For $X, Y \subseteq I$, if $X \subset Y$, then $O_B(X) \geq O_B(Y)$.

Lemma 2. [10] Let $X_1, X_2 \in M(B)$. If $X_1 \subset X_2$, then $O_B(X_1) > O_B(X_2)$.

Lemma 3. [10] Let $X_1, X_2 \in M(B)$. If $X_1 \cap X_2 \neq \emptyset$, $X_1 - X_2 \neq \emptyset$ and $X_2 - X_1 \neq \emptyset$, then there exists $Y \in M(B)$ such that $X_1 \cap X_2 \subseteq Y$ and $O_B(Y) \geq O_B(X_1) + O_B(X_2)$.

Lemma 4. [10] Let B_1 and B_2 be two databases. Assume $B_1 \subseteq B_2$. If $X \in M(B_1)$, then $X \in M(B_2)$. That is if X is a maximal itemset in current database, it is also a maximal itemset after new transactions are inserted.

Definition 3. A maximal itemset Y is said to be a supreme cover of itemset X in database B if $O_B(X) = O_B(Y)$ and $X \subseteq Y$.

Lemma 5. [10] If $t \in B$, then $t \in M(B)$. That is, every transaction in a database is a maximal itemset.

Lemma 6. [10] For each itemset X in database B , there is one and only one supreme cover for X .

Theorem 1. [10] *The number of the occurrences of itemset X can be determined as follows:*

$$O_B(X) = \begin{cases} \max_{Y \in M(B) \text{ and } X \subseteq Y} O_B(Y) & \text{if } X \text{ occurs in } B \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 2. *The set of supreme covers in database B induces an equivalence relation on the set of itemsets that occur in database B and each supreme cover is a representative of an equivalence class in the induced equivalence relation.*

Proof: Let U be the set of itemsets that occur in database B . The induced equivalence relation γ can be defined as follows: For $X, Y \in U$, $X\gamma Y$ if the supreme cover of X equals the supreme cover of Y . Following lemma 6, every itemset in U has a unique supreme cover. Therefore, γ is reflexive, transitive and symmetric and hence γ is an equivalence relation. □

Lemma 7. [10] *If $X \in M(B)$ and $t \subseteq I$, then $X \cap t \in M(B \cup \{t\})$. That is, $X \cap t$ is a maximal itemset in $B \cup \{t\}$.*

Assume the maximal itemsets of database B , denoted as $M(B)$ and the number of the occurrences of the maximal itemsets in B are known. If a new transaction t is inserted into B , the maximal itemsets of database $B \cup \{t\}$, $M(B \cup \{t\})$, can be derived by the following procedure:

Maximal_insert;

$$B' = B \cup \{t\};$$

$$M(B') = M(B);$$

$$\text{Let } S = \{t \cap X \mid X \in M(B)\}.$$

For each $Z \in S$ do

$$O_{B'}(Z) = 1 + O_B(Z);$$

$$M(B') = M(B) \cup \{Z\}$$

If $t \notin M(B')$ do

$$M(B') = M(B') \cup \{t\} \text{ and } O_{B'}(t) = 1;$$

End of Maximal_insert;

Theorem 3. *The Maximal_insert procedure maintains the maximal itemset of database $B \cup \{t\}$ and the number of the occurrences of the maximal itemsets in $B \cup \{t\}$ correctly. The time of updating is linear to the number of maximal itemsets.*

Proof: Following lemma 4, lemma 5 and lemma 7, the Maximal_insert procedure adds maximal itemsets in $M(B)$, itemsets in S and t into $M(B')$. To show that no maximal itemset is missed in $M(B')$, we assume Y is a maximal itemset in B' but not recorded in $M(B')$ by the procedure.

If Y occurs in B but is not maximal in B , Y must occur in t so that Y becomes maximal in B' . That is $Y \subset t$. Let W be the supreme cover of Y in B . Let $Z = t \cap W$. Then $Z \in S$ and $Y \subseteq Z$. If $Y = Z$, Y would not be missed by

the procedure. The only chance that Y will be missed is that $Y \subset Z$. Assume that

$$Y \subset Z. \tag{1}$$

Since W is the supreme cover of Y in B , we have

$$O_B(Y) = O_B(W). \tag{2}$$

Since $Z \subseteq W$, we have

$$O_B(Z) \geq O_B(W). \tag{3}$$

Following (2) and (3), we have

$$O_B(Z) \geq O_B(Y). \tag{4}$$

By assumption (1), we have

$$O_B(Y) \geq O_B(Z). \tag{5}$$

(4) and (5) yield

$$O_B(Y) = O_B(Z). \tag{6}$$

Hence

$$O_{B'}(Y) = O_{B'}(Z). \tag{7}$$

However, if $Y \subset Z$, (7) violates the condition on which Y to be a maximal itemset in B' . Hence Y is not a maximal itemset in B' .

If Y does not occur in B , Y will be a maximal itemset in B' only when $Y = t$. Under such case, Y will not be missed.

Following the above arguments, no maximal itemset will be missed by the Maximal_insert procedure. And it is obvious the time complexity of the procedure is linear. \square

To construct the maximal itemsets of a database, we can start with an empty database and repeatedly insert each transaction by the procedure Maximal_insert. As an illustrative example, the maximal itemsets of Table 1 identified by repeatedly applying Maximal_insert are shown in Table 2.

Table 2. The heap of the maximal itemsets in the sample database

Maximal itemsets	The number of occurrences
{Bread}	4
{Bread, PeanutButter}	3
{Beer}	2
{Milk}	2
{Beer, Milk}	1
{Beer, Bread}	1
{Bread, Jelly, PeanutButter}	1
{Bread, Milk, PeanutButter}	1

3 Generating Large Itemsets

Having maximal itemsets determined by the procedure in the preceding section, we build a heap of maximal itemsets that maintains the maximal itemsets in descending order of the occurrence numbers at the cost of $O(M)$, where M is the number of maximal itemsets. With the heap of maximal itemsets and a given minimum support, the following algorithm can be employed to generate large itemsets satisfying the minimum support.

Large_itemsets;

```

//input:  $M(B)$  and minimum support  $s$ 
//output: large itemsets  $L$ 
//Assume the heap  $H(M(B))$  maintains maximal itemsets  $M(B)$  in
//descending order of the number of occurrences of maximal itemsets.
 $L = \emptyset$ ;
 $X = Top(H(M(B)))$ 
While ( $O(X) \geq s * |B|$ )
    For each  $Y \subseteq X$  and  $Y \neq \emptyset$  do
         $L = L \cup Y$ ;
        RemoveTop( $H(M(B))$ )
     $X = Top(H(M(B)))$ 

```

End of Large_itemsets;

Theorem 4. *The algorithm **Large_itemsets** runs in time $O(M'2^N + M' \log M)$, where N is the maximum number of items in a maximal itemset, M' is the number of the maximal itemsets with minimum support greater than the required support, and M is the number of the maximal itemsets.*

Proof: The "While loop" iterates $O(M')$ times and each time the heap is maintained at the cost of $O(\log M)$. Therefore, the total overhead spent in the heap is $O(M' \log M)$. Since the maximum number of a maximal itemset is N , the maximum number of subsets of a maximal itemset is 2^N . Therefore, the total time spent in the "For loop" is $O(M'2^N)$. The time complexity of the algorithm is $O(M'2^N + M' \log M)$. Note that N is usually small. □

Given the sample database in Table 1, the maximal itemsets are identified by the algorithm **Maximal_insert**, which takes time $O(M)$. The heap of the maximal itemsets is shown in Table 2 Large itemsets of the sample database in Table 1 with minimum support $s = 0.3$ are shown in the third column of Table 3. Maximal itemsets in Table 2 are screened so that only those maximal itemsets with

Table 3. Large itemsets of the sample database with minimum support $s = 30\%$

Maximal itemsets	The number of occurrences	Large itemsets
{Bread}	4	{Bread}
{Bread, PeanutButter}	3	{Bread},{PeanutButter},{Bread,PeanutButter}
{Beer}	2	{Beer}
{Milk}	2	{Milk}

minimum support greater than or equal to 0.3 are left in the first column of Table 3. Large itemsets are the nonempty subsets of the maximal itemsets in the first column of Table 3.

4 Conclusions

In this paper, some properties of maximal itemset such as the equivalence relation of supreme covers are investigated. Once the maximal itemsets are identified, the large itemsets for a given minimum support can be generated in time $O(M'2^N + M' \log M)$.

Acknowledgement

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC94-2416-H-019-006-.

References

1. R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in large database," ACM SIGMOD conference, pp. 207-216, Washington DC, USA, 1993.
2. R. Agrawal, T. Imielinski and A. Swami, "Database mining: a performance perspective," IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, pp. 914-925, 1993.
3. R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," ACM International conference on Very Large Data Bases, pp. 487-499, 1994.
4. R. Agrawal and R. Srikant, "Mining sequential patterns," IEEE International Conferences on Data Engineering, pp. 3-14, 1995.
5. S. Brin, R. Motwani, J.D. Ullman and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," ACM SIGMOD Conference, pp. 255-264, Tucson, Arizona, USA, 1997.
6. D.W. Cheung, J. Han, V.T. Ng and C.Y. Wong, "Maintenance of discovered association rules in large databases: an incremental updating approach," IEEE International Conference on Data Engineering, pp. 106-114, 1996.
7. D.W. Cheung, S.D. Lee, and B. Kao, "A general incremental technique for maintaining discovered association rules," In Proceedings of Database Systems for Advanced Applications, pp. 185-194, Melbourne, Australia, 1997.
8. R. Feldman, Y. Aumann, A. Amir, and H. Mannila, "Efficient algorithms for discovering frequent sets in incremental databases," ACM SIGMOD Workshop on DMKD, pp. 59-66, USA, 1997.
9. T.P. Hong, C.Y. Wang and Y.H. Tao, "A new incremental data mining algorithm using pre-large itemsets," International Journal on Intelligent Data Analysis, 2001.
10. H.-S. Lee, "Incremental association mining based on maximal itemsets", Lecture Notes in Computer Science 3681 (2005) 365-371.
11. N.L. Sarda and N.V. Srinivas, "An adaptive algorithm for incremental mining of association rules," IEEE International Workshop on Database and Expert Systems, pp. 240-245, 1998.

A Fuzzy Multiple Criteria Decision Making Model for Airline Competitiveness Evaluation

Hsuan-Shih Lee¹ and Ming-Tao Chou²

¹ Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan

² Department of Aviation and Maritime Management
Chang Jung Christian University
Taiwan 711, Taiwan

Abstract. This paper presents a fuzzy multiple criteria decision making model to the evaluation of airline competitiveness over a period. The evaluation problem is formulated as a fuzzy multiple criteria decision making problem and solved by our strength-weakness based approach. After the strength and weakness matrices for airlines are derived, the weights of criteria, strength matrix and weakness matrix can be aggregated into strength indices and weakness indices for airlines, by which each airline can identify his own strength and weakness. The strength and weakness indices can be further integrated into an overall performance indices, by which airlines can identify their competitiveness ranking.

1 Introduction

Airline competitiveness can be measured by a range of efficiency and effectiveness performance measures across a number of distinct dimensions that can reflect the capabilities and offerings of airlines in serving their customers. The performance evaluation of airlines can be measured in terms of some key competitiveness measures, such as cost [1], operational performance [2,3], cost and productivity [4,5], price and productivity [6], price and service quality [7], productivity and efficiency [8,9,10,11], profitability [7,10], safety [12], service quality [13,14], and service quality and productivity [15]. However, these single measures alone do not reflect the overall airline competitiveness. In this paper, we assume the key performance measures used are cost (C_1), productivity (C_2), service quality (C_3), price (C_4), and management (C_5) [16].

An airline competitiveness evaluation problem can be formulated as a multiple criteria decision making (MCDM) problem in which alternatives are the airlines to be evaluated and criteria are the performance measures of airlines under consideration. In traditional MCDM, performance rating and weights are measured in crisp numbers. To evaluate competitiveness of airlines in a specific year, traditional MCDM methods may suffice, since all performance ratings are crisp. However, if we want to evaluate the competitiveness of airlines over a period, say 5 years, traditional MCDM methods may be inadequate. We can

not represent the performance of an airline under a specific measure by a crisp number, since the performance may vary within a range in 5 years. One way to represent the overall performance over a period is to represent the performance by a fuzzy number. Therefore, fuzzy multiple criteria decision making (FMCDM) is introduced to evaluate the performance of airlines over a period. A FMCDM for m airlines and n criteria can be modeled as follows:

$$D = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & \dots & \tilde{A}_{1n} \\ \tilde{A}_{21} & \tilde{A}_{22} & \dots & \tilde{A}_{2n} \\ \tilde{A}_{m1} & \tilde{A}_{m2} & \dots & \tilde{A}_{mn} \end{bmatrix}$$

and

$$W = [\tilde{W}_1 \tilde{W}_2 \dots \tilde{W}_n]$$

where \tilde{A}_{ij} is the fuzzy number representing the performance of i th airline under j th criterion and \tilde{W}_j is the fuzzy number representing the weight of j th criterion over a period.

In dealing with fuzzy numbers, ranking fuzzy number is one of the important issues. Lee [18] has proposed a new fuzzy ranking method based on fuzzy preference relation satisfying all criteria proposed by Yuan [21]. In [19], we extended the definition of fuzzy preference relation and proposed an extended fuzzy preference relation which satisfies additivity and is easy to compute. Based on previous result, in this paper, we are going to propose a new method for evaluating competitiveness of airlines over a period.

2 Preliminaries

Definition 1. The α -cut of fuzzy set \tilde{A} , \tilde{A}^α , is the crisp set $\tilde{A}^\alpha = \{x \mid \mu_{\tilde{A}}(x) \geq \alpha\}$. The support of \tilde{A} is the crisp set $Supp(\tilde{A}) = \{x \mid \mu_{\tilde{A}}(x) > 0\}$. \tilde{A} is normal iff $\sup_{x \in U} \mu_{\tilde{A}}(x) = 1$, where U is the universe set.

Definition 2. \tilde{A} is a fuzzy number iff \tilde{A} is a normal and convex fuzzy subset of real number.

Definition 3. A triangular fuzzy number \tilde{A} is a fuzzy number with piecewise linear membership function $\mu_{\tilde{A}}$ defined by

$$\mu_{\tilde{A}}(x) = \begin{cases} \frac{x-a_1}{a_2-a_1}, & a_1 \leq x \leq a_2, \\ \frac{a_3-x}{a_3-a_2}, & a_2 \leq x \leq a_3, \\ 0, & \text{otherwise,} \end{cases}$$

which can be denoted as a triplet (a_1, a_2, a_3) .

Definition 4. Let \tilde{A} be a fuzzy number. Then \tilde{A}_α^L and \tilde{A}_α^U are defined as $\tilde{A}_\alpha^L = \inf_{\mu_{\tilde{A}}(z) \geq \alpha} (z)$ and $\tilde{A}_\alpha^U = \sup_{\mu_{\tilde{A}}(z) \geq \alpha} (z)$ respectively.

Definition 5. [20] An extended fuzzy preference relation R on fuzzy numbers is an extended fuzzy subset of the product of fuzzy numbers with membership function $-\infty \leq \mu_R(\tilde{A}, \tilde{B}) \leq \infty$ being the preference degree of fuzzy number \tilde{A} to fuzzy number \tilde{B} .

1. R is reciprocal iff $\mu_R(\tilde{A}, \tilde{B}) = -\mu_R(\tilde{B}, \tilde{A})$ for all fuzzy numbers \tilde{A} and \tilde{B} .
2. R is transitive iff $\mu_R(\tilde{A}, \tilde{B}) \geq 0$ and $\mu_R(\tilde{B}, \tilde{C}) \geq 0 \Rightarrow \mu_R(\tilde{A}, \tilde{C}) \geq 0$ for all fuzzy numbers \tilde{A} , \tilde{B} and \tilde{C} .
3. R is additive iff $\mu_R(\tilde{A}, \tilde{C}) = \mu_R(\tilde{A}, \tilde{B}) + \mu_R(\tilde{B}, \tilde{C})$
4. R is a total ordering iff R is reciprocal, transitive and additive.

In [20], we have defined an extended fuzzy preference relation on fuzzy numbers, which is reciprocal, transitive and additive. Some results in [20] are as follows.

Definition 6. [20] For any fuzzy numbers \tilde{A} and \tilde{B} , we define the extended fuzzy preference relation $F(\tilde{A}, \tilde{B})$ by the membership function

$$\mu_F(\tilde{A}, \tilde{B}) = \int_0^1 ((\tilde{A} - \tilde{B})_\alpha^L + (\tilde{A} - \tilde{B})_\alpha^U) d\alpha \tag{1}$$

Lemma 1. [20] F is reciprocal, i.e., $\mu_F(\tilde{B}, \tilde{A}) = -\mu_F(\tilde{A}, \tilde{B})$.

Lemma 2. [20] F is additive, i.e., $\mu_F(\tilde{A}, \tilde{B}) + \mu_F(\tilde{B}, \tilde{C}) = \mu_F(\tilde{A}, \tilde{C})$.

Lemma 3. [20] F is transitive, i.e., $\mu_F(\tilde{A}, \tilde{B}) \geq 0$ and $\mu_F(\tilde{B}, \tilde{C}) \geq 0 \Rightarrow \mu_F(\tilde{A}, \tilde{C}) \geq 0$.

Lemma 4. [20] Let $\tilde{A} = (a_1, a_2, a_3)$ and $\tilde{B} = (b_1, b_2, b_3)$ be two triangular fuzzy numbers. Then $\mu_F(\tilde{A}, \tilde{B}) = (a_1 + 2a_2 + a_3 - b_1 - 2b_2 - b_3)/2$.

3 The Proposed Method

In this section, we propose a method for the performance of airlines over a period based on the concepts proposed in [20]. For clarity, some results in [20] are reiterated here. The preference function of one fuzzy number \tilde{A}_{ij} over another number \tilde{A}_{kj} is defined as follows:

$$P(\tilde{A}_{ij}, \tilde{A}_{kj}) = \begin{cases} \mu_F(\tilde{A}_{ij}, \tilde{A}_{kj}) & \text{if } \mu_F(\tilde{A}_{ij}, \tilde{A}_{kj}) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Let J be the set of benefit criteria and J' be the set of cost criteria where

$$J = \{1 \leq j \leq n \text{ and } j \text{ belongs to the benefit criteria}\}$$

$$J' = \{1 \leq j \leq n \text{ and } j \text{ belongs to the cost criteria}\},$$

and

$$J \cup J' = \{1, \dots, n\}.$$

The strength matrix $S = (S_{ij})$ is given by letting

$$S_{ij} = \begin{cases} \sum_{k \neq i} P(\tilde{A}_{ij}, \tilde{A}_{kj}) & \text{if } j \in J \\ \sum_{k \neq i} P(\tilde{A}_{kj}, \tilde{A}_{ij}) & \text{if } j \in J'. \end{cases} \tag{2}$$

Similarly, the weakness matrix $I = (I_{ij})$ is given by letting

$$I_{ij} = \begin{cases} \sum_{k \neq i} P(\tilde{A}_{kj}, \tilde{A}_{ij}) & \text{if } j \in J \\ \sum_{k \neq i} P(\tilde{A}_{ij}, \tilde{A}_{kj}) & \text{if } j \in J'. \end{cases} \tag{3}$$

The fuzzy weighted strength matrix $\tilde{S} = (\tilde{S}_i)$ can be obtained by

$$\tilde{S}_i = \sum_j S_{ij} \tilde{W}_j \tag{4}$$

and the fuzzy weighted weakness matrix $\tilde{I} = (\tilde{I}_i)$ can be obtained by

$$\tilde{I}_i = \sum_j I_{ij} \tilde{W}_j, \tag{5}$$

where $1 \leq i \leq m$. Now we are ready to present our competitiveness evaluation method.

Step 1: Assume the evaluation interval is a period of T years. Let p_{ij}^t be the performance value of i -th airline under j -th criterion in year $1 \leq t \leq T$. Let triangular fuzzy number \tilde{A}_{ij} be the fuzzy performance of i -th airline under j -th criterion and denoted as $(a_{ij1}, a_{ij2}, a_{ij3})$. Define fuzzy performance matrix $D = (\tilde{A}_{ij})$ of airlines over a period by letting

$$a_{ij1} = \min_{1 \leq t \leq T} p_{ij}^t, \tag{6}$$

$$a_{ij2} = \frac{\sum_{t=1}^T p_{ij}^t}{T}, \tag{7}$$

and

$$a_{ij3} = \max_{1 \leq t \leq T} p_{ij}^t. \tag{8}$$

Step 2: Calculate the strength matrix by (2).

Step 3: Calculate the weakness matrix by (3).

Step 4: Calculate the fuzzy weighted strength indices by (4).

Step 5: Calculate the fuzzy weighted weakness indices by (5).

Step 6: Derive the strength index S_i from the fuzzy weighted strength and weakness indices by

$$S_i = \sum_{k \neq i} P(\tilde{S}_i, \tilde{S}_k) + \sum_{k \neq i} P(\tilde{I}_k, \tilde{I}_i) \tag{9}$$

Step 7: Derive the weakness index I_i from the fuzzy weighted strength and weakness indices by

$$I_i = \sum_{k \neq i} P(\tilde{S}_k, \tilde{S}_i) + \sum_{k \neq i} P(\tilde{I}_i, \tilde{I}_k) \tag{10}$$

Step 8: Aggregate the strength and weakness indices into total performance indices by

$$t_i = \frac{S_i}{S_i + I_i} \tag{11}$$

Step 9: Rank airlines by total performance indices t_i for $1 \leq i \leq m$.

4 Numerical Example

Assume there are three airlines to be evaluated under 5 criteria (cost, productivity, service quality, price, and management) over a period of 5 years. Performance ratings of airlines in 5 years are given in five-point Likert scales and converted into fuzzy numbers by (6), (7), and (8) as shown in table 1. Assume fuzzy weights of criteria given by experts as also shown in table 1. The competitiveness ranking of airlines is solved as follows:

Table 1. The fuzzy decision matrix and fuzzy weights

	C_1	C_2	C_3	C_4	C_5
A_1	(3,4,5)	(3,3.6,4)	(3,3.4,4)	(3,3.8,4)	(2,2.6,4)
A_2	(1,1.8,3)	(4,4.4,5)	(4,4.2,5)	(4,4.4,5)	(4,4.6,5)
A_3	(3,3.4,4)	(3,3.4,4)	(3,3.6,4)	(3,3.2,4)	(3,3.6,4)
Weight	(0.7,0.9,1)	(0.9,1,1)	(0.77,0.93,1)	(0.9,1,1)	(0.43,0.63,0.83)

Step 1: The fuzzy performance of airlines and fuzzy weights of criteria are shown in table 1

Step 2: The strength matrix derived by (2) is shown in table 2.

Step 3: The weakness matrix derived by (3) is shown in table 3.

Step 4: The fuzzy weighted strength indices of airlines derived by (4) are shown in table 4.

Step 5: The fuzzy weighted weakness indices of airlines derived by (5) are shown in table 5.

Step 6: The strength indices of airlines derived by (9) are shown in table 6.

Step 7: The weakness indices of airlines derived by (10) are shown in table 7.

Step 8: The total performance indices aggregated by (11) are shown in table 8.

Step 9: The rank of airlines by total performance indices are shown in table 9.

Table 2. The strength matrix

	C_1	C_2	C_3	C_4	C_5
A_1	0	0.2	0	0.6	0
A_2	7.3	3.8	3.4	3.8	5.5
A_3	1.1	0	0.2	0	1.5

Table 3. The weakness matrix

	C_1	C_2	C_3	C_4	C_5
A_1	5.3	1.8	2	1.6	5
A_2	0	0	0	0	0
A_3	3.1	2.2	1.6	2.8	2

Table 4. The fuzzy weighted strength indices of airlines

	fuzzy weighted strength index
A_1	(0.72, 0.8, 0.8)
A_2	(16.933, 20.797, 22.865)
A_3	(1.569, 2.121, 2.545)

Table 5. The fuzzy weighted weakness indices of airlines

	fuzzy weighted weakness index
A_1	(10.46, 13.18, 14.85)
A_2	(0, 0, 0)
A_3	(8.762, 10.538, 11.36)

Table 6. The strength indices of airlines

	strength index
A_1	0
A_2	122.088
A_3	7.854

Table 7. The weakness indices of airlines

	weakness index
A_1	72.825
A_2	0
A_3	57.117

Table 8. The total performance indices of airlines

total performance index	
A_1	0
A_2	1
A_3	0.12088

Table 9. The rank of airlines based on total performance indices

rank	
A_1	3
A_2	1
A_3	2

5 Conclusions

In this paper, we have presented a FMCDM for airline performance over a period. With our method, two matrices are constructed. Namely, they are the strength matrix and weakness matrix from which the strength and weakness indices are derived. With strength and weakness indices, airlines can identify their strength and weakness under the performance measures taken into consideration. Airlines can identify their competitive positions by the overall performance indices obtained by aggregating the strength and weakness indices.

Acknowledgment

This research work was supported by the National Science Council of the Republic of China under grant No. NSC94-2416-H-019-006-.

References

1. T.H. Oum, C. Yu, Cost competitiveness of major airlines: an international comparison, *Transportation Research A* 32(6) (1998) 407-422.
2. Bureau of Industry Economics. Aviation: international performance indicators. Research report 59. Canberra: Australian Government Publishing Service, 1994.
3. M. Schefczyk, Operational performance of airlines: an extension of traditional measurement paradigms, *Strategic Management Journal* 14 (1993) 301-317.
4. D. Encaoua, Liberalizing European airlines: cost and factor productivity evidence, *International Journal of Industrial Organization* 9 (1991) 109-124.
5. R. Windle, The world's airlines: a cost and productivity comparison, *Journal of Transport Economics and Policy* 25(1) (1991) 31-49.
6. D.H. Good, E.L. Rhodes, Productive efficiency, technological change and the competitiveness of U.S. airlines in the Pacific Rim, *Journal of the Transportation Research Forum* 31(2) (1991) 347-358.
7. Bureau of Transportation and Communications Economics, The progress of aviation reform, Research report 81, Canberra: Australian Government Publishing Service (1993).

8. D.H. Good, M.I. Nadiri, L.H. Roller, R.C. Sickles, Efficiency and productivity growth comparisons of European and U.S. airlines: a first look at the data, *The Journal of Productivity Analysis* 4 (1993) 115-125.
9. D.H. Good, L.H. Roller, R.C. Sickles, Airline efficiency differences between Europe and the US: implications for the pace of EC integration and domestic regulation, *European Journal of Operational Research* 80(1) (1995) 508-518.
10. T.H. Oum, C. Yu, A productivity comparison of the world's major airlines, *Journal of Air Transport Management* 2(3/4) (1995) 181-195.
11. R. Windle, M. Dresner, A note on productivity comparisons between air carries, *Logistics and Transportation Review* 31(2) (1995) 125-134.
12. M. Janic, An assessment of risk and safety in civil aviation, *Journal of Air Transport Management* 6 (2000) 43-50.
13. Y.H. Chang, C.H. Yeh, A survey analysis of service quality for domestic airlines, *European Journal of Operational Research* 139(1) (2002) 166-177.
14. C. Young, C. Lawrence, M. Lee, Assessing service quality as an effective management tool: the case fo the airline industry, *Journal of Marketing Theory and Practice* 2(2) 1994 76-96.
15. L.J. Truitt, R. Haynes, Evaluating service quality and productivity in the regional airline industry, *Transportation Journal* 33(4) 1994 21-32.
16. Y.-H. Chang, C.-H Yeh, Evaluating airline competitiveness using multiattribute decision making, *Omega* 29 (2001) 405-415.
17. C.L. Hwang, K. Yoon, *Multiple Attributes Decision Making Methods and Applications*, Springer, Berlin Heidelberg, 1981.
18. H.-S. Lee, A new fuzzy ranking method based on fuzzy preference relation, *2000 IEEE International Conference on Systems, Man And Cybernetics* (2001) 3416-3420.
19. H.-S. Lee, An extended fuzzy preference relation for comparison of fuzzy numbers, *The 6h World Multi-Conference on Systemics, Cybernetics and Informatics*, July 14-18, 2002, Orlando, USA, XI 76-79.
20. H.-S. Lee, A fuzzy multi-criteria decision making model for the selection of distribution center, *Lecture Notes in Computer Science* 3612 (2005) 1290-1299.
21. Y. Yuan, Criteria for evaluating fuzzy ranking methods, *Fuzzy Sets and Systems* 44, 139-157 (1991).

Goal Programming Methods for Constructing Additive Consistency Fuzzy Preference Relations

Hsuan-Shih Lee¹ and Wei-Kuo Tseng²

¹ Department of Shipping and Transportation Management
National Taiwan Ocean University

² Department of Logistics Management
China College of Marine Technology and Commerce
Department of Shipping and Transportation Management
National Taiwan Ocean University

Abstract. Decision makers may present their preferences over alternatives as fuzzy preference relations. Usually, there exist inconsistencies in the preference relation given by decision makers. In this paper, we propose methods based on goal programming to obtain fuzzy preference relations that satisfy additive consistency from the subjective preference relations given by decision makers.

1 Introduction

Decision-making process usually consists of multiple individuals interacting to reach a decision. Each decision maker (expert) may have unique motivations or goals and may approach the decision process from a different angle, but have a common interest in reaching eventual agreement on selecting the best options. To do this, experts have to express their preferences by means of a set of evaluations over a set of alternatives. It has been common practice in research to model decision-making problems in which all the experts express their preferences using the same preference representation format. However, in real practice this is not always possible because each expert has his unique characteristics with regard to knowledge, skills, experience and personality, which implies that different experts may express their evaluations by means of different preference representation formats. In fact, this is an issue that recently has attracted the attention of many researchers in the area of decision-making, and as a result different approaches to integrating different preference representation formats have been proposed [1,2,8,10,28,29]. In these research papers, many reasons are provided for fuzzy preference relations to be chosen as the base element of that integration. Group fuzzy decision making models deal with problems in which performance ratings of alternatives and weights of the attributes may be given in fuzzy numbers. Different models have been proposed for group fuzzy decision making problems [15-18].

Another important issue to bear in mind when information is provided by experts is that of “consistency” [3,4,11]. Due to the complexity of most decision-making problems, experts’ preferences may not satisfy formal properties that fuzzy preference

relations are required to verify. Consistency is one of them, and it is associated with the transitivity property.

Many properties have been suggested to model transitivity of fuzzy preference relations and consequently, consistency may be measured according to which of these different properties are required to be satisfied. Some of these suggested properties are as follows:

- (1) Triangle condition [11,19]
- (2) Weak transitivity [11,23]
- (3) Max-min transitivity [11,27]
- (4) Max-max transitivity [7,11,27]
- (5) Restricted max-min transitivity [11,23]
- (6) Restricted max-max transitivity [11,23]
- (7) Additive transitivity [11,19,23]
- (8) Multiplicative transitivity [11,22,23,25,26]

Amongst these properties two of them attract more attentions in recent research [11,25,26], which are additive transitivity and multiplicative transitivity.

Many methods have been proposed to draw consistent preferences from a multiplicative preference relation, such as the eigenvector method [22], the least square method [12], gradient eigenvector method [5], Logarithmic least square method [6], generalized chi square method [24], etc. When using fuzzy preference relations, some priority methods have been given using what have been called choice functions or degrees [1,9,13,14,20,21]. In this paper, we propose two least deviation methods for constructing additive consistent fuzzy preference relations from fuzzy preference relations.

2 Preliminaries

For simplicity, we let $N = \{1, 2, \dots, n\}$.

Definition 2.1. Let $R = (r_{ij})_{n \times n}$ be a preference relation, then R is called a fuzzy preference relation [2,13,23], if

$$r_{ij} \in [0,1], \quad r_{ij} + r_{ji} = 1, \quad r_{ii} = 0.5 \quad \text{for all } i, j \in N .$$

Definition 2.2. Let $R = (r_{ij})_{n \times n}$ be a fuzzy preference relation, then R is called an additive consistent fuzzy preference relation, if the following additive transitivity (given by Tanino [23]) is satisfied:

$$r_{ij} = r_{ik} - r_{jk} + 0.5, \quad \text{for all } i, j, k \in N$$

Let $w = (w_1, w_2, \dots, w_n)$ be the priority vector of the additive preference relation

$R = (r_{ij})_{n \times n}$, where $w_i > 0$, $i = 1, 2, \dots, n$, $\sum_{i=1}^n w_i = 1$. If $R = (r_{ij})_{n \times n}$ is an

additive consistent preference relation, then such a preference relation is given by

$$r_{ij} = \frac{w_i - w_j}{2} + 0.5, \quad i, j = 1, 2, \dots, n. \tag{1}$$

3 Methods Based on Least Deviation

Let $w = (w_1, w_2, \dots, w_n)$ be the priority vector of the fuzzy preference relation

$R = (r_{ij})_{n \times n}$, where $w_i > 0, i = 1, 2, \dots, n, \sum_{i=1}^n w_i = 1$. If $R = (r_{ij})_{n \times n}$ is an

additive consistent fuzzy preference relation, then such a preference relation must satisfy (1):

$$r_{ij} = \frac{w_i - w_j}{2} + 0.5, \quad i, j = 1, 2, \dots, n$$

However, in general, (1) does not hold. We relax (1) into (2):

$$r_{ij} \approx \frac{w_i - w_j}{2} + 0.5, \quad i, j = 1, 2, \dots, n. \tag{2}$$

We want to find a priority vector $w = (w_1, w_2, \dots, w_n)$ such that r_{ij} is as close as

possible to $\frac{w_i - w_j}{2} + 0.5$. Thus we can construct the following multiple objective

programming model:

$$\begin{aligned} \text{(MOP1)} \quad & \min |r_{ij} - (\frac{w_i - w_j}{2} + 0.5)|, \quad i, j \in N \\ \text{s.t.} \quad & w_i \geq 0, \quad i \in N, \quad \sum_{i=1}^n w_i = 1. \end{aligned} \tag{3}$$

The problem of finding a priority vector can also be formulated as the following programming model:

$$\begin{aligned} \text{(P1)} \quad & \min \sum_{i=1}^n \sum_{j=1}^n |r_{ij} - (\frac{w_i - w_j}{2} + 0.5)| \\ \text{s.t.} \quad & w_i \geq 0, \quad i \in N, \quad \sum_{i=1}^n w_i = 1. \end{aligned} \tag{4}$$

The model (MOP1) can be transformed into the following multiple objective programming model:

$$\begin{aligned}
 & \text{(MOP2) } \min (p_{ij} + q_{ij}), \quad i, j \in N \\
 & \text{s.t.} \quad r_{ij} - \left(\frac{w_i - w_j}{2} + 0.5\right) = p_{ij} - q_{ij} \\
 & \quad p_{ij}, q_{ij} \geq 0, \quad i, j \in N \\
 & \quad w_i \geq 0, \quad i \in N, \quad \sum_{i=1}^n w_i = 1.
 \end{aligned} \tag{5}$$

The model (MOP2) can be approached with the following fuzzy multiple objective programming model:

$$\begin{aligned}
 & \text{(LP1) } \max \lambda \\
 & \text{s.t.} \quad 1 - (p_{ij} + q_{ij}) \geq \lambda, \quad i, j \in N \\
 & \quad r_{ij} - \left(\frac{w_i - w_j}{2} + 0.5\right) = p_{ij} - q_{ij}, \quad i, j \in N \\
 & \quad p_{ij}, q_{ij} \geq 0, \quad i, j \in N \\
 & \quad w_i \geq 0, \quad i \in N, \quad \sum_{i=1}^n w_i = 1.
 \end{aligned} \tag{6}$$

The model (P1) can be transformed into the following linear programming model:

$$\begin{aligned}
 & \text{(LP2) } \min \sum_{i=1}^n \sum_{j=1}^n (p_{ij} + q_{ij}) \\
 & \text{s.t.} \quad r_{ij} - \left(\frac{w_i - w_j}{2} + 0.5\right) = p_{ij} - q_{ij}, \quad i, j \in N \\
 & \quad p_{ij}, q_{ij} \geq 0, \quad i, j \in N \\
 & \quad w_i \geq 0, \quad i \in N, \quad \sum_{i=1}^n w_i = 1.
 \end{aligned} \tag{7}$$

Let $(w_1^*, w_2^*, \dots, w_n^*)$ be the priority vector obtained from the model (LP1) or the model (LP2). Then an additive consistent fuzzy preference relation $R^* = (r_{ij}^*)_{n \times n}$ can be obtained from $R = (r_{ij})_{n \times n}$ by

$$r_{ij}^* = (w_i^* - w_j^*)/2 + 0.5 \tag{8}$$

4 Numerical Examples

In this section, two numerical examples are presented to illustrate the models proposed by us.

Example 4.1. For a decision-making problem, there are three alternatives under consideration. The decision maker provides his/her preferences over these three alternatives in the following fuzzy preference relation:

$$R = \begin{bmatrix} 0.5 & 0.58 & 0.2 \\ 0.35 & 0.5 & 0.1 \\ 0.7 & 0.85 & 0.5 \end{bmatrix}.$$

By applying the model (LP1) to the fuzzy preference relation, we have the priority vector:

$$(w_1^*, w_2^*, w_3^*) = (0.25, 0, 0.75).$$

Following (8), we have the following additive consistent fuzzy preference relation:

$$R^* = \begin{bmatrix} 0.5 & 0.625 & 0.25 \\ 0.375 & 0.5 & 0.125 \\ 0.75 & 0.875 & 0.5 \end{bmatrix}$$

Example 4.2. For a decision-making problem, assume there are three alternatives under consideration. The decision maker provides his/her preferences over these three alternatives in the following fuzzy preference relation:

$$R = \begin{bmatrix} 0.5 & 0.6 & 0.2 \\ 0.35 & 0.5 & 0.1 \\ 0.7 & 0.85 & 0.5 \end{bmatrix}.$$

By applying the model (LP2) to the fuzzy preference relation, we have the priority vector:

$$(w_1^*, w_2^*, w_3^*) = (0.233333, 0.0333333, 0.733334).$$

Following (8), we have the following additive consistent fuzzy preference relation:

$$R^* = \begin{bmatrix} 0.5 & 0.6 & 0.25 \\ 0.4 & 0.5 & 0.15 \\ 0.75 & 0.85 & 0.5 \end{bmatrix}$$

5 Conclusion

In light of fuzzy preference relations, the problem of constructing consistent fuzzy preference relations has been addressed. In the process of decision making, decision

makers may provide their preferences over alternatives under consideration as fuzzy preference relations. However, there may exist inconsistency within the preference relation provided by decision makers. In this paper, we have proposed two models for constructing fuzzy preference relations that satisfy additive consistency. Numerical examples are also presented to illustrate the application of our models.

Acknowledgement

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC94-2416-H-019-006-.

References

1. F. Chiclana, F. Herrera, E. Herrera-Viedma, Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations, *Fuzzy Sets and Systems* 97 (1998) 33-48.
2. F. Chiclana, F. Herrera, E. Herrera-Viedma, Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations, *Fuzzy Sets and Systems* 122 (2001) 277-291.
3. F. Chiclana, F. Herrera F. E. Herrera-Viedma, Reciprocity and consistency of fuzzy preference relations, In: B. De Baets, J. Fodor (Eds.), *Principles of Fuzzy Preference Modelling and Decision Making*, Academia Press (2003) 123-142.
4. F. Chiclana, F. Herrera F. E. Herrera-Viedma, Rationality of induced ordered weighted operators based on the reliability of the source of information in group decision-making, *Kybernetika* 40 (2004) 121-142.
5. K.O. Cogger, P.L. Yu, Eigenweight vectors and least-distance approximation for revealed preference in pairwise weight ratios, *Journal of Optimization Theory and Application* 46 (1985) 483-491.
6. G. Crawford, C. Williams, A note on the analysis of subjective judgement matrices, *Journal of Mathematical Psychology* 29 (1985) 387-405.
7. D. Dubois, H. Prade, *Fuzzy Sets and Systems: Theory and Application*, New York: Academic Press (1980).
8. Z.-P. Fan, S.-H. Xial, G.-F. Hu, An optimization method for integrating tow kinds of preference information in group decision-making, *Computers & Industrial Engineering* 46 (2004) 329-335.
9. F. Herrera, E. Herrera-Viedma, J.L. Verdegay, A sequential selection process in group decision-making with linguistic assessment, *Information Sciences* 85 (1995) 223-239.
10. F. Herrera, L. Martinez, P.J. Sanchez, Managing non-homogeneous information in group decision making, *European Journal of Operational Research* 166 (2005) 115-132.
11. E. Herrera-Viedma, F. Herrera, F. Chiclana, M. Luque, Some issues on consistency of fuzzy preference relations, *European Journal of Operational Research* 154 (2004) 98-109.
12. R.E. Jensen, An alternative scaling method for priorities in hierarchical structures, *Journal of Mathematical Psychology* 28 (1984) 317-332.
13. J. Kacprzyk, Group decision making with a fuzzy linguistic majority, *Fuzzy Sets and Systems* 18 (1986) 105-118.

14. J. Kacprzyk, M. Roubens, *Non-Conventional Preference Relations in Decision-Making*, Berlin, Springer (1988).
15. H.-S. Lee, Optimal consensus of fuzzy opinions under group decision making environment, *Fuzzy Sets and Systems* 132(3) (2002) 303-315.
16. H.-S. Lee, On fuzzy preference relation in group decision making, *International Journal of Computer Mathematics* 82(2) (2005) 133-140.
17. H.-S. Lee, A Fuzzy Method for Measuring Efficiency under Fuzzy Environment, *Lecture Notes in Computer Science* 3682 (2005) 343-349.
18. H.-S. Lee, A Fuzzy Multi-Criteria Decision Making Model for the Selection of the Distribution Center, *Lecture Notes in Artificial Intelligence* 3612 (2005) 1290-1299.
19. R.D. Luce, P. Suppes, Preference utility and subject probability, In R.D. Luce (ed.) et al. *Handbook of Mathematical Psychology*, pp. 249-410, Vol. III New York: Wiley (1965).
20. S.A. Orlovvsky, Decision making with a fuzzy preference relation, *Fuzzy Sets and Systems* 1 (1978) 155-167.
21. M. Roubens, Some properties of choice functions based on valued binary relations, *European Journal of Operational Research* 40 (1989) 309-321.
22. T. L. Saaty, *The Analytic Hierarchy Process*, New York: McGraw-Hill (1980).
23. T. Tanino, Fuzzy preference orderings in group decision-making, *Fuzzy Sets and Systems* 12 (1984) 117-131.
24. Z.S. Xu, Generalized chi square method for the estimation of weights, *Journal of Optimization Theory and Applications* 107 (2002) 183-192.
25. Z.S. Xu, Two methods for ranking alternatives in group decision-making with different preference information, *Information: An International Journal* 6 (2003) 389-394.
26. Z.S. Xu, Q.L. Da, An approach to improving consistency of fuzzy preference matrix, *Fuzzy Optimization and Decision Making* 2 (2003) 3-12.
27. H.J. Zimmermann, *Fuzzy Set Theory and Its Applications*, Dordrecht: Kluwer (1991).
28. Q. Zhang, J.C.H. Chen, Y.-Q He, J. Ma, D.-N. Zhou, Multiple attribute decision making: approach integrating subjective and objective information, *International Journal of Manufacturing Technology and Management* 5(4) (2003) 338-361.
29. Q. Zhang, J.C.H. Chen, P.P. Chong, Decision consolidation: criteria weight determination using multiple preference formats, *Decision Support Systems*, 38 (2004) 247-258.

A Multiple Criteria Decision Making Model Based on Fuzzy Multiple Objective DEA

Hsuan-Shih Lee¹ and Chen-Huei Yeh²

¹ Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan

² Department of Shipping and Transportation Management
National Taiwan Ocean University
Yang Ming Marine Transport Corporation

Abstract. In multiple criteria decision making (MCDA) problems, a decision maker often needs to select or rank alternatives that are associated with non-commensurate and conflicting criteria. This paper formulates a multiple criteria decision making problem as a fuzzy multiple objective data envelopment analysis model where inputs correspond to cost criteria and outputs correspond to benefit criteria. The fuzzy multiple objective data envelopment analysis model is different from the traditional DEA model in that a common set of weights is determined so that the efficiencies of all the DMUs are maximized simultaneously by maximizing the fuzzy degree of all efficiencies.

1 Introduction

In multiple criteria decision making (MCDM) problems, a decision maker often needs to select or rank alternatives that are associated with noncommensurate and conflicting criteria [3,6], which can be represented as follows:

$$D = \begin{matrix} & C_1 & C_2 & \dots & C_n \\ \begin{matrix} A_1 \\ A_2 \\ \vdots \\ A_m \end{matrix} & \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{matrix}$$

and

$$W = [w_1 \ w_2 \ \dots \ w_n]$$

where A_1, A_2, \dots, A_m are the alternatives to be evaluated, C_1, C_2, \dots, C_n are performance measures against which performance of alternatives are measured, a_{ij} is the performance rating of i th alternative against j th criterion, and w_j is the weight of j th criterion.

MCDM problems arise in many real-world situations [3,4,6,8], for example, in production planning problems, criteria such as production rate, quality, and

cost of operations are considered in the selection of the satisfactory plan. A lot of research work has been conducted on MCDM.

One of the hot research topics in MCDM is the determination of the weights of the criteria. Criteria importance is a reflection of the decision maker’s subjective preference as well as the objective characteristics of the criteria themselves [12]. The subjective preference is usually assigned by the decision makers based on their own experiences, knowledge and perception of the problem via a preference elicitation technique such as the analytic hierarchy process (AHP) [9]. This process of assigning subjective preferences to the criteria is referred to as subjective weighting. Such methods can be found in [1,5,10]. The objective preference can be drawn from the performance ratings by methods such as Shannon’s entropy concept [11].

In this paper, we propose a method to elicit objective preferences to criteria from performance ratings based on data envelopment analysis (DEA) [2], which is a methodology that has been widely used to measure relative efficiencies within a group of decision making units (DMUs) that utilize several inputs to produce a set of outputs. It has been applied to evaluate schools, hospitals and various organizations with multiple inputs and outputs. It also can be applied to the problems where the inputs and outputs are uncertain [7]. However, in the philosophy of DEA, each DMU (alternative) is evaluated by choosing the weights of inputs and outputs (criteria) which are the best for the evaluated DMU. That is, all DMUs are not evaluated with common weights and hence the discrimination power of DEA is usually low. To enhance the discrimination power of DEA, we propose a fuzzy multiple objective model for DEA so that all DMUs are evaluated with a common set of weights.

2 The Data Envelopment Analysis Model

DEA is a mathematical model that measures the relative efficiency of DMUs with multiple inputs and outputs with no obvious production function to aggregate the data in its entirety. Relative efficiency is defined as the ratio of total weighted output over weighted input. By comparing n units with s outputs denoted by y_{rk} , $r = 1, \dots, s$ and m inputs denoted by x_{ik} , $i = 1, \dots, m$, the efficiency measure for DMU k is

$$\begin{aligned}
 h_k = & \text{Max } \sum_{r=1}^s u_r y_{rk} \\
 \text{s.t. } & \sum_{i=1}^m v_i x_{ik} = 1, \\
 & \sum_{i=1}^m v_i x_{ij} - \sum_{r=1}^s u_r y_{rj} \geq 0 \text{ for } j = 1, \dots, n, \\
 & u_r \geq 0 \text{ for } r = 1, \dots, s, \\
 & v_i \geq 0 \text{ for } i = 1, \dots, m.
 \end{aligned} \tag{1}$$

Model (1), often referred to as the input-oriented CCR (Charnes Cooper Rhodes) model [2], assumes that the production function exhibits constant returns-to-scale.

From model (1), we can formulate the following multiple objective linear programming model:

$$\begin{aligned}
 &Max \sum_{r=1}^s u_r y_{r1} \\
 &Max 1 + \sum_{r=1}^s u_r y_{r1} - \sum_{i=1}^m v_i x_{i1} \\
 &Max \sum_{r=1}^s u_r y_{r2} \\
 &Max 1 + \sum_{r=1}^s u_r y_{r2} - \sum_{i=1}^m v_i x_{i2} \\
 &\vdots \\
 &Max \sum_{r=1}^s u_r y_{rm} \\
 &Max 1 + \sum_{r=1}^s u_r y_{rm} - \sum_{i=1}^m v_i x_{im} \\
 &s.t. \quad \sum_{i=1}^m v_i x_{ij} \leq 1, \text{ for } j = 1, \dots, n, \\
 &\quad \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} \leq 0 \text{ for } j = 1, \dots, n, \\
 &\quad u_r \geq 0 \text{ for } r = 1, \dots, s, \\
 &\quad v_i \geq 0 \text{ for } i = 1, \dots, m.
 \end{aligned} \tag{2}$$

Since the objective functions in (2) satisfy that

$$0 \leq \sum_{i=1}^m v_i x_{ij} \leq 1, j = 1, \dots, n$$

and

$$0 \leq 1 + \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij}, j = 1, \dots, n,$$

we formulate (2) as the following fuzzy multiple objective linear programming:

$$\begin{aligned}
 &Max \alpha \\
 &s.t. \quad 1 + \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} \geq \alpha, \text{ for } j = 1, \dots, n, \\
 &\quad \sum_{i=1}^m v_i x_{ij} \geq \alpha \text{ for } j = 1, \dots, n, \\
 &\quad \sum_{i=1}^m v_i x_{ij} \leq 1, \text{ for } j = 1, \dots, n, \\
 &\quad \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} \leq 0 \text{ for } j = 1, \dots, n, \\
 &\quad u_r \geq 0 \text{ for } r = 1, \dots, s, \\
 &\quad v_i \geq 0 \text{ for } i = 1, \dots, m.
 \end{aligned} \tag{3}$$

3 The Multiple Criteria Decision Model

Assume the alternatives are known. Let $S = \{S_1, S_2, \dots, S_n\}$ denote a discrete set of n possible alternatives. We also assume the criteria are known. Let $R = \{R_1, R_2, \dots, R_m\}$ denote the set of criteria. Let B denote the set of benefit criteria and C denote the set of cost criteria. That is $B \cup C = R$. Let $A = [a_{ij}]_{n \times m}$ denote the decision matrix where $a_{ij} (\geq 0)$ is the consequence with a numerical value for alternative S_j with respect to criterion R_i , where $i = 1, \dots, m$ and $j = 1, \dots, n$. The multiple criteria decision making model based on (3) can be formulated as follows:

$$\begin{aligned}
 & \text{Max } \alpha \\
 & \text{s.t.} \\
 & \quad 1 + \sum_{i \in B} w_i a_{ij} - \sum_{i \in C} w_i a_{ij} \geq \alpha, \text{ for } j = 1, \dots, n, \\
 & \quad \sum_{i \in B} w_i a_{ij} - \sum_{i \in C} w_i a_{ij} \leq 1, \text{ for } j = 1, \dots, n, \\
 & \quad \sum_{i \in B} w_i a_{ij} - \sum_{i \in C} w_i a_{ij} \leq 0 \text{ for } j = 1, \dots, n, \\
 & \quad w_i \geq 0 \text{ for } i = 1, \dots, m.
 \end{aligned} \tag{4}$$

The overall value (rating) of alternative S_j can be expressed as

$$d_j = \frac{\sum_{i \in B} w_i^* a_{ij}}{\sum_{i \in C} w_i^* a_{ij}} \tag{5}$$

where w_i^* is the optimal solution of (4).

4 Conclusions

In multiple criteria decision making (MCDA) problems, a decision maker often needs to select or rank alternatives that associated with non-commensurate and conflicting criteria consisting of benefit criteria and cost criteria. DEA may be employed to solve a multiple criteria decision making problem by treating the benefit criteria as outputs and the cost criteria as inputs when the weights of the criteria are unknown. However the main drawback of DEA lies in its low discrimination power. To overcome this chief shortcoming, we have proposed a fuzzy multiple objective data envelopment analysis model for multiple criteria decision making problems. The weights of criteria are determined so that the efficiencies of all DMUs are maximized simultaneously by maximizing the fuzzy degree of all efficiencies.

Acknowledgement

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC94-2416-H-019-006-.

References

1. F.H. Barron, B.E. Barrett, Decision quality using ranked attribute weights, *Management Science* 42 (1996) 1515-1523.
2. A. Charnes, W.W. Cooper, E. Rhodes, Measuring the efficiency of decision-making units, *European Journal of Operational Research* 2 (1978) 429-444.
3. S.-J. Chen, C.-L. Hwang, *Fuzzy Multiple Attribute Decision Making: Methods and Applications*, Springer, New York, 1992.
4. W.D. Cook, M. Kress, A multiple-criteria composite index model for quantitative and qualitative data, *European Journal of Operational Research* 78 (1994) 367-379.
5. B.F. Hobbs, A comparison of weighting methods in power plant citing, *Decision Sciences* 11 (1978) 725-737.

6. C.-L. Hwang, K. Yoon, *Multiple Attribute Decision Making: Methods and Applications*, Springer, Berlin, 1981.
7. H.-S. Lee, A fuzzy method for measuring efficiency under fuzzy environment, *Lecture Notes in Computer Science* 3682 (2005) 343-349.
8. J. Ma, Z.-P. Fan, L.-H. Huang, A subjective and objective integrated approach to determine attribute weights, *European Journal of Operational Research* 112 (1999) 397-404.
9. T.L. Saaty, *Decision Making for Leaders*, 3rd ed. McGraw-Hill, New York, 1995.
10. P.J.H. Schoemaker, C.D. Waid, An experimental comparison of different approaches to determining weights in additive utility models, *Management Science* 28 (1982) 182-196.
11. C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, Urbana: The University of Illinois Press, 1947.
12. M. Zeleny, *Multiple Criteria Decision Making*, McGraw-Hill, New York, 1982.

A Fuzzy Multiple Objective DEA for the Human Development Index

Hsuan-Shih Lee¹, Kuang Lin¹, and Hsin-Hsiung Fang^{1,2}

¹ Department of Shipping and Transportation Management
National Taiwan Ocean University

² Department of Shipping and Business Management
China College of Marine Technology and Commerce

Abstract. Traditionally, the HDI is calculated under the assumption that all component indices are given the same weights. Although this assumption has been supported in the Human Development Reports, it has met also considerable criticism in the literature. In this paper, we present a new model to determine the weights of component indices in the light of data envelopment analysis (DEA). We develop a fuzzy multiple objective DEA model to assess the relative performance of the countries in terms of human development by using optimal weights for the component indices of the HDI.

1 Introduction

Human development is about much more than the rise or fall of national incomes. It is about creating an environment in which people can develop their full potential and lead productive, creative lives in accord with their needs and interests. This way of looking at development, often forgotten in the immediate concern with accumulating commodities and financial wealth, is not new. Philosophers, economists and political leaders have long emphasized human wellbeing as the purpose, the end, of development. As Aristotle said in ancient Greece, “Wealth is evidently not the good we are seeking, for it is merely useful for the sake of something else.” In attempt to consider different aspect of life when measuring human development, the United Nations Development Program introduced in 1990 the Human Development Index (HDI), which is a composite index calculated on the basis of three socioeconomic indicators that reflect three major dimensions of human development: longevity, educational attainment and standard of living.

Since its establishment, the HDI has met considerable criticism [4,9,10]. A comprehensive overview of the literature on alternative computational methods for calculating the HDI can be found in [8]. Despotis [6] relaxed the constraint of HDI that equal weights are given to its component indices by a two-stage DEA-like index-maximizing model to assess the relative performance of the countries in terms of human development.

In this paper, we address the HDI problem with a fuzzy multiple objective DEA model by using optimal common weights for the component indices of the HDI. In Section 2, the results in [6] are reviewed. In Section 3, our fuzzy multiple objective DEA model is proposed and comparison of 27 countries with our model is also presented.

2 A DEA Approach to the HDI

A critical issue in estimating the human development index is the fact that equal weights are assumed for its three component indices. This affects to some extent the relative position of the countries in the HDI ranking. With regards to this issue, Mahlberg and Obersteiner [7] introduced the idea of using the DEA (data envelopment analysis) [2,3] approach to assess the relative performance of the countries in terms of human development, as this notion is defined and on the basis of the data given in the Human Development Report of 1998. In line with the HDI, in which the component indices are all considered to contribute positively in the HDI, the authors suggested an output-oriented DEA model by assuming constant returns to scale. Despotis [6] revisited Mahlberg and Obersteiner’s basic formulation to present a simplified index-maximizing LP (linear programming) model, which Despotis used to estimate an ideal value of the composite index for each one of the countries in the region of Asia and The Pacific. Then Despotis extended the calculations through a goal-programming model to derive a new measure of human development. This new measure is developed under the same assumptions as the original HDI, except that of the equal weights given to the three major component indices: live expectancy at birth (LEI), educational attainment (EDI) and GDP per capita (GDPI). For clarity, we first reiterate the DEA method and Despotis’s results [6].

3 The Basic DEA Model

Given a set of n units, each operating with m inputs and s outputs, let y_{rj} be the amount of the r th output from unit J , and x_{ij} be the amount of the i th input to the J th unit. According to the classical DEA model, the relative efficiency of a particular units J_o is obtained by the optimal value of the objective function in the following fractional linear program (primal, constant returns to scale –input-oriented DEA model):

$$\max h_{j_o}(u, v) = \frac{\sum_{r=1}^s u_r y_{rj_o}}{\sum_{i=1}^m v_i x_{ij_o}} \tag{1}$$

subject to

$$\sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} \leq 0, \quad j = 1, \dots, n$$

$$u_r, v_i \geq \varepsilon \quad \forall r, i$$

The decision variables $u = (u_1, \dots, u_s)$ and $v = (v_1, \dots, v_m)$ are respectively the weights given to the s outputs and to the m inputs. To obtain the relative efficiencies of all the units, the model is solved n times, for one unit at a time. Model (1) allows for great weight flexibility, as the weights are only restricted by the requirement that they should not be zero (the infinitesimal \mathcal{E} ensures that) and they should not make the efficiency of any unit greater than one. Model (1) can be solved as a linear program by letting the denominator in the objective function equal to some constant (1 for example) and then maximizing its numerator as shown in the following model:

$$\begin{aligned}
 \max h_{j_o}(u, v) &= \sum_{r=1}^s u_r y_{rj_o} \\
 \text{subject to} \\
 \sum_{i=1}^m v_i x_{ij_o} &= 1 \\
 \sum_{r=1}^s u_r y_{rj} - \sum_{i=1}^m v_i x_{ij} &\leq 0, \quad j = 1, \dots, n \\
 u_r, v_i &\geq \mathcal{E} \quad \forall r, i
 \end{aligned} \tag{2}$$

Let (u^j, v^j) be the optimal weighting structure for unit j and $h_j^* = h_j(u^j, v^j)$ be its efficiency score. According to their efficiency scores, the units are classified into two groups: the DEA-efficient units (those attaining $h_j^* = 1$), A DEA-efficient unit is actually self-rated efficient by choosing the set of weights that shows it in the best advantage. This sort of efficiency is not always incontestable, as it may be achieved by unbalanced weights, namely by over-weighting some particular inputs and outputs against the others.

3.1 An Index-Maximizing Model

Let C be the set of the countries under evaluation, $j \in C$ stands for any country in C and j_o stands for the evaluated country. Let also w_{LEI} , w_{EDI} and w_{GDPI} be the unknown weights of the three indicators LEI, EDI and GPDI, respectively. Assume there is one dummy input of 1 for all the countries and the three indicators are outputs. Model (2) is equivalent to the following program:

$$\begin{aligned}
 \max h_{j_o} &= w_{LEI} LEI_{j_o} + w_{EDI} EDI_{j_o} + w_{GDPI} GDPI_{j_o} \\
 \text{subject to} \\
 w_{LEI} LEI_j + w_{EDI} EDI_j + w_{GDPI} GDPI_j &\leq 1, \quad j \in C \\
 w_{LEI}, w_{EDI}, w_{GDPI} &\geq \mathcal{E}
 \end{aligned} \tag{3}$$

Let h_j^0 be the optimal value of the objective function when the model (3) is solved for country j . In accordance with the HDI, the values h_j^0 ($j \in C$) are bounded in the interval $[0,1]$. Countries that achieve a score of $h_j^0 = 1$ are in correspondence to the so called “efficient decision making units” in the DEA terminology. Respectively, if the score is $h_j^0 < 1$, the country j might be considered as “inefficient”. However, “efficiency” has no special meaning in this case, as no kind of transformation of inputs to outputs is assumed.

3.2 A Fair Assessment of the HDI Based on Common Weights

To further discriminate the countries that achieve a DEA score of 1, Despotis [5] dealt only with globally efficient countries. These are the countries that maintain their 100% efficiency score under a common weighting structure. Depotis [6] suggested for this purpose the following model with parameter t :

$$\begin{aligned} & \max t \frac{1}{27} \sum_{j=1}^{27} d_j + (1-t)z \\ & \text{subject to} \\ & w_{LEI}LEI_j + w_{EDI}EDI_j + w_{GDPI}GDPI_j + d_j = h_j^0, \quad j \in C \quad (4) \\ & d_j - z \leq 0, \quad j \in C \\ & w_{LEI}, w_{EDI}, w_{GDPI} \geq \varepsilon \\ & z \geq 0, d_j \geq 0, \quad j \in C. \end{aligned}$$

The first term of the objective function, when considered solely (for $t = 1$) represents the mean deviation (the L_1 norm) between the DEA scores and the adjusted global efficiency scores for all the countries. The second term, when considered solely (for $t = 0$) represents, through the non-negative variable z , the maximal deviation (the L_1 norm) between the above efficiency scores and the model is reduced to a minmax goal-programming model. Varying the parameter t between these two extreme values, Despotis provided the model with the flexibility to “compromise” between the two norms and to explore different sets of common weights, beyond the extreme ones that minimized the maximal and the mean deviation, respectively.

4 An Assessment of the HDI Based on Fuzzy DEA

A multiple objective model trying to maximize the efficiency scores of all the countries is as follows:

$$\max h_j \quad j \in C$$

subject to

$$w_{LEI}LEI_j + w_{EDI}EDI_j + w_{GDPI}GDPI_j = h_j, \quad j \in C \tag{5}$$

$$h_j \leq 1, \quad j \in C$$

$$w_{LEI}, w_{EDI}, w_{GDPI} \geq \varepsilon$$

The best efficiency score a country can achieve is 1 and the lowest efficiency score of a country is 0. The imprecise goal for each of the objective functions can be modeled by the following linear membership function:

$$\mu_j = \frac{h_j - 0}{1 - 0}$$

where 0 is the worst acceptable level for h_j and 1 is the totally desirable level for h_j . After determining the membership functions for each of the objective functions, if we adopt the maximizing decision proposed by Bellman and Zadeh [1] the resulting problem to be solved is:

$$\max \min_i \mu_j$$

subject to

$$w_{LEI}LEI_j + w_{EDI}EDI_j + w_{GDPI}GDPI_j \leq 1, \quad j \in C$$

$$w_{LEI}LEI_j + w_{EDI}EDI_j + w_{GDPI}GDPI_j = h_j, \quad j \in C \tag{6}$$

$$\mu_j = \frac{h_j - 0}{1 - 0}, \quad j \in C$$

$$w_{LEI}, w_{EDI}, w_{GDPI} \geq \varepsilon$$

Namely, the problem is to maximize the minimum membership function value. As is well known, this problem is equivalent to solving the following problem:

$$\max \alpha$$

subject to

$$w_{LEI}LEI_j + w_{EDI}EDI_j + w_{GDPI}GDPI_j \leq 1, \quad j \in C$$

$$w_{LEI}LEI_j + w_{EDI}EDI_j + w_{GDPI}GDPI_j = h_j, \quad j \in C \tag{7}$$

$$\mu_j = \frac{h_j - 0}{1 - 0}, \quad j \in C$$

$$\mu_j \geq \alpha, \quad j \in C$$

$$w_{LEI}, w_{EDI}, w_{GDPI} \geq \varepsilon$$

Table 1. Data and optimization results for the countries of Asia and The Pacific

HDI Reg. Rank	HDI	Country	Life exp. at birth	LEI	Adult literacy rate	Comb. Gross enrol. Ratio	EDI	GDP	GDPI	DEA score	FMO DEA
27	0,461029	Bangladesh	58,6	0,56	40,1	36	0,387333	1361	0,435754	0,627	0,560323
25	0,483092	Bhutan	61,2	0,603333	42	33	0,39	1536	0,455943	0,675	0,590322
4	0,848176	Brunei Darussalam	75,7	0,845	90,7	72	0,844667	16765	0,854863	0,976	0,973619
23	0,511384	Cambodia	53,5	0,475	65	61	0,636667	1257	0,422487	0,67	0,624831
13	0,705694	China	70,1	0,751667	82,8	72	0,792	3105	0,573416	0,89	0,883472
6	0,769353	Fiji	72,9	0,798333	92,2	81	0,884667	4231	0,62506	0,968	0,959537
2	0,872187	Hong Kong China (SAR)	78,6	0,893333	92,9	64	0,832667	20763	0,89056	1	1
20	0,563102	India	62,9	0,631667	55,7	54	0,551333	2077	0,506305	0,707	0,687515
15	0,670566	Indonesia	65,6	0,676667	85,7	65	0,788	2651	0,547032	0,841	0,831952
12	0,708641	Islamic Republic of Iran	69,5	0,741667	74,6	69	0,727333	5121	0,656924	0,849	0,846665
3	0,853924	Republic of Korea	72,6	0,793333	97,5	90	0,95	13478	0,818438	1	0,99015
24	0,483949	Lao People's Democratic Republic	53,7	0,478333	46,1	57	0,497333	1734	0,47618	0,563	0,560323
5	0,771182	Malaysia	72,2	0,786667	86,4	65	0,792667	8137	0,734212	0,912	0,908717
10	0,725261	Maldives	65	0,666667	96	75	0,89	4083	0,619117	0,937	0,875546
16	0,628828	Mongolia	66,2	0,686667	83	57	0,743333	1541	0,456485	0,824	0,815846
19	0,58509	Myanmar	60,6	0,593333	84,1	56	0,747333	1199	0,414602	0,787	0,755962
26	0,473328	Nepal	57,8	0,546667	39,2	61	0,464667	1157	0,408651	0,612	0,588615
22	0,522558	Pakistan	64,4	0,656667	44	43	0,436667	1715	0,474341	0,735	0,648134
21	0,542407	Papua New Guinea	58,3	0,555	63,2	37	0,544667	2359	0,527554	0,636	0,634182
8	0,743779	Philippines	68,6	0,726667	94,8	83	0,908667	3555	0,596005	0,956	0,923707
11	0,71162	Samoa (Western)	71,7	0,778333	79,7	65	0,748	3832	0,608528	0,883	0,880146
1	0,881065	Singapore	77,3	0,871667	91,8	73	0,855333	24210	0,916195	1	0,997029
18	0,614417	Solomon Islands	71,9	0,781667	62	46	0,566667	1940	0,494916	0,875	0,793331
9	0,732945	Sri Lanka	73,3	0,805	91,1	66	0,827333	2979	0,566501	0,942	0,935504
7	0,745278	Thailand	68,9	0,731667	95	61	0,836667	5456	0,66675	0,901	0,893051
17	0,623073	Tuvalu	67,7	0,711667	64	47	0,583333	3120	0,57422	0,797	0,756352
14	0,671486	Viet Nam	67,8	0,713333	92,9	63	0,829333	1689	0,471791	0,886	0,875114

Model (7) can be further simplified as:

$$\max \alpha$$

subject to

$$w_{LEI} LEI_j + w_{EDI} EDI_j + w_{GDPI} GDPI_j \leq 1, \quad j \in C \tag{8}$$

$$w_{LEI} LEI_j + w_{EDI} EDI_j + w_{GDPI} GDPI_j \geq \alpha, \quad j \in C$$

$$w_{LEI}, w_{EDI}, w_{GDPI} \geq \epsilon$$

The data of the 27 countries of the regional aggregate of Asia and The Pacific adopted in [6] is fed into model (8). The result is shown in the last column of Table 1. As shown in Table 1, three out the 27 countries achieve the highest score of 1 by traditional DEA. However, these three countries can be further differentiated by our method.

5 Conclusion

In light of data envelopment analysis, the human development index is revisited. A new measure of human development is proposed with a model based on fuzzy multiple objective DEA. The superiority of the new measure is based on the fact the weights assumed for the component indicators are less arbitrary and contestable. All countries are compared with the same criteria in a DEA context. In stead of evaluating countries by assuming all component indices are given the same weights, we assess the relative performance of the countries by using optimal common weights for the component indices of the HDI.

Acknowledgement

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC94-2416-H-019-006-.

References

1. R.E. Bellman, L.A. Zadeh, Decision making in a fuzzy environment, *Management Science* 17 (1970) 141-164.
2. A. Charnes, W.W. Cooper, E. Rhodes, Measuring the efficiency of decision making units, *European Journal of Operational Research* 2 (1978) 429-444.
3. W.W. Cooper, L.M. Seiford, K. Tone, *Data Envelopment Analysis*, Dordrecht: Kluwer Academic Publisher (1999).
4. M. Desai, Human development: concepts and measurement, *European Economical Review* 35 (1991) 350-357.
5. D.K. Despotis, Improving the discriminating power of DEA: focus on globally efficient units, *Journal of the Operational Research Society* 53(3) (2002) 314-323.
6. D.K. Despotis, Measuring human development via data envelopment analysis: the case of Asia and the Pacific, *Omega* 33 (2005) 385-390.
7. B. Mahlberg, M. Obersteiner, Remeasuring the HDI by data envelopment analysis, IIASA interim report IR-01-069, Luxemburg (2001).
8. E. Neumayer, The human development index—a constructive proposal, *Ecological Economics* 39 (2001) 101-114.
9. F. Noorbakhsh, A modified human development index, *World Development* 26 (1998) 517-528.
10. A.D. Sagar, A. Najam, The human development index: a critical review, *Ecological Economics* 25 (1998) 249-264.

Visualization Architecture Based on SOM for Two-Class Sequential Data

Ken-ichi Fukui¹, Kazumi Saito², Masahiro Kimura³, and Masayuki Numao¹

¹ The Institute of Scientific and Industrial Research, Osaka University, Japan
fukui@ai.sanken.osaka-u.ac.jp

<http://www.ai.sanken.osaka-u.ac.jp>

² NTT Communication Science Laboratories, Japan

³ Department of Electronics and Informatics, Ryukoku University, Japan

Abstract. In this paper, we propose a visualization architecture that constructs a map suggesting clusters in sequence that involve classification utilizing the class label information for the display method of the map. This architecture is based on Self-Organizing Maps (SOM) that are to create clusters and to arrange the similar clusters near within the low dimensional map. This proposed method consists of three steps, firstly the winner neuron trajectories are obtained by SOM, secondly, connectivity weights are obtained by a single layer perceptron based on the winner neuron trajectories, finally, the map is visualized by reversing the obtained weights into the map. In the experiments using time series of real-world medical data, we evaluate the visualization and classification performance by comparing the display method by the number of sample ratio for classes belonging to each cluster.

1 Introduction

Kohonen's Self-Organizing Maps (SOM)[1] is an unsupervised neural-network that generates a map reflecting similarity among input data. SOM generalizes similar samples as a reference vector corresponding to a neuron. Where, the nearest reference vector to a sample is called a "Winner Neuron" or a "Best Matching Unit (BMU)". Samples are clustered into winner neurons. Moreover, since neurons' topology is predefined, reference vectors are updated so that similar reference vectors are arranged close to each other in a low dimensional topology space usually two or three dimensions. This feature is different from classical techniques that put high dimensional data into low dimensional space without clustering such as the Multi-dimensional Scaling (MDS)[2] or Sammon's Mapping[3]. Therefore, SOM is a suitable technique for data mining and visualizing large scale data.

Since conventional SOM is unsupervised learning, it would not be meaningful for it to use class label information. Consequently, supervised learning models for SOM have been proposed (e.g.[1,4,5,6]). These models try to improve classification performance by taking into account class label information when updating the reference vectors. Namely, supervised SOM try to create clusters that classify labels correctly. On the contrary, unsupervised SOM try to create clusters

based on the similarity of input data. Classification performance will improve by using supervised SOM, however, the nature of unsupervised SOM that mines similar samples will fade.

In this paper, we adopt conventional unsupervised SOM to acquire the feature of the data as clusters, and then utilize the class label to create the meaning of the clusters. At this time, the typical display methodology of SOM is to represent each cluster in varying gray levels on the cluster samples' density[7]. Even if class labels are given, it is not enough to make use of class label information since majority decision is used to determine the representative class of the cluster[6,8].

We are proposing a visualization architecture for a two-class problem. This architecture consists of two learning stages, namely, learning the winner neuron using SOM and learning the connection weights using a single layer perceptron. When the data is given in a series such as time series, winner neurons' trajectory is obtained by SOM. Afterwards, utilizing this feature and the class label, construct a discriminant function of a single layer perceptron where a winner neurons' trajectory as an instance. Finally, each cluster is created meaning as a node value by means of reversing connection weights into the map nodes. It is expected that this visualization scheme can suggest clusters related to classification among latent clusters of specific period in the series data. Note that our proposed architecture requires sequential and two-class data.

To validate the proposed methodology, we conducted experiments using time series in medical data, specifically blood test data of hepatitis patients. We compared visualization and classification performance to a simple display methodology which is a number of sample ratio for classes belonging to a cluster.

2 Visualization Architecture

2.1 SOM Learning Model

SOM is an unsupervised neural-network that generates a map reflecting similarity among input data. Let V dimension of N input data be $\mathbf{x}_n = (x_{n,1}, \dots, x_{n,V})$, ($n = 1, \dots, N$). The map is normally constructed by neuron nodes arranged into a two dimensional grid at regular intervals. Reference vectors are updated as follows:

$$\mathbf{m}_j^{new} = \mathbf{m}_j + h_{c(\mathbf{x}),j}[\mathbf{x} - \mathbf{m}_j]. \quad (1)$$

$$c(\mathbf{x}) = \arg \min_i \|\mathbf{x} - \mathbf{m}_i\|. \quad (2)$$

where the coefficient $h_{c(\mathbf{x}),j}$ represents a neighborhood function and its first index $c(\mathbf{x})$ indicates the winner neuron obtained using equation(2). As a neighborhood function, a Gaussian distribution-based definition is used:

$$h_{c(\mathbf{x}),j} = \alpha \exp\left(-\frac{\|\mathbf{r}_j - \mathbf{r}_{c(\mathbf{x})}\|^2}{2\sigma^2}\right). \quad (3)$$

where \mathbf{r}_j represents a coordinate of the j^{th} neuron arranged within the grid, and α and σ are parameters that control the learning. Monotonically decreasing strategy is taken for the learning parameters.

The objective function of the SOM learning model[9] can also be described as:

$$E_{SOM} = \sum_{i=1}^M \sum_{\mathbf{x}_n \in C_i} \sum_{j=1}^M h_{i,j} \|\mathbf{x}_n - \mathbf{m}_j\|^2. \tag{4}$$

where M is the number of neuron nodes, C_i is a set of samples that satisfy $c(\mathbf{x}) = i$.

2.2 SBSOM

Conventional SOM does not provide any absolute meaning to the axes within the map. In order to increase the instinctive interpretability of the map, we adopted a modification called Sequence-Based SOM (SBSOM), so that the sequential data arranged into a specific axis while preserving the property of SOM[8]. Since the proposed method is for sequential data, SBSOM can be applied. Note, however, that it is not necessary to employ SBSOM to the proposed method.

In SBSOM, input and reference vectors are extended as (\mathbf{x}_n, t_n) and (\mathbf{m}_j, s_j) , respectively. t_n and s_j are indexes of a sequence such as time-tag assuming the same discretization. The distance definition is modified as follows:

$$c(\mathbf{x}) = \arg \min_j \delta(t_n, s_j) \|\mathbf{x}_n - \mathbf{m}_j\|. \tag{5}$$

$$\delta(t_n, s_j) = \begin{cases} 1 & \text{if } t_n = s_j, \\ \infty & \text{otherwise.} \end{cases} \tag{6}$$

For instance, if the reference vector indexes s_j those neuron nodes are in the same row are set to the same value and in order of column direction, the map will have the meaning of sequential order in its horizontal axis. However, owing to the neighborhood function, similar samples are arranged into the column direction as well effected by forward and backward of the sequence. Therefore, SBSOM adds the sequential meaning to the axis of the map while making use of the property of SOM.

2.3 Visualization Architecture

The proposed visualization architecture is illustrated in Fig.1. Let k be an index of K sequential data, and n_k be the number of elements in the k^{th} sequential data. The k^{th} sequential data is represented by $\{(\mathbf{x}_r^{(k)}, t_r^{(k)}) : r = 1, \dots, n_k\}$, where $\sum_{k=1}^K n_k = N$. The visualization architecture consists of three steps:

1. Train the reference vectors using SOM or SBSOM as $\{(\mathbf{x}_r^{(k)}, t_r^{(k)}) : r = 1, \dots, n_k, k = 1, \dots, K\}$ are input data. For instance, the first sequence of the first data $(\mathbf{x}_1^{(1)}, t_1^{(1)})$ is an input. As a consequence, the winner neuron trajectories are obtained within the visualization layer.
2. Construct a discriminant function by a single layer perceptron utilizing class labels as the winner neuron trajectories are input. To be more concrete, let

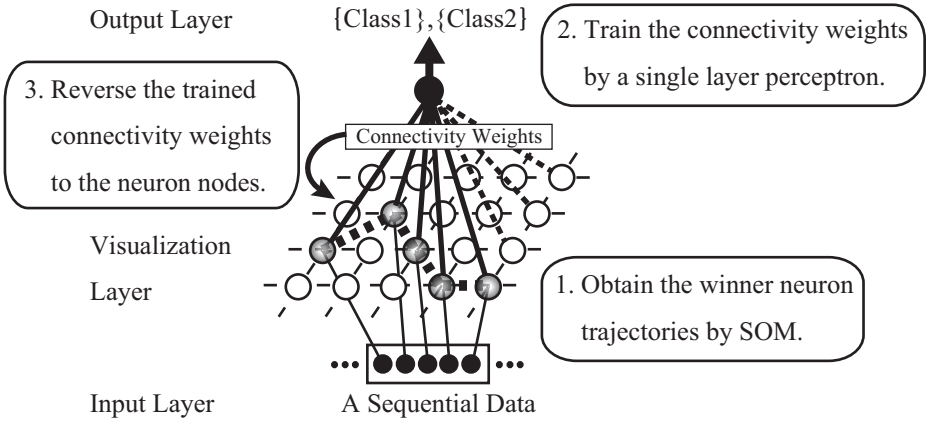


Fig. 1. Visualization Architecture. This illustrates only one example of a sequential data.

an input vector be $\mathbf{y}_k = (y_{k,1}, \dots, y_{k,M})$. Set $y_{k,m}$ corresponding to the m^{th} neuron node ($m = 1, \dots, M$) for the k^{th} sequence as follows:

$$y_{k,m} = \begin{cases} 1/n_k & \text{if } \exists r \mathbf{x}_r^{(k)} \in C_m, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Let the connectivity weights be $\mathbf{w} = (w_1, \dots, w_M)$. The discriminant function is represented by $\hat{z}_k = \mathbf{w} \cdot \mathbf{y}_k$. Let $z_k = \{-1, 1\}$, ($k = 1, \dots, K$) be class labels. Obtain the optimal weights by minimizing the following objective function:

$$E_{pct} = \sum_{k=1}^K (z_k - \hat{z}_k)^2. \quad (8)$$

When equation (8) is minimized by steepest decent method, connectivity weights \mathbf{w} are updated by the following equation:

$$w_m^* = w_m + \beta \sum_{k=1}^K \{(z_k - \hat{z}_k)y_{k,m}\}, \quad (9)$$

where β is a learning rate.

3. Finally, the map is visualized by reversing the obtained weights into the nodes of the visualization layer and setting the weights as contracting density.

The benefits of this methodology are to derive the interpretation of the discriminant function for the map via the perceptron, and to suggest the cluster in a sequence that involves classification through SOM. Note that the assumption is that the higher the weight the more important it is for classification.

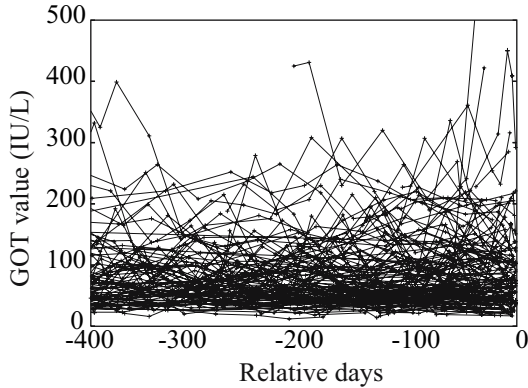


Fig. 2. Actual GOT value. The horizontal axis indicates the relative days from IFN administration, where the right edge is the administration day.

3 Experiment

In order to evaluate the proposed methodology, we conducted experiments using the hepatitis patients' data set by comparing visualization and classification performance to a simple display method.

3.1 Data Set

We used time series of blood test data of 137 hepatitis C patients for one year as collected from one of the Japanese hospitals¹. We used 11 common inspection items indicating hepatic functions, namely, GOT, GPT, TP, ALB, T-BIL, D-BIL, I-BIL, TTT, ZTT, CHE, and T-CHO. The interval of their inspections is about one month. The medical data changes in a large way as shown in Fig.2 for example and is a multi-dimensional (multi-inspections) data.

3.2 Problem Setting and Objective

In recent years, interferon (IFN) is used as medicine to cure hepatitis. However, IFN is expensive and has strong side-effects. Moreover, recovery rate of IFN on patients is only about 30% in average. It is crucial to predict the effectiveness of IFN based on the inspection results before it is administered. This relieves the patient of any physical, mental, as well as cost burdens. The data is classified a priori into two classes, namely, 55 positive examples and 82 negative examples in terms of existence of the hepatitis virus through a HCV-RNA inspection² before and after IFN administration.

¹ The data was provided by the hospital affiliated to the medical faculty of Chiba University, Japan.

² HCV-RNA inspection is known to be highly reliable which verifies existence of the hepatitis C virus by PCR (Polymerase Chain Reaction) amplification.

The objective of this experiment is to construct a map that suggests periods in which patients' groups giving the same inspection trends whom IFN effect or not. The result validates the proposed methodology by comparing a simple display method by the ratio of positive and negative examples in each cluster.

3.3 Preprocessing

In order to avoid bias among attributes (inspections), the attribute values are standardized based on the index presented by a doctor. All the attribute values are standardized in the same continuous value from 1 to 6, the actual value that is in the range of normal is set as 3, and cut off extremely high and low values as well. At the same time, since the inspection intervals are different in each patient, the discretized points are set to the same by linear interpolation with one week interval. Since the inspection intervals are almost one month, this interpolation interval is sufficient.

3.4 Visualization Results

In this section, two display methods are compared using the same results obtained by SBSOM (52×15 nodes), namely, the same winner neuron trajectories. The one shown in Fig.3 is displayed by the ratio that is (positive examples)/(positive and negative examples) belonging to the cluster represented as a neuron node. Note that a confidence weight is multiplied depending on the number of cluster elements. The one shown in Fig.4 is displayed by connectivity weights obtained by the perceptron. In both maps, the horizontal axis indicates the relative days from IFN administration, where the right edge is the administration day. The vertical axis does not have an absolute meaning, however, it has relative meaning preserving similarity between clusters which is the property of

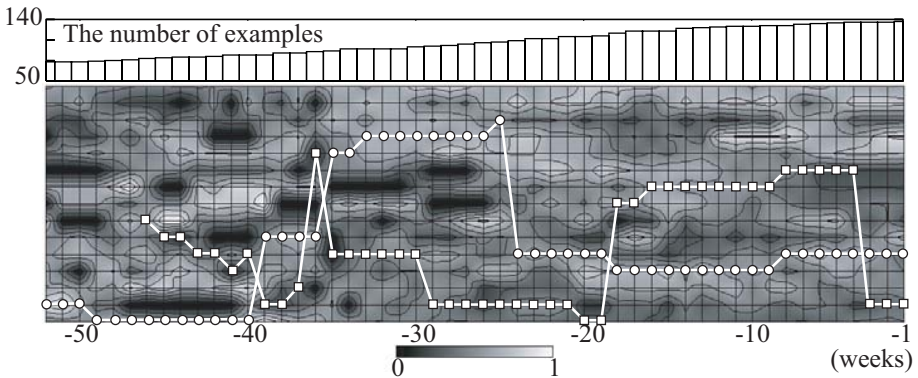


Fig. 3. Displayed by ratio of the positive and negative examples for each cluster. The winner neuron trajectories of two patients are illustrated together, where circle is positive patient and square is negative one for example. The histogram on top indicates the number of examples at the time.

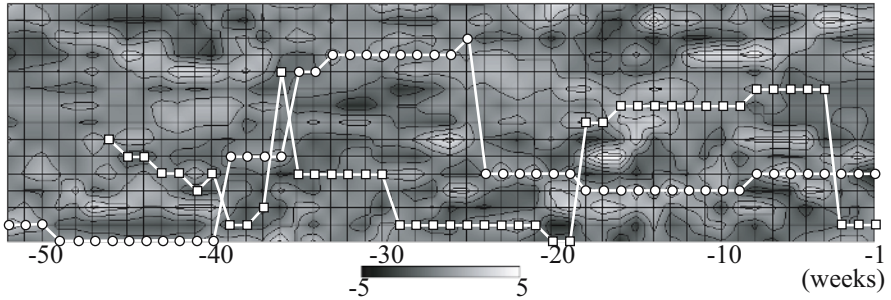


Fig. 4. Displayed by connectivity weights. The same winner neuron trajectories trained by SBSOM is used.

SOM. This map is available for visual data mining in the sense that if the winner neuron trajectory passes through a high or low node, that means that the cluster has relationship to IFN effectiveness. For example, an expert can refer the other patients’ case in the same cluster.

Comparing Fig.3 and Fig.4, since there are clusters whose elements may be only one or two before 30 weeks as inspection data is not much. Their node values tend to zero or one by the ratio display method. Therefore, there is a bias to the minority elements’ clusters. In contrast, there is no such bias by connectivity weights display method. The node values appear well balanced as they spread across the whole period.

3.5 Classification Accuracy

We validated that the node values indicate correct classification, in other words, the higher the value, the more effective the IFN is, the lower the value the less effective. We evaluated classification performance based on the winner neuron trajectory. In the case of ratio display method, if the average of the nodes that winner neuron trajectory passes through is more than 0.5 then the prediction is positive, otherwise prediction becomes negative. In the case of connectivity weights display method, the discriminant function is used as a criterion.

The results of a 10-fold cross validation are listed in Table 1. The average of the maximum and the minimum percentage of correct answers by the ratio display method is 47.53% for the test set. The result is equivalent to having a random scheme, hence, classification cannot generalize such that it cannot

Table 1. The result of 10-fold cross validation

DISPLAY METHOD	Ratio	Connectivity Weights
CRITERION	Average	Discriminant Function
Training Set	79.78±4.78	99.60±0.40
Test Set	47.53±16.76	64.29±21.43

predict the effectiveness given unknown patients. On the other hand, our method gives almost perfect prediction for known data and 64.24% for unknown ones, albeit with a large variance.

4 Conclusion

In this paper, we proposed the visualization architecture utilizing a single layer perceptron as a method to display the results obtained from using SOM. The advantage of the method is that it constructs the map suggesting the cluster in a sequence that involves classification by utilizing the class label in the display method. The initial experiment using real world medical data validated the following two points by comparing the simple display method by the number of samples ratio for classes.

- There is no unnecessary bias to periods that have less data, there may be clusters which has few element.
- It was confirmed that generalized classification performance was acquired, though it can not be said that the accuracy is sufficient.

So as to improve generalization performance, re-investigation of preprocessing, consideration of time warping or gradient of the attributes, and constraint of over fitting can be considered. Furthermore, the comments from a medical doctor are needed to verify the validity of visualization.

References

1. Kohonen, T.: Self-Organizing Maps. Springer-Verlag, Heidelberg (1995)
2. Kruskal, J.B., Wish, M.: Multidimensional scaling. Number 07-011 in Paper Series on Quantitative Applications in the Social Sciences (1978)
3. Sammon, J.: A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* **c-18** (1969) 401–09
4. Hecht-Nielsen, R.: Counterpropagation networks. In: *IEEE First international Conference on Neural Networks*. (1987) 19–32
5. Fritzke, B.: Growing cell structures: A selforganizing networks for unsupervised and supervised learning. *Neural Networks* **7** (1994) 1441–1460
6. Fukuda, N., Saito, K., Matsuo, S., Ishikawa, M.: Som learning model using teacher information. Technical Report 732, IEICE (2004)
7. Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Honkela, J., Paatero, V., Saarela, A.: Self organization of a massive document collection. *IEEE Transaction on Neural Networks* **11(3)** (2000) 574–585
8. Fukui, K., Saito, K., Kimura, M., Numao, M.: Visualizing dynamics of the hot topics using sequence-based self-organizing maps. *Lecture Notes in Computer Science* **3684** (2005) 745–751
9. Kohonen, T.: Comparison of som point densities based on different criteria. *Neural Computation* **11** (1999) 2081–2095

Approximate Solutions for the Influence Maximization Problem in a Social Network

Masahiro Kimura¹ and Kazumi Saito²

¹ Department of Electronics and Informatics, Ryukoku University
Otsu, Shiga 520-2194, Japan

² NTT Communication Science Laboratories, NTT Corporation
Seika-cho, Kyoto 619-0237, Japan

Abstract. We address the problem of maximizing the spread of information in a large-scale social network based on the *Independent Cascade Model (ICM)*. When we solve the *influence maximization problem*, that is, the optimization problem of selecting the most influential nodes, we need to compute the expected number of nodes influenced by a given set of nodes. However, an exact calculation or a good estimate of this quantity needs a large amount of computation. Thus, very large computational quantities are needed to approximately solve the influence maximization problem based on a natural greedy algorithm. In this paper, we propose methods to efficiently obtain good approximate solutions for the influence maximization problem in the case where the propagation probabilities through links are small. Using real data on a large-scale blog network, we experimentally demonstrate the effectiveness of the proposed methods.

1 Introduction

A social network is the network of relationships and interactions among social entities such as individuals, organizations and groups. Examples include email networks, hyperlink networks of web sites, trackback networks of blogs, and scientific collaboration networks. With the recent availability of large data sets of real social networks, there has been growing interest in social network analysis [10,5,4,1,6,8,11].

Information, ideas, and influence can propagate through a social network in the form of “word-of-mouth” communications. Thus, it is important to investigate the problem of maximizing the spread of information in the underlying network in terms of sociology and marketing. In particular, Domingos and Richardson [2], Richardson and Domingos [12], and Kempe *et al.* [5,6] studied the *influence maximization problem*, that is, the problem of choosing a set of k nodes to target for initial activation such that it yields the largest expected spread of information, where k is a given integer. Although models for diffusion processes on a network have been studied in various fields including epidemiology, sociology, marketing and physics [5,4], one of the conceptually simplest models is the *Independent Cascade Model (ICM)* used by Goldenberg [3], Kempe *et al.*

[5,6], and Gruhl *et al.* [4]. In this paper, we consider the influence maximization problem in a social network based on the ICM.

The ICM is a stochastic process model in which information propagates from a node to its neighboring nodes at each time-step according to a probabilistic rule. Therefore, we need to compute the expected number $\sigma(A)$ of nodes influenced by a given set A of nodes to solve the influence maximization problem. Here, we call $\sigma(A)$ the *influence* of target set A . It is an open question to compute the influence $\sigma(A)$ exactly by an efficient method, and so good estimates were obtained by simulating the random process 10,000 times [5]. However, such computations become very heavy for a large-scale social network.

Kempe *et al.* [5] proposed a natural greedy algorithm to obtain a good approximate solution for the influence maximization problem, and presented its theoretical approximation guarantee. However, very large computational quantities were needed to approximately solve the influence maximization problem in a large-scale social network based on the greedy algorithm. In this paper, we propose methods to efficiently obtain good approximate solutions for the influence maximization problem for the ICM in the case where the propagation probabilities through links are small. Using real data from a large-scale blog network, we experimentally demonstrate the effectiveness of the proposed methods.

2 Influence Maximization Problem

Based on the work of Kempe *et al.* [5], we recall the definition of the ICM, and briefly explain a natural greedy algorithm to approximately solve the influence maximization problem in this model.

2.1 Independent Cascade Model

We consider the ICM for the spread of a certain information through a social network represented by a directed graph. First, we call nodes *active* if they have accepted the information. We assume that nodes can switch from being inactive to being active, but cannot switch from being active to being inactive. When node u first becomes active at step t , it is given a single chance to activate each currently inactive *child* v , and succeeds with probability $p_{u,v}$, where $p_{u,v}$ is a constant that is independent of the history of the process, and node v is called a *child* of node u and node u is called a *parent* of node v if there is a directed link from u to v . If u succeeds, then v will become active at step $t + 1$. If multiple parents of v first become active at step t , then their activation attempts are sequenced in an arbitrary order, but done at step t . Whether or not u succeeds, it cannot make any further attempts to activate v in subsequent rounds. The process terminates if no more activations are possible.

2.2 Greedy Algorithm

We consider the influence maximization problem in the ICM. Namely, for a given positive integer k , we consider finding a set A_k^* of k nodes such that $\sigma(A_k^*) \geq$

$\sigma(B)$ for any set B of k nodes in the ICM. For this problem, Kempe *et al.* [5] proposed the following natural greedy algorithm:

1. Start with $B = \emptyset$.
2. **for** $i = 1$ to k **do**
3. Choose a node v_i maximizing $\sigma(B \cup \{v_i\}) - \sigma(B)$.
4. Set $B \leftarrow B \cup \{v_i\}$.
5. **end for**

Let B_k denote a set of k nodes obtained by this algorithm. Then, Kempe *et al.* [5] proved that $\sigma(B_k) \geq (1 - 1/e) \sigma(A_k^*)$, that is, they presented an approximation guarantee for this algorithm. Their proof relies on the theory of *submodular functions* [9]. Here, for a function f that maps a subset of a finite ground set U to a nonnegative real number, f is called *submodular* if $f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T)$ for any $u \in U$ and any pair (S, T) of subsets of U with $S \subset T$. They proved the result of the approximation guarantee by showing that the function σ is submodular for the ICM.

However, for a naive implementation of this greedy algorithm, we need to compute the influence $\sigma(A)$ for each target set A . Since it is not clear how to evaluate $\sigma(A)$ exactly by an efficient method, Kempe *et al.* [5] obtained a good estimate by simulating the random process 10,000 times for each target set. They argued that the quality of approximation after 10,000 iterations is comparable to that after 300,000 or more iterations.

3 Proposed Methods

Now, we propose two methods to efficiently obtain good approximate solutions for the influence maximization problem in the ICM when the propagation probabilities through links are small. The methods use the information diffusion models in a social network proposed by Kimura and Saito [7], and compute approximate solutions in these models based on the greedy algorithm defined in Section 2.2.

Below, we define these information diffusion models, and see that the influence $\sigma(A)$ of target set A can be exactly and efficiently computed for these models. Therefore, the approximate solutions for the influence maximization problem can be efficiently computed under the proposed methods. Moreover, we see that an approximation guarantee for the greedy algorithm can be obtained in these models.

3.1 Information Diffusion Models

We define two natural special cases of the ICM. Let A be an initial set of active nodes in the network, that is, the set of nodes that first become active at step 0. For nodes u and v in the network, let $d(u, v)$ denote the graph distance from u to v , and let $d(A, v)$ denote the graph distance from A to v , that is, $d(A, v) = \min_{u \in A} d(u, v)$. When there is not a path from u to v , we set $d(u, v) = \infty$. Note that the value $d(A, v)$ can be efficiently computed by graph theory [10].

First, we define the *Shortest-Path Model (SPM)*. The SPM is a special case of the ICM such that each node v has the chance to become active only at step $t = d(A, v)$. In other words, each node is activated only through the shortest paths from an initial active set. Namely, the SPM is a special type of the ICM where only the most efficient information spread can occur.

Next, we slightly generalize the SPM within the ICM, and define the *SP1 Model (SP1M)*. In the SP1M, each node v has the chance to become active only at steps $t = d(A, v)$ and $t = d(A, v) + 1$. In other words, v cannot be activated excluding the paths from A to v whose length are equal to $d(A, v)$ or $d(A, v) + 1$.

Since the SPM and SP1M are such approximations to the ICM that disregard the information propagation via long-length paths from an initial active set, the proposed methods can be expected to provide good approximate solutions for the influence maximization problem in the ICM when the propagation probabilities through links are small.

3.2 Exact Computation of Influence

We consider computing efficiently the exact value of $\sigma(A)$ of target set A for the SPM and SP1M. Let V be the set of all nodes in the network, N the number of elements of V , and V_A the set of nodes v such that $d(A, v) < \infty$. For any $v \in V$, let $P_t(v; A)$ denote the probability that v first becomes active at step t , and let $PA(v)$ denote the set of all parent nodes of v . Here, note that $P_t(v; A) = 0$ for any $t \geq 0$ if $v \notin V_A$.

We begin with the SPM. We consider computing $\sigma(A)$ from a computation of $P_t(v; A)$ for any $t \geq 0$ and $v \in V$. Note first that for any $v \in V_A$, $P_t(v; A) = 0$ if $t \neq d(A, v)$. Thus, we can focus on $t = d(A, v) (< \infty)$. Then, it is easily shown that $P_t(v; A)$ is computed by

$$P_t(v, A) = \sum_{W \subset PA(v)} P_{t-1}(W|PA(v); A) P_t(W \rightarrow v), \tag{1}$$

where the summation is taken over all subsets of $PA(v)$, $P_{t-1}(W|PA(v); A)$ denotes the probability that subset W first becomes active at step $t - 1$ in $PA(v)$, and $P_t(W \rightarrow v)$ denotes the probability that v is infected from W at step t . From Equation 1, we can prove the following theorem.

Theorem 1. (Kimura and Saito [7]) *With the SPM, the exact value of $\sigma(A)$ is computed in the following way.*

$$\sigma(A) = \sum_{v \in V_A} P_{d(A,v)}(v; A),$$

where for each $v \in V_A$,

$$P_t(v; A) = 1 - \prod_{u \in PA(v)} (1 - p_{u,v} P_{t-1}(u; A)),$$

if $t = d(A, v)$, and $P_t(v; A) = 0$ if $t \neq d(A, v)$.

Theorem 1 implies that the exact value of $\sigma(A)$ can be efficiently computed in the SPM.

Next, we consider the SP1M. In this case, for any $v \in V_A$, $P_t(v; A) = 0$ if $t \neq d(A, v)$, $d(A, v) + 1$. Thus, we focus on $t = d(A, v)$ and $t = d(A, v) + 1$ for $v \in V_A$. Then, it is easily shown that $P_t(v; A)$ is computed by

$$P_t(v, A) = (1 - P_{t-1}(v; A)) \sum_{W \subset PA(v)} P_{t-1}(W|PA(v); A) P_t(W \rightarrow v). \quad (2)$$

From Equation 2, we can prove the following theorem.

Theorem 2. (Kimura and Saito [7]) *With the SP1M, the exact value of $\sigma(A)$ is computed in the following way.*

$$\sigma(A) = \sum_{v \in V_A} \left(P_{d(A,v)}(v; A) + P_{d(A,v)+1}(v; A) \right),$$

where for each $v \in V_A$,

$$P_t(v; A) = (1 - P_{t-1}(v; A)) \left\{ 1 - \prod_{u \in PA(v)} (1 - p_{u,v} P_{t-1}(u; A)) \right\},$$

if $t = d(A, v)$, $d(A, v) + 1$, and $P_t(v; A) = 0$ if $t \neq d(A, v)$, $d(A, v) + 1$.

Theorem 2 implies that the exact value of $\sigma(A)$ can be efficiently computed in the SP1M.

3.3 Approximation Guarantees

For the SPM and SP1M, we consider the influence maximization problem, and investigate an approximation guarantee for the greedy algorithm defined in Section 2.2. We fix an integer k ($1 \leq k < N$). Let A_k^* be a set that maximizes the value of σ over all k -element subsets of V , and let B_k be a k -element set obtained by the greedy algorithm. Then, we can prove the following theorem by using the theory of submodular functions.

Theorem 3. (Kimura and Saito [7]) *In the SPM and SP1M, we have the following approximation guarantee for the greedy algorithm:*

$$\sigma(B_k) \geq \left(1 - \frac{1}{e} \right) \sigma(A_k^*).$$

4 Experimental Evaluation

Using real data on a large-scale social network, we experimentally investigate the effectiveness of the proposed methods.

4.1 Blog Network Data

We describe the details of the data set used in our experiments.

Blogs are personal on-line diaries managed by easy-to-use software packages, and they have spread rapidly through the World Wide Web [4]. Compared with ordinary web sites, one of the most prominent features of blogs is the existence of *trackbacks*. Unlike a hyperlink, a blogger can construct a link from another blog b to his own blog by putting a trackback on b . Bloggers discuss various topics and do mutual communications by putting trackbacks on their blogs each other. Here, we regard a link created by a trackback as a bidirectional link, and investigate a trackback network of blogs as an example of social network.

We collected large data of such a blog network in the following way. We exploited the blog “Theme salon of blogs” (<http://blog.goo.ne.jp/usertheme/>), where a blogger can recruit trackbacks of other bloggers by registering an interesting theme. By tracing ten steps ahead the trackbacks from the blog of the theme “JR Fukuchiyama Line Derailment Collision” in “Theme salon of blogs”, we collected a large connected trackback network in May, 2005. Here, the total numbers of blogs and trackbacks were 12,047 and 39,960, respectively. We call this data set the BN data.

4.2 Evaluation Method

For a positive integer k ($< N$), let B_k be the optimal target set with k elements that is obtained by the greedy algorithm in the ICM, and let A_k be an approximate solution for the optimal target set with k elements. We evaluate the performance of the approximate solution A_k for the influence maximization problem in the ICM by F -measure $F(k)$,

$$F(k) = |A_k \cap B_k| / k,$$

where $|A_k \cap B_k|$ denotes the number of elements in the set $A_k \cap B_k$.

4.3 Experimental Results

In our experiments, we assigned a uniform probability of p to each directed link in the network for each information diffusion model, that is, $p_{u,v} = p$ for any directed link (u, v) . As regards small propagation probabilities, we investigated $p = 1\%$.

According to Kempe *et al.* [5], we estimated the influence $\sigma(A)$ of target set A in the ICM as follows: We started the process by initially activating set A , and counted the number of active nodes at the end of the process. We then used the empirical mean obtained by simulating the stochastic process 10,000 times as the estimate. However, such estimates needed very heavy computations. For example, for the BN data, the estimates of $\sigma(v)$ ($v \in V$) for the ICM needed about one hour. Here, all our experimentation was undertaken on a single Dell PC with an Intel 3.4GHz Xeon processor, with 2GB of memory. Incidentally, it took about one and three minutes, respectively, to obtain the exact computations

of $\sigma(v)$ ($v \in V$) for the SPM and SP1M. Thus, we investigated the approximate solution based on 100 simulations in the ICM for reference purposes.

On the other hand, Kempe *et al.* [5] investigated a natural special case of the ICM called the “*Only-Listen-Once*” Model (OLOM). In the OLOM, each node v has a parameter p_v so that the parent of v that first attempts to activate v succeeds with probability p_v , and all subsequent attempts to activate v deterministically fail. In other words, v only listens to the first parent that tries to activate it. We also investigated the approximate solution based on the OLOM for reference purposes. Here, we used 10,000 simulations to estimate the influence $\sigma(A)$ of target set A in the OLOM according to Kempe *et al.* [5].

Figure 1 displays the F -measures $F(k)$ of the approximate solutions based on the SPM, the SP1M, 100 simulations in the ICM, and the OLOM at target set size k ($1 \leq k \leq 15$). Here, circles, squares, triangles, and diamonds indicate the results for the SPM, the SP1M, 100 simulations in the ICM, and the OLOM, respectively. It is observed that the proposed methods significantly outperformed the method based on 100 simulations in the ICM and the method based on the OLOM. In particular, the method based on the SP1M achieved an extremely high performance. These results show the effectiveness of the proposed methods.

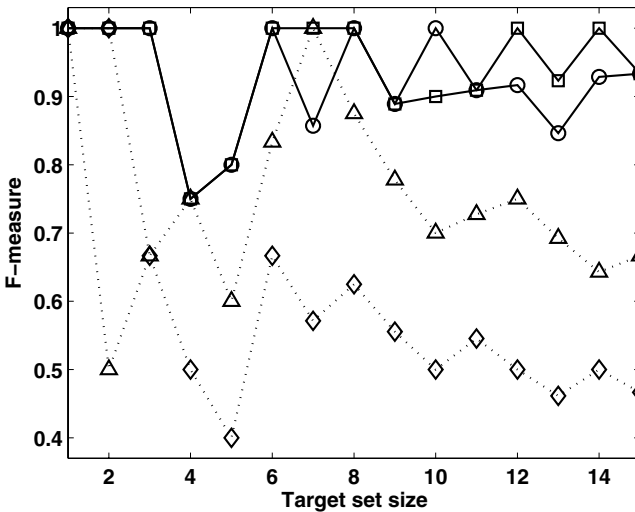


Fig. 1. Performance of approximate solutions in the BN data (“o”: SPM. “□”: SP1M. “△”: 100 simulations. “◇”: OLOM.)

5 Conclusions

We have considered the influence maximization problem in a large-scale social network for the ICM. Although we must compute the influence $\sigma(A)$ of target set A to solve the influence maximization problem, good estimates of $\sigma(A)$ needed

a large amount of computation. Thus, very large computational quantities were needed to approximately solve the influence maximization problem based on the natural greedy algorithm.

We have proposed methods to efficiently obtain good approximate solutions for the influence maximization problem in the case where the propagation probabilities through links are small. The proposed methods exploit natural models for information diffusion in a social network called the SPM and the SP1M. These models are natural special cases of the ICM. For these models, the influence $\sigma(A)$ of each target set A can be exactly and efficiently computed, and the provable performance guarantee for the natural greedy algorithm can be obtained. Using real data on a large-scale blog network, we have experimentally demonstrated that the proposed methods work well and the method based on the SP1M is especially effective.

References

1. Domingos, P., Mining social networks for viral marketing, *IEEE Intelligent Systems*, **20** (2005) 80-82.
2. Domingos, P., and Richardson, M., Mining the network value of customers, *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2001) 57-66.
3. Goldenberg, K. J., Libai, B., and Muller, E., Talk of the network: A complex systems look at the underlying process of word-of-mouth, *Marketing Letters*, **12** (2001) 211-223.
4. Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A., Information diffusion through blogspace, *Proceedings of the 13th International World Wide Web Conference* (2004) 491-501.
5. Kempe, D., Kleinberg, J., and Tardos, E., Maximizing the spread of influence through a social network, *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003) 137-146.
6. Kempe, D., Kleinberg, J., and Tardos, E., Influential nodes in a diffusion model for social networks, *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming* (2005) 1127-1138.
7. Kimura, M., and Saito, K., Tractable models for information diffusion in social networks, *Submitted for Publication* (2006).
8. McCallum, A., Corrada-Emmanuel, A., and Wang, X., Topic and role discovery in social networks, *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (2005) 786-791.
9. Nemhauser, G. L., and Wolsey, L. A., *Integer and Combinatorial Optimization*, Wiley, New York, 1988.
10. Newman, M. E. J., Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, *Physical Review E*, **64** (2001) 016132.
11. Palla, G., Derényi, I., Farkas I., and Vicsek, T., Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, **435** (2005) 814-818.
12. Richardson, M., and Domingos, P., Mining knowledge-sharing sites for viral marketing, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002) 61-70.

Improving Convergence Performance of PageRank Computation Based on Step-Length Calculation Approach

Kazumi Saito¹ and Ryohei Nakano²

¹ NTT Communication Science Laboratories
2-4 Hikaridai, Seika, Soraku, Kyoto 619-0237 Japan
saito@cslab.kecl.ntt.co.jp

² Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555 Japan
nakano@ics.nitech.ac.jp

Abstract. We address the task of improving convergence performance of PageRank computation. Based on a step-length calculation approach, we derive three methods, which respectively calculates its step-length so as to make the successive search directions orthogonal (orthogonal direction), minimize the error at the next iteration (minimum error) and make the successive search directions conjugate (conjugate direction). In our experiments using a real Web network, we show that the minimum error method is promising for this task.

1 Introduction and Background

Methods for computing the relative rank of Web pages based on link (hyperlink) structure play an important role for modern Web ranking systems. One successful and well-publicized link-based ranking system is PageRank, which is used by the Google search engine [2]. In comparison with other Web ranking methods such as HITS [5] or social network analysis methods such as betweenness centrality [8], some attractive features of PageRank include its query-independence, its potential scalability and its virtual immunity to spamming [6].

Although we can obtain the PageRank scores by computing the stationary vector of a Markov chain, the sheer size of the matrix makes this problem highly challenging. Originally, Brin and Page [2] used power iterations for obtaining the PageRank vector. In order to improve its convergence performance, an approach to solving a linear system has been studied extensively (e.g., [3,6]). On the other hand, we believe that a step-length calculation approach, which has been widely used in a number of fields such as nonlinear optimization [7] and neural computation [1], must be promising for this purpose. However, this approach has not yet been explored to our knowledge.

In this paper, we address the task of improving convergence performance of PageRank computation based on a step-length calculation approach. In Section 2,

after reviewing the standard PageRank method, we describe three methods for computing PageRank scores of a given Web network. In Section 3, after explaining our experimental settings, we report the convergence performance of these methods. In Section 4, we discuss some related work and future directions.

2 PageRank Calculation Methods

In this section, after reviewing the standard PageRank method, we propose three methods for computing a PageRank vector of a given Web network.

2.1 Standard PageRank Method

For a given Web network (graph), let $S = \{1, \dots, N\}$ be a set of Web pages (vertices), and \mathbf{A} be its adjacency matrix. Namely, the (i, j) component of the adjacency matrix, denoted by $a_{i,j}$, is set to 1 if there exists a hyperlink (directed edge) from pages i to j ; otherwise 0. Let l_i be the number of out-links from Web page i ($l_i = \sum_{j=1}^N a_{i,j}$), then we can consider the row-stochastic transition matrix \mathbf{P} , each of whose elements is defined by

$$p_{i,j} = \begin{cases} a_{i,j}/l_i & (l_i > 0) \\ v_j & (l_i = 0) \end{cases} \tag{1}$$

where \mathbf{v} is some probability distribution over pages, i.e., $v_i \geq 0$ and $\sum_{i=1}^N v_i = 1$. This model means that from dangling Web pages without out-links ($l_i = 0$), a random suffer jumps to page j with probability v_j . The vector \mathbf{v} is referred to as a personalized vector because we can define \mathbf{v} according to user's preference.

Let \mathbf{x} denote a vector representing the PageRank scores over pages, where $x_i \geq 0$ and $\sum_{i=1}^N x_i = 1$. Then the PageRank vector \mathbf{x} is defined as a limiting solution of the following iterative process,

$$\mathbf{x}_{k+1} = (c\mathbf{P}^T + (1 - c)\mathbf{v}\mathbf{e}^T)\mathbf{x}_k = c\mathbf{P}^T\mathbf{x}_k + (1 - c)\mathbf{v}, \tag{2}$$

where $c \in [0, 1]$ is a teleportation coefficient, $\mathbf{e} = (1, \dots, 1)^T$, and \mathbf{P}^T stands for a transposed matrix of \mathbf{P} . This model means that with the probability $1 - c$, a random suffer also jumps to some page according to the probability distribution \mathbf{v} . The matrix $(c\mathbf{P}^T + (1 - c)\mathbf{v}\mathbf{e}^T)$ is referred to as a Google matrix. The standard PageRank method calculates its solution by directly iterating Equation (2), after initializing \mathbf{x}_0 adequately. One measure to evaluate its convergence performance is defined by

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_{L1} \equiv \sum_{i=1}^N |x_{k+1,i} - x_{k,i}|. \tag{3}$$

Note that any method can give almost the same PageRank scores if it makes Equation (3) almost zero. This is because the unique solution of Equation (2) is guaranteed.

2.2 Orthogonal Direction Method

From Equation (2), we can define the following error vector \mathbf{g}_k over pages at the k -th iteration,

$$\mathbf{g}_k = -(\mathbf{x}_{k+1} - \mathbf{x}_k) = \mathbf{Q}\mathbf{x}_k - (1 - c)\mathbf{v}, \tag{4}$$

where $\mathbf{Q} \equiv \mathbf{I} - c\mathbf{P}^T$ and \mathbf{I} stands for the N -dimensional identity matrix. Let α_k be some positive step-length, then we can consider the following form of general update formula.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k \tag{5}$$

Clearly, by setting $\alpha_k = 1$ at each iteration, we can obtain the standard PageRank method equivalent to Equation (2). On the other hand, Equation (5) is in the form similar to the standard gradient descent method [7]. However, we cannot directly define an objective function because the matrix \mathbf{Q} is not symmetric. Whereas, by focusing on the property that the gradient descent method makes the successive search directions \mathbf{g}_k and \mathbf{g}_{k+1} orthogonal, we can derive a method for calculating step-length α_k at the k -th iteration. Hereafter, this method is referred to as an orthogonal direction method.

From Equations (4) and (5), we can update the error vector \mathbf{g}_{k+1} at the $(k + 1)$ -th iteration as follows:

$$\mathbf{g}_{k+1} = \mathbf{g}_k - \alpha_k \mathbf{Q}\mathbf{g}_k. \tag{6}$$

Thus by using the following step-length α_k ,

$$\alpha_k = \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{Q}\mathbf{g}_k}, \tag{7}$$

we can make the error vectors \mathbf{g}_k and \mathbf{g}_{k+1} orthogonal, i.e., $\mathbf{g}_k^T \mathbf{g}_{k+1} = 0$. Below we summarize the orthogonal direction method:

- OD1.** Initialize \mathbf{x}_0 , and set $k = 0$ and $\mathbf{g}_0 = \mathbf{Q}\mathbf{x}_0 - (1 - c)\mathbf{v}$;
- OD2.** If $\|\mathbf{g}_k\|_{L1} < \epsilon$, then terminate;
- OD3.** Calculate a vector: $\mathbf{r}_k = \mathbf{Q}\mathbf{g}_k$;
- OD4.** Calculate the step-length: $\alpha_k = (\mathbf{g}_k^T \mathbf{g}_k) / (\mathbf{g}_k^T \mathbf{r}_k)$;
- OD5.** Update the PageRank vector: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k$;
- OD6.** Update the error vector: $\mathbf{g}_{k+1} = \mathbf{g}_k - \alpha_k \mathbf{r}_k$;
- OD7.** Set $k = k + 1$, and go to step **OD2**.

Note that since $\mathbf{e}^T \mathbf{g}_k = 0$, from Equation (5) we can guarantee $\mathbf{e}^T \mathbf{x}_k = 1$ for any α_k at any iteration.

We summarize the computational complexity of the orthogonal direction method. In **OD3**, we need to calculate a product of the matrix \mathbf{Q} and vector \mathbf{g}_k , but this amount of computation is also required in the standard PageRank method. More specifically, let L be the total number of hyperlinks, then by using a sparse network representation, we can calculate such a product within $O(L)$ complexity. In the other steps, we can perform their computation within $O(N)$ complexity. Note that $N \ll L \ll N^2$ in most Web networks. Therefore, the order of computational complexities of the standard PageRank and orthogonal direction are equivalent.

2.3 Minimum Error Method

As our second method, we consider calculating step-length α_k so as to minimize the error vector \mathbf{g}_{k+1} in terms of the L2 norm. Namely, we define its objective function as follows:

$$\|\mathbf{g}_{k+1}\|_{L2}^2 = \|\mathbf{g}_k - \alpha_k \mathbf{Q}\mathbf{g}_k\|_{L2}^2 = (\mathbf{g}_k - \alpha_k \mathbf{Q}\mathbf{g}_k)^T (\mathbf{g}_k - \alpha_k \mathbf{Q}\mathbf{g}_k). \tag{8}$$

Then we can easily obtain the following optimal step-length α_k that minimizes equation (8).

$$\alpha_k = \frac{\mathbf{g}_k^T \mathbf{Q}\mathbf{g}_k}{\mathbf{g}_k^T \mathbf{Q}^T \mathbf{Q}\mathbf{g}_k}. \tag{9}$$

Hereafter, this method is referred to as a minimum error method.

Clearly we can obtain the minimum error method just by replacing **OD4** of the orthogonal direction method with Equation (9). Below we summarize the minimum error method:

- ME1.** Initialize \mathbf{x}_0 , and set $k = 0$ and $\mathbf{g}_0 = \mathbf{Q}\mathbf{x}_0 - (1 - c)\mathbf{v}$;
- ME2.** If $\|\mathbf{g}_k\|_{L1} < \epsilon$, then terminate;
- ME3.** Calculate a vector: $\mathbf{r}_k = \mathbf{Q}\mathbf{g}_k$;
- ME4.** Calculate the step-length: $\alpha_k = (\mathbf{g}_k^T \mathbf{r}_k) / (\mathbf{r}_k^T \mathbf{r}_k)$;
- ME5.** Update the PageRank vector: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k$;
- ME6.** Update the error vector: $\mathbf{g}_{k+1} = \mathbf{g}_k - \alpha_k \mathbf{r}_k$;
- ME7.** Set $k = k + 1$, and go to step **ME2**.

As described earlier, we can also guarantee $e^T \mathbf{x}_k = 1$ at any iteration. Its order of computational complexity is equivalent to those of the standard PageRank and orthogonal direction methods.

In the case of the minimum error method, we can guarantee its convergence in terms of the L2 norm. By noting that

$$\frac{\|\mathbf{g}_{k+1}\|_{L2}^2}{\|\mathbf{g}_k\|_{L2}^2} = 1 - \frac{(\mathbf{g}_k^T \mathbf{r}_k)^2}{\|\mathbf{g}_k\|_{L2}^2 \|\mathbf{r}_k\|_{L2}^2} \tag{10}$$

we can see that the minimum error method reduces the error at each iteration unless $\mathbf{g}_k^T \mathbf{r}_k = 0$. In addition, we can show that $\mathbf{g}_k^T \mathbf{r}_k \neq 0$ as follows. Let $\{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ be a set of right eigenvectors of the stochastic transition matrix \mathbf{P}^T , then we can express the error vector by $\mathbf{g}_k = \sum_i a_i \mathbf{u}_i$ where a_i is some adequate coefficient.

$$\mathbf{g}_k^T \mathbf{r}_k = \mathbf{g}_k^T (\mathbf{I} - c\mathbf{P}^T)\mathbf{g}_k = \sum_i a_i^2 (1 - c\lambda_i) \tag{11}$$

Here we usually set $c < 1$, and in general $\lambda_i \leq 1$ because \mathbf{P} is a Markov transition matrix. Therefore since $\mathbf{g}_k^T \mathbf{r}_k \neq 0$, we can guarantee that the minimum error method always converges in terms of the L2 norm.

2.4 Conjugate Direction Method

As our third method, we consider calculating step-length α_k based on the standard conjugate direction method [7]. Since it is not easy to make all search directions Q-orthogonal, we focus only on the following two conditions,

$$\mathbf{g}_{k+1}^T \mathbf{d}_k = 0, \quad \mathbf{d}_{k+1}^T \mathbf{Q} \mathbf{d}_k = 0, \tag{12}$$

where \mathbf{d}_k stands for the search direction at the k -th iteration. The step-length α_k is calculated as follows:

$$\alpha_k = \frac{\mathbf{g}_k^T \mathbf{d}_k}{\mathbf{g}_k^T \mathbf{Q} \mathbf{g}_k}. \tag{13}$$

Hereafter, this method is referred to as a conjugate direction method.

Below we summarize the conjugate direction method.

- CD1.** Initialize \mathbf{x}_0 , and set $k = 0$, $\mathbf{g}_0 = \mathbf{Q} \mathbf{x}_0 - (1 - c) \mathbf{v}$ and $\mathbf{d}_0 = -\mathbf{g}_0$;
- CD2.** If $\|\mathbf{g}_k\|_{L1} < \epsilon$, then terminate;
- CD3.** Calculate a vector: $\mathbf{r}_k = \mathbf{Q} \mathbf{d}_k$;
- CD4.** Calculate the step-length: $\alpha_k = (\mathbf{g}_k^T \mathbf{d}_k) / (\mathbf{d}_k^T \mathbf{r}_k)$;
- CD5.** Update the PageRank vector: $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$;
- CD6.** Update the error vector: $\mathbf{g}_{k+1} = \mathbf{g}_k + \alpha_k \mathbf{r}_k$;
- CD7.** Calculate coefficient: $\beta_k = (\mathbf{g}_{k+1}^T \mathbf{r}_k) / (\mathbf{d}_k^T \mathbf{r}_k)$;
- CD8.** Update search direction: $\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k$;
- CD9.** Set $k = k + 1$, and go to step **CD2**.

Note that since $\mathbf{e}^T \mathbf{d}_k = 0$ for any β_k due to **CD8**, we can again guarantee $\mathbf{e}^T \mathbf{x}_k = 1$ at any iteration. Its order of computational complexity is also equivalent to those of the standard PageRank, orthogonal direction and minimum error methods.

3 Evaluation by Experiments

After explaining our experimental settings, we report the convergence performance of the methods described in the previous section.

3.1 Experimental Settings

In our experiments, we used an English version of the Wikipedia article network consisting of 762,311 Web pages and 15,745,863 hyperlinks¹. Here we collected the network as of November 2005.

As for our experimental settings, we set the initial and personalized vectors of each method to $\mathbf{x}_0 = \mathbf{v} = (1/N, \dots, 1/N)^T$. The parameter ϵ that controls the termination criterion was set to $\epsilon = 10^{-12}$. In addition to the evaluation measure defined in Equation (3) based on the L1 norm, we also evaluated each method by using the following measure based on the L2 norm.

¹ <http://en.wikipedia.org/wiki/Wikipedia>

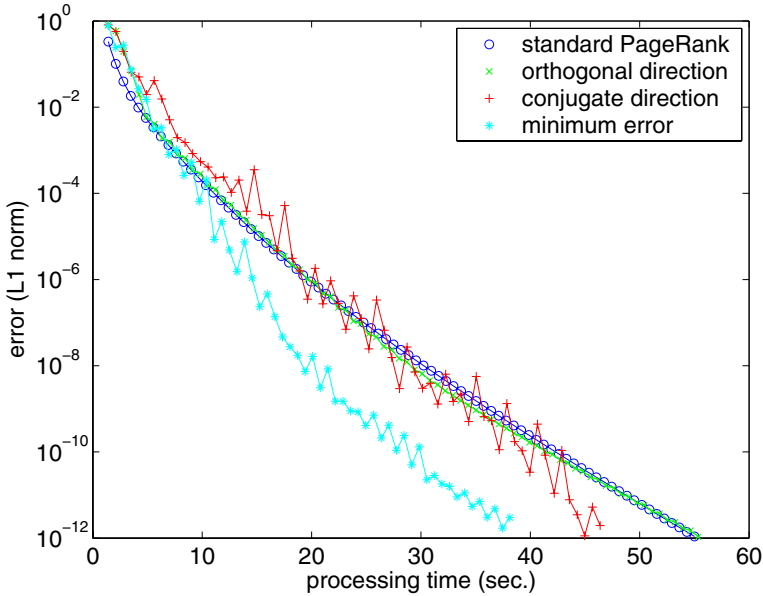


Fig. 1. Convergence performance based on L1 norm

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_{L2}^2 \equiv \sum_{i=1}^N (x_{k+1,i} - x_{k,i})^2 \tag{14}$$

All our experiments were done by using a Dell PC with an Intel 3.4GHz Xeon processor with 2GB of memory.

3.2 Experimental Results

Figure 1 shows experimental results based on the L1 norm. This figure shows that the performances of the standard PageRank and orthogonal direction methods were almost the same, while the error reduction curve of the conjugate direction method was somewhat unstable. Among these methods, we can see that the minimum error method worked most efficiently.

Figure 2 shows experimental results based on the L2 norm. We can observe that the orthogonal direction method worked slightly better than the standard PageRank in the middle of iterations, and the conjugate direction method was still somewhat unstable. Again we can see that the minimum error method worked most efficiently among these methods.

Table 1 compares the processing times until convergence, the total numbers of iterations and the average one-iteration times of these methods. From this table, we can see that the one-iteration times were almost comparable among these methods. Our experimental results indicate that the approach based on the step-length calculation is promising, and we believe that it is possible to produce better methods by elaborating step-length calculation techniques.

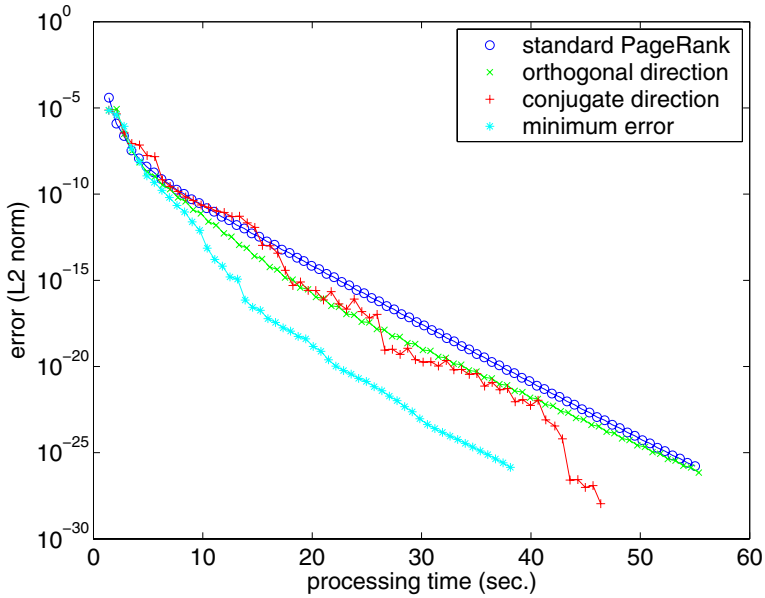


Fig. 2. Convergence performance based on L2 norm

Table 1. Processing time, number of iterations, and one iteration time

	PageRank	orthogonal	conjugate	min. error
processing time (sec.)	55.01	55.35	46.39	38.13
number of iterations	79	78	65	54
one iteration time (sec.)	0.6963	0.7096	0.7137	0.7061

4 Related and Future Research

The standard PageRank method was formalized as a power method for calculating the principal eigenvector with respect to the Google matrix $(cP^T + (1 - c)\mathbf{v}\mathbf{e}^T)$. Although there exist a large number of established methods such as LU factorization [4], it is not easy to directly apply such methods due to the size of the matrix. Thus power iterations have still been widely used to calculate the PageRank scores.

As mentioned earlier, work on solving a linear system $\mathbf{Q}\mathbf{x} = (1 - c)\mathbf{v}$ with some techniques including the Jacobi, Gauss-Seidel, SOR, and Krylov subspace methods has been studied by a number of researchers [3,6]. However, one group of techniques require an adequate preprocessing to the matrix \mathbf{Q} , and another group of techniques substantially increase one iteration processing time. For instance, one Krylov subspace method (biconjugate gradient) needs to calculate both $\mathbf{P}^T\mathbf{x}$ and $\mathbf{P}\mathbf{x}$ at each iteration, making one-iteration time twice as much.

Compared with these conventional techniques, our approach focused on the following type of general update formula.

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k. \quad (15)$$

In our experiments, we showed that without substantial increases of one-iteration processing time, the minimum error method worked efficiently. In addition, we have shown theoretically that this method always converges in terms of the L2 norm. However, in order to clarify its relative strength and weakness, we need to perform further experiments using a wider variety of networks.

5 Conclusion

We addressed the task of improving convergence performance of PageRank computation. Based on a step-length calculation approach, we derived three methods, which respectively calculates its step-length so as to make the successive search directions orthogonal (orthogonal direction), minimize the error at the next iteration (minimum error) and make the successive search directions conjugate (conjugate direction). In our experiments using a real Web network, we showed that the minimum error method is promising for this task.

References

1. C. M. Bishop, *Neural networks for pattern recognition*, Clarendon Press, (1995).
2. S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine, In *Proceedings of the Seventh International World Wide Web Conference* (1998) 107–117.
3. D. Gleich, L. Zhukov, and P. Berkhin, Fast parallel PageRank: a linear system approach In *Proceedings of the 14th International World Wide Web Conference* (2004).
4. G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, (1989).
5. J. Kleinberg, Authoritative sources in a hyperlinked environment, In *Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms* (1998) 668–677.
6. A. N. Langville and C. D. Meyer, Deeper inside PageRank, *Internet Mathematics*, **1:3** (2005) 335–380.
7. D. G. Luenberger, *Linear and nonlinear programming*. Addison-Wesley (1984).
8. M.E.J. Newman, The structure and function of complex network, *SIAM Review*, **45:2** (2003) 167–256.

Prediction of the *O*-glycosylation Sites in Protein by Layered Neural Networks and Support Vector Machines

Ikuko Nishikawa, Hirotaka Sakamoto, Ikue Nouno, Takeshi Iritani,
Kazutoshi Sakakibara, and Masahiro Ito

College of Information Science and Engineering,
Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan
nishi@ci.ritsumei.ac.jp

Abstract. *O*-glycosylation is one of the main types of the mammalian protein glycosylation, which is serine or threonine specific, though any consensus sequence is still unknown. In this paper, a layered neural network and a support vector machine are used for the prediction of *O*-glycosylation sites. Three types of encoding for a protein sequence within a fixed size window are used as the input to the network, that is, a sparse coding which distinguishes all 20 amino acid residues, 5-letter coding and hydrophathy coding. In the neural network, one output unit gives the prediction whether a particular site of serine or threonine is glycosylated, while SVM classifies into the 2 classes. The performance is evaluated by the Matthews correlation coefficient. The preliminary results on the neural network show the better performance of the sparse and 5-letter codings compared with the hydrophathy coding, while the improvement according to the window size is shown to be limited to a certain extent by SVM.

1 Introduction

Glycosylation is one of the main topics in the post-genome era, and plays an important role in life phenomena, binding with the protein or the lipid. *O*-glycosylation is one of the main types of mammalian protein glycosylation and known to recognize serine or threonine specific. However, not all serine and threonine residues are glycosylated, and no consensus sequence is obtained.

In this paper, a preliminary study of the prediction of *O*-glycosylation sites in the protein is reported. Layered neural networks were used in Ref.[1] for the prediction. We examine three ways of encoding for the input data of a protein sequence using the same data set with the reference. Here, a support vector machine with radial basis function is used, as well as a 3-layered neural network, with various sizes of the window for the input sequences. The paper is organized as follows. In Section 2, the description is given on the data used for the training and evaluation, followed by three encoding methods of the input data for the prediction. Then Section 3 shows the prediction method using a layered feedforward neural network, while a support vector machine is used in Section 4. Some extensions on the encoding are studied in Section 5, and Section 6 summarizes the present results.

2 Protein Sequences as the Input Data

2.1 Mammalian Protein Sequences with the Glycosylation Annotation

The data used for the prediction of *O*-glycosylation sites in the protein are chosen to be the same with Ref.[1]. That is, from their glycosylation database OGlycbase[3], 85 protein sequences are selected, which contain a total number of 422 experimentally verified *O*-glycosylation sites. The structural information is registered in the Protein Data Bank (PDB) for 14 sequences among them[1]. 85 protein sequences are divided into 3 sets, which are denoted as dataset 1, 2 and 3 in the following. As *O*-glycosylation is known to be serine or threonine specific, we focus on these two kinds of amid acid residues in the protein sequences. Each protein sequence in the three datasets contains some serine and threonine residue sites which are annotated experimentally as being glycosylated, together with other serine and threonine sites which have no such annotations. Let us call the former a positive site, while the latter a negative site. The total numbers of the positive sites, both of serine(S) and threonine(T) residues, contained in the 3 datasets are shown in Table 1. The negative sites of either serine(S) or threonine(T) residue in 85 protein sequences are chosen with equal probability, and the numbers are also shown in the table. The total numbers of positive and negative sites are 828 for all 3 datasets. The above data is selected to be the same as in Ref.[1] as far as which are described in Ref.[2].

Table 1. The number of positive and negative sites in 3 datasets

Dataset	1	2	3
No. of positive S sites	51	34	73
No. of positive T sites	87	74	103
No. of negative S or T sites	690	720	652
Total	828	828	828

Among the 3 sets of S and T sites, 2 sets are used for the training while the rest 1 set is used for the evaluation, for 3 fold cross-validation of the neural network training.

2.2 Encoding of the Protein Sequences

The goal of the present research is to predict whether individual serine or threonine site is glycosylated under an appropriate condition. Therefore, the input to the prediction system is a certain length of a protein sequence with a serine or threonine site at the center, which is a target of the prediction. The length of the protein sequence, or a window size W_s , is one of the parameters in the prediction. The length W_s sequence is composed of 20 amino acid residues, together with some vacancies if the target S or T site is closer from the sequence terminal than $(W_s - 1)/2$.

Each amino acid is encoded by the following 3 coding methods.

Sparse coding. 21- binary sequence is used to code one site of amino acid or a vacancy.

As the center of the length W_s protein sequence is either S or T, therefore, the total sequence is coded by $(W_s - 1) \times 21 + 2$ binary codes. This rapid increase of the input

data size according to the window size W_s causes the difficulty for a layered neural network, because of the limited number of the available data with annotations as are described in the previous subsection.

5-letter coding. In order to decrease the binary length of the above sparse coding, 20 kinds of amino acids are classified into 5 groups depending on their biochemical characteristics[2], namely, aliphatic, charged, polar, cyclic and the other. Both S and T are included in the polar group. Then, the sparse coding based on this reduced 5-letter alphabet needs only 6-binary sequence for each site of amino acid or a vacancy. And the total sequence is coded by $(W_s - 1) \times 6 + 2$ binary codes, which reduces the difficulty of a large input data size, compared with the original sparse coding.

Hydropathy coding. Each amino acid is coded by a real value which expresses its hydropathy index, which ranges from -4.5 to $+4.5$. A vacancy outside the terminal is coded by 0.0 . As the result, the input data is expressed by a length W_s sequence of real numbers.

3 Prediction by a Feedforward Neural Network

3.1 3-Layered Neural Network

Feedforward neural network with 3 layers is used for the prediction. The input to the network is the encoded protein sequence of the length W_s . Therefore, the number of input units is the same as the length of the encoded data sequence of either binary or real number. There is only one output unit, whose output value corresponds to whether the center of the input sequence is glycosylated or not. The number of hidden units is determined through the experiments. The activation function of the hidden and output units is a sigmoid function with the common parameter value ($\beta = 0.9$ in the following). The back-propagation with a momentum term is used for the training. In the computer experiments in the next subsection, the parameters are set as learning coefficient $\eta = 0.01$ and momentum coefficient $\alpha = 0.1$ or 0.2 .

Because of the limited number of the available data with the annotation, 3 fold cross-validation is used as is already mentioned in Sec.2. And again because of the small number of the positive data compared with the negative data, the following Matthews correlation coefficient[1] is used for the evaluation, instead of a conventional square mean root error.

$$c = \frac{t_p t_n - f_p f_n}{\sqrt{(t_n + f_n)(t_n + f_p)(t_p + f_n)(t_p + f_p)}}, \quad (1)$$

where t_p and t_n are the numbers of correctly predicted as the positive and negative sites, respectively, while f_p and f_n are the numbers of falsely predicted as the positive and negative sites, respectively.

3.2 Computer Experiments

3 methods of encoding are used. The window size is set as $W_s = 3, 5, 7$ and 9 for sparse coding, while $W_s = 11, 21$ and 31 for both 5-letter coding and hydropathy coding. 10 trials with different initial conditions are executed for each input data set.

The results of each coding are shown in Table 2 and Fig.1. The highest Matthews coefficient values c for the evaluation data set with 828 sites after the training by 2 sets with 1656 sites are shown for various W_s .

Table 2. Prediction results using the layered neural networks are shown by Matthews coefficient c for the evaluation data with 828 S and T sites

Window size W_s	3	5	7	9	11	21	31
Sparse coding	0.313	0.303	0.491	0.464			
5-letter coding					0.304	0.361	0.394
Hydropathy coding					0.162	0.151	0.154

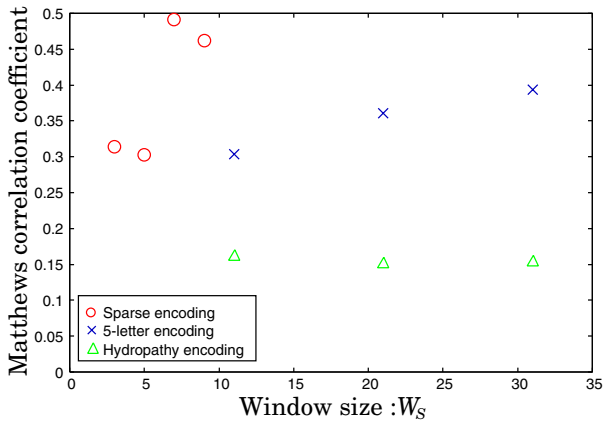


Fig. 1. Prediction results using the layered neural networks for the various window size W_s

For the sparse coding, W_s is limited to the small values in order to keep the number of the parameters (weights) reasonable under the limited number of data. On the other hand, the neural network is unable to learn sufficiently with $W_s = 3, 5$ even for the training data. The reason is clarified that several input data of such short protein sequences contradict each other for the prediction, that is, some are positive and the others are negative with the same input sequence. $W_s = 7$ does not suffer this problem and gives the successful result with $c = 0.491$. For the 5-letter coding, larger W_s are taken, and the better results are obtained according to the increase of W_s , though c remains low compared with the sparse coding. Hydropathy coding shows the poor results up to $W_s = 31$, which may indicate the hydropathy is less essential for the glycosylation.

Based on the above results, the sparse and the 5-letter codings with larger W_s are examined using a support vector machine in the next section.

4 Prediction by a Support Vector Machine

Support vector machine (SVM)[4] is used for the prediction. Then, window size W_s of the input data is able to be extended to larger values even for the sparse coding, as the

system size and the number of variable parameters are independent of the dimension of the input data in SVM.

A famous open source package of SVM, SVM-light[5] is used for the experiments. Radial basis function is chosen as a kernel function, then γ is a unique parameter in this kernel function $\exp(-\gamma||\cdot||^2)$. Another parameter for the SVM training is the margin size C . The following experiments are done with the parameter range of $C = 0.1, 0.25, 0.5, 1.0, 2.5, 5.0, 10, 25, 50, 100$ and $\gamma = 1.0 \times 10^{-4} \sim 1.0 \times 10^0$.

The results of the sparse and the 5-letter codings with each W_s are shown in Table 3 and Fig.2 by the Matthews coefficient for the evaluation data with 828 sites after the training by 1656 sites. The highest coefficient value is shown for each W_s among the results obtained from the above parameter ranges of C and γ . Fig.3 shows the

Table 3. Prediction results using SVM is shown by Matthews coefficient c for the test data with 828 S and T sites

Window size W_s	3	5	7	11	21	31	41	51
Sparse coding	0.240	0.223	0.468	0.344	0.389	0.380	0.313	0.358
5-letter coding	0.044	0.130	0.181	0.221	0.271	0.380	0.406	0.405

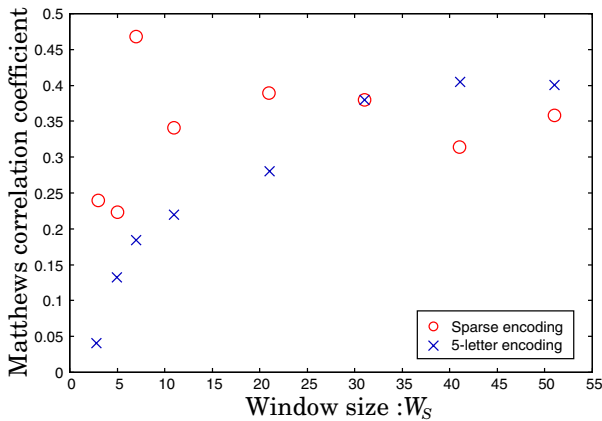


Fig. 2. Prediction results using SVM for the various window size W_s

Matthews coefficients for $W_s = 21$ with various values of C and γ . As is shown there, the results are not much dependent on the margin size C , but largely affected by γ . The coefficients show peak values around the small values of γ , while they remain 0.0 for the larger $\gamma (\geq 0.25)$. Zero correlation is the result of the SVM output, by which all data are classified as negative sites for the evaluation data instead of the perfect classification for the training data. This tendency is commonly seen with the other values of W_s , and the peak moves to smaller γ for larger W_s , as the dimension of the input data becomes higher.

These preliminary results are summarized as follows. First, the results for $W_s = 3, 5$ and 7 are slightly worse but show almost the same tendency and the values with the

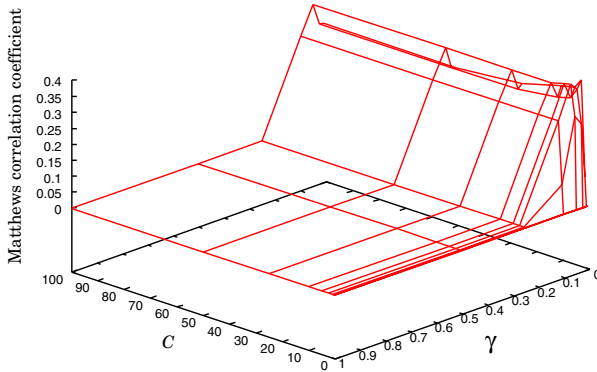


Fig. 3. Prediction results using SVM for $W_s=21$ with various values of C and γ

results obtained by the layered neural network in the previous section, while the classifications of the training data are perfect and therefore better by SVM. This could imply an appropriate selection of the kernel function. Second, the results are not improved by the increase of W_s for the sparse coding. The performance of the sparse coding is overtaken by the 5-letter coding for $W_s > 31$. This implies the detailed information of each residue is necessary or significant only on the several nearest neighbors of a target site. In addition, the high performances obtained only at $W_s = 7$ both by the layered neural network and by SVM seem rather singular, and could be specific to the dataset used for the present experiments. The overall results are still comparable level with what are reported in Ref.[1] for the sparse and 5-letter codings. The better results obtained in [1] are not by the site-specific coding but by the combination of the average quantities over the window size. On the contrary, the goal of our study is to show that the site-specific information is still necessary and not only the average quantity decides the *O*-glycosylation process to take place.

The results of both previous and present sections imply that some improvement is expected by using multiple site-specific codings, for example, a sparse coding for the sites near the center and a reduced letter coding for the outer sites. Therefore in the next section, a few extensions on the encoding are studied. In any coding method, SVM has the advantage to cope with the higher dimensional input data, as the size of feedforward neural networks is unable to be extended owing to the limited number of currently available positive data.

5 Some Extensions on the Encoding

5.1 Multiple Encoding

A Combination of 2 different coding methods is examined in this subsection. That is, the sparse coding is used in the neighborhood of the center, which is a target of the prediction, while the 5-letter coding is used for the outer sequence. Let us denote the

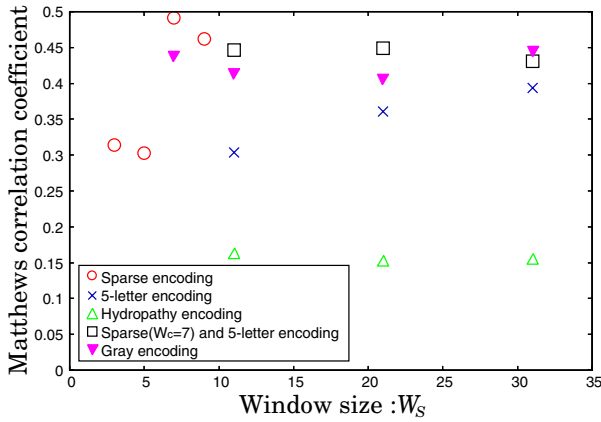


Fig. 4. Prediction results using the layered neural networks with the various encodings

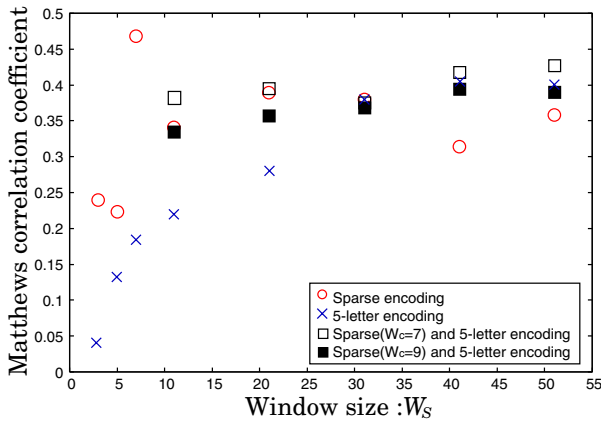


Fig. 5. Prediction results using SVM with the various encodings

window size of the center region as W_c . Then, the length $(W_s - W_c)/2$ sequences are coded by the 5-letter coding in both outer sides.

The results of the multiple coding are shown in Fig.4 and Fig.5 by the Matthews coefficient. Fig.4 is the results obtained by the layered neural network with $W_c = 7$, and $W_s = 11, 21$ and 31 . The coefficient values obtained in Sec.3 are also indicated for the comparison. The multiple coding outperforms the 5-letter coding at all 3 values of W_s . However, the results are slightly worse compared with the sparse coding at $W_s = 7$ and 9 . This could be caused by the high performance of the sparse coding at $W_s = 7$, which was mentioned in the previous section.

Fig.5 is the results obtained by SVM with $W_c = 7, 9$, and $W_s = 11, 21, 31, 41$ and 51 . The coefficient values obtained in Sec.4 are also indicated for the comparison. The multiple coding outperforms both the sparse coding and the 5-letter coding at all 5 values of W_s .

5.2 Gray Encoding

The 5-letter coding reduces the length of a binary code sequence by reducing the number of the alphabet. In this section, the gray coding is used to reduce the binary length to 5. Namely, 00001 to 10110 are used to code 20 kinds of amino acid residues, and 00000 is used for a vacancy. The order of 20 amino acids which corresponds to the ascending order in the gray coding is taken as, Arg, Lys, Asn, Asp, Gln, Glu, His, Pro, Tyr, Ser, Trp, Thr, Gly, Ala, Met, Cys, Phe, Leu, Val and Ile, to consider the biochemical characteristics to some extent. The results at $W_s = 7, 11, 21$ and 31 are shown in Fig.4. The higher performance are obtained compared with the 5-letter coding at all values of W_s .

6 Summary

The paper is a preliminary study on the prediction of *O*-glycosylation sites in the mammalian protein using both layered neural networks and support vector machines. The basic idea is to utilize the site-specific information of the residues around a glycosylation site without the explosion of the system size. The improvement is expected by additional information such as a structural information around the target site and a composition of the protein, and also by a support vector machine with an appropriate space and a kernel.

References

1. Julenius, K., Molgaard, A., Gupta, R. and Brunak, S.: Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites, *Glycobiology* **15** No.2 (2004) 153–164
2. Julenius, K., Molgaard, A., Gupta, R. and Brunak, S.: Supplementary material on Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites (2004)
3. <http://www.cbs.dtu.dk/databases/oglycbase/>
4. Cristianini, N., and Taylor, J.S.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge Univ. Press, (2000)
5. <http://svmlight.joachims.org/>

A Bayesian Approach to Emotion Detection in Dialogist's Voice for Human Robot Interaction

Shohei Kato, Yoshiki Sugino, and Hidenori Itoh

Dept. of Computer Science and Engineering, Graduate School of Engineering,
Nagoya Institute of Technology,
Gokiso-cho Showa-ku Nagoya 466-8555 Japan
{shohey, y-sugino, itoh}@ics.nitech.ac.jp

Abstract. This paper proposes a method for sensitivity communication robots which infer their dialogist's emotion. The method is based on the Bayesian approach: by using a Bayesian modeling for prosodic features. In this research, we focus the elements of emotion included in dialogist's voice. Thus, as training datasets for learning Bayesian networks, we extract prosodic feature quantities from emotionally expressive voice data. Our method learns the dependence and its strength between dialogist's utterance and his emotion, by building Bayesian networks. Bayesian information criterion, one of the information theoretical model selection method, is used in the building Bayesian networks. The paper finally proposes a reasoner to infer dialogist's emotion by using a Bayesian network for prosodic features of the dialogist's voice. The paper also reports some empirical reasoning performance.

1 Introduction

Recently, robotics research has been shifting from industrial to domestic application, and several domestic, human centered robots, aimed at communicating expressively with human, have been developed (e.g., [6,7,8,11,18]). To live with people, robots need to understand people's instruction through communication with them. Communication, even if it is between robots and human, should involve not only conveying messages or instructions but also psychological interaction, such as comprehending mutual sentiments, sympathizing with the other person, and enjoying conversation itself. To communicate this way, a robot requires several mechanisms that enable the robot to recognize human emotions, to have emotions, and to express emotions.

On the other hand, Bayesian networks, one of the eminently practical probabilistic reasoning techniques for reasoning under uncertainty, are becoming an increasingly important area for research and application in the entire field of Artificial Intelligence (e.g., [21,16,3,1]).

In this paper, we propose a method for sensitivity communication robots which infer their dialogist's emotion. The method is based on the Bayesian approach: by using a Bayesian network which models prosodic features of dialogist's voice.

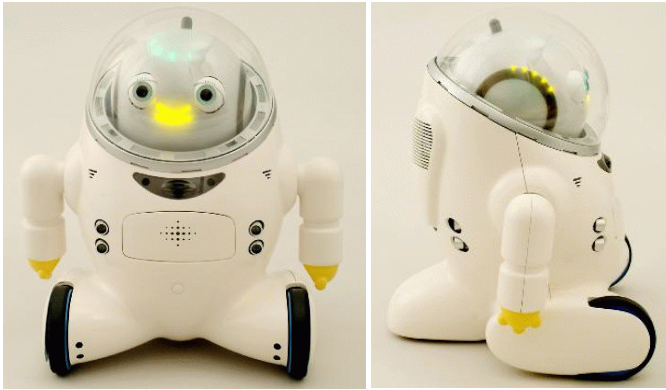


Fig. 1. Ifbot appearance

2 A Sensitivity Communication Robot Ifbot

In this research, A novel robot, Ifbot, which can communicate with human by joyful conversation and emotional facial expression has been developed by our industry-university joint research project [17,15]. Figure 1 shows Ifbot appearance. With two arms, wheels instead of legs and with an astronaut's helmet, Ifbot is 45-centimeter-tall, seven-kilogram robot. Ifbot is able to converse with a person by fundamental voice recognition and synthesis engines. Ifbot is also able to communicate with a person, showing its “emotions” through facial expression mechanisms and gestures. The mechanism for controlling Ifbot's emotional facial expressions [12] have 10 motors and 101 LEDs. The motors actuate Ifbot's neck (2 DOFs), both sides of the eyes (2 DOFs for each), and both side of the eyelids (2 DOFs for each). Ifbot is also capable of recognizing 10 persons by CCD cameras and vision system [14], and has a vocabulary of thousands of words and adapts its conversation to the habits and personalities of different people.

The target of this research is to enable Ifbot to communicate heartfully and expressively with human. In addition to the above mechanisms, Ifbot requires a mechanism which recognizes human emotions. This is a very important for the target. In this paper, we propose a Bayesian approach to the dialogist's emotion recognition mechanism. Figure 2 shows the overview of the human robot conversation system focused on emotion processing, which we intend to implement in Ifbot. This paper describes a voice analysis based method of building a Bayesian network for emotion detection.

3 Bayesian Networks

A Bayesian network (BN) is a graphical structure that allows us to represent and reason about an uncertain domain [16]. The graph structure is constrained to be a directed acyclic graph (or simply dag). The node in a Bayesian network

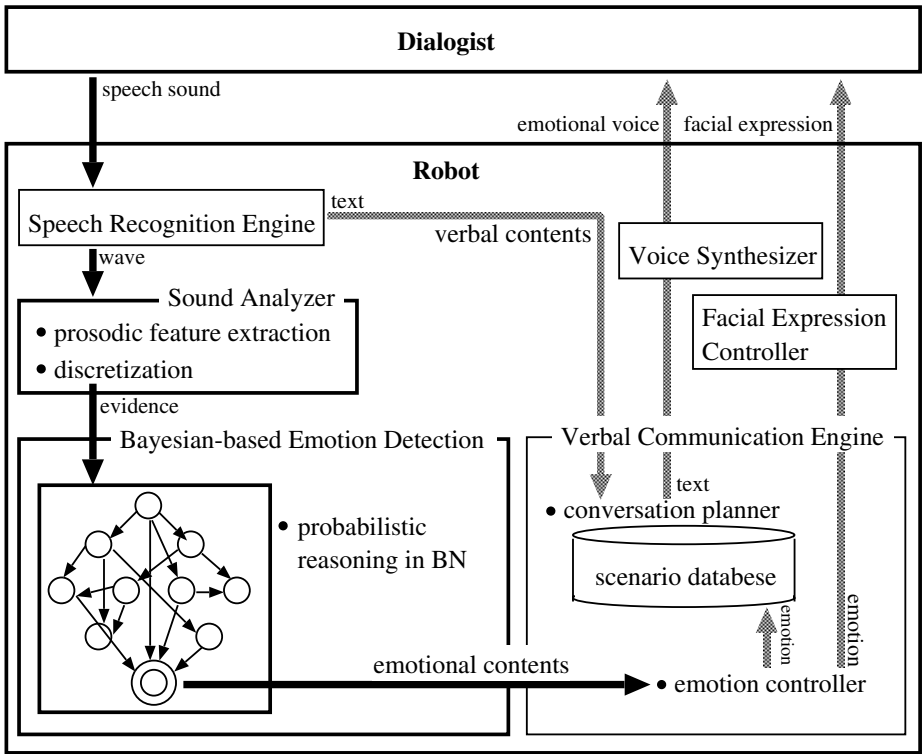


Fig. 2. Bayesian-based emotion detection system for sensitivity communication robot

represent a set of random variables from the domain. A set of directed arcs (or links) connects pairs of nodes, representing the direct dependencies between variables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node.

Most commonly, BNs are considered to be representations of joint probability distributions. Consider a BN containing the n nodes, X_1 to X_n , taken in that order. A particular value in the joint distribution $P(X_1 = x_1, \dots, X_n = x_n)$ is calculated as follows:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(x_i | Parents(X_i)), \tag{1}$$

where $Parents(X_i) \subseteq \{x_1, \dots, x_{i-1}\}$ is a set of parent nodes of X_i . This equation means that node X_i is dependent on only $Parents(X_i)$ and is conditionally independent of nodes except all nodes preceding X_i .

Once the topology of the BN is specified, the next step is to quantify the relationships between connected nodes. Assuming discrete variables, this is done by specifying a conditional probability table (CPT). Consider that node X_i has

Table 1. The Conditional Probability Table for X_i

$p(X_i = y_1 Parents(X_i) = x_1)$	$\cdots \cdots$	$p(X_i = y_1 Parents(X_i) = x_m)$
$\cdots \cdots \cdots \cdots \cdots \cdots$	$\cdots \cdots$	$\cdots \cdots \cdots \cdots \cdots \cdots$
$p(X_i = y_n Parents(X_i) = x_1)$	$\cdots \cdots$	$p(X_i = y_n Parents(X_i) = x_m)$

n possible values y_1, \dots, y_m and its parent nodes $Parents(X_i)$ have m possible combinations of values x_1, \dots, x_m . The conditional probability table for X_i is composed of shown in Table 1.

Once the topology of the BN and the CPT are given, we can do the probabilistic inference in the BN by computing the posterior probability for a set of query nodes, given values for some evidence nodes. Belief propagation (BP) proposed in [21] is a well-known inference algorithm for singly-connected BNs, which has simple network structure called a polytree. Assume X is a query node, and there is some set of evidence nodes \mathbf{E} (not including X). The task of BP is to update the posterior probability of X by computing $P(X|\mathbf{E})$.

In the most general case, the BN structure is a multiply-connected network, where at least two nodes are connected by more than one path in the underlying undirected graph, rather than simply a tree. In such networks, BP algorithm does not work, and then, several enhanced algorithms, such as junction tree [10], logic sampling [9] and loopy BP [20], are proposed as an exact or approximate inference method in multiply-connected networks.

In this research, we intend to reduce the scale of knowledge-base and computational cost drastically, by utilizing Bayesian networks for knowledge representation for emotion detection. Reasoning from partial evidence is forte of Bayesian networks. This is a great advantage of the implementation in domestic robots.

4 Learning of Emotion Detection Engine

In this paper, we focus on the prosodic features of dialogist’s voice as a clue to what emotion the dialogist expresses. The section describes a BN modeling for this problem.

4.1 Training Data

Speech data for learning should be expressive of emotions. In this research, we used numbers of segments of voice samples, which are spoken emotionally by actors and actresses, from films, TV dramas, and so on. In segmentation, transcriber selects six emotional labels (anger, disgust, sadness, fear, surprise or happiness). This label is a goal attribute for emotion detection. It should be noticed that all voice samples in the training data are labeled with any of six emotions, that is, the training data has no voice with neutral emotion. We think that a BN learned from this training data can detect neutral emotion by giving nearly flat probabilities to all emotions.

4.2 Feature Extraction for BN Modeling

Voice has three components (prosody, tone and phoneme). It becomes obvious from past several researches that prosodic component is the most relative to emotional voice expressions (e.g., [22,4]). In this research, as attributes of training data, we adopt three prosodic events: energy, fundamental frequency and duration as acoustic parameters for BN modeling. Prosodic analysis is done for 11 ms frames passed through Hamming extracted from voice waveforms sampled at 22.05 kHz.

The attributes concerning energy, maximum energy (PW_{MAX}), minimum energy (PW_{MIN}), mean energy (PW_{MEAN}) and its standard deviation (PW_S) are determined by the energy contours for the frames in a voice waveform. The attributes concerning fundamental frequency, maximum pitch ($F0_{MAX}$), minimum pitch ($F0_{MIN}$), mean pitch ($F0_{MEAN}$) and its standard deviation ($F0_S$) are determined by short time Fourier transforms for the frames in a voice waveform. As an attribute concerning duration, we measure the duration per a single mora (Tm).

In this paper, the goal attribute ($EMOT$) and the above nine prosodic feature values are assigned to the nodes of a BN model. For learning the discrete causal structure of a BN model, all prosodic features are converted to discrete values.

4.3 Learning Causal Structure of the BN model

The section describes how to specify the topology of the BN model for emotion detection and to parametrize CPT for connected nodes. The emotion detection BN modeling is to determine the qualitative and quantitative relationships between the output node containing goal attribute (emotions) and nodes containing prosodic features. In this paper, we adopt a model selection method based on the Bayesian information criterion (BIC), which has information theoretical validity and is able to learn a high prediction accuracy model by avoidance of over-fitting to training data. Let M be a BN model, $\hat{\theta}_M$ be the maximum likelihood (ML) estimate of the parameter representing M , and d be the number of parameters. A BN model M is evaluated by BIC of M defined as

$$BIC(\hat{\theta}_M, d) = \log_{\hat{\theta}_M}^N P(D) - \frac{d \log N}{2}, \quad (2)$$

where D is training samples and N is the number of the samples. In case that D is partially observed, expectation maximization (EM) algorithm [5] is utilized for estimating θ asymptotically with incomplete data in the training samples.

In this research, as knowledge for emotion detection, we install a BN model, which maximizes BIC for his dialogist's voice data, in the robot. We adopt K2 [2,3] as the search algorithm. K2 needs the pre-provided variable order. We, thus, consider every possible permutation of three node groups: PW , $F0$ and Tm .

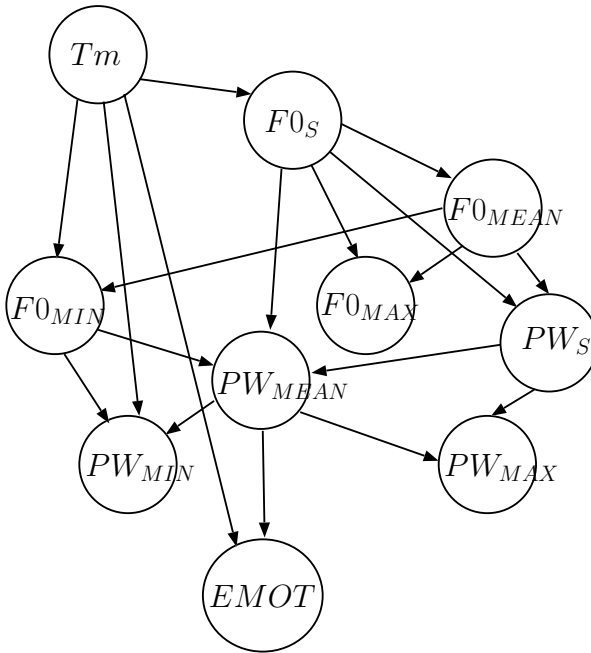


Fig. 3. A Bayesian network for emotion detection

5 Empirical Results

The section describes an empirical results of our Bayesian approach. Firstly, we collected 550 segments of voice waveform and labeled any of six emotional contents as mentioned in Section 4.1, and then extracted nine prosodic features from each of the segments and assigned them to the attributes as mentioned in Section 4.2. In the prosodic analysis, we used the Snack sound toolkit [13]. We then modeled BNs for randomly selected 500 samples with changing six variable orders by Bayes Net Toolbox [19]. Figure 3 shows one of the results with the variable order $Ts \prec F0 \prec PW \prec EMOT$.

We, then, examined the inference performance of the BN model by the emotion detection test for the rest 50 samples of collected segments. In reasoning with the BN, we used junction tree [10], which is well-known as an exact inference algorithm in multiply connected BNs.

For comparison, we also examined by principal component analysis (PCA) and classification based on Mahalanobis distance in four PC space. Table 2 shows the results. The results indicates that the BN has largely acceptable accuracy rates for emotions except sadness and surprise. In this experiment, BN was not able to acquire the knowledge concerning surprise, for lack of voice waveforms which is labeled surprise. We will dedicate to the collection of voice segments with surprise and other various emotional contents.

Table 2. The Accuracy Rates

Emotion	BN (%)	PCA (%)
Anger	66.6	39.2
Disgust	50.0	39.7
Sadness	27.2	29.1
Fear	80.0	41.7
Happiness	42.8	33.3

6 Conclusion

This paper proposed a Bayesian method for sensitivity communication robots which infer their dialogist's emotion. The method provided a Bayesian modeling of prosodic features of dialogist's voice for emotion detection. Considering practical human robot communication, unfortunately, accurate realtime speech recognition and voice analysis is not fully guaranteed. In such case, Bayesian approach proposed in this paper gives much benefit to emotion detection by probabilistic inference from partial evidence and reasoning under uncertainty.

We consider that emotional contents are conveyed not only by voice but also by verbal and facial expression and gesture. In future works, we will dedicate to the BN modeling of the verbal information for emotion detection during conversation, and describe the Bayesian mixture approach: by using a mixture of voice and verbal Bayesian networks in the forthcoming papers.

Acknowledgment

Ifbot was developed as part of an industry-university joint research project among the Business Design Laboratory Co., Ltd., Brother Industries, Ltd., A.G.I. Inc, ROBOS Co., and the Nagoya Institute of Technology. We are grateful to all of them for their input. This work was supported in part by the Kayamori Foundation for Informational Science Advancement, and by a Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research under grant #17500143.

References

1. T. Akiba and H. Tanaka. A Bayesian approach for user modelling in dialog systems. In *15th International Conference of Computational Linguistics*, pages 1212–1218, 1994.
2. G. F. Cooper and E. Herskovits. A Bayesian method for constructing Bayesian belief networks from databases. pages 86–94, 1991.
3. G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

4. R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.
5. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.
6. G. Endo, J. Nakanishi, J. Morimoto, and G. Cheng. Experimental studies of a neural oscillator for biped locomotion with QRIO. In *IEEE International Conference on Robotics and Automation (ICRA2005)*, pages 598–604, 2005.
7. M. Fujita. Development of an Autonomous Quadruped Robot for Robot Entertainment. *Autonomous Robots*, 5:7–18, 1998.
8. M. Fujita, H. Kitano, and T. Doi. *Robot Entertainment*. A. Druin and J. Hendler, eds., Robots for kids: exploring new technologies for learning, Morgan Kaufmann, Ch.2, pp. 37–70, 2000.
9. M. Henrion. Propagating uncertainty in Bayesian networks by logic sampling. *Uncertainty in Artificial Intelligence*, 2:149–163, 1988.
10. F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.
11. S. Kanda, Y. Murase, and K. Fujioka. Internet-based Robot: Mobile Agent Robot of Next-generation (MARON-1). volume 54, pages 285–292, 2003. (in Japanese).
12. M. Kanoh, S. Kato, and H. Itoh. Facial expressions using emotional space in sensitivity communication robot “ifbot”. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*, pages 1586–1591, 2004.
13. K. Sjölander. *The Snack Sound Toolkit*. <http://www.speech.kth.se/snack/>.
14. S. Kato, S. Ohsiro, K. Watabe, H. Itoh, and K. Kimura. A domestic robot with sensitive communication and its vision system for talker distinction. In *Intelligent Autonomous Systems 8*, pages 1162–1168. IOS Press, 2004.
15. S. Kato, S. Ohshiro, H. Itoh, and K. Kimura. Development of a communication robot ifbot. In *IEEE International Conference on Robotics and Automation (ICRA2004)*, pages 697–702, 2004.
16. K. B. Korb and A. E. Nicholson. *Bayesian Artificial Intelligence*. Chapman & Hall/CRC, 2003.
17. Business Design Laboratory Co. Ltd. *The Extremely Expressive Communication Robot, Ifbot*. <http://www.business-design.co.jp/en/product/001/index.html>.
18. Y. Murase, Y. Yasukawa, K. Sakai, and et al. Design of a compact humanoid robot as a platform. In *Proc. of the 19-th conf. of Robotics Society of Japan*, pages 789–790, 2001. (in Japanese), <http://pr.fujitsu.com/en/news/2001/09/10.html>.
19. K. P. Murphy. *Bayes Net Toolbox*. <http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>.
20. K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: an empirical study. pages 467–475, 1999.
21. J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.
22. K. R. Scherer, T. Johnstone, and G. Klasmeyer. *Vocal expression of emotion*. R. J. Davidson, H. Goldsmith, K. R. Scherer eds., Handbook of the Affective Sciences (pp. 433–456), Oxford University Press, 2003.

Finding Nominally Conditioned Multivariate Polynomials Using a Four-Layer Perceptron Having Shared Weights

Yusuke Tanahashi¹, Kazumi Saito², Daisuke Kitakoshi¹, and Ryohei Nakano¹

¹ Nagoya Institute of Technology

Gokiso-cho, Showa-ku, Nagoya 466-8555 Japan

{tanahasi, kitakosi, nakano}@ics.nitech.ac.jp

² NTT Communication Science Laboratories, NTT Corporation

2-4 Hikaridai, Seika, Soraku, Kyoto 619-0237 Japan

saito@cslab.kecl.ntt.co.jp

Abstract. We present a method for discovering nominally conditioned polynomials to fit multivariate data containing numeric and nominal variables using a four-layer perceptron having shared weights. A polynomial is accompanied with the nominal condition stating a subspace where the polynomial is applied. To get a succinct neural network, we focus on weight sharing, where a weight is allowed to have one of common weights. A near-zero common weight can be eliminated. Our method iteratively merges and splits common weights based on 2nd-order criteria, escaping from local optima. Moreover, our method selects the optimal number of hidden units based on cross-validation. The experiments showed that our method can restore the original sharing structure for an artificial data set, and discovers rather succinct rules for a real data set.

1 Introduction

Finding a numeric relationship such as polynomials from data is a key research issue of data mining. Given multivariate data containing numeric and nominal variates, we consider solving piecewise multivariate polynomial regression, where each polynomial is accompanied with the nominal condition stating a subspace where the polynomial is applied. Such a nominally conditioned polynomial is called a *rule*. The RF6.4 [2] finds a set of rules by using a four-layer perceptron.

In data mining using neural networks, a crucial research issue is to find a succinct network from data. To achieve this aim, we focus on *weight sharing* [1], where weights are divided into clusters, and weights within the same cluster have the same value called a *common weight*. A common weight very close to zero can be removed, called *weight pruning*. Recently, a weight sharing method called *BCW (bidirectional clustering of weights)* was proposed [4]. The BCW iteratively merges and splits common weights based on second-order criteria, escaping from local optima through bidirectional operations.

Sections 2 and 3 explain the RF6.4 and BCW1.2[5] respectively; the latter enhanced the original BCW[4] in a couple of points. We combine the BCW1.2 with the RF6.4 and evaluate its performance in Section 4.

2 Piecewise Multivariate Polynomial Regression: RF6.4

This section explains the basic framework and rule restoring of the RF6.4 [2].

Basic Framework. Let $(q_1, \dots, q_{K_1}, x_1, \dots, x_{K_2}, y)$ or $(\mathbf{q}, \mathbf{x}, y)$ be a vector of variables, where q_k and x_k are nominal and numeric explanatory variables, and y is a numeric dependent variable. For each q_k we introduce a *dummy variable* q_{kl} defined as follows: $q_{kl} = 1$ if q_k matches the l -th category, and $q_{kl} = 0$ otherwise. Here $l = 1, \dots, L_k$, and L_k is the number of distinct categories appearing in q_k .

As a true model governing data, we consider the following set of I^* rules.¹

$$\text{if } \bigwedge_k \bigvee_{q_{kl} \in Q_k^i} q_{kl} \quad \text{then } y = \phi(\mathbf{x}; \mathbf{w}^i), \quad i = 1, \dots, I^*, \quad (1)$$

where Q_k^i and \mathbf{w}^i denote a set of q_{kl} and a parameter vector respectively used in the i -th rule. As a class of numeric equations $\phi(\mathbf{x}; \mathbf{w}^i)$, we consider the following multivariate polynomial, whose power values are not restricted to integers.

$$\phi(\mathbf{x}; \mathbf{w}^i) = w_0^i + \sum_{j=1}^{J^i} w_j^i \prod_{k=1}^{K_2} x_k^{w_{jk}^i} = w_0^i + \sum_{j=1}^{J^i} w_j^i \exp\left(\sum_{k=1}^{K_2} w_{jk}^i \ln x_k\right). \quad (2)$$

Here, \mathbf{w}^i is composed of w_0^i , w_{jr}^i and w_{jk}^i , while J^i is the number of terms.

Equation (1) can be represented by the following single numeric function and it can be learned by using a single *four-layer perceptron* as shown in Fig. 1.

$$f(\mathbf{q}, \mathbf{x}; \boldsymbol{\theta}) = c_0 + \sum_{j=1}^J c_j \exp\left(\sum_{k=1}^{K_2} w_{jk} \ln x_k\right), \quad (3)$$

$$c_0 = \sum_{r=1}^R v_{0r} \sigma_r, \quad c_j = \sum_{r=1}^R v_{jr} \sigma_r, \quad \sigma_r = \sigma \left(\sum_{k=1}^{K_1} \sum_{l=1}^{L_k} v_{rkl} q_{kl} \right), \quad (4)$$

where $\boldsymbol{\theta}$ is composed of v_{0r} , v_{jr} , v_{rkl} , and w_{jk} . The optimal number J^* is found through cross-validation.

Rule Restoring. As the first step, coefficient vectors $\mathbf{c}^\mu = (c_0^\mu, c_1^\mu, \dots, c_{J^*}^\mu)$ for all samples $\mu = 1, \dots, N$ are quantized into I representatives $\{\mathbf{a}^i = (a_0^i, a_1^i, \dots, a_{J^*}^i) : i = 1, \dots, I\}$. For vector quantization we employ the k-means [3] due to its simplicity, to obtain I disjoint subsets $\{G_i : i = 1, \dots, I\}$ where the distortion d_{VQ} is minimized. The optimal number I^* is found through cross-validation process.

$$d_{VQ} = \sum_{i=1}^I \sum_{\mu \in G_i} \|\mathbf{c}^\mu - \mathbf{a}^i\|^2, \quad \mathbf{a}^i = \frac{1}{N_i} \sum_{\mu \in G_i} \mathbf{c}^\mu. \quad (5)$$

As the second step, the final rules are obtained by solving a simple classification problem whose training samples are $\{(\mathbf{q}^\mu, i(\mathbf{q}^\mu)) : \mu = 1, \dots, N\}$, where $i(\mathbf{q}^\mu)$ indicates the representative label of the μ -th sample. Here we employ the C4.5 decision tree generation program [6] due to its wide availability.

¹ Here a rule “if A then B” means “when A holds, apply B.”

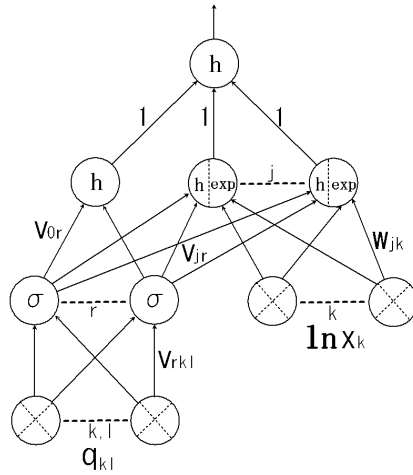


Fig. 1. Four-Layer Perceptron for RF6.4

3 Weight Sharing Method: BCW1.2

This section explains the basic framework and procedure of the BCW1.2 [5].

Notation. Let $E(\mathbf{w})$ be an error function to minimize, where \mathbf{w} denotes a vector of weights $(w_1, \dots, w_d, \dots, w_D)$. Then, we define a set of disjoint clusters $\Omega(G) = \{S_1, \dots, S_g, \dots, S_G\}$, such that $S_1 \cup \dots \cup S_G = \{1, \dots, D\}$. Also, we define a vector of common weights $\mathbf{u} = (u_1, \dots, u_g, \dots, u_G)^T$ associated with a cluster set $\Omega(G)$ such that $w_d = u_g$ if $d \in S_g$. Note that $\hat{\mathbf{u}}$ is obtained by training a neural network whose structure is defined by $\Omega(G)$. Now we consider a relation between \mathbf{w} and \mathbf{u} . Let \mathbf{e}_d^D be the D -dimensional unit vector whose elements are all zero except for the d -th element which is equal to unity. Then the weights \mathbf{w} can be expressed by using a $D \times G$ transformational matrix \mathbf{A} as follows.

$$\mathbf{w} = \mathbf{A}\mathbf{u}, \quad \mathbf{A} = \left[\sum_{d \in S_1} \mathbf{e}_d^D, \dots, \sum_{d \in S_G} \mathbf{e}_d^D \right]. \tag{6}$$

Bottom-up Clustering. A one-step bottom-up clustering transforms $\Omega(G)$ into $\Omega(G - 1)$ by a merge operation; i.e., clusters S_g and $S_{g'}$ are merged into a cluster $\tilde{S}_g = S_g \cup S_{g'}$. We want to select a suitable pair so as to minimize the increase of $E(\mathbf{w})$ as defined below. Here $\mathbf{H}(\mathbf{w})$ denotes the Hessian of $E(\mathbf{w})$.

$$DisSim(S_g, S_{g'}) = \frac{(\hat{u}_g - \hat{u}_{g'})^2}{(\mathbf{e}_g^G - \mathbf{e}_{g'}^G)^T (\mathbf{A}^T \mathbf{H}(\hat{\mathbf{w}}) \mathbf{A})^{-1} (\mathbf{e}_g^G - \mathbf{e}_{g'}^G)}. \tag{7}$$

This is regarded as the second-order criterion for merging S_g and $S_{g'}$, called the *dissimilarity*. We select a pair of clusters which minimizes $DisSim(S_g, S_{g'})$ and

merge the two clusters. After the merge, the network with $\Omega(G - 1)$ is retrained. This is the *one-step bottom-up clustering with retraining*.

Top-down Clustering. A one-step top-down clustering transforms $\Omega(G)$ into $\Omega(G + 1)$ by a split operation; i.e., a cluster S_g is split into two clusters S'_g and S_{G+1} where $S_g = S'_g \cup S_{G+1}$. In this case, we want to select a suitable cluster and its partition so as to maximize the decrease of the error function.

Just after the splitting, we have a $(G + 1)$ -dimensional common weight vector $\tilde{\mathbf{v}} = (\hat{\mathbf{u}}^T, \hat{\mathbf{u}}_g)^T$, and a new $D \times (G + 1)$ transformational matrix \mathbf{B} defined as

$$\mathbf{B} = \left[\begin{array}{cccc} \sum_{d \in S_1} e_d^D, & \dots, & \sum_{d \in S'_g} e_d^D, & \dots, & \sum_{d \in S_G} e_d^D, & \sum_{d \in S_{G+1}} e_d^D \end{array} \right]. \tag{8}$$

Then, we define the general *utility* as follows. The utility values will be positive, and the larger the better. Here $\mathbf{g}(\mathbf{w})$ denotes the gradient of $E(\mathbf{w})$.

$$\text{GenUtil}(S_g, S_{G+1}) = \kappa^2 \mathbf{f}^T (\mathbf{B}^T \mathbf{H}(\mathbf{B}\tilde{\mathbf{v}})\mathbf{B})^{-1} \mathbf{f}. \tag{9}$$

$$\kappa = \mathbf{g}(\mathbf{B}\tilde{\mathbf{v}})^T \sum_{d \in S_{G+1}} e_d^D, \quad \mathbf{f} = \mathbf{e}_{G+1}^{G+1} - \mathbf{e}_g^{G+1}, \tag{10}$$

The original BCW [4] employs a very constricted splitting such as splitting into only one element and the others. Here we remove the constraint. Consider the criterion (9). When a cluster g to split is unchanged, $\mathbf{f}^T (\mathbf{B}^T \mathbf{H}(\mathbf{B}\tilde{\mathbf{v}})\mathbf{B})^{-1} \mathbf{f}$ won't be significantly changed. Since κ is the summation of gradients over the members of a cluster $G + 1$, the gradients to add together should have the same sign if you want a larger κ^2 . Thus, the gradients of the cluster are sorted in ascending order and examined is only splitting into smaller-gradients and larger-gradients. Examining all such candidates, we select the cluster to split and its splitting which maximize the criterion (9). After the splitting, the network with $\Omega(G + 1)$ is retrained. This is the *one-step top-down clustering with retraining*.

BCW1.2. The procedure of BCW1.2 is shown below. The BCW1.2 always converges since the number of different \mathbf{A} is finite.

- step 1:** Get the initial set $\Omega(D)$ through learning. Perform scalar quantization for $\Omega(D)$ to get $\Omega_1(2)$. Remember the matrix $\mathbf{A}^{(0)}$ at $\Omega_1(2)$. $t \leftarrow 1$.
- step 2:** Perform repeatedly the one-step top-down clustering with retraining from $\Omega_1(2)$ to $\Omega(2+h)$. Update the best performance for each G if necessary.
- step 3:** Perform repeatedly the one-step bottom-up clustering with retraining from $\Omega(2+h)$ to $\Omega_2(2)$. Update the best performance for each G if necessary. Remember $\mathbf{A}^{(t)}$ at $\Omega_2(2)$.
- step 4:** If $\mathbf{A}^{(t)}$ is equal to one of the previous ones $\mathbf{A}^{(t-1)}, \dots, \mathbf{A}^{(0)}$, stop. Output the best performance of $\Omega(G)$ for each G as the final result. Otherwise, $t \leftarrow t + 1$, $\Omega_1(G) \leftarrow \Omega_2(G)$ and go to step 2.

4 Evaluation by Experiments

Artificial Data Set We consider the following rules.

$$\begin{cases} \text{if } q_{11} \wedge q_{21} & \text{then } y = 2 + 2x_1x_2^{0.5}x_3^{0.5} + 3x_3^{1.5}x_4x_5, \\ \text{if } q_{12} & \text{then } y = 3 + 1x_1x_2^{0.5}x_3^{0.5} + 1x_3^{1.5}x_4x_5, \\ \text{else} & y = 4 + 4x_1x_2^{0.5}x_3^{0.5} + 2x_3^{1.5}x_4x_5. \end{cases} \quad (11)$$

Here we have 15 numeric and 3 nominal explanatory variables with $L_1 = L_2 = 2$ and $L_3 = 3$. Variables q_3, x_6, \dots, x_{15} are irrelevant. A sample is randomly generated with each x_k in the range of $(0, 1)$, and y is calculated to follow Eq. (11) with Gaussian noise $\mathcal{N}(0, 0.1)$ added. We did a trial with $N = 300$ and $R = 5$.

The learning is terminated when each element of the gradient is less than 10^{-5} . Weight sharing was applied only to weights w_{jk} . We set the width of bidirectional clustering as $h = 10$. The number of hidden units was changed from one to four; $J = 1, \dots, 4$. For model selection we employ 10-fold cross-validation. Note that $J^* = 2$ and $G^* = 4$ for our data.

Table 4 shows 10-fold CV error E_{CV} of RF6.4 + BCW1.2 for artificial data. A number in bold type indicates the best for each J . E_{CV} was minimized at $J = 2$ and $G = 4$, which coincide with $J^* = 2$ and $G^* = 4$.

Table 1. E_{CV} of RF6.4 + BCW1.2 for Artificial Data

G	J = 1	J = 2	J = 3	J = 4
2	8.0861e-02	2.5691e-02	3.2313e-02	2.0596e-02
3	6.3194e-02	1.4182e-02	1.3174e-02	1.4946e-02
4	4.6682e-02	1.2069e-02	1.3436e-02	1.3248e-02
5	4.4039e-02	1.2258e-02	1.3514e-02	1.4922e-02
6	4.4394e-02	1.2123e-02	1.2522e-02	1.4090e-02
7	4.6495e-02	1.2463e-02	1.2664e-02	1.3120e-02
8	4.7449e-02	1.2675e-02	1.2972e-02	1.3639e-02
9	4.8975e-02	1.2819e-02	1.3107e-02	1.4816e-02
10	4.8931e-02	1.2969e-02	1.3512e-02	1.5504e-02

Figure 2 shows how training error E and CV error E_{CV} changed through BCW1.2 learning for a certain segment under the model of $J = 2$. The bidirectional clustering was repeated twice until convergence. Training error E changed monotonically, while E_{CV} was minimized at $G=4$, which is reasonable since a smaller G causes insufficient capability and a larger G results in overfitting.

Table 2 shows the results of rule restoring. Cross-validation error E_{VQCV} was minimized when $I = 3$, indicating the optimal number of rules is 3. Since all the nominal variables are ignored for $I = 1$, we can see that nominal variables played a key role in improving generalization performance. Moreover, when we compare two tables, we find the final rule set has better generalization than the best perceptron. We pruned the near-zero ($|u| < 0.01$) common weight and re-trained to get the final common weights: $u_1=1.0480$, $u_2=0.4889$, $u_3=1.5105$, and

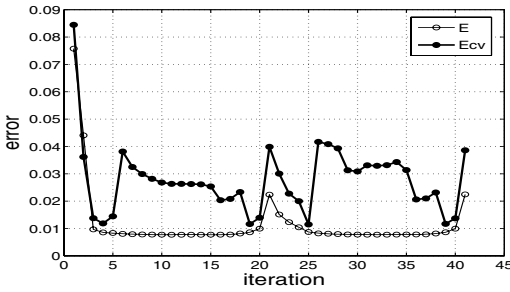


Table 2. Rule Set Comparison for Artificial Data ($J = 2$)

models	EvQCV
$I = 1$	0.82878
$I = 2$	0.13164
$I = 3$	0.01165
$I = 4$	0.01221
$I = 5$	0.01221
$I = 6$	0.01248

Fig. 2. Bidirectional Clustering for Artificial Data

$u_4=0$. By applying the C4.5 program, the following rule set is straightforwardly obtained. We can see that the rules almost equivalent to the original were found.

$$\begin{cases}
 \text{if } q_{11} \wedge q_{21} \\
 \text{then } y = 2.003 + 1.976x_1^{1.048}x_2^{0.489}x_3^{0.489} + 3.060x_3^{1.511}x_4^{1.048}x_5^{1.048} \\
 \text{if } q_{12} = 1 \\
 \text{then } y = 3.010 + 0.962x_1^{1.048}x_2^{0.489}x_3^{0.489} + 0.970x_3^{1.511}x_4^{1.048}x_5^{1.048} \\
 \text{else } y = 4.017 + 4.045x_1^{1.048}x_2^{0.489}x_3^{0.489} + 1.965x_3^{1.511}x_4^{1.048}x_5^{1.048}
 \end{cases} \quad (12)$$

PBC Data Set. The PBC data set ² contains data on the trials in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984, where survival time is to be predicted. It has 10 numeric and 7 nominal explanatory variables and one dependent variable (survival time). Numeric variables x_1, \dots, x_{10} correspond to Z_2, Z_8, \dots, Z_{16} , and nominal variables q_1, \dots, q_7 correspond to $Z_1, Z_3, \dots, Z_7, Z_{17}$, where Z_k is defined in the description of data. The numbers of categories are $L_1 = L_2 = L_3 = L_4 = L_5 = 2$, and $L_6 = L_7 = 3$. Before the analysis, the variables were normalized as follows: $\tilde{y} = (y - \text{mean}(y))/\text{std}(y)$, $\ln \tilde{x}_k = (\ln x_k - \text{mean}(\ln x_k))/\text{std}(\ln x_k)$. We did a trial with $N = 110$, $R = 3$,

Table 3. E_{LOO} of RF6.4 + BCW1.2 for PBC Data

G	J = 1	J = 2	J = 3	J = 4
2	8.4420e+05	5.1295e+05	6.2619e+05	7.4895e+05
3	9.1566e+05	4.4733e+05	1.0712e+06	8.5356e+05
4	1.1695e+06	4.6286e+05	1.0034e+06	8.3001e+05
5	2.1965e+06	4.6778e+05	1.0899e+06	1.1724e+06
6	3.1032e+06	7.5186e+05	1.1889e+06	1.2824e+06
7	7.7539e+06	1.4735e+06	1.3250e+06	1.6739e+06
8	2.2190e+07	1.2533e+07	2.0442e+06	2.8298e+06
9	2.4465e+07	1.5570e+07	4.4900e+06	3.0468e+06
10	4.1510e+07	1.5908e+07	6.0462e+06	4.3691e+06

² The data set was taken from the database of Waikato University, New Zealand.

$J = 1, \dots, 4$. For model selection we employ the leave-one-out. Weight sharing was applied only to weights w_{jk} , and we set $h = 8$.

Table 3 shows leave-one-out error E_{LOO} of RF6.4 + BCW1.2 for the PBC data. E_{LOO} was minimized at $J = 2$ and $G = 3$. Figure 3 shows how E and E_{LOO} changed through the BCW1.2 learning for a certain segment under the model of $J = 2$. The bidirectional clustering was repeated three times until

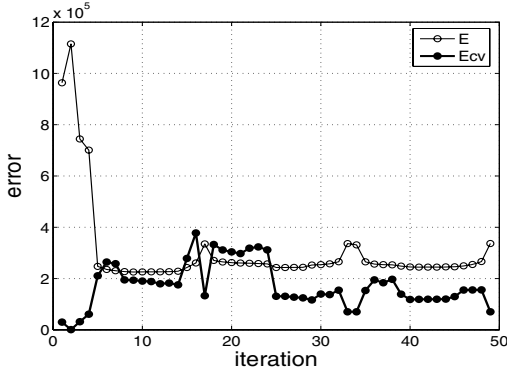


Fig. 3. Bidirectional Clustering for PBC Data

Table 4. Rule Set Comparison for PBC Data ($J = 2, G = 3$)

models	E_{vQcv}
I = 1	8.9106e+06
I = 2	1.0729e+06
I = 3	8.3613e+05
I = 4	7.9142e+05
I = 5	4.3684e+05
I = 6	4.1275e+05
I = 7	4.1374e+05
I = 8	4.2011e+05
I = 9	4.2119e+05
I = 10	4.1564e+05

Table 5. The Final Rules for PBC Data

class	nominal variables							polynomial
	q1	q2	q3	q4	q5	q6	q7	
1	1	*	*	1	1	1	2,3	$\tilde{f} = 0.5382 + 1.7327 \times \tilde{x}_7^{u_1} \tilde{x}_8^{u_3} \tilde{x}_{10}^{u_1} - 1.7251 \times \tilde{x}_2^{u_3} \tilde{x}_8^{u_3}$
	1	2	*	2	1	1	3	
	1	1	*	2	1	1	2	
	1	*	*	1	2	1	2	
	1	1	*	*	*	*	1	
	2	*	2	*	*	*	1	
2	2	2	*	1	1	1	1,2	$\tilde{f} = 1.5977 + 0.5981 \times \tilde{x}_7^{u_1} \tilde{x}_8^{u_3} \tilde{x}_{10}^{u_1} - 1.7300 \times \tilde{x}_2^{u_3} \tilde{x}_8^{u_3}$
	2	2	*	2	1	1	1	
3	1	*	*	2	2	1	2	$\tilde{f} = 2.0980 + 4.1885 \times \tilde{x}_7^{u_1} \tilde{x}_8^{u_3} \tilde{x}_{10}^{u_1} - 4.4901 \times \tilde{x}_2^{u_3} \tilde{x}_8^{u_3}$
	1	2	*	*	*	*	1	
4	1	*	*	*	*	2,3	2,3	$\tilde{f} = -0.5927 - 0.0279 \times \tilde{x}_7^{u_1} \tilde{x}_8^{u_3} \tilde{x}_{10}^{u_1} + 0.0044 \times \tilde{x}_2^{u_3} \tilde{x}_8^{u_3}$
	1	2	*	2	1	1	2	
	1	1	*	2	1	1	3	
	1	*	*	*	2	1	3	
	2	*	*	*	*	2,3	*	
	2	2	*	2	1	1	2	
5	2	*	*	*	2	1	1,2	$\tilde{f} = 4.1600 - 0.4422 \times \tilde{x}_7^{u_1} \tilde{x}_8^{u_3} \tilde{x}_{10}^{u_1} - 3.5448 \times \tilde{x}_2^{u_3} \tilde{x}_8^{u_3}$
	2	1	*	*	1	1	1,2	
	2	*	1	*	1	1	3	
	2	1	1	*	2	1	3	
6	2	2	1	*	1	1	3	$\tilde{f} = 1.6839 - 0.0422 \times \tilde{x}_7^{u_1} \tilde{x}_8^{u_3} \tilde{x}_{10}^{u_1} - 1.4975 \times \tilde{x}_2^{u_3} \tilde{x}_8^{u_3}$

Table 6. Performance Comparison with Other Methods for PBC Data

method	E_{LOO}
RF6.4 with weight decay ($J = 2, \lambda = 100$)	1.2389e+06
RF6.4 + BCW1.2 (perceptron: $J = 2, G = 3$)	4.4733e+05
RF6.4 + BCW1.2 (rules: $J = 2, G = 3, I = 6$)	4.1275e+05
Multiple regression	9.4969e+05
Quantification theory (Type 1)	1.0413e+06

convergence. Table 4 shows the results of rule restoring. Cross-validation error E_{VQCV} was minimized when $I = 6$, indicating that the optimal number of rules is 6. We pruned the near-zero ($|u| < 0.01$) common weight and retrained to get the final common weights: $u_1=0.6709$, $u_2=0$, and $u_3=0.2317$. Table 5 shows the final rules obtained in respect of normalized variables. Table 6 compares E_{LOO} of RF6.4 + BCW1.2 with those of other methods. When we combine the BCW1.2 with the RF6.4, the best performance is obtained.

5 Conclusion

We combine the weight sharing method BCW1.2 with the piecewise polynomial regression method RF6.4. Our experiments showed the proposed method worked well. In the future we will do further experiments to evaluate the method.

References

1. S. Haykin. *Neural Networks, 2nd Edition*. Prentice-Hall, 1999.
2. Y. Tanahashi, K. Saito and R. Nakano. Piecewise multivariate polynomials using a four-layer perceptron. *Proc. KES'04, LNAI 3214*, pp.602-608, 2004.
3. S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Information Theory*, IT-28(2):129-137, 1982.
4. K. Saito and R. Nakano. Structuring neural networks through bidirectional clustering of weights. In *Proc. 5th Int. Conf. on Discovery Science*, pp. 206-219, 2002.
5. Y. Tanahashi, K. Saito and R. Nakano. Model selection and weight sharing of multi-layer perceptron. In *Proc. KES'05, LNAI 3684*, pp.716-722, 2005.
6. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

Development of Know-How Information Sharing System in Care Planning Processes – Mapping New Care Plan into Two-Dimensional Document Space

Kaoru Eto¹, Tatsunori Matsui², and Yasuo Kabasawa¹

¹ Nippon Institute of Technology, Faculty of Engineering, Miyashiro Gakuendai 4-1, Saitama 345-8501, Japan

{eto, kabasawa}@nit.ac.jp

² Waseda University, Faculty of Human Sciences, Tokorozawa Mikajima 2-579-15, Saitama 359-1192, Japan

matsui-t@waseda.jp

Abstract. The purpose of this study is to develop a computer support system for educating personnel who are involved in care management. We wish to develop a system in which know-how information can be shared. We consider that visualizing and showing care plans drawn up by experts in various forms allows a beginner to see the differences between their plans and an expert plan. Sharing know-how information is possible by recording, accumulating, and giving titles to what has been noticed in comparing documents. This function can visualize the similarities among documents that interpreted the results of an assessment, and can flexibly change different viewpoints. In order to promote user awareness, we mapped a user's new document into a two-dimensional document space, and confirmed that the results of this mapping were appropriate.

1 Introduction

Since April 2000, with the establishment of the Long-Term Care Insurance System in Japan, duty-of-care planning has been imposed. It is aimed at improving the quality of long-term care services. However, since there are many simplified and stereotyped plans, it has become apparent that clients' true care needs are not being extracted. The main factor in this issue is a lack of care managers' ability. In order to solve problems stemming from care managers' insufficient abilities, it is urgent that we establish a method of care manager training that involves care management. There are four important processes in care management. This system supports the process in which the result of an assessment is interpreted.

A skillful care manager not only has a good wealth of knowledge, but also can use specific know-how information depending on the situation. We cannot handle know-how information in the same way as other kinds of knowledge. As a method, we decided to assist a beginner in noticing weaknesses in their care plans

by visualizing and showing a care plan drawn up by an expert in various forms. Differences between a beginner's plan and an expert's appear most typically in their different viewpoints to interpret the results of an assessment. Making these differences noticeable is one method of extracting know-how information. As a function, the similarities between documents, which interpreted the assessment results, are visualized as are the results that change similarities using a flexibly changing viewpoint. Differences are made noticeable using this function. In this paper, in order to promote user awareness, we visualized similarities among a user's document and an expert's document, and achieved the function of flexibly changing viewpoints. We also confirmed that these results were appropriate.

2 Know-How Information

Know-how information is tacit knowledge. It is said that there are two kinds of human knowledge: explicit knowledge, which is knowledge expressed verbally; and tacit knowledge, which is knowledge expressed non-verbally. A lot of knowledge is tacit knowledge and this is considered to be a very important element in group behavior.

2.1 The Definition of the Know-How Information

Know-how information has mainly been studied as part of knowledge management in the field of business administration [1]. A lot of research has focused on administration jobs. For example, focusing on knowledge that an office worker acquires when performing a task [2], and informal information generated during work processes [3]. Such know-how information is contained within routine work procedures, and it can be easily extracted. Recently, research on knowledge management has also been done in the field of medical treatment and welfare [4],[5].

In this study, we define know-how information as information that assists in understanding a client's daily life. This know-how information includes heuristic information, such as cases experienced in the past, sequences for interpreting the results of an assessment, and how to decide on a viewpoint. To define it concretely, it is information that grasps the relevance among assessment items, especially information that shows the strength of that relevance.

2.2 The Method and Function of Know-How Information Extraction

Documents are a mix of general knowledge, which refers to explicit knowledge such as theories and rules, and specific knowledge, which refers to tacit knowledge based on experiences such as original viewpoints, original patterned knowledge, and conceptualization. The extraction of know-how information involves separating specific knowledge and general knowledge.

3 Support Functions

We consider the support functions in which a viewpoint can be flexibly changed and the results can be visualized. We extract know-how information using these functions.

3.1 Support Functions

The method for making general knowledge and specific knowledge separate involves creating differences between a beginner's and an expert's documents using repetitions that involve things such as operations, classifying the document for every viewpoint change using a concept-base, and visualizing and presenting the results. Furthermore, our method uses a KOMI(Kanai Original Modern Instrument) chart, "graphical recording sheet", that represents the origins of a document. That is, both qualitative data (the document) and quantitative data (the KOMI chart) are shown, and differences are highlighted by fusing together both data sets. We believe that these differences create a trigger for separation. This is stated in detail below.

- (1) A user refers to the statistic values of an original idea or a KOMI chart, and then changes the viewpoints.
- (2) Based on these viewpoint changes, the user moves vertically and horizontally in the hierarchy of a concept-base, and calculates the degree of similarity each time.
- (3) The system then classifies the document from the calculation results.
- (4) Next, the system visualizes the classification results in a two-dimensional document space displayed on the computer.
- (5) By clicking on the document number according to its classification, the KOMI chart that originates the document is shown.
- (6) Finally, the user records, accumulates, and gives a title to what has been noticed.

By seeing many documents of the high expert of similarity, the difference between oneself and other persons can be seen and know-how information is accumulated. We think that these differences separate the specific knowledge and general knowledge that are used as the basis of the documents. We consider this separated specific knowledge to be know-how information. This separation process is considered to be a way of extracting know-how information. This idea is shown in Fig.1.

4 A Concept-Base and Changing Viewpoints

A concept-base is a knowledge-base that expresses concepts with its own and other conceptual sets. In this paper, we constructed a concept-base that is specialized for the field of nursing and welfare. This concept-base has a tree structure with six levels and constructed by a thesaurus [6]. Viewpoints are changed by moving up and down the levels of this concept-base.

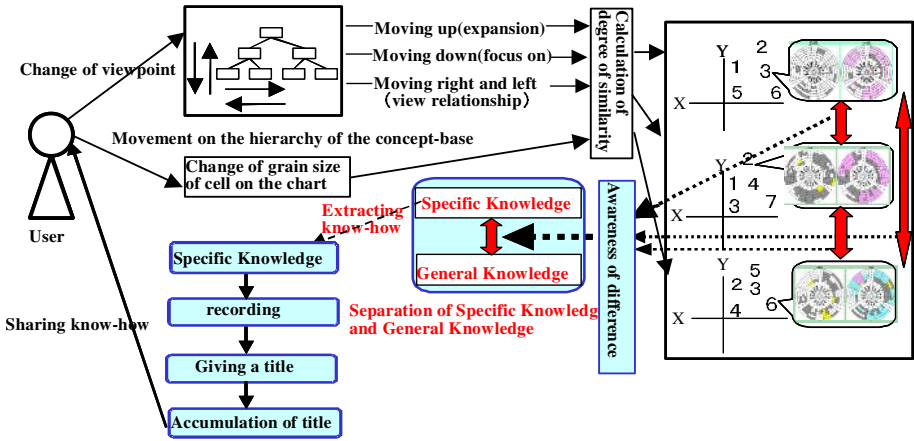


Fig. 1. The model for extracting and sharing the know-how information

4.1 Construction of Concept-Base

(1) Extracting concepts

The concept-base contains about 4300 terms that consist of keywords extracted from the documents that includes the KOMI chart, the assessment items, and the textbook for the KOMI chart.

(2) Encoding a concept

The extracted keywords were encoded based on a Japanese language thesaurus [7]. Keywords that did not exist in the thesaurus, such as the names of diseases, were encoded in our own system [8]. The code is used to identify the position

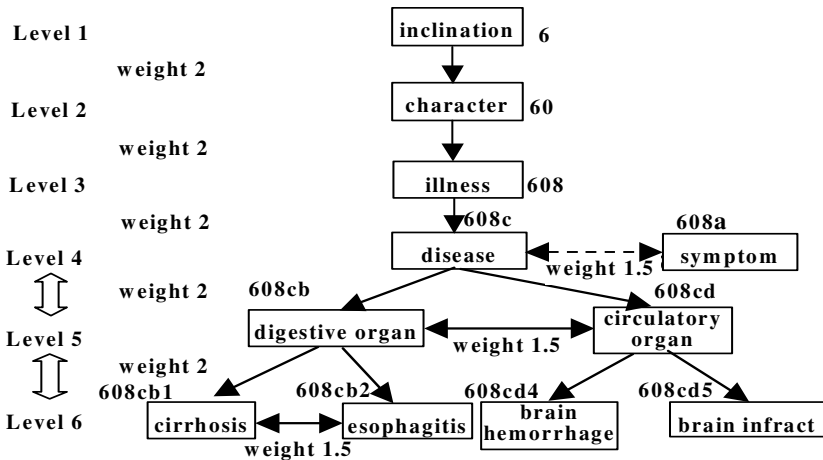


Fig. 2. Hierarchic structure of concept-base and encoding of a concept

on the hierarchy of the concept-base. Especially the length of a digit means the level on a hierarchy. A part of concept-base is shown in Fig.2.

4.2 Changing Viewpoints

Changing viewpoints refers to expanding and narrowing viewpoints to see the degrees of relevance. Whenever a viewpoint is changed, the distance between documents is calculated, and the difference is made noticeable by visualizing the results. This function supports the extraction of differences in the viewpoints of an expert and a beginner. In this paper, we achieved this function by moving the level of the concept-base. Changing viewpoints requires five kinds of selection section, and three kinds of movement on the concept-base; there are 15 ways in total to do this.

The hierarchy of the concept-base is movable on three levels, from level 5 and level 4 one-by-one from the six lowest levels. As shown in Fig. 2, level 6 is the most concrete concept. If the level is moved up, the degree of abstraction becomes large.

5 Calculation of the Degrees of Similarity and Visualization

The system extracts keywords from each document. The distance between keywords is determined by the nearness of the concept between keywords. We consider sibling relationships (adjacent keywords on the same level) in the thesaurus to be nearer than child-parent relationships (hierarchical relationships).

5.1 Similarities Between Documents

The distance between keywords is defined as the sum of the weight of the branch of the node that reaches other keywords from one keyword. The weight of the branch of the node of a parent-child relationship in the tree is set to 2, and the sibling relationship is set to 1.5. Therefore, the distance between keywords changes when the viewpoint changes.

The distance between documents is determined by the distance between the keyword sets. Let the viewpoint on concept-base be L . The distance between keyword a and b in the viewpoint L is expressed as $d(a, b|L)$. Let keyword set be A, B , The distance between $\forall a \in A$ and set B in viewpoint L is defined by

$$D_w(a, B|L) = \min_b \{d(a, b|L)\}. \quad (1)$$

The distance from keyword set A to keyword set B in the viewpoint L is defined by

$$D(A \rightarrow B|L) = \frac{1}{|A|} \sum_{a \in A} D_w(a, B|L), \quad (2)$$

where the symbol $|A|$ stands for the elements of the set A . The distance between keyword set A and B on the viewpoint L is defined by

$$D(A, B|L) = D(B, A|L) = \max\{D(A \rightarrow B|L), D(B \rightarrow A|L)\}. \quad (3)$$

5.2 Visualizing Similarities

Even if the distance between keyword sets (document) is obtained as a numeric value, it is difficult to intuitively see the difference between many documents. We considered mapping each document as a point into the display of a personal computer. In practice, we tried to map documents on a plane (a two-dimensional space) using Kruskal's method [9], which is specifically aimed at multidimensional scaling.

6 Mapping the Document of a New Plan into Two-Dimensions

Expert's documents mapped into two-dimension space by Kruskal's method previously. Beginner's new document map in same space. We expected this function to highlight the effects that promoted an awareness of the differences between a user and an expert.

6.1 Calculations for Mapping New Case into Two-Dimensions

We wish to determine the two-dimension coordinate for a new case. The step is shown below.

- (1) Calculate the distace between expert' documents and beginner's new document.
- (2) Choose three cases that is the nearest to the new case.
- (3) Determine the coordinate of new case by distances and coordinates of three cases. For example,when we determine the coordinate $P(x, y)$, which is at a distance of l_1, l_2, l_3 , from the three points $P_1(x_1, y_1), P_2(x_2, y_2)$, and $P_3(x_3, y_3)$ on the plane. However, generally there is no point P on this plane. In Kruskal's method,the space is expanded and/or contracted locally. When a new case is mapped, the degree of expansion and/or contraction is assumed to be uniform in small area. Therefore, we can be determined the point of this plane by scaling the three distances with the same coefficient k .

6.2 Discussion

The upper left of Fig.3 shows the results of having added the new case to the two-dimensional space based on the similarities among all documents for 35 cases. Each expert' document is identified by a number. Three cases that were the nearest to the new case from the similarity calculation results were No. 33, 4, and 29. The two-dimensional position coordinates obtained from the calculation method described above were not mapped on the position that was necessarily nearest to the three examples. However, new cases were mapped into an appropriate range. These results were based on the properties of the multidimensional scaling, By

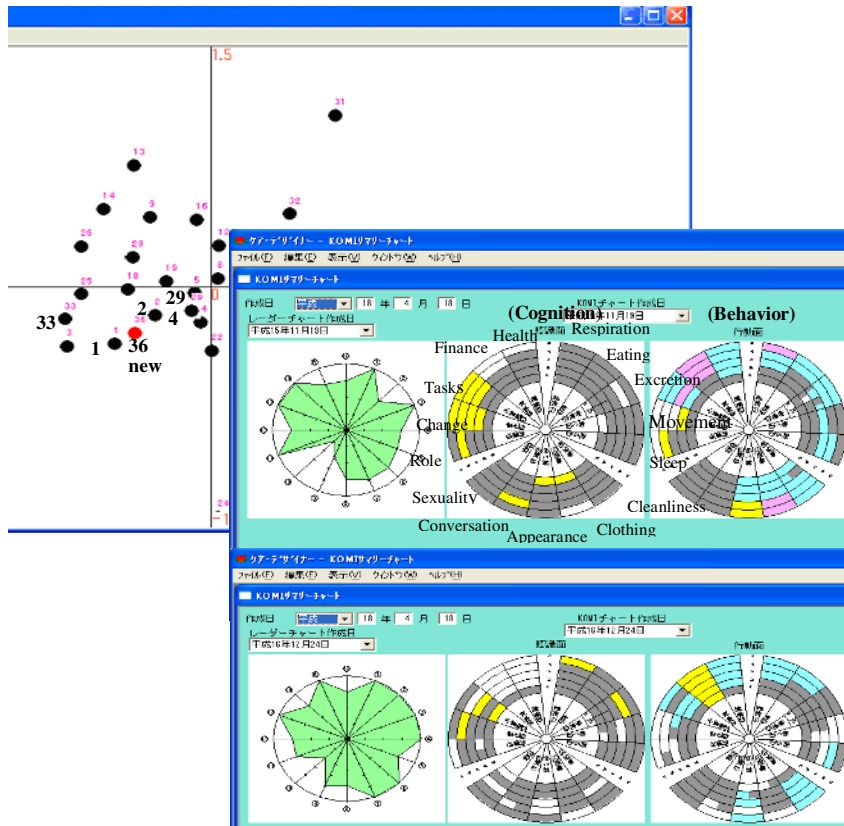


Fig. 3. Mapping into two-dimensions added the new case and displaying of the chart of cases near to a new case

increasing the number of cases, the similarity calculation results and the results of mapping into two-dimensions became closer. Two cases, "No.1, No.2" were near to the new case. The clients in three cases suffered from dementia, cerebral palsy, or senility and could not move freely, there was a common feature that all these clients used a wheelchair. The results of analyzing the keywords in the new case and the keywords in the existing two cases are as follows: (1) The same keywords found in the new case were found on two occasions in case 1, and on five occasions in case 2 ; and (2) Keywords sharing a sibling relationship were found on four occasions in case 1 and on five occasions in case 2. From these results, we concluded that two-dimensional mapping was appropriate. It's possible to display the chart of a case that is near to a new case, as in the lower right of Fig.3. Using this function, by seeing the original chart, a user can understand why the new document has been mapped into that particular position. When a user compares an expert's chart with his/her own chart, the similarities and

differences are clearly visible. If the user's chart is similar, it is appropriate, and if it differs, the user can make the necessary adjustments to improve his/her document.

7 Conclusion

We have developed the system to make a beginner notice the difference between beginner's care plan and an expert's by visualizing a care plan with various forms that were drawn up by an expert. As a function, the similarities among documents that interpret the results of an assessment were calculated using a concept-base, and the results were then mapped into two-dimensions. Furthermore, to promote user awareness, we created a function that shows the position of the user's new plan in relation to an expert's plan group in a two-dimensional document space. We confirmed that the results of mapping these documents into two-dimensions was appropriate.

We expect that this function will affect the way in which users rearrange their documents in relation to how experts document the same kinds of cases in a two-dimensional document space.

References

1. Nonaka I., Takeuchi H.: *The Knowledge Creating Company*. Oxford University Press, New York (1995)
2. Seki Y., Yamakami T., Shimizu A.: FISH: Information Sharing and Handling System. *Trans. IEICE*, Vol. J76D-U.No.6, (1993) 1223–1231 (in Japanese)
3. Shikida M., Kadowaki C., Kunifuji S.: A Proposal of an Informal Information Sharing Method by Linking to Flow in Group Work Processes. *Recent Trends and Developments. J.IPSJ*, Vol. 41. No.10, (2000) 2731–2741 (in Japanese)
4. Umemoto K., Kanno M., Moriwaki K., Kamata G.: *Knowledge Management of the field of Medical treatment and Welfare*. Nissoken, Nagoya (2003) (in Japanese)
5. Special Edition: *Knowledge Management of Nursing administration*. J. Nursing Management Tokyo (2002) (in Japanese)
6. Eto K., Kabasawa Y., Matsui T., Okamoto T.: *Development of Know-how Information Sharing System Using KOMI Chart in Care Planning Processes. -Classification of Document as a Results of Assessment- Technical Report IEICE, ET2003-70*, (2003) 59–64 (in Japanese)
7. Ono S., Hamanishi M.: *Thesaurus for Japanese Language* Kadokawa, Tokyo (1985)
8. *Japan Medical Abstracts Society: Thesaurus for Medical and Health related Terms* Tokyo (2002) (in Japanese)
9. Okada A., Imaizumi T.: *Multidimensional scaling using Personal Computer* Kyoritsu Shuppan, Tokyo (1994) (in Japanese)

Educational Evaluation of Intelligent and Creative Ability with Computer Games: A Case Study for e-Learning

Yukuo Isomoto

Graduate School of Economics and Information, Gifu Shotoku Gakuen University
1-38, Nakauzura, Gifu City, Gifu Prefecture, Japan, 500-8288
yisomoto@gifu.shotoku.ac.jp

Abstract. In information ages, a computer is our active intelligent collaborator to work and learn something, so that learner's intelligent ability must be taken into consideration even in education of information technologies (IT). Aim of this paper is to propose quantitative evaluation of the intelligent ability in computer usage. The author discusses some elements in the intelligent ability (such as quickness, logics, accuracy, memory, discernment, patience, decision, and inference), and formulates quantitative evaluation method of the elements with computer games. A tutor of IT education will be able to apply result of the research to computer assisted education for supporting learner's intelligent activities.

1 Introduction

In the present information society, advanced information technologies are drastically enhancing the quality of human intelligent ability in collaboration with a computer. Actually, we often realize that a computer is a powerful assistant in human intelligent activities in learning or working. The more we understand our intelligent ability by ourselves, the more efficiently we can work actively our intelligence. Really for instance in education, if a tutor has enough information about learner's intelligent ability, he/she can effectively assist a learner. Aim of this paper is to discuss quantitative evaluation method of the intelligent ability with computer games to understand the feature of intelligent ability.

Game playing is intellectually so attractive and popular that computer games have been applied to computer assisted training software for working or learning [1],[2],[3],[4],[5],[6]. In this paper, the author also designs quantitative evaluation method of intelligent ability with the use of computer games, that get the action data of player's intellectual activities. We observe the player's challenge to the computer games, and analyze his/her intellectual action data. In chapter 2, we survey briefly the intelligent ability and computer games. In chapter 3, we formulate statistical analyses of quantitative intelligent ability. Chapter 4 concludes our discussion.

2 Intelligent Ability on Computer Games

When a player challenges a computer game, he/she may use even unconsciously his/her intelligent ability within the rules of individual games. To analyze details of

the intelligent ability, the author classifies intelligent ability into eight elements, and moreover arranges ten computer games as a testing condition, in which the player’s actions are observed to get the action data related to intelligent ability.

2.1 Elements of Intelligent Ability

Until now, even though intelligent ability has been analyzed with Kraepelin’s intelligence test, the analyzed results are not suitable to know a computer user’s intelligent ability. In this paper for concrete modeling, the author supposes that a computer user works eight elements in the intelligent ability as follows (see Table 1);

- (1) Quickness, (2) Logics, (3) Accuracy, (4) Memory,
- (5) Discernment, (6) Patience, (7) Decision, (8) Inference.

A dictionary says “Intelligent” as follows;

- 1. Having a high level of ability to learn, understand, communicate, and think about things, or showing this ability.

And also the word “ability” means;

- 1. Something that you are able to do, especially because you have a particular mental or physical skill.

Table 1. Elements of intelligent ability

Elements	Meanings of the words (referred to a dictionary)
(1) Quickness	1. Fast, or done in a very short amount of time. 2. Able to learn and understand things fast. 3. A repair to something or answer to a problem that happens quickly for a short time.
(2) Logics	1. A set of sensible and correct reasons, or reasonable thinking. 2. The science or study of careful REASONING using formal method.
(3) Accuracy	1. The ability to do something in an exact way without making a mistake. 2. The quality of being correct or true.
(4) Memory	1. Ability to remember things, places, experiences etc. 2. Something that you remember from the past about a person, place, or experience.
(5) Discernment	1. The ability to make good judgments about people, style and things
(6) Patience	1. The ability to wait calmly, accept delays, or continue doing something difficult for a long time, without becoming angry or anxious. 2. The ability to accept trouble and other people’s annoying behavior without complaining or becoming angry.
(7) Decision	1. The quality of trying to do something even when it is difficult. 2. The act of deciding something officially.
(8) Inference	1. Something that you think is probably true, based on information that you already know. 2. The act of inferring something.

2. Someone's, especially a student's, level of intelligence or skill, especially in school or college work.
3. To do something as well as you can.

The author arranges computer games to evaluate the intelligent ability in accordance with these explanations.

2.2 Computer Games for Intelligence Test

The arranged computer games in Table 2 are designed to get player's action data, which consist of the n -th player's m -th challenge to the i -th game as follows:

p_{mni} ; positive point (or success point), f_{mni} ; negative point (or fault point),
 l_{mni} ; difficulty level, t_{mni} ; playing time or thinking time.

Table 2. Games referred in this paper

Name of Games	Rules of the Score	Evaluated Elements
(1) Polygon Game	When you put correctly polygons as closely as you can, you get the better score.	(3) Accuracy (5) Discernment
(2) Labyrinth Game	A popular game [1],[2],[3]. As you arrive at the exit in shorter time, you get the better score	(5) Discernment (8) Inference
(3) Life game	Conway's life game [4],[5]. Cells appear or disappear according to population density.	(2) Logics (8) Inference
(4) L-System	Lindenmayer's system [6]. The more blossoms you generate, the more score you can get.	(2) Logics
(5) Bounding Ball	The more a cannon ball you shoot, bounds on wall, ceiling, and pole, the more your score is.	(2) Logics (3) Accuracy
(6) Capture Balls	You get score when you can capture two small balls within a big ball by collision.	(1) Quickness (8) Inference
(7) Escaping Game	Various sizes of bars are falling down. You get the better score, when you prevent a ball from crashing with a bar as longer time as you can.	(1) Quickness (7) Decision
(8) Billiards	You knock a white ball, and balls successively collide for 60 seconds. The number of dropped balls is your point.	(3) Accuracy (8) Inference
(9) Passing through Game	You hit sequentially balls, which go through moving shields. The more points you can hit accurately, the more score you get.	(1)Quickness (5)Discernment
(10) Spot Number Game	You memorize figures on cards. The more you remember figures on cards in correct series, the better score you get	(4) Memory

For common basis among the games, we define an “achievement score s_{mni} ”:

$$s_{mni} = F(p_{mni}, F_{mni}, I_{mni}, t_{mni}) \tag{1}$$

Eq.(1) is formulated in accordance with the individual games (see Appendix).

2.3 Player’s Action and Intelligent Ability

The author proposes empirically quantitative relation a_{ik} (see Table 3) among the elements of intelligent ability (at the top line) and the score in computer games (at the left column). For instance, “Polygon game” is mainly related to the ability of “Discernment” and “Accuracy”. And also “Spot Number game” is mainly related to the ability of “Memory”. Values of a_{ik} are given within the interval $0 \leq a_{ik} \leq 5$ (“0” means unrelated and “5” does the closest relation). Through checking its adaptability to our experience, the value of a_{ik} may be modified.

3 Evaluation of Intelligent Ability

A game player uses his/her intelligent ability to enjoy solving the problems encountered in various phases of a game. By analysis of the player’s score, we can evaluate the player’s intelligent ability. In this chapter, the quantitative evaluation method of the elements of intelligent ability is formulated.

Table 3. Relation of the i-th game to the k-th element of intelligent ability: $0 \leq a_{ik} \leq 5$

	k-th Elements of intelligent ability							
	(1) Quick-ness	(2) Log-ics	(3) Accur-acy	(4) Mem-ory	(5) Discern-ment	(6) Pa-tience	(7) Deci-sion	(8) Infer-ence
(1) Polygon	0	0	5	0	5	0	0	2
(2) Labyrinth	0	0	0	3	5	0	0	5
(3) Life game	0	5	0	0	0	0	0	5
(4) L-System	0	5	0	0	0	0	0	0
(5) Bounding	0	5	5	0	0	0	0	0
(6) Capture	5	0	3	0	0	2	0	5
(7) Escaping	5	0	0	0	0	4	5	0
(8) Billiards	0	0	5	0	0	0	0	5
(9) Passing thr	5	0	2	0	5	0	0	0
(10) Spot No.	0	0	0	5	2	0	0	0

3.1 Statistical Analyses for the Evaluation

In general, the player’s action data are not always stable but rather fluctuate by player’s condition and/or phases of a computer game. Therefore, we average his/her “achievement score” to get stable factors of the action data as follows:

$$S_{ni} = \frac{\sum_{m=1}^M S_{mni}}{M} \tag{2}$$

M is the frequency of n-th player's challenge. The S_{ni} has the information about the n-th player's action data in his/her challenge to the i-th game. Analysis of the S_{ni} gives us the player's intelligent ability.

The S_{ni} ($n=1,2,\dots,N$) is different from player to player. To understand the difference among players, the author estimates S_{ni} in a statistical method. Then the average of the S_{ni} among the players $n=1 \sim N$ is estimated as follows:

$$S_{0i} = \frac{\sum_{n=1}^N S_{ni}}{N} \quad (3)$$

N is the total number of players.

Suppose that the S_{ni} forms uniform distribution, and calculate distribution D_i of the S_{ni} among the players as follows:

$$D_i = \frac{\sum_{n=1}^N (S_{ni} - S_{0i})^2}{N} \quad (4)$$

And also the standard deviation σ_i of the S_{ni} is calculated as follows:

$$\sigma_i = [D_i]^{1/2} \quad (5)$$

According to Eq.(2), (3), and (5), the n-th player's success grade G_{ni} of the i-th game is estimated as "deviation value" for the i-th game as follows,

$$G_{ni} = 50 + \frac{100 \times (S_{ni} - S_{0i})}{\sigma_i} \quad (6)$$

The G_{ni} in Eq.(6) is the n-th player's deviation value for the i-th game among the players. The G_{ni} is quantitatively related to the elements of intelligent ability through the a_{ik} (see Table 3). Then the n-th player's k-th element Q_{nk} of intelligent ability is evaluated as follows:

$$Q_{nk} = \sum_{i=1}^I G_{ni} \times a_{ik} \quad (7)$$

"I" is the number of the arranged games.

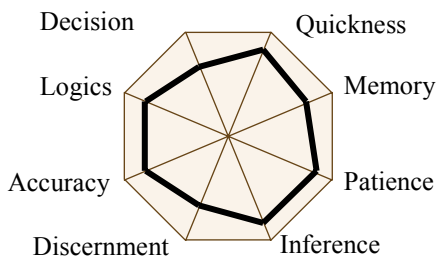


Fig. 1. Radar Chart showing the elements of intelligent ability

Fig.1 shows an example of a radar chart that visualizes the Q_{nk} for the n-th player's intelligent ability. Each radius denotes the value of Q_{nk} for k, which is plotted with the

bold line. When many persons committed to the intelligence test, the radar chart gives us the information in more reliability. From the viewpoint of e-Learning, the radar chart gives a tutor the information about learner's characteristics of intelligent ability. For instance, observing the form of the bold line, a tutor knows the balance, the strong point and the weak point of learner's intelligent ability.

3.2 Empirical Determination of "a_{ik}"

Even though we have no empirical scale to measure intelligent ability, the parameters a_{ik} theoretically correlate the G_{ni} to the Q_{nk} (see Eq.(7)). Table 3 is temporarily given in accordance with our intuition. If we have well known sample data of the G_{ni} and the Q_{nk}, we can calculate values of the a_{ik} mathematically. By analyzing enough data, the values of a_{ik} will be determined well in statistical way, and we can refine the table 1, 2, and 3.

4 Discussion and Conclusion

At the present time, we have quantitatively very few information about intelligent ability. The intelligence test proposed in this paper will give us more precise method to obtain information about attitude of intelligent ability more than the past method. The author thinks the method will give reliable attitude of intelligent ability by analyzing abundant action data, which we will be able to gather through the internet near future (URL "http://tsplaza.jp/game").

In not only education of computer science but also e-Learning world, the result of the intelligence test gives a tutor the educational information how to assist a learner in his/her intelligent and creative activity. Moreover, when you want to realize efficient collaboration with a computer, it is preferable for you to know your intelligent ability by yourself. The intelligence test proposed here is the initial stage of our research for such educational support.

Finally, the author should acknowledge the financial support of Japanese Ministry of Education, Culture, Sports and Technology to this research.

References

1. Ivailo Ivanov, Vesela Ilieva: ToolKID language based software package for children, <http://eurologo2005.oeiizk.waw.pl/PDF/E2005IvanovIlieva.pdf>
2. Avgi Valsami: THE LABYRINTH GAME, A speech application primarily intended for blind or visual impaired people, <http://www.dcs.shef.ac.uk/internet/teaching/projects/archive/msc2002/pdf/m1av.pdf>
3. The labyrinth Society: <http://www.labyrinthsociety.org/>
4. Martin Gardner : MATHEMATICAL GAMES The fantastic combinations of John Conway's new solitaire game "life", http://ddi.cs.uni-potsdam.de/HyFISCH/Produzieren/lis_projekt/proj_gamelife/ConwayScientificAmerican.htm
5. Mike Matton: A Flexible Simulation Tool for Cellular Automata. <http://www.cs.kuleuven.be/~tomdw/educqation/files/2002-2003MikeMatton.pdf>
6. Christian Jacob: Modeling Growth with L-System & Mathematica: <http://pages.cpsc.ucalgary.ca/~jacob/Publications/ModelingGrowth.mapdf>

Appendix. Achievement Score of Computer Games

Even though individual computer games (see Table 2) are designed to get player's action data; p_{nmi} , f_{nmi} , l_{nmi} , t_{nmi} in Eq.(1), meanings of the action data depend on the design and the rules of individual games (see <http://tsplaze.jp/game>). For equivalent estimation among the games, the author defines an achievement score s_{mni} ($i=1, 2, \dots, 10$) to evaluate success grade in the i -th game.

(1) Polygon Game

In the initial condition, a yellow polygon is set on the center of the panel. You choose one of the eight type polygons from the template, and put it on the panel one by one. An edge of each polygon must adjoin to one of its neighbors at least, but not overlapped. When the polygons fill the panel, the game is over.

$$s_{mn1} = (p_{mn1} - f_{mn1}) \times l_{mn1} / ((p_{mn1} + f_{mn1}) \times t_{mn1}). \quad (A1)$$

p_{mn1} : The total points of correctly placed polygons.

f_{mn1} : The total points of misplaced polygons.

l_{mn1} : Level, $1 \leq l_{mn1} = 4 / (\text{Polygon size}) \leq 4$. t_{mn1} : Playing time.

(2) Labyrinth Game [1],[2],[3]

You start from entrance, and goes along the path forward direction to the exit. When you arrive at the exit, the game is over.

$$s_{mn2} = [(p_{mn2} - p'_{mn2} - f_{mn2}) / (p_{mn2} + p'_{mn2})] \times (l_{mn2} / 5)^2 / (f_{mn2} \times t_{mn2}). \quad (A2)$$

p_{mn2} : Frequency of forward steps.

p'_{mn2} : Frequency of backward steps

f_{mn2} : Frequency of error steps.

l_{mn2} : Scale of the game, $5 \leq l_{mn2} \leq 50$.

t_{mn2} : Playing time.

(3) Life Game (Conway's Life Game) [4],[5]

Through whole processes of the game, the number of cells increases or decreases in accordance with the generating rule. We estimate the initial number, the maximum number, and the final number.

$$s_{mn3} = (t_{mn3} / t'_{mn3}) \times p_{mn3} \times (l_{mn3} / f_{mn3}). \quad (A3)$$

p_{mn3} : The maximum number of cells through the whole of propagating process..

f_{mn3} : The initial number of cells.

l_{mn3} : The final number of cells.

t_{mn3} : The last generation.

t'_{mn3} : The player's thinking time

(4) L-System (Lindenmayer's system) [6]

In L system, a tree grows up in accordance with a given generative grammar. You compose rules of a generative grammar, and a tree grows up and blooms within 5 steps. The more blossoms the tree has, the higher the score is.

$$s_{mn4} = 100 \times (p_{mn4} / l_{mn4}) / [(f_{mn4} - 4) \times (t_{mn4} + 6)]. \quad (A4)$$

p_{mn4} : The number of blossoms.

f_{mn4} : The number of conditioning characters.

l_{mn4} : Growing steps.

t_{mn4} : Player's thinking time.

(5) Bounding Ball

You fire a cannon ball to shoot the target. When the ball touches on the target, you success. If the ball falls on green floor, it is your fault. You get higher point when the ball bounds on walls, ceiling, pole, and board as many times as possible.

$$s_{mn5} = 20 \times p_{mn5} \times (1 - 0.5 \times f_{mn5}) / t_{mn5}. \tag{A5}$$

p_{mn5} : Final point. f_{mn5} : Success (0) or Fail (1).
 l_{mn5} : No meaning. t_{mn5} : player's thinking time.

(6) Capture Balls

Two small balls are moving. You move a big ball to collide it with the small balls with a mouse. When you capture two small moving balls inside of a big ball by collision, you win successfully.

$$s_{mn6} = p_{mn6} \times 600 \times (1 - 0.5 \times f_{mn6}) / (t_{mn6} + 10). \tag{A6}$$

p_{mn6} : The number of captured balls 0, 1, or 2. f_{mn6} : Clash (1).
 l_{mn6} : No meaning. t_{mn6} : Playing time.

(7) Escaping Game

Various sizes of bars are continuously falling down from ceiling, and you must move a blue ball to prevent from clashing against the bars. As a bar has arrived on the floor, you get a point. When you put the ball on the ceiling, the game is successfully over. When a ball clashes a bar, the game is over in fault.

$$s_{mn7} = [(1 - f_{mn7} \times 3/4) \times p_{mn7}] \times [(1 - f_{mn7}/2) \times f_{mn7}] \times t_{mn7} \times (l_{mn7}/10). \tag{A7}$$

p_{mn7} : Positive point. f_{mn7} : Success (0) or fault (1).
 l_{mn7} : Speed of falling bars. t_{mn7} : Playing time.

(8) Billiards

Ten colored balls are set on a green board. You knock a white ball, and it runs and collides with other balls. The balls collide continuously with each other for 1 minute.

When a ball falls down in one of holes, you get points.

$$s_{mn8} = (p_{mn8} / (p_{mn8} + f_{mn8})) \times (10 / (t_{mn8} + 9)). \tag{A8}$$

p_{mn8} : The number of fallen balls. f_{mn8} : The number of left balls.
 l_{mn8} : No meaning. t_{mn8} : Player's thinking time.

(9) Passing Through Game

Shields are moving up and down on the right hand side. On the left hand side, blue balls appear and you hit successively the balls, which go through gaps of the moving shields. When a ball goes through the moving shields, you get a point. When a ball crashes the shield, it is your fault.

$$s_{mn9} = 20 \times p_{mn9} / ((p_{mn9} + f_{mn9}) \times t_{mn9}). \tag{A9}$$

p_{mn9} : The number of passing balls. f_{mn9} : The number of crashing balls.
 l_{mn9} : No meaning. t_{mn9} : Playing time (1 minute).

(10) Spot Number Game

You memorize all figures on displayed cards. As you click start button, the numbers are erased. You click correctly the cards in serial way corresponding to the figures, and get points.

$$s_{mn10} = 250 \times [(5 \times p_{mn10} - f_{mn10}) / (p_{mn10} + f_{mn10})] \times l_{mn10} / (5 \times t_{mn10} + t'_{mn10}).$$

(A10)

p_{mn10} : The number of success.

f_{mn10} : The number of mistake.

l_{mn10} : The total number of cards.

t_{mn10} : Memorizing time.

t'_{mn10} : Playing time.

The Supporting System for Distant Learning Students of Shinshu University

Hisayoshi Kunimune¹, Mika Ushiro², Masaaki Niimura¹, and Yasushi Fuwa²

¹ Faculty of Engineering, Shinshu University

² Graduate School of Science and Technology, Shinshu University,
4-17-1 Wakasato, Nagano, 380-8553 Japan

{kunimune, ushiro, niimura, fuwa}@cs.shinshu-u.ac.jp

Abstract. E-Learning gives the opportunity to students to learn at any time and any place. Over 250 students have enrolled in Shinshu University Graduate School on the Internet (SUGSI), from which almost 90% are working professionals. Working professionals are usually highly motivated when they enroll in the distance learning program. However, an important problem is that their motivation might decrease and learning could stagnate. Currently, we are implementing a supporting program which can rapidly find out learners' decreased motivation. In this paper, we describe the distant student support system we are deploying and give some initial results on its evaluation.

1 Introduction

Students can learn at any time and any place by using self-paced e-Learning courses. Working professionals are especially utilizing e-Learning courses as distant students for recurrent education.

We established SUGSI in 2002 for recurrent education of working professionals. We have developed 20 self-paced e-Learning course materials for SUGSI and various learning support systems, such as drill examinations, on-line discussions, and so on[1][2][3]. Since its creation, over 250 students have enrolled in SUGSI.

The development and continuous improvement of course materials and learning support systems have been important activities in order to provide an appropriate educational environment for distant learners. However, we think that a more detailed support for individual students is needed to enrich even more our e-Learning program.

Working professionals are usually highly motivated when they enroll in the distance learning program. However, an important problem is that their motivation might decrease and learning could stagnate. This usually happens, for example, when the distant students become busier than expected at their working place or there are concepts and lessons difficult to understand by themselves.

In order to give a better e-learning environment for our distant students and increase their chances of success, we are trying to establish a distant student support system which can rapidly find out learners' decreased motivation in SUGSI. In this paper, we describe the distant student support system we are deploying and give some initial results on its evaluation.

2 The Learning Courses of SUGSI

Currently, SUGSI offers 20 lectures on the Internet. Lecturers can create materials of their courses using multimedia. We provide several types of web-based examinations in every section of a course text based on the contents of the section. For instance, we offer drill type examinations and a system for the submissions of electronic reports.

In drill type examinations, the drill system randomly selects questions from a pool of questions and students can repeatedly take the examination. Hereby, students acquire knowledge from these drills[2][3].

All web-based examination systems including report systems have a database to record students' progress. When a student passes an examination, the examination system commits information such as "ID number", "time elapsed", and "number of tries before passing" to the database. It is important to note that a lecturer can know the learning progress of students from these data. Because a lecturer gives the credit of a lecture when a student passed all of examinations of the lecture.

Since SUGSI was opened in 2002, the number of enrolled students has been 81, 73, 71, and 54 for the years 2002, 2003, 2004, and 2005, respectively. Here, we include a brief analysis of the students in SUGSI. Fig. 1 shows the age composition and Fig. 2 shows the working-status composition of students. From these figures, we can see that almost 80% of the students are between their 30's to 40's and that 88% of the students are full-time working professionals. We can see that the students of SUGSI tend to start their learning after they finish their work. We prepare the long-term program for workers who are busy their business. Students can attend the master's course for 4 years maximum with payment for 2 years by using this program.

Fig. 3 shows the number of students who have passed examinations sorted by hour in weekdays and holidays. This figure shows that a relatively large number of students take examinations and there are students taking examinations at all hours.

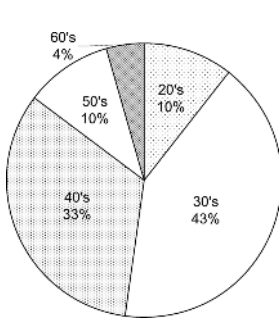


Fig. 1. Age composition of students

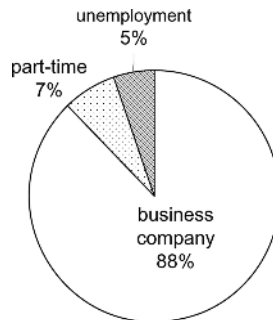


Fig. 2. Working-status composition of students

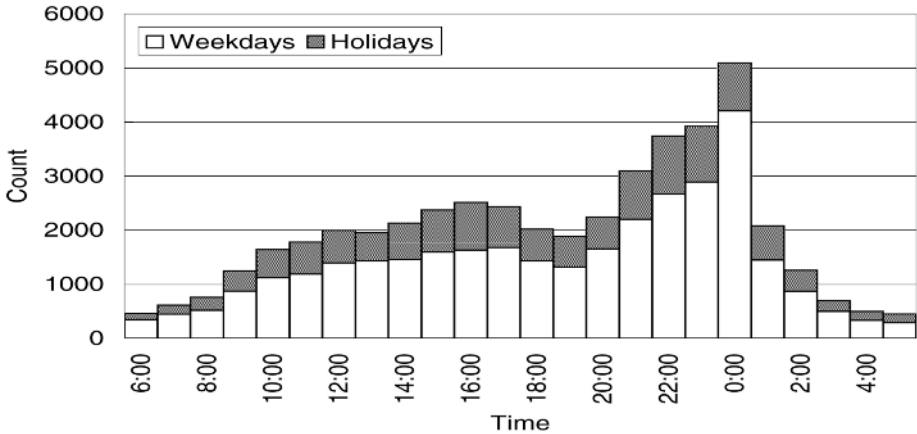


Fig. 3. Hourly distribution of students passing examinations

3 Trial of Support for Students

We started the trial of support for distant learning students from October 2003. As mentioned above, we record the students’ progress by means of the web-based examination systems. In addition, we developed a system to display the learning progress of students and make available this information to students and lecturers using different views. Fig. 4 shows a page displaying progress data for lecturers. Lecturers can confirm how many examinations each student has passed from this page.

In SUGSI, there are over 360 examinations, and a minimum of two hundred examinations have to be passed to complete the master program. Also, the standard number of years to complete the program is two. In our initial trial of the support system, based on the remaining years to completion and the number of passed test by a student, we decided three conditions to consider a student will be at risk of decreasing his/her learning motivation, as illustrated in Table 1. These conditions are meant to help students at all stages of their learning program and indicate that a student should have passed at least 25%, 50%, and 75% of the examinations by the first, second, and third quarter of the program, respectively. If a student within any of the three conditions made no progress for the last two weeks and had no contact with the teachers, we start to support the student.

Support mail is very effective to maintain learning motivation of learning[4][5]. Based on these reports and our experiences, we decided the procedure of support for students is as follows. Following procedures are worked out by the staff of SUGSI.

- (1) Send an email notifying the current progress data and asking the student’s situation.
- (2) Wait for a reply for two weeks and confirm the progress data.

- (3) Counsel the student about his/her curriculum design, if the student replies.
- (4) Try to contact the student using telephone and facsimile, if the student does not reply.



Fig. 4. Page displaying progression data for lecturers

Table 1. The conditions for considering at-risk students

condition #	remaining year(s) before completion	total passed tests
1	1.5	≤ 50
2	1	≤ 100
3	0.5	≤ 150

4 Evaluation of Support

We tried the support system for 27 students whose profile corresponded with the conditions mentioned in the previous section from October 9th 2003. Fig. 5 shows the result of the trial support. From these figures, we can see that 23 students (85%) replied and that 12 students (44%) improved their learning progress. Also, it can be seen that 5 of them completed the master’s course, 6 improved their curriculum by applying to a long-term program, and 1 dropped out the master’s course. On the other hand, 15 students (56%) did not improve their learning progress, and 3 of them quitted the course. However, 8 of them continue to enroll by applying to a long-term program.

Fig. 6 shows how the learning progress of 5 students that completed the program changed after sending support mails. In Fig. 6, the horizontal axis indicates the weeks passed relative to the day the support mail was sent, where 0 means the week the mail was sent and minus means the weeks before sending

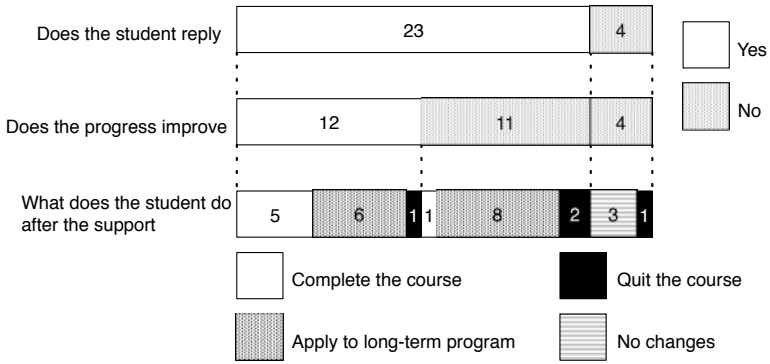


Fig. 5. The result of support mail

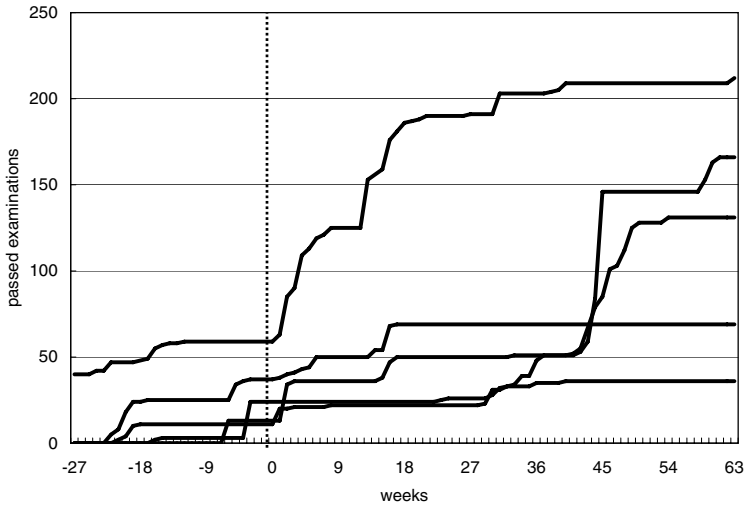


Fig. 6. Changing of learning progressions after sending support mails

the mail. The vertical axis indicates the total number of sections the student passed.

From these figures, it is clear that the students were inspired by the support mails to increase their learning motivations, except for one student who improved his/her learning progress before he/she received the mail.

5 Supporting System

In response to the results of the first trial, we improved the method of support and continue the support for students. In the trial, the condition to consider

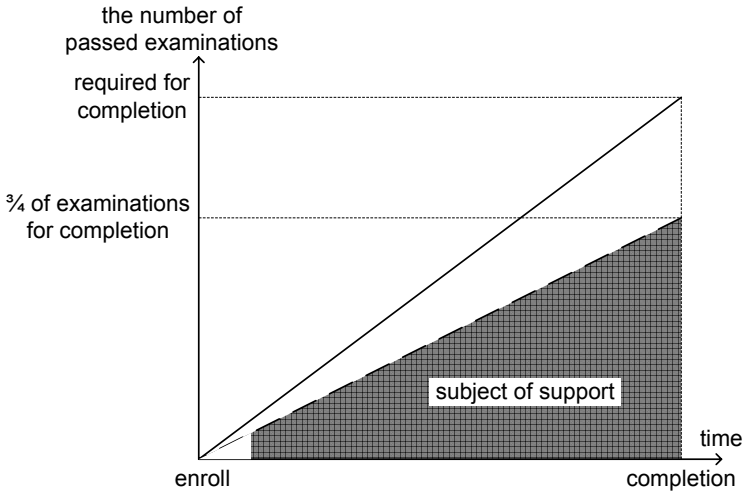


Fig. 7. The improved condition

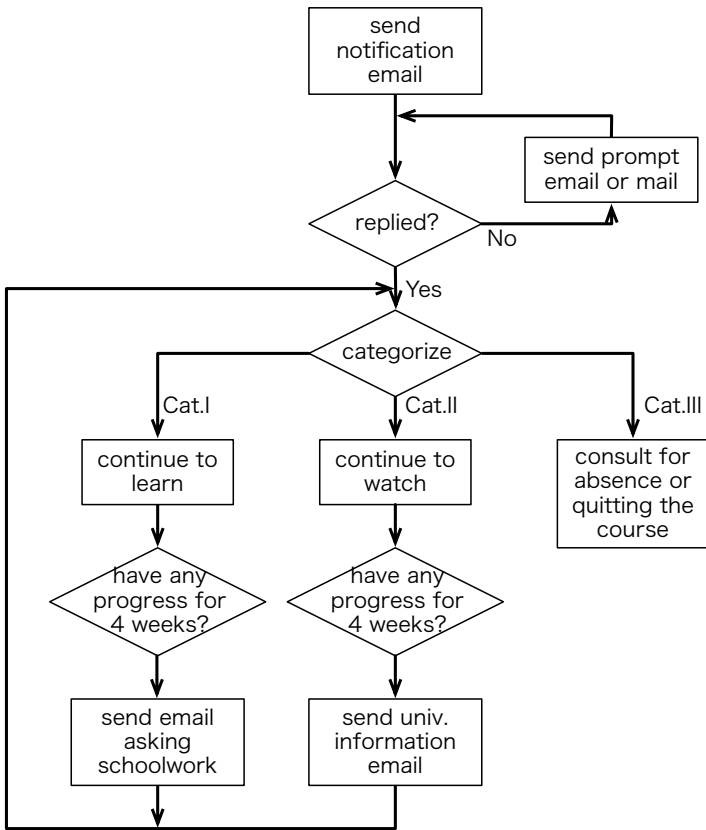


Fig. 8. The improved procedure

a student at-risk was discrete. We improved the condition as shown by Fig. 7 to implement a more detailed support. Also, we improved the procedure of the support as shown by Fig. 8.

The improved procedure is as follows:

- (1) Send the email to ask the student’s situation.
- (2) Send the email or mail to prompt a response, if the student does not reply.
- (3) Classify the student condition in one of the following categories based on the analysis of the response from the student, if the student replies.
 - I. The student can continue to learn.
 - II. The student has the motivation of learning, but he/she has some problems or some issues on their jobs which impede his/her learning.
 - III. The student loses the motivation of leaning.
- (4) Continue to watch the progress for “Category I student”. Send the email to ask his/her school work, if he/she has no progress for 4 weeks.
- (5) Continue to watch the progress for “Category II student”. Send an email with University information such as announcements of events or other information about Shinshu University to notice that we care about the student, if he/she has no progress for 4 weeks.
- (6) Consult for taking a leave of absence or quitting the course for “Category III student”.

We started the improved support from September 15th 2005, and 36 students corresponded the improved condition. Fig. 9 shows the result of the improved support.

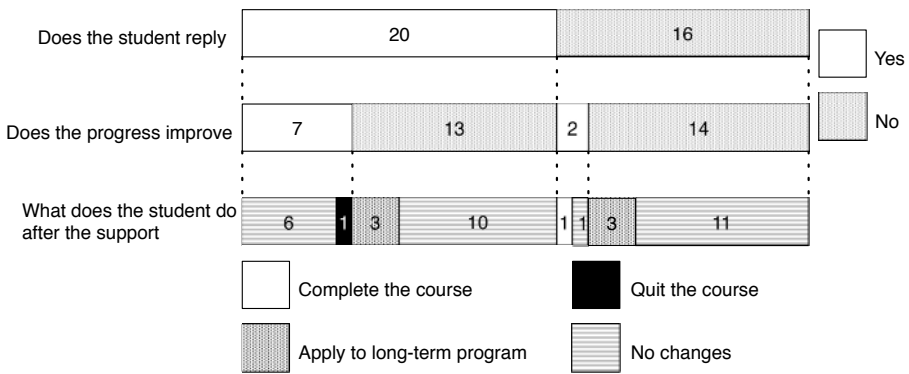


Fig. 9. The result of the improved support mail

We have some issues to be considered on the method of student support.

- The interval of support.

It is very difficult to decide the interval of sending support mails. Students feel pressured by the mails at short intervals, and students feel neglected by the mails at long intervals. Now, we set the interval up as four weeks based on our experiences.

- Support activities for “Category II student”.

Now, we send an email with University information for “Category II students”. We will examine the adequacy of that support and will continue to improve the procedure of support.

The rate of completion of SUGSI students is about 60%. This is higher than the completion rate of almost of the other correspondence schools. We think the support system helps to increase the completion rate.

6 Conclusion

E-Learning courses have an advantage of being able to understand the learning progress of the students for the teachers. By using the advantage of e-Learning courses, we try to support distant learning students in order to keep their motivation of learning.

In Shinshu University, besides the course material offered in SUGSI, over 100 e-Learning course materials have been developed and the support system will become increasingly essential to enrich the educational environment for all kinds of distant learners.

In the future, we would like to continue improving the e-Learning course materials and the method of student support.

References

1. Kunimune, H., Niimura, M., Wasaki, K., Fuwa, Y., Shidama, Y., and Nakamura, Y.: The Learning System of Shinshu University Graduate School of Science and Technology on the Internet. Proceedings of KES 2005. Part III, (2005) 1296-1302
2. Fuwa, Y., Shidama, Y., Wasaki, K., and Nakamura, Y.: The Plan of the Shinshu University Internet Graduate School. JSiSE. 19, 2 (2002) 112-117
3. Fuwa, Y., Nakamura, Y., Yamazaki, H., Oshita, S.: Improving University Education using a CAI System on the World Wide Web and its Evaluation. JSiSE. 20, 1 (2003) 27-38
4. Visser, L., Plomp, T., Kuiper: Development research applied to improve motivation in distance education. Association for Educational Communications and Technology, Houston, TX (1999)
5. Gabrielle, D.M.: The effects of technology-mediated instructional strategies on motivation, performance, and self directed learning. Proceeding of ED-Media (2003) 2569-2575

Reflection by Knowledge Publishing

Akihiro Kashihara and Yasuhiro Kamoshita

Dept. of Information and Communication Engineering,
The University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo, 182-8585, Japan
{kashihara, kamoshita}@ice.uec.ac.jp

Abstract. Self-directed learning often finishes with incomplete knowledge. It is hard for learners to be aware of the knowledge incompleteness. The main issue addressed in this paper is how to enhance their awareness. Our approach to this issue is to enable learners to publish knowledge obtained from their learning process to review it. The most important point towards knowledge publishing is how to represent knowledge so that the learners can reflect on it from another perspective. However, such knowledge representation would be troublesome for the learners. This paper discusses how to promote learners' knowledge publishing by automatically generating a hypertext with TOC (Table Of Contents) as representation of knowledge they have learned in their learning process. Publishing the hypertext enables the learners to gain not only their own reviews but also the reviews from the peers.

1 Introduction

Learning in a self-directed way has been more and more important along with increasing opportunities for learning with Web contents [1], [2]. On the other hand, learners often finish it with incomplete knowledge [3]. It is also hard for them to reflect on the knowledge incompleteness [4],[5]. This is one of the most crucial issues addressed in self-directed learning.

Our approach to this issue is to enable learners to publish knowledge obtained from their self-directed learning process and review it. Such knowledge publishing would enhance their awareness of the incompleteness of knowledge they have learned [5]. The most important point toward knowledge publishing is how to represent knowledge they have learned so that they can review it. Such knowledge representation requires them to put knowledge in a proper order, which can guide them in reviewing their knowledge. It also needs clarifying relationships among pieces of their knowledge and an overview of their knowledge. However, it is difficult for the learners to publish their knowledge since the knowledge representation process would impose a cognitive overload on them [1].

The main issue addressed in this paper is how to help learners publish knowledge they have acquired in self-directed learning with existing hypertext-based contents. The self-directed learning process involves navigating pages and making semantic relationships among the contents learned at the navigated pages to construct

knowledge [4],[6]. Such navigation with knowledge construction is called navigational learning.

In this paper, we describe a framework for knowledge publishing. It enables learners to learn hypertext-based contents with Interactive History (IH for short), which we have developed for scaffolding self-directed navigational learning [2],[7]. IH generates their navigational learning history and knowledge map as representation of knowledge they have constructed. In order to enable learners to reflect on their constructed knowledge from another perspective, the framework automatically generates a hypertext with TOC (Table Of Contents) from the knowledge map, which can be browsed by Web browser. The TOC suggests in which order the learners review their knowledge they have constructed in navigational learning. The hyperspace map and hyperlinks between the pages included in the hypertext suggest the semantic relationships among pieces of the learners' knowledge. The hypertext can be also published to the peers after the learners edit it. In this way, knowledge publishing is viewed as hypertext publishing.

2 Navigational Learning and Interactive History

Let us first reconsider self-directed navigational learning with hypertext-based contents, and then introduce IH that scaffolds the navigational learning process.

Hypertext-based contents provide learners with hyperspace, which consists of the pages and their hyperlinks. The learners generally start navigating the pages for achieving a learning goal. The movement between the various pages is often driven by a local goal called navigation goal to search for the page that fulfills it. Such navigation goal is also regarded as a sub goal of the learning goal. The navigational learning process includes producing and achieving a number of navigation goals. We currently classify navigation goals into six: Supplement, Elaborate, Compare, Justify, Rethink, and Apply. We refer to the process of fulfilling a navigation goal as primary navigation process [7]. This is represented as a link from the starting page where the navigation goal arises to the terminal page where it is fulfilled. Navigation goal signifies how to improve the domain knowledge learned at the starting page.

The navigation process can be modeled as a number of primary navigation processes [7]. In each primary navigation process, learners would integrate the contents learned at the starting and terminal pages. For instance, a learner may search for the meaning of an unknown term to supplement what he/she has learned at the current page or look for elaboration of the description given at the current page. Carrying out several primary navigation processes, learners would construct knowledge from the contents they have integrated in each primary navigation process. Each learner would construct his/her own knowledge even when the same content is learned with the same learning goal.

In order to scaffold such knowledge construction process, we have developed IH that provides learners with two functions: annotated navigation history and knowledge map. IH first enables learners to annotate a navigation history, which includes the pages sequenced in order of time they have visited, with primary navigation processes [2],[7].

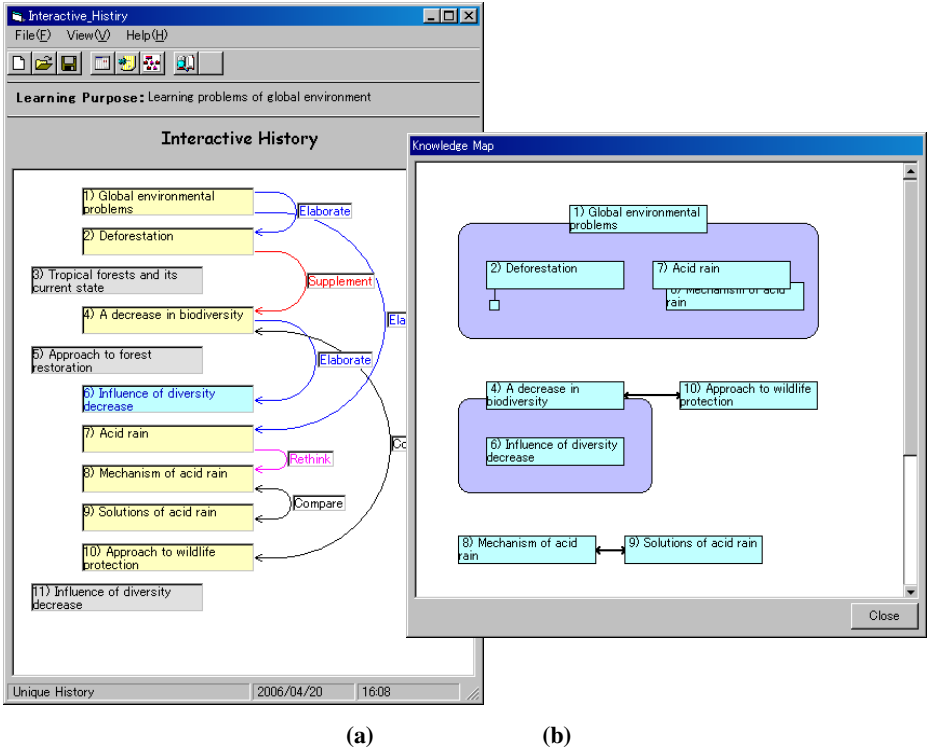


Fig. 1. Annotated navigation history and knowledge map generated with IH

Fig. 1(a) shows an example of annotated navigation history. IH monitors learners' navigation in the Web browser to generate the navigation history in the *Annotated Navigation History* window. Each node corresponds to the page visited. The learners can annotate the navigation history with the primary navigation processes that they have carried out. They can also take a note about the contents learned at the starting or terminal pages, which could be copied and pasted from the corresponding portions of the pages. The note is linked to the node in the annotated navigation history. The learners can look at the note on their demand by mouse-clicking the corresponding node.

When the annotated navigation history includes more primary navigation processes, the learners have more difficulty in understanding the semantic relationships among the pages included in these primary navigation processes and in constructing their knowledge. When the primary navigation processes are overlapped each other, in particular, it is hard to understand the relationships among the pages included. In order to reduce such difficulty, IH generates a knowledge map by transforming primary navigation processes into visual representation. (See [2] for detailed generation mechanism.)

Fig. 1(b) shows an example of knowledge map. The knowledge map is generated from the annotated navigation history shown in Fig. 1(a). Viewing this map, for example, the learner can visually understand that he/she elaborated the contents of *Global environmental problems* by referring to *Deforestation* and *Acid rain*

rain, and so on. By mouse-clicking a page in the knowledge map, they can look at the note they have taken about the page.

The knowledge map generally consists of several islands including some primary navigation processes. We call them Knowledge Islands (KIs). The knowledge map shown in Fig. 1(b) consists of three KIs, which include four primary navigation processes, two primary navigation processes, and one navigation process.

The results of the case study, which we have had with IH, suggest that it can promote knowledge construction in self-directed navigational learning process [8].

3 Knowledge Publishing

Although IH can help learners reflect on what they have constructed so far in hyperspace, it is still hard for them to be aware of the insufficiency or unsuitableness of knowledge. How they gain the awareness is a key issue addressed in self-directed learning.

Knowledge publishing is a promising approach to this issue. In publishing knowledge learners have constructed in hyperspace, it is necessary for them to represent knowledge so that they can review it. Such knowledge representation requires the learners to put constructed knowledge in a proper order, which can guide them in reviewing the constructed knowledge. It also needs clarifying relationships among pieces of their knowledge and drawing the whole structure.

Such knowledge representation process provides the learners with opportunities for reflecting on the constructed knowledge from another perspective. In addition, it is possible to obtain reviews from the peers after publishing. The peer reviews also enable the learners to gain their awareness of the insufficiency and unsuitableness of constructed knowledge, which they could not be aware of by themselves.

On the other hand, knowledge publishing is not easy for learners. Representing their knowledge for publication after navigational learning process would be troublesome for them, and requires them to make much cognitive efforts [1].

In order to resolve this issue, we have developed a framework for knowledge publishing, in which learners are encouraged to author a hypertext as representation of knowledge they have constructed with IH in their navigational learning process. It is quite hard to author the hypertext from scratch. The framework accordingly provides the learners with functionality that automatically generates the hypertext from the knowledge map obtained from IH. The hypertext includes the pages that the knowledge map includes. The learners are then allowed to edit it. The edited hypertext can be published.

The important point in the hypertext authoring is to author the hypertext understandable for the learners or peers. The learners are accordingly provided with a TOC of the hypertext, which is also automatically generated. Fig. 2 shows an example of TOC, which is generated from the knowledge map shown in Fig. 1(b). The TOC suggests in which order the learners should review their knowledge. They are also provided with a map of hyperspace provided by the hypertext and hyperlinks between the pages included in the hypertext, which suggest the whole structure and the semantic relationships among pieces of the constructed knowledge. The TOC,

hyperspace map, and pages of the hypertext are generated as html files, which can be browsed by Web browser.

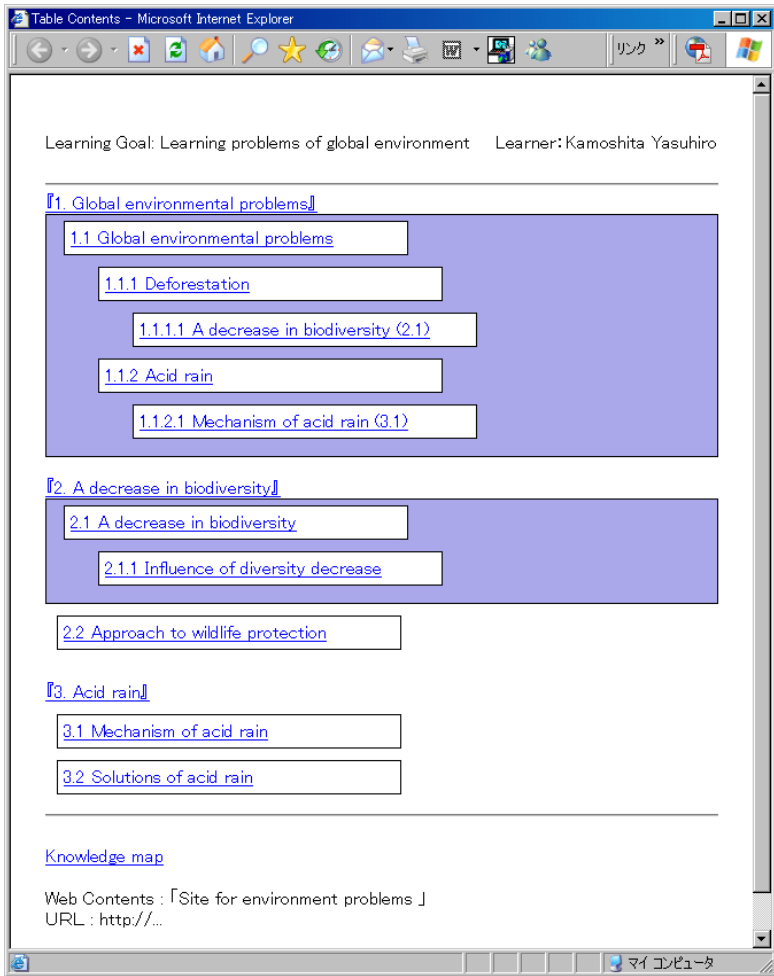



Fig. 2. TOC

The hyperspace map is represented as the knowledge map in a page, whose nodes are clickable and linked to the corresponding pages. Fig. 3 also shows an example of page included in the hypertext, which describes the contents learned about the corresponding page, and which embeds links to the related pages. The hypertext is published after the learners edit it. In our framework, knowledge publishing is viewed as hypertext authoring. The hypertext authoring process provides learners with opportunities for reflecting on their knowledge incompleteness from a perspective different from the one they have drawn in their navigational learning process.

1.1.2 Acid rain

Report of existence of the first acid rain was in industrial city Manchester in Britain of the Industrial Revolution age. In 1852, **Smith in Britain** analyzed the chemical composition of precipitation in the factory area, and pointed out the relation between air pollution and acid rain. It is this time that word "Acid rain" appeared for the first time. It is also famous for "Fog of London" to take SO_x , to become the fog of sulfuric acid, and to have given big damage.



Acid rain is currently seen in large cities of Latin America, Eastern European nations, and China. In these nations, the main cause for acid rain is air pollution, which is caused by the delay of applying regulations of [the coal use, car increase, and exhaust emissions \[Rethink\]](#).

[1.1.2.1 Mechanism of acid rain \[Rethink\]](#)

Back	Home	Next
1.1.1.1 A decrease in biodiversity (2.1)	Table of Contents Hyperspace Map	1.1.2.1 Mechanism of acid rain (3.1)

ページが表示されました マイコンピュータ

Fig. 3. Page contents authored

4 Hypertext Authoring

TOC is generated from the knowledge map as follows. Each section of TOC is first created according to each KI included in the knowledge map. Each section has its title that is selected from representative page in the corresponding KI, and includes several items indicating the starting and terminal pages of primary navigation processes in the KI. It also has a hierarchical structure where the terminal page item of the primary navigation process becomes a subsection of the starting page item. Each page item is numbered according to the hierarchical structure, and has a link to the corresponding page content published.

The TOC shown in Fig. 2 consists of three sections corresponding to the KIs in the knowledge map shown in Fig. 1(b). In the KI of *Global environmental problems*, for

example, there are four primary navigation processes as shown in Fig. 1(b). The page of *Global environmental problems* is the starting one of the two primary navigation processes whose terminal pages are *Deforestation*, and *Acid rain*. The page item of *Global environmental problems*, which is also selected as the first section title numbered as 1., is accordingly numbered as 1.1, and items of the other pages are also numbered as 1.1.1, and 1.1.2. The pages of *A decrease in biodiversity* and *Mechanism of acid rain* are the terminal ones of the remaining primary navigation process whose starting pages are *Deforestation* and *Acid rain*. The page items are accordingly numbered as 1.1.1.1, and 1.1.2.1.

Learners are allowed to edit it when the TOC is presented to them. The edit operations allowed are changing the title of the section, the order of the sections, and the order of the page items in the section. In Fig.2, for example, the title of the third section, which is originally defined as 3. *Mechanism of acid rain and Solutions of acid rain*, is changed to 3. *Acid rain*.

The edited TOC represents the whole structure of the hypertext published. It also illustrates the order in which the learners should review knowledge they have constructed. The peers are also expected to follow the hierarchical numbers attached to the page items to review the hypertext contents.

After the TOC is generated, hyperspace of the hypertext is constructed as follows. Each page, which is linked from the TOC, is first generated. The content of the page is obtained from the note learners have taken about the corresponding page in the knowledge map. The links from the page to the terminal pages are also embedded under the page contents. The anchors of the links are described as the following form: hierarchical number, title of the terminal page item, and navigation goal annotated between the page and the terminal page. The navigation goals attached to the anchors encourage the learners to grasp the semantic relationships between the page and the terminal pages before reviewing the terminal pages. In the bottom of the page, in addition, navigation links (Back, Home, and Next) are prepared for helping them follow the hierarchical order indicated in the TOC.

For example, the content of the page shown in Fig. 3 is generated from the note taken about the page of *Acid rain*, which corresponds to the page 1.1.2 in the first section of the TOC shown in Fig. 2. Under the content of the page, one link from the page to the terminal pages is anchored as *1.1.2.1 Mechanism of acid rain [Rethink]*. *[Rethink]* means the navigation goal of the primary navigation process from the page of *Acid rain* to the terminal page of *Mechanism of acid rain*. This navigation goal description helps the learner grasp the semantic relationships between the page of *Acid rain* and the terminal pages before reviewing the terminal page.

Learners are allowed to edit the pages included in the hypertext when the hypertext is presented to them. The edit operations allowed are changing the contents of the pages by referring to the corresponding pages, and embedding the link anchors involving the navigation goals into the contents. In the original content of the page shown in Fig. 3, the link is embedded as the anchors of *the use of coal, the rapid increase of car, and exhaust emissions [Rethink]*. The reason why the link to the page of *Mechanism of acid rain* is anchored to the words *the use of coal, car increase, and exhaust emissions* is that the learner thinks these are representative causes for acid rain. The embedded link intends to help the learner review the relationships to the terminal pages in the context of the current page.

5 Conclusion

This paper has proposed reflection by knowledge publishing as a promising solution to the issue of how to enhance awareness of incomplete knowledge learners have constructed in self-directed navigational learning with hypertext-based contents. The most important point towards knowledge publishing is to represent their constructed knowledge so that they can review it.

This paper has also demonstrated a framework for knowledge publishing, which provides learners with functionality that generates a hypertext with TOC as knowledge representation for publication. It is obtained from the knowledge map learners have generated with IH in their learning process. The hypertext enables the learners to reflect on their constructed knowledge from another perspective. It also enables them to publish their knowledge with less cognitive efforts. In addition, they can get the reviews from the peers.

In future, we need to ascertain whether learners can promote knowledge publishing, and whether they can get instructive reviews from themselves and from the peers. We will then refine the framework according to the results.

References

1. Jonassen, D.H.: *Computers as Mindtools for Schools*, 2nd ed., Merrill Prentice Hall (2000).
2. Kashihara, A., and Hasegawa, S.: *LearningBench: A Self-Directed Learning Environment on the Web*, Proc. of ED-MEDIA2003 (2003) 1032-1039.
3. Hammond, N.: *Learning with Hypertext: Problems, Principles and Prospects*, in McKnight, C., Dillon, A., and Richardson, J. (eds): *HYPERTEXT A Psychological Perspective*, Ellis Horwood Limited (1993) 51-69.
4. Thuring, M., Hannemann, J., and Haake, J.M.: *Hypermedia and cognition: Designing for comprehension*, Communication of the ACM, Vol. 38, No.8 (1995) 57-66.
5. Kashihara, A., and Hasegawa, S.: *Unknown Awareness in Navigational Learning on the Web*, Proc. of ED-MEDIA2004 (2004) 1829-1836.
6. Cunningham, D.J., Duffy, T.M., and Knuth, R.A.: *The Textbook of the Future*, in McKnight, C., Dillon, A., and Richardson, J. (eds): *HYPERTEXT A Psychological Perspective*, Ellis Horwood Limited (1993) 19-49.
7. Kashihara, A., Hasegawa, S., and Toyoda, J.: *An Interactive History as Reflection Support in Hyperspace*, Proc. of ED-MEDIA2000 (2000) 467-472.
8. Kashihara, A., and Hasegawa, S.: *A Model of Meta-Learning for Web-based Navigational Learning*, International Journal of Advanced Technology for Learning, Vol.2, No.4, ACTA Press (2005) 198-206.

Self-learning System Using Lecture Information and Biological Data

Yurie Iribe¹, Shuji Shinohara², Kyoichi Matsuura², Kouichi Katsurada²,
and Tsuneo Nitta²

¹ Toyohashi University of Technology, Information and Media Center
1-1 Hibirigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441- 8580, Japan
iribe@imc.tut.ac.jp

² Toyohashi University of Technology, Graduate School of Engineering
1-1 Hibirigaoka, Tempaku-cho, Toyohashi-shi, Aichi 441- 8580, Japan
{shinohara, matsuura}@vox.tutkie.tut.ac.jp,
{katurada, nitta}@tutkie.tut.ac.jp
<http://www.vox.tutkie.tut.ac.jp/>

Abstract. One of today's hot topics in the field of education is the learning support system. With the progress of networks and multimedia technologies, various types of web-based training (WBT) systems are being developed for distance- and self-learning. Most of the current learning support systems synchronously reproduce lecture resources such as videos, slides, and digital-ink notes written by the teacher. However, from the perspective of support for student learning, these systems provide only keyword retrieval. This paper describes a more efficient learning support system that we developed by introducing lecture information and student arousal levels extracted from biological data. We also demonstrate the effectiveness of the proposed system through a preliminary experiment.

1 Introduction

Many studies have been conducted focusing on the technology of learning support systems. With the progress of networks and multimedia technologies, various types of web-based training (WBT) systems are being developed for self-learning [1] [2]. The main advantage of WBT is that students are able to learn at their own pace at university or at home under a network environment. On the other hand, however, some students may find it difficult to maintain their concentration in an isolated environment [3] [4]; therefore, it is important that WBT effectively holds the students' interest.

Most of the current support systems for web-based learning synchronously reproduce lecture resources such as videos and Microsoft PowerPoint slides that the teacher uses during the lecture [5] [6]. These systems are very useful for confirming and reviewing any parts of the lecture that may have been missed. However, they do not take into consideration the student's level of attentiveness during the lecture. Furthermore, different students may find different slides difficult to understand.

We are developing a self-learning support system to solve the above problems. The system will help students to review lectures more efficiently by introducing the level

of arousal, concentration, and understanding extracted from each student's biological data (i.e. EOG (electrooculogram), EEG (electroencephalogram), and ECG (electrocardiogram)) and lecture information (i.e. questions/answers). In this paper, we describe a lecture view system as part of the self-learning support system that offers lecture resources and various supports suited to each student.

2 System Outline

The proposed lecture view system supports individual-based self-learning (mainly for reviewing lectures) by using each student's lecture information (questions and replies, etc.) and biological data obtained during the lecture. The student's lecture information is acquired using the ILS (Interactive Lecture System) developed in our group [7]. Fig.1 shows an outline of the proposed system. In the following sections, the ILS and the lecture view system are described.

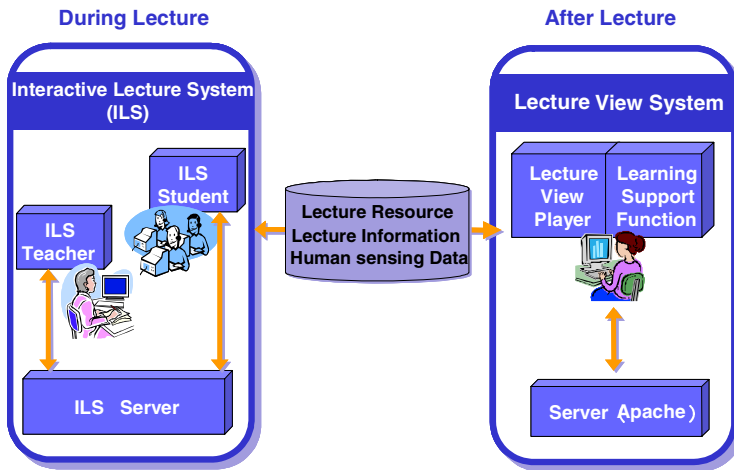


Fig. 1. Outline of the proposed system

2.1 Interactive Lecture System (ILS)

We developed a lecture support system called ILS (Interactive Lecture System) through which the students can ask questions and express their wishes during a lecture (Fig. 2). The teacher can easily determine the level of understanding from the students' reactions and subsequently ensure that the lecture matches the students' needs. The system is equipped with a function whereby the students can input their questions easily and quickly. In addition, the system encourages questions between students using a function whereby students can answer each other's questions (answer promotion function).

Fig.2 shows an ILS window. The system is composed of the teacher terminal, the student terminals, and the lecture server, all connected through the network. The teacher can use digital ink to freely draw lines and characters on slides that are being displayed and the data is sent to the student terminals through the lecture server. ILS displays the lecture resources on both the teacher and student PC terminals. When the teacher changes a slide or draws anything, the same screens are synchronously displayed on the students' PCs. When a student inputs the contents of a question using the keyboard, the question is displayed on the teacher terminal. In addition, ILS provides a function to acknowledge a question in order to determine what the student wants to know, and the question and acknowledgement number are displayed on the teacher terminal and each student terminal through the lecture server. However, the teacher can only find out what a student wants to know through the question function, and some students hesitate to ask questions or are unable to express a vague question in a sentence. Therefore, we added an SOS button on the ILS so that students can easily tell the teacher what they want.

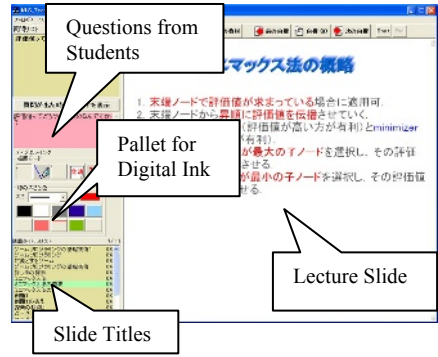


Fig. 2a. Window of ILS – teacher terminal

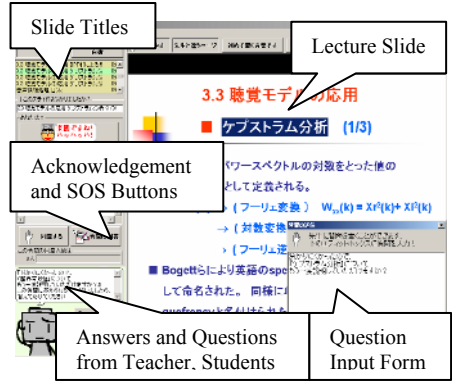


Fig. 2b. Window of ILS – student terminal

2.2 Lecture View System

This section describes the lecture view system proposed in this paper. Fig.3 shows an outline of the system.

2.2.1 Extraction of Lecture Information and Biological Response Data

The lecture view system consists of a lecture view player that faithfully reproduces the lecture, and a learning support function that uses each student's lecture information and arousal level during the lecture (Fig. 4).

The lecture view system extracts the lecture slide switch data, digital-ink data from the ILS that records the teacher's operation data during a lecture. The system acquires question contents and the time that various events (question button, SOS button and acknowledgement button) occur. This data is described using XML, and is accumulated in the database. Moreover, the lecture resources that include the lecture video

and slides are generated using ILS and the digital video camera and are converted into HTML.

Our group (working under a 21st Century COE Program Grant for “Intelligent Human Sensing”) developed a sensing device that extracts certain human biological information. The device is portable and includes a wearable micro sensor chip. A special chair for the student to sit in with an installed sensing device has been completed and our group is presently conducting experiments using it.

In the experiment, analysis

is performed on the level of arousal, understanding, and concentration by extracting biological information such as EOG (electrooculogram), EEG (electroencephalogram), and ECG (electrocardiogram) from the student listening to the lecture. The proposed lecture view system uses the arousal level for biological data.

2.2.2 Lecture View Player

In the current e-learning for viewing a lecture, the mainstream is to offer the students lecture resources such as slides and videos on the Web so that they can review the lecture. Therefore, the lecture view player that we built synchronously displays the lecture video, and the lecture slides are converted from PowerPoint into HTML. It displays them based on the slide switch time acquired from ILS, and the lecture is reproduced on the PC.

To watch a lecture video, students can control the play speed (from half to double), clearly catching the teacher’s voice by using Variable Player [8] (Fig. 4①). Thus, the students can efficiently hear the lecture video.

Moreover, the slides and digital-ink notes are displayed in synchronization with the lecture video (Fig. 4②). Slide titles are lined up along a vertical bar representing the time axis (Fig. 5). By clicking a title or sliding a handle on the bar to an arbitrary position, students can start lecture contents from the desired position. Thus, the students can learn the lecture resources in an efficient manner.

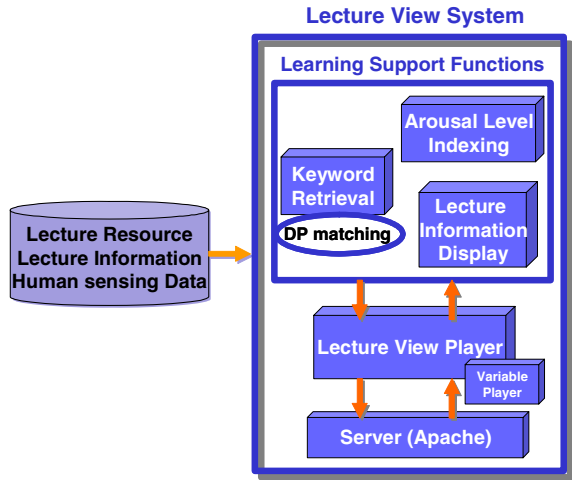


Fig. 3. Lecture View System

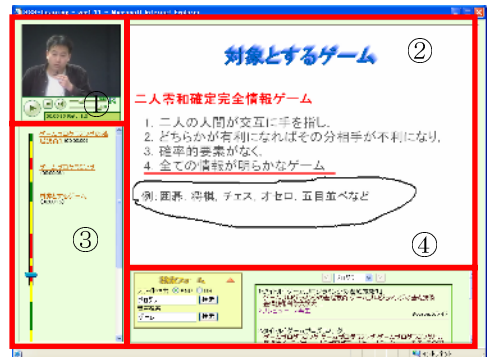


Fig. 4. Window of Lecture View System

The lecture view player is implemented as a Web application by Perl/CGI and JavaScript, and uses Internet Explorer. Therefore, students can use the system any time through a Web browser under the network environment.

2.2.3 Learning Support Function

The lecture view player synchronously displays the lecture video and the slides. However, it is not sufficient to merely reproduce the teacher's explanation. In order to provide appropriate learning support for the students, it is necessary to consider the student's state during the lecture. Therefore, our proposed lecture view system contains the following three functions (Fig. 3).

(1) Utilization of Arousal Level Data

This function uses arousal level data to support the students. The arousal level represents the sleeping state (the low arousal level) and waking state (the high arousal level) and the other state (It is somnolent state or cannot be judged to be asleep even if waking) of each learner in the class. We distinguish the sleeping state (the low arousal level) and waking state (the high arousal level) according to the somnolent state. The somnolent state is characterized by prolonged blink intervals and blink duration. The somnolence is revealed as a gradual increase of both parameters. The EOG signal is quantified as both standardized blink interval and standardized blink duration [9].

The arousal levels are distinguished by three stages, and expressed by the color of the slide bar according to the level (Fig. 4③ and Fig. 5). In cases where a student may not have heard part of the lecture during a low arousal level, the missed part can be efficiently found and studied by indexing the arousal level. On the other hand, students presumably listen intensively when they are in the high arousal level; therefore, the lecture view player automatically plays those parts at two or three times the speed. The students can also manually change the play speed and either fast-forward through parts where the arousal level is high or repeat parts where the concentration level was low. As a result, the students can review the lecture efficiently and effectively.

(2) Display of Student's Lecture Information

Lecture information that students questioned or found difficult to understand requires careful review. Therefore, the learning support function displays student's lecture information (Fig. 4③ and Fig. 5), which consists of the following three elements acquired from the ILS.

- Question/answer
- Acknowledgement buttons to question/answer
- SOS buttons

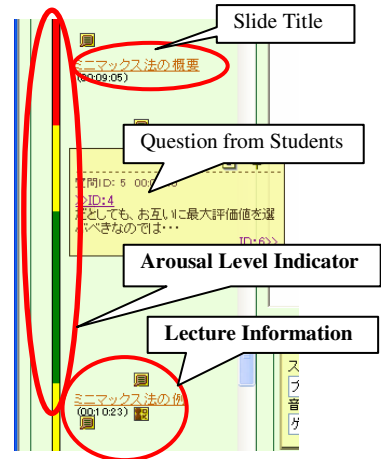


Fig. 5. Information in the lecture

Questions and answers during the lecture are useful for the review because students may not remember everything from the lecture afterwards. Moreover, the system displays question/answer icons corresponding to the time of the lecture video and slides for easy recognition. When a question/answer icon is clicked, the content of the question/answer is displayed in a pop-up box (Fig. 5).

The students can press the SOS button when there is something they do not understand. The teacher can analyze the parts where a lot of SOS buttons were pressed, and then prepare additional resources to improve the lecture.

(3) Retrieval of Keywords

One of the factors behind a student not understanding a lecture is unfamiliarity with the terms used in the teacher's explanation. Easy retrieval of those terms would be very useful when the student is reviewing the lecture. Therefore, we built a function for retrieving the lecture resources corresponding to the keywords input by the student.

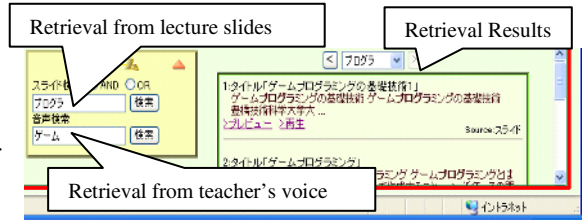


Fig. 6. Retrieval of keywords in text and speech

Our system includes not only slide retrieval but also voice retrieval (Fig. 4 ④ and Fig. 6). In retrieval of the teacher's voice, the phoneme parts that correspond to the keyword are retrieved from the voice data by DP matching. DP matching can detect a part near the keyword even if the phoneme row is substituted for a similar phoneme row, because it is based on the distance between the phoneme distinction feature vectors. When the student selects a retrieval result, the lecture resources are played from the vicinity where the teacher uttered the keyword. As for retrieval of the lecture slides, the part corresponding to the keyword from the character string in the HTML file (PowerPoint) is specified, and the slide that contains the part is displayed as the retrieval result. In addition, the keyword retrieval that we propose gives priority to the results in order of low arousal level because the student may not have understood the terms during that time. As a result, the student can efficiently retrieve their terms.

3 Preliminary Experiment

In this chapter, we describe the experiment to demonstrate the effectiveness of our proposed system.

3.1 Data Sets

We presented a 30-minute lecture on “Game Programming” using the ILS. As a result, 30 minutes of lecture resources were generated. In addition, we cut out five sections (about 2 or 3 minutes each) from among the lecture resources (30 minutes) in Fig.7. The five sections each were cut out as meaningful section. We joined the remaining sections and created lecture resources (Part A) totaling about 15 minutes in order to produce a time zone with high arousal level. We assumed the parts that we

cut out (Part B) as being in a time zone when the arousal level was low. In contrast, Part A was assumed to be in a time zone when the arousal level is high. In addition, we created five problems from Part A (Problem A), and five problems from Part B (Problem B).

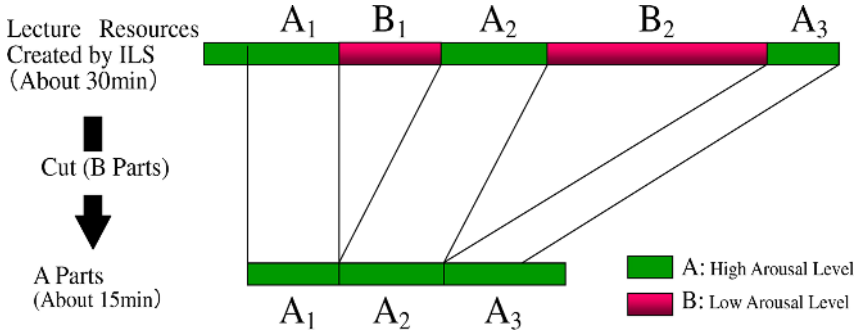


Fig. 7. Data Sets

3.2 Experimental Setup

To ensure efficient review of the lecture, the system contains three support functions. Therefore, it is necessary to measure the efficiency of those support functions. Concretely, we compared the following two systems through the experiment. System A is the system that we propose.

- System 1 (The lecture view system): Lecture view player + three support functions
- System 2: Lecture view player (The lecture video and the lecture slide are simply played)

Seven students who had not attended the lecture on "Game Programming" participated in the experiment. We divided them into Group 1 (four students) and 2 (three students). Group 1 learned the lecture by using System 1, and Group 2 learned the lecture by using System 2. The order of the experiment was as follows:

1. We explained Group 1 and 2 how to use the system.
2. Group 1 and 2 watched Part A.
3. Group 1 and 2 answered Problem A and we measured the response time.
4. Group 1 and 2 solved Problem B by using System 1 and 2, respectively. We measured the answer time. Contents used for each system are the lecture resources of 30 minutes (Part A and Part B are included).

3.3 Experimental Results

Fig.8 shows the experimental results. Problem A and Problem B was simple, most of the students were able to get 100 points. There is a significant difference between the response time for Problem A and the response time for Problem B. Group 1 is P =

$0.035 < 0.05$, and Group 2 is $P = 0.028 < 0.05$. It is clear that considerable time was needed to review the part where the arousal level was low. Therefore, carefully reviewing those parts of the lecture during a low arousal level leads to an efficient review of the lecture. Moreover, there is a difference of 106 seconds in the time taken for Group 1 and Group 2 to answer Problem B ($P = 0.136$). We can see from the results that the learning time for parts during the low arousal level was reduced by using our system. Therefore, it follows that the lecture view system is useful for achieving an efficient and effective review. However, there is not a significant difference between the response time for Group 1 and Group 2 to answer Problem B. The cause is that sample numbers are small. Therefore, we plan to conduct experiments by many students.

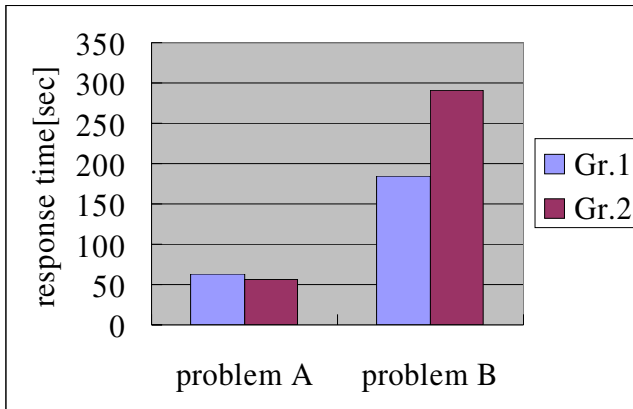


Fig. 8. Response time to problems: Gr.1 (with review support functions), Gr.2 (without review support functions) Problem A (from Part A), Problem B (from Part B)

4 Conclusions

We developed a unique lecture view system that contains three support functions using students' arousal level and lecture information and we evaluated the system through a preliminary experiment. The learning time was shortened at the low arousal level period, confirming that the proposed system is effective as a lecture review system for self-learning. In the future, we plan to improve the system through regularly conducted experiments. In addition, we intend to evaluate the user interface of our system.

Acknowledgments

This research was supported by a 21st Century COE Program Grant for "Intelligent Human Sensing" and the Matsushita Education Foundation.

References

1. Lee, Y. and Geller, J.: A Collaborative and Sharable Web-Based Learning System. *International Journal on E-Learning*. Vol. 2, No. 2, Norfolk, VA: AACE (2003) 35-45
2. Helic, D., Krottmaier, H., Maurer, H., and Scerbakov, N.: Enabling Project-Based Learning in WBT Systems. *International Journal on E-Learning*. Vol. 4, No. 4, Norfolk VA: AACE (2005) 445-461.
3. e-Learning Hakusyo 2002/2003, ALIC, Ohmsha, Tokyo (2003)
4. Tamaki, M., Kuwabara, T., Yamada, K., Muto, M., and Shimura, A.: Technology Trends in Human Interaction Conscious e-Learning, Vol. 86, No. 11, The Institute of Electronics, Information and Communication Engineers (2003) 826-833
5. Producer 2003, Microsoft, <http://www.microsoft.com/office/powerpoint/producer/prodinfo/features.msp>
6. EZ-presenter, Hitachi Advanced Digital, <http://www.hitachi-ad.co.jp/ezplat/index.html>
7. Ikeya, H., Sato, K., Yamada, H., and Nitta, T.: Activating questions and answers between a teacher and students using a web-based lecture system. 66th Information Processing Society of Japan (2004) 401-402
8. Takagi, T. and Miyasaka, E.: A Speech Prosody Conversion System with a High Quality Speech Analysis-Synthesis Method, *EUROSPEECH93*, Vol. 31, No. 3 (1993) 995-998
9. Sekiguchi, Y., Suzuki, N., Aono, M., Shinohara, S., Nakauchi, S., Horihata, S. and Yasuda, Y.: Prototype System for Intelligent Human Sensing: Using EEG, ECG, EOG and TIP to Detect the State of Mental Concentration and Somnolence, *Proceedings of the Fourth Symposium on Intelligent Human Sensing IHSS2006*, (2006) 27-30

Hybrid Approach of Augmented Classroom Environment with Digital Pens and Personal Handhelds

Motoki Miura and Susumu Kunifuji

School of Knowledge Science
Japan Advanced Institute of Science and Technology
{miuramo, kuni}@jaist.ac.jp
<http://css.jaist.ac.jp/~miuramo/>

Abstract. We have been developing a system *AirTransNote*, a computer-mediated learning system that employs digital pen to realize paper-centric augmented classroom. Although the approach was sophisticated, it restricted feedback effects which potentially improve learning. To maximize the feedback effects of the system, we present a hybrid approach of augmented classroom environment. We classified the type of feedback loops in terms of the hybrid approach and improved the system to accomplish the functions by adding (1) note browsing interface for handheld, (2) worksheet editor for teachers, (3) handwriting character recognition engine for versatile use, and (4) HTTP embedded server function for flexible reference of collective results and notes. We also conducted a feasibility study and investigated the effectiveness of the immediate feedback under the hybrid approach.

1 Introduction

Mobile and wireless networking technologies accelerate the activities of computer-mediated learning and CSCL as stated by Roschelle and Pea[1]. The potential of these technologies is recognized and investigated[2], and many systems [3],[4],[5],[6] have been proposed to improve the effect of collaborative learning by sharing student notes and annotations in a classroom environment. Most of the systems employ digital devices such as tablet PCs and handheld computers to capture notes.

In contrast, we have developed a system *AirTransNote* (ATN) for reducing the additional effort for students. ATN employs paper-centric approach[7] to augment the learning and teaching activities in the traditional classroom by sharing notes. ATN can capture a student's note written on a regular paper with a digital pen device and transmit the note to the teacher's computer immediately. Moreover, our system provides a remote for teacher to encourage the natural interaction in a physical classroom space[8]. The remote attaches a RFID reader to provide intuitive selection of student's note by touching it. Typical scenario of learning with ATN is also described in [8].

Though ATN employs handhelds for transmitting student notes, we have not fully considered the effect of the handhelds because we mainly focused on the paper-centric approach. However, we have noticed that the effect of ATN system can be improved when we make full use of the handhelds. When the full support of handhelds is realized, ATN can be used to activate collaborative and cooperative learning among students as well as the automatic personal reply. In this paper we discuss the characteristics of ATN in terms of hybrid approach which utilizes both paper and handheld as media of CSCL activities. Additional functions developed toward the hybrid approach are also described.

2 AirTransNote as Hybrid Approach

We considered that the paper-centric approach[7] was epoch, but “transmitting student’s note to a teacher” was insufficient for evolution of the learning environment since some feedback loops were not established. To magnify the effect of the note sharing function, the learning system should afford the feedback loops as much as it can. We enumerate the type of the feedback loops and summarize advantages of student handhelds we employed as follows.

(1) *Student Self Feedback Loop.* Maruyama and colleagues have adopted Anoto-based pens¹ to collect students’ activities in English as second language class[9]. In order to synchronize notes, the student taps on a “send” region of paper or puts the pen on the cradle. Using Anoto alone is fairly simple, but students neither confirm nor fix the note captured and transmitted. Handheld can supply the students functions not only revision of notes but also reflection, because the note displayed on the handheld device (different visual to the note itself) might somewhat arouse consciousness with meta perspective. In either case, the revision function will ease students’ anxiety about publicizing notes.

(2) *Between System-student Feedback Loop.* With the handheld the student can get individual feedbacks from the system. Even if the feedback is simple like correct/incorrect messages and sounds, the students will be encouraged to progress. The feedback is similar to self-checking of the result with answer, but a feeling of satisfaction for automatic checking will be higher than that of self-checking.

(3) *Between Teacher-student Feedback Loop.* The between teacher-student feedback loop is established by guidance from the teacher to the students. Before the guidance, the teacher should recognize the status of the students. AirTransNote provides note browser and summarization list. From the note browser, the teacher can annotate the notes. The annotation is transferred and displayed to the student handhelds. From the summarization list, the teacher can grasp the progress of the students. Thus the teacher can guide students by transmitting advisory annotations as well as contacts directly.

¹ <http://www.anotofunctionality.com/navigate.asp>

(4) *Between Students Feedback Loop.* The handhelds for student have great potential to freely exchange the notes among the students and to score them. The collaborative learning activities will work effectively if the teacher controls the transactions of the note sharing activities properly.

3 System Improvement

In order to maximize the number of available feedback loops and to enhance the learning effect, we have improved our ATN system.

3.1 For Student Self Feedback Loop

ATN Manager, software which runs on a teacher’s PC, can overlay students’ notes on page images, whereas ATN Transmitter (prior software which runs on a student handheld to send notes) displays the notes without page images. As a matter of course, showing page images is more appropriate for the student to recognize the note. Therefore we have developed ATN Mediator as a successor of the transmitter. Figure 1 shows the snapshot of the handheld. A student can see both notes and annotations with the page image. Thus the student can grasp the context of notes and recognize what the student has sent. The view can also be used to confirm whether a calibration process goes well or not. The mediator scrolls the note view to make visible the area where the student writes or taps on the paper. Thus the trouble of view control can be reduced.

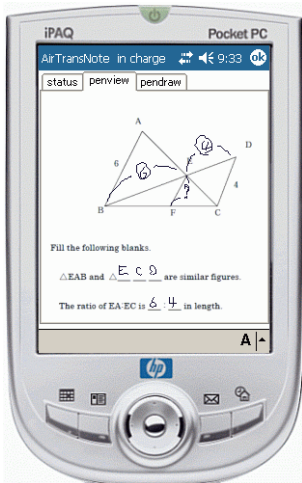


Fig. 1. Note browsing interface for handheld (ATN Mediator)

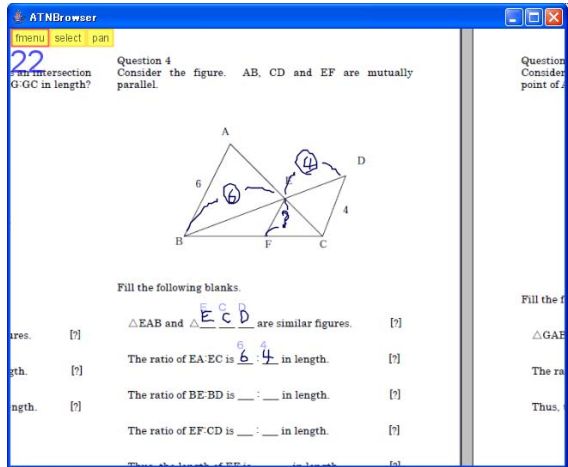


Fig. 2. Note browsing interface for teacher’s PC (ATN Browser)

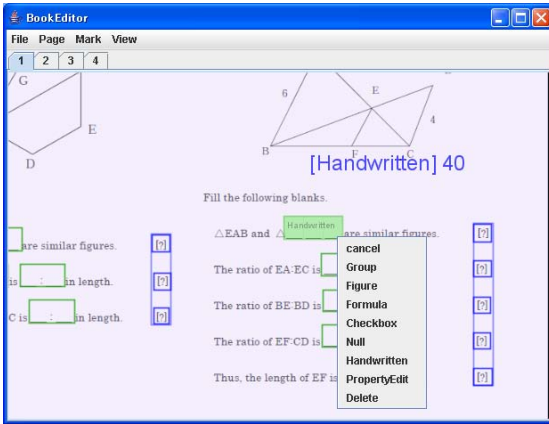


Fig. 3. Worksheet Editor for teacher

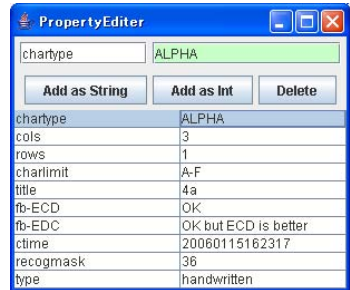


Fig. 4. Region Property Editor

3.2 For Between System-Student Feedback Loop

The between system-student feedback loop is significant for computer-mediated learning environment since it can provide immediate responses to the students. We have already implemented a mechanism which recognizes simple check-boxes. But in order to amplify the effect of note, we have embedded a handwritten character recognition engine developed at Nakagawa laboratory, Tokyo University of Agriculture and Technology[10] with our system. ATN Manager can interpret student's handwritten notes and return feedback to the student's handheld. The result of recognition can also be shown on the screen (Figure 2).

To make the system recognize notes on a paper, the teacher should prepare a worksheet which contains specifications on how the notes are processed. We have developed a page editor to design the worksheet. Figure 3 shows the interface of the page editor. The teacher first creates a new page by dragging a worksheet image file (jpeg, png, and gif) and dropping it to the window. Then the teacher draws rectangles to set regions on the page and specifies the type of recognizer for the regions. After that, the teacher enters properties of the region at the editor window (Figure 4). The region properties consist of title, preferences for recognizer, and feedbacks. When the result matches with the properties beginning with 'fb-' prefix, pre-defined feedback is generated. In substitution for the editor, the region properties can be specified by loading a text file.

3.3 For Between Teacher-Student Feedback Loop

The teacher through the teacher's remote as well as the PC should refer a collective result of recognition. We could implement an original browsing interface for the collective result from the remote PDA, but we decided to utilize a standard web browser to watch the result. The standard web browser is suitable for watching a result whose structure may dynamically change. For that reason,

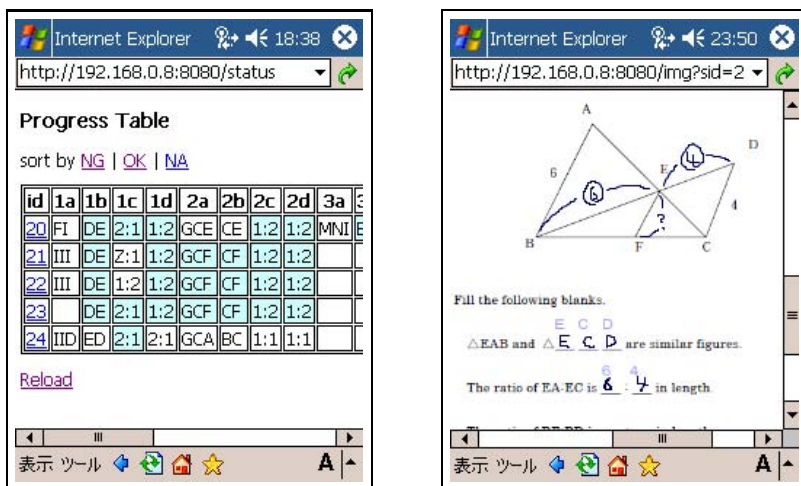


Fig. 5. Collective result (left) and note image (right) rendered by standard web browser on Pocket PC

we have combined a web server with AirTransNote. We chose OOWeb² HTTP server written in Java to embed. The embedded web server responses latest collective results and note images in png format when the teacher requests from a browser (see Figure 5). The teacher can find the students in need of guidance by sorting the order of the table rows by number of incorrect or blank, and then browse the note image from the remote.

3.4 For Between Students Feedback Loop

The students can browse other students' notes as the same HTTP way through their handheld. In addition to this, the system can send a URL to the handheld to open a web-page as feedback. Therefore the students can share their notes or ideas and encourage them by voting on each others.

4 Related Works

EduClick[11] employs wireless IR remotes for collecting student's answers in class. The simplicity of EduClick will contribute casual and wider use, but its shortcoming is that it only accepts selection-based responses. Singh et al. [12] developed a collaborative note taking system in which the students could reuse the text in teacher's slide or another member's note by tapping to reduce the typing. Our system employs PDA as a device for transmitting note. Though the real estate problem even exists with our system in displaying comments, the digital pen can alleviate the efforts of inputting. Tallyn et al. [13] also pointed out the advantages of augmented paper media in educational settings.

² <http://ooweb.sourceforge.net/>

A Tablet PC is similar to a digital pen device in collecting time-stamped notes. The Tablet PC is more flexible than the digital pen in that it can modify the notes and its properties such as colors and thickness. Also the display can reflect the feedback with overlays. Thus many systems have been developed [14],[4],[2]. However, Tablet PC is large, heavy and expensive. Moreover, it does not provide the natural feeling of taking notes, which is significantly different from taking notes on real paper. We consider that familiarity is more important than flexibility especially in early stage of the computer-supported lecture.

5 Feasibility Study

We conducted a feasibility study to collect the comments of the ATN system with sound feedback and direct touching interface to select note by teacher's remote. We collected 20 volunteer participants all of them were graduate students. We prepared two worksheets, geometry test and IQ quiz, with replies according to the answer. When the answer is correct, the mediator plays a "ding" sound; otherwise, it plays a low buzz sound immediately. The teacher utilized the remote to close up to the student's answer during a reviewing session after the each test. We spent 15 minutes for test session and 5 minutes for reviewing session for each worksheet. Figure 6 (a) and (b) show scenes of the study.



Fig. 6. Scene of the feasibility study

Lessons Learned. Figure 7 shows the questionnaire items and the results. Regarding the immediate transferring function of notes, 75% of participants were satisfied (Q1). The result of Q2 shows many participants agreed with the effectiveness of sharing notes. Regarding the stress in sharing, in comparison to our previous experiment on mathematical class on 40 high school students[7], few participants mentioned stress in this feasibility study. We can pose a hypothesis that the following four factors would affect either aggravation(+) or alleviation(-) of the stress: (1+) Difficulty and formality of the problem, (2+) Smaller answer area which prevents rewriting, (3+) No confidence without feedback, and (4-) Physical note selection by direct touching operation. We have to investigate the effect of each factor in further experiments.

We had adopted sounds for main feedback in this study. 55% of participants thought that the feedback is important (Q3) because it helps confirm the result.

But some participants said that the system is very noisy, so they cannot concentrate. Considering the higher acceptance rate in (Q1), the real-time feedback itself is supported. However, we should carefully design the feedback, especially volume levels in sound playback.

From the teacher’s view, the remote was useful for closing up to a student’s note immediately. The teacher could easily select a student and let the student explain the answer for the class. Though the teacher tended to select students who sat on aisle seats, this tendency is covered by relaxed seat arrangement and teacher’s mind.

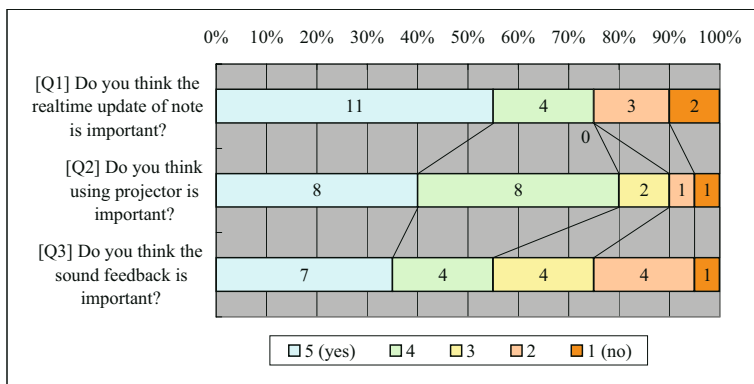


Fig. 7. Result of the questionnaire

6 Conclusion

The paper-centric design of ATN is effective since it frees one from PC operation. However, the power of the devices introduced is underutilized while the system strictly applies the design criterion. In order to maximize the power and effects, we propose a hybrid approach which utilizes both a digital pen and a handheld for collaborative learning. We have considered four feedback loop types as criteria for hybrid design and improved the ATN system in terms of the criteria. Moreover, we conducted a feasibility study to clarify the effectiveness of the immediate feedback. Though the activity of the study only includes publicizing notes on a large shared screen, we will evaluate the effects of collaborative learning activities including a “between students” feedback loop. We will also apply the hybrid system to problem-posing style learning activity because the personal handheld can support the advanced problem-posing works.

Acknowledgment

I would like to express my heartfelt gratitude to Mr. Hideto Oda and Prof. Masaki Nakagawa at Tokyo University of Agriculture and Technology for offering

their skills and software. Part of this work was supported by a grant-in-aid for Scientific Research (15020216, 17011028).

References

1. Roschelle, J., Pea, R.: A walk on the WILD side: How wireless handhelds may change CSCL. In: Proc. of CSCL 2002. (2002) 51–60
2. Liu, T.C., Wang, H.Y., Liang, J.K., Chan, T.W., Yang, J.C.: Applying Wireless Technologies to Build a Highly Interactive Learning Environment. In: IEEE Int. Workshop on Wireless and Mobile Technologies in Education (WMTE'02). (2002) 63–70
3. Anderson, R.J., Hoyer, C., Wolfman, S.A., Anderson, R.: A Study of Digital Ink in Lecture Presentation. In: Proc. of CHI 2004. (2004) 567–574
4. Kam, M., Wang, J., Iles, A., Tse, E., Chiu, J., Glaser, D., Tarshish, O., Canny, J.: Livenotes: A System for Cooperative and Augmented Note-Taking in Lectures. In: Proc. of CHI 2005. (2005) 531–540
5. Davis, R.C., Landay, J.A., Chen, V., Huang, J., Lee, R.B., Li, F., Lin, J., III, C.B.M., Schleimer, B., Price, M.N., Schilit, B.N.: NotePals: Lightweight Note Sharing by the Group, for the Group. In: Proc. of CHI '99. (1999) 338–345
6. Yoshino, T., Munemori, J.: SEGODON: Learning Support System that can be Applied to Various Forms. In Ghaoui, C., ed.: E-Education Applications: Human Factors and Innovative Approaches, Information Science Publishing (2004) 132–152
7. Miura, M., Kunifuji, S., Shizuki, B., Tanaka, J.: Augmented Classroom: A Paper-Centric Approach for Collaborative Learning System. In: Proc. of 2nd Int. Symposium on Ubiquitous Computing Systems (UCS2004), LNCS 3598. (2004) pp.104–116
8. Miura, M., Kunifuji, S., Shizuki, B., Tanaka, J.: AirTransNote: Augmenting Classrooms with Digital Pen Devices and RFID Tags. In: Proc. of IEEE Int. Workshop on Wireless and Mobile Technologies in Education (WMTE2005). (2005) 56–58
9. <http://www.cec.or.jp/e2a/other/04PDF/b1.pdf> (in Japanese; the URL of the project page is <http://www.hitachi-ks.co.jp/cec/index.html>).
10. Nakagawa, M., Akiyama, K., Tu, L.V., Homma, A., Higashiyama, T.: Robust and Highly Customizable Recognition of On-Line Handwritten Japanese Characters. In: Proc. of the 13th Int. Conf. on Pattern Recognition (ICPR'96). Volume 3. (1996) 269–273
11. Huang, C.W., Liang, J.K., Wang, H.Y.: EduClick: A Computer-Supported Formative Evaluation System with Wireless Devices in Ordinary Classroom. In: Proc. of Int. Conference on Computers in Education. (2001) 1462–1469
12. Singh, G., Denoue, L., Das, A.: Collaborative Note Taking. In: 2nd IEEE Int. Workshop on Wireless and Mobile Technologies in Education (WMTE'04). (2004) 163–167
13. Tallyn, E., Frohlich, D., Linketscher, N., Signer, B., Adams, G.: Using paper to support collaboration in educational activities. In: Proc. of CSCL 2005. (2005)
14. Iwayama, N., Akiyama, K., Tanaka, H., Tamura, H., Ishigaki, K.: Handwriting-Based Learning Materials on a Tablet PC: A Prototype and Its Practical Studies in an Elementary School. In: Proc. of Ninth Int. Workshop on Frontiers in Handwriting Recognition (IWFHR04). (2004) 553–538

An Interactive Multimedia Instruction System: IMPRESSION for Multipoint Synchronous Online Classroom Environment

Yuki Higuchi¹, Takashi Mitsuishi¹, and Kentaro Go²

¹ Graduate School of Educational Informatics, Tohoku University
27-1 Kawauchi, Aoba-ku, Sendai, Miyagi 989-8576, Japan
{yukix, takashi}@ei.tohoku.ac.jp

² Center for Integrated Information Processing, University of Yamanashi
4-3-11 Takeda, Kofu, Yamanashi 400-8511, Japan
go@yamanashi.ac.jp

Abstract. In order to perform effective synchronous online classes with multimedia educational materials, we have developed interactive instruction system named IMPRESSION, which facilitates interactive presentation of various multimedia materials provided by web server on the Internet. In this paper, we clarify the design concept of our system, and show the design and implementation of it. And we discuss the effectiveness of our system through the practical use in the online class for high school students with high-bandwidth networks.

1 Introduction

In order to develop sufficient educational environments, many instructional theories and methodologies have been proposed and various ITs (Information Technologies) have been introduced into classroom. There also exist many studies proposing the tools facilitate a presentation via the networks. Such recent developments of ITs provide opportunities to design and use new forms of education such as e-learning or online classes. They afford us the flexibility of time and place. It has the possibility of enhancing the students' educational opportunities and outcomes. However, it is far inferior to provide sufficient instructions in actual online environments –in this paper, we especially discuss the capabilities of sharing the digital ink and materials among multiple users– because of the design concept or the function limitations of existing tools.

In order to provide the sufficient capabilities, we have developed an interactive instruction system named IMPRESSION facilitates interactive presentation using educational materials provided on the networks. In this paper, we clarify the design concept of our system, describe design and implementation of it, and discuss the effectiveness of it through the practical use in an online class and comparisons with related works.

2 Instruction System for Online Classroom Environment

In this section, we describe existing tools and clarify the requirement for flexible instructions from instructional and technological viewpoints. We also show the design and implementation of the interactive instruction system we have developed.

2.1 Requirement for Online Classroom Environment

There are several tools providing the online classroom environment [1], [2]. However, most existing tools have some difficulties to conduct flexible class such as existing face-to-face classes in which teachers have used chalk and blackboard. Since most tools are supposed the class in which teachers present educational materials –prepared in slide-sheets form beforehand– sequentially one by one, they may restrict the flexibility of teachers' activities (e.g., handwriting, presentation of materials, operation of them). Anderson et al. said “existing tools used in online classroom environments make extemporaneous teaching difficult in discussion situation, thereby it incurs monotony of class” [3]. We think that above mentions are attributable to the design concept of existing tools, which are supposing that the problems of plans or materials clarified in a class are revised after the class is over. On the other hand, we can forecast that the unexpected events occur during classes enough; students can not understand the class, students ask unexpected questions, and so on. However, these existing tools do not suppose to handle such events appropriately during classes in the design concept.

Requirement from Instructional Viewpoint: Alley et al. said “the traditional presentation slide design oversimplifies the subject matter, and it quickly becomes monotonous for audiences”. They insisted that new slide design should present supporting evidence of the subject matter in a visual way with images, graphs, or visual arrangements of text in their proposing guidelines [4]. Debert wrote about the strategies for revisions that are made in instructional materials (e.g., add illustrations, rearrange the sequence, or change the instructional medium) after formative evaluation [5]. We proposed an instructional design process model taking into account extemporary feedback during classes [6]. Based on these above mentions, the ideal tools are required to allow teachers to use multimedia educational materials for visual aids for the subject matter. Additionally it is required to modify the materials or annotations on demand through formative evaluations during classes. Furthermore, for reflections, it is required to record their instructions during the classes and review them after the class is over.

Requirement from Technological Viewpoint: There are several problems underlying the construction of distributed systems. For example, one is traffic cost. It is popular to employ multicast solution. However, we assume an Internet based online classroom environment. We must employ an application-based multicast solution, rather than the one of IP-based. Next one is consistency maintenance of operations from multiple users. Most systems adopt conflict prevention approach based on locking method. It is also important for users to be able to understand own status –rights or role for operation–.

2.2 Interactive Instruction System: IMPRESSION

In order to conduct effective instruction in face-to-face or synchronous online classes, we have developed an interactive instruction system: IMPRESSION (Interactive Multimedia PREsentation System for Shared Instructional Objects on the Networks) [7].

This system consists of a *terminal for operator*, several *terminals for attendee* and a *lecture server*, as shown in Fig. 1. Each terminal is implemented on Windows XP by Visual Basic .NET 2003 and the server is on Solaris8 by Java2SE 1.5.0_05-b05. Each

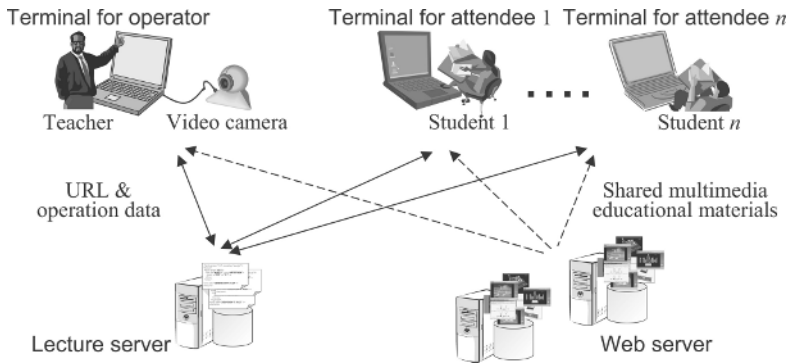


Fig. 1. IMPRESSION system structure

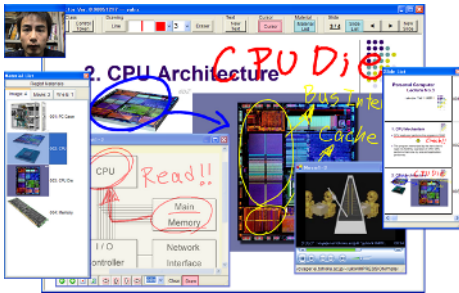


Fig. 2. Snapshot of the terminal for operator

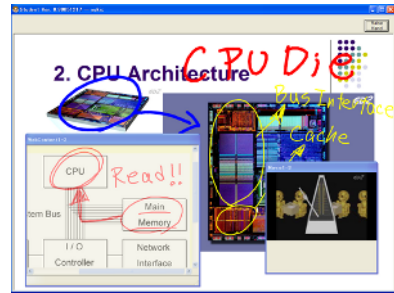


Fig. 3. Snapshot of the terminal for attendee

terminal uses shared multimedia materials provided by several public web servers on the Internet.

Using the terminal for operator, instructors can incorporate multimedia materials (e.g., pictures, video-clips, web pages) from web servers. This can be done anytime: before or during the class. During class, while responding to students' reactions, instructors can select and present appropriate materials from the registered ones to the main display, draw annotations on them, turn over the slide-sheets –registered as a set of pictures– sequentially on the display, and indicate arbitrary spots on the display using a cursor as shown in Fig. 2. Additionally, it can record the whole scene of the class using a microphone and a video camera connected to it –not for video streaming–.

As instructors' operations, both URLs of the used materials and the operation data (e.g., coordinates, width and color of annotations, position and size of picture) are transmitted via the lecture server to the terminal for attendee. It captures materials from the specified web servers and executes operations for them based on the transmitted data. In other words, the data and operations on the terminal for operator synchronize with the terminal for attendee. Therefore, the students can participate in the class while watching the same screen as shown in Fig. 3. The students can also operate the displayed materials and draw annotations on them if the instructor permits them.

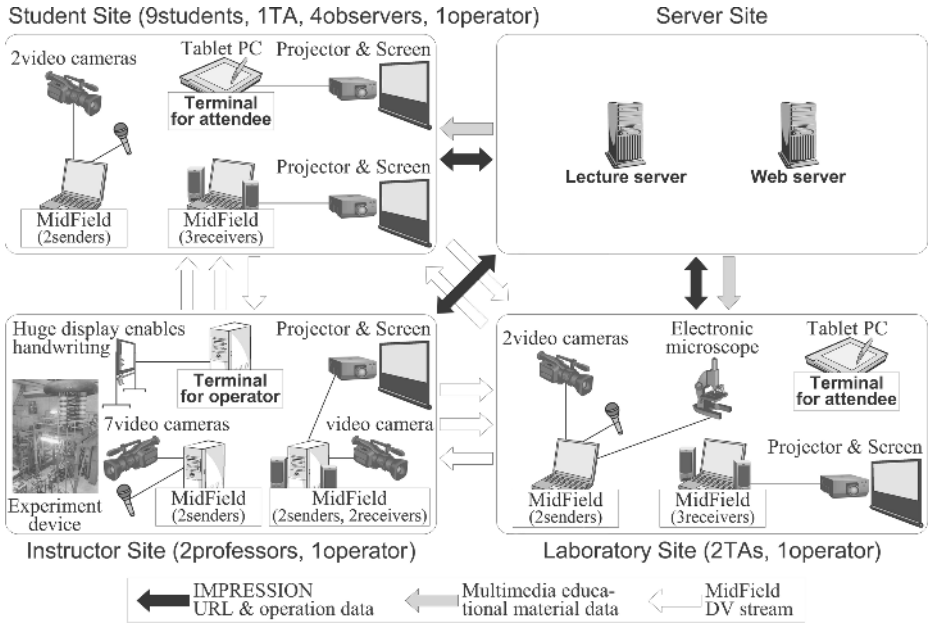


Fig. 4. Used equipment and applications in the class

The permissions consist of three modes: *instructor*, *student*, and *cooperation*. Instructor mode permits instructor using terminal for operator to execute all operations. Student mode permits one of students using terminal for attendee to execute operations without registering and presenting materials. Cooperation mode permits all participants to operate cursor and draw annotation without eraser operation –other operations are locked–. These modes are mutually exclusive relations, and switched by iconic interface on terminal for operator. These limitations of operations are to prevent the conflict.

Transmitted URLs and operation data are recorded sequentially and stored as lecture data onto the lecture server in XML form. With the lecture data, each terminal can receive used multimedia materials from the web servers and recorded video data from the terminal for operator, reproduce the operations along with timeline anytime.

In summary, the IMPRESSION system allows instructors to leverage timely multimedia data during the class. The instructors can react quickly and effectively to modify an instructional plan based on formative evaluation in class. Moreover, the instructors can conduct reflections of their instructions after class. The students can also review the class data to enrich their understanding of the class contents.

3 Experiment

We performed the experiment of the actual online class with implemented system and investigated the effectiveness of it. In this section, we describe the outline of the experiment and show the result about network traffic and the comments from participants.



Fig. 5. A classroom view in instructor site



Fig. 6. A classroom view in student site

3.1 Outline of the Experiment

The class was a high school–university cooperative class, which assumes using a huge device for physical experiments installed at Tohoku University Japan. Fig. 4 depicts the location of participants, used equipments, and applications. We connected among four sites –three were *instructor*, *laboratory*, and *server site* in a university at Miyagi prefecture. Last one was *student site* at Tokyo where is about 300km away from Miyagi– with JGN-II (Japan Gigabit Network-II) [8]. The participants were two university professors, nine high school students, three TAs (teaching assistants), four observers –they are high school teachers–, and three operators. We used IMPRESSION in order to share the digital ink and materials among these sites. We also used videoconference application named MidFiled [9] with DV (digital video) format.

3.2 Result of the Experiment

The class of 180 minutes has been performed. Fig. 5 shows the scene in instructor site where the professor is annotating on the screen of IMPRESSION using 50-inch diagonal display enable handwriting. Fig. 6 shows the scene in student site where images of IMPRESSION and MidField are projected on the two screens, in which the left one is of IMPRESSION, the right one is of MidField –in which are instructor, experiment devices from instructor site and the scene from laboratory site–.

The main objective of this class was brushing-up the students' knowledge already learned. Thus, the professors and TAs considered it difficult to prepare a complete content set of its educational materials beforehand. Therefore, they decided to design the class as a discussion class: the class had progressed while asking some questions to the student promptly to determine their comprehension level of the subject and modify the class plan during the class.

Network Traffic: In order to confirm an ideal network environment of IMPRESSION, we calculated the network traffic between laboratory site and server site using network protocol analyzer: Ethereal [10] at laboratory site. Fig. 7 shows the traffic both of up and down stream of all data by IMPRESSION and MidField. Fig. 8 shows the traffic both of up and down stream of URL and operation data by IMPRESSION.

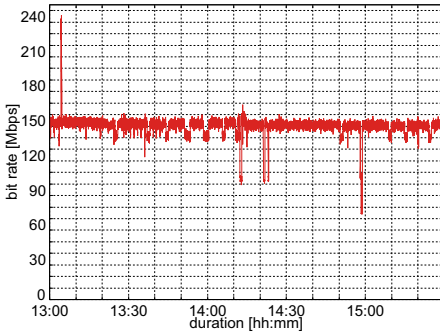


Fig. 7. Network traffic of all used applications

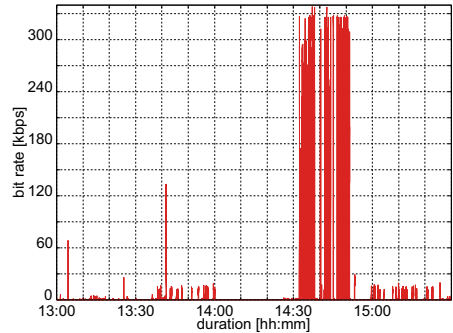


Fig. 8. Network traffic of IMPRESSION

The scale of the graph shown in Fig. 7 is *Mbps*. The peak just after the class started is attributable to that the terminal for attendee in laboratory site captured materials from web server based on transmitted URL. The scale of the graph shown in Fig. 8 is *kbps*. The peak around middle of class is attributable to that the TAs in laboratory site operated the cursor of IMPRESSION. In other part of the class, participants also did several operations, but because of communication protocol, the cursor function needs more bandwidth than the other functions. In other words, we need about 300kbps bandwidth when we operate cursor function very often, but we need about only 10kbps otherwise.

On the other hand, we did not evaluate network delay with quantitative method, but the qualitative method from participants confirms that there is no problem about it.

Comments from Participants: After the class, we asked participants several open-ended questions. We collected negative comments for IMPRESSION such that “the operations for Macromedia Flash components contained in web page materials does not share with other site” by an operators and “it is unavailable that the annotation to the surface of video-clip materials” by a teacher. On the other hand, we also collected positive comments from participants for the functions of IMPRESSION allow us an interactive presentation using several multimedia materials and annotation for them. Furthermore, in online classroom environment, it is not understandable where instructor explain of, but participants were able to understand it enough by shared cursor functions.

Pointed out the usability problems of some functions, we were able to confirm that IMPRESSION is usable in actual online classes. However, we also collect comments from a observer such that “we were not able to participate to the online class with same way as face-to-face class, thereby it is necessary to study instructional methodology for online class environment”. It is consistent with students’ comment such that “we have not been accustomed to participate to online class, we were not able to communicate well with the persons on other site”. Thus, we should explore the instructional methodology and reassess the capabilities of used systems. For example, in face-to-face class environment, participants turn their gaze on the arbitrary speaker by themselves. On the other hand, in the online class, instructors should turn participants’ gaze on the speaker. Thus, we need a seamless method to turn their gaze; it is left for future work.

4 Related Works

There are related tools such as Classroom Presenter, Ubiquitous Presenter, and InkBoard which have similar functions with ones of our IMPRESSION. In this section, we clarify the differences between these tools and our system.

First ones are Classroom Presenter [11] and its derivative system: Ubiquitous Presenter [12]. These are developed for both face-to-face and synchronous online classes using slide-sheet materials and annotations on them. Classroom Presenter also facilitates participants note-taking, which are sent back to the instructor if necessary. Ubiquitous Presenter inherits functions of it, and is used by participants using Tablet PCs. For non-Tablet users, the annotations and control are able to be viewed using public web browsers, but drawing annotation or the control are not available from it.

Second one is InkBoard [13] which provides brainstorming environment by multiple users in co-design artifact context. It facilitates not only annotations on prepared pictures, but also distinction of participants' annotations and replay the stroke of them individually on demand along with the timeline. Furthermore, InkBoard and above Classroom Presenter provide teleconference capability by the videoconference application: ConferenceXP [14]. InkBoard uses the APIs. On the other hand, Classroom Presenter is selectable as a presentation module of it.

Above tools are useful for cases in which instructors use only slide-sheets materials. However, in actual classes, instructors often use picture or video-clip materials if necessary. Therefore, participants using above tools must not only use additional tools for such materials in class, but also prepare all materials to use in class beforehand. Thus, instructors can not present suitable materials immediately through formative evaluation during class. In addition, they can not conduct reflections correctly by play backing the class including picture or video-clip operations. On the other hand, with IMPRESSION, instructors can use several multimedia materials, they can review the class correctly integrated with recorded audio and video. However, several functions –note-taking or viewing them using web browser– have not been provided yet; it is left for future work.

When we handle multimedia materials, network traffic problem usually arises. For example, it may be thereby incurred by the server system that provides the materials. However the terminal of IMPRESSION receives materials to be used from several web servers, which are stored as caches, network traffic does not cost any one certain server system. In addition, materials themselves are not transmitted among terminals, thereby is required not so much bandwidth between the terminal and lecture server.

5 Conclusion

In this paper, we clarified the design concept of our interactive instruction system named IMPRESSION, showed design and implementation of it. It facilitates interactive presentation using shared multimedia educational materials provided by public web servers on the Internet. It also facilitates play backing the class after it is over. In addition, we discussed the effectiveness of it through the practical use in a multipoint synchronous online class and the comparisons with related works.

The practical use confirms that it had several problems yet to be used flexibly or smoothly in online classes. However, the function of interactive presentation using

shared materials is peculiar feature of our system. The drawback of our system is that it has no videoconferencing capability. In order to realize the smoothness and effectiveness of online class environment, our future work is to enhance our system to integrate it with videoconference system.

Acknowledgment

We appreciate Prof. Shin Iwasaki of Tohoku University, Japan for providing us the opportunity to evaluate our system, and the participants of the experimental class. We are also grateful to Prof. Kazuo Nagano of the University of the Sacred Heart, Tokyo, Japan for permitting our use of materials. Materials we used for snapshots included in this paper are provided at the web site “Jyohokiki To Jyohosyakai No Sikumi Sozaiyu” (<http://www.kayoo.org/home/mext/joho-kiki/> in Japanese). This study was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. J. A. Brotherton, and G. D. Abowd: Lessons learned from eClass: assessing automated capture and access in the classroom. *ACM Trans. on CHI*, 11(2) (2004) 121–155
2. E. W. Johnson, D. Tougaw, J. D. Will, and A. Kraft: Distance learning: teaching a course from a remote site to an on-campus classroom. *Proc. of ASEE/IEEE FIE2005*, (2005) F1H-1–6
3. R. Anderson, J. Beavers, T. VanDeGrift, and F. Videon: Videoconferencing and presentation support for synchronous distance learning. *Proc. of ASEE/IEEE FIE2003*, (2003) F3F-13–18
4. M. Alley, M. Schreiber, and H. Muffo: Pilot testing of a new design for presentation slides to teach science and engineering. *Proc. of ASEE/IEEE FIE2005*, (2005) S3G-7–12
5. C. B. Weston: Formative evaluation of instructional materials: an overview of approaches. *Canadian Journal of Educational Communication*, 15(1) (1986) 5–17
6. Y. Higuchi, T. Mitsuishi, and K. Go: A methodology and a system for scenario-based instructional design of interactive instruction with multimedia educational materials. *Proc. of HCI 2005*, (2005) CD-ROM
7. Y. Higuchi, M. Oyamada, T. Mitsuishi, and S. Iwasaki: Design and evaluation of interactive instruction system: IMPRESSION for a distance class. *Proc. of IEEE ICICS 2005*, (2005) 841–845
8. JGN2 Home Page: <http://www.jgn.nict.go.jp/e/>
9. K. Hashimoto, and Y. Shibata: Extendable media stream mechanisms for heterogeneous communication environments. *Proc. of IEEE ICITA'2005*, (2) (2005) 751–754
10. Ethereal: A Network Protocol Analyzer: <http://www.ethereal.com/>
11. R. Anderson, R. Anderson, B. Simon, S. A. Wolfman, T. VanDeGrift, and K. Yasuhara: Experiences with a tablet PC based lecture presentation system in computer science courses. *Proc. of ACM SIGCSE'04*, (2004) 56–60
12. M. Wilkerson, W. Griswold, and B. Simon: Ubiquitous presenter: increasing student access and control in a digital lecturing environment. *Proc. of ACM SIGCSE'05*, (2005) 116–120
13. H. Ning, J. R. Williams, A. H. Slocum, and A. Sanchez: InkBoard – tablet PC enabled design-oriented learning. *Proc. of IASTED CATE 2004*, (2004) 154–160
14. Microsoft Research ConferenceXP Project: <http://www.conferencexp.net/>

A System Framework for Bookmark Sharing Considering Differences in Retrieval Purposes

Shoichi Nakamura¹, Maiko Ito², Hirokazu Shirai², Emi Igarashi²,
Setsuo Yokoyama³, and Youzou Miyadera³

¹ Fukushima University, Department of Computer Science and Mathematics,
Kanayagawa 1, Fukushima, 960-1296, Japan
nakamura@sss.fukushima-u.ac.jp

² Tohoku Gakuin University, Department of Computer Science,
2-1-1, Tenjinzawa, Izumi, Sendai, 981-3193, Japan

³ Tokyo Gakugei University, Division of Natural Science,
4-1-1, Nukui-Kita, Koganei, Tokyo, 148-8501, Japan
{miyadera, yokoyama}@u-gakugei.ac.jp

Abstract. Bookmarking is a popular way to manage retrieval results when searching through web documents, and sharing bookmarks among users can increase information retrieval efficiency. However, a crude sharing style that makes no distinction between retrieval purposes will hinder effective bookmark sharing and any information retrieval based on it. To prevent this problem, we have developed a support system for bookmark sharing which considers differences in search purposes and user groups. We also provide a method for visualizing the relationships among bookmarks.

1 Introduction

Search engines are used to find useful documents from among the enormous number available through the Internet, and bookmarking is an effective method for managing the result of such information retrieval. Moreover, the importance of cooperative information searches and the sharing of results is increasing with the spread of collaborative intellectual endeavors in a network environment. Thus, there is a growing need for effective support of bookmark sharing.

Previous research related to bookmark sharing has been reported. Nakajima et al. developed a system which supports the sharing of web browsing histories through the process of generating bookmarks [1]. Although the sharing process contributes to efficient retrieval, this system can be applied only within a limited scope as its implementation policy is for a specific retrieval purpose. Support systems to automatically collect and share bookmarks [2] [3] and a commercial bookmark sharing service [4] have been also developed. The shortcoming of these systems is that any bookmarks are shared at all times regardless of retrieval purposes, but personal bookmarks usually include information for many different retrieval purposes. A more useful method would be to extract only bookmarks relevant for the retrieval purpose and share these among the interested users.

We have developed a support system to extract useful bookmarks according to the retrieval purpose and share these within an appropriate user scope. Moreover, we

have developed a method for visualizing the relations among bookmarks. This visualization method is implemented as a function in the support system because it is important to grasp the relations among bookmarks to allow efficient information retrieval and practical use of the retrieval results. Thus, this support system enables users to retrieve useful documents and share results more effectively than with existing methods.

2 Bookmark Sharing Support Considering the Retrieval Scene

2.1 Retrieval Scene

First, we define what constitutes a bookmark for the purpose of this research. A bookmark consists of a title, a URL, and attributes. Attributes currently include the retrieval purpose, user group, registration date, data on the most recent referral, number of referrals, and link information. Link information means the hyperlinks described in the document and the total number of links.

Next, we need a way to regulate the features of information retrieval activities and bookmark sharing to avoid the current problems in bookmark sharing. Generally, the retrieval purposes of each user differ depending on the situation. The presentation of excessive bookmarks irrelevant to the retrieval purpose will hinder effective retrieval activities. This problem becomes more serious when many users want to share bookmarks. Moreover, public display of irrelevant bookmarks to other users may lead to problems concerning privacy. Therefore, proper control of the user scope for sharing is important. In information retrieval, each situation can be thought of as a scene that consists of a search purpose and a user group. In this research, we focus on this scene, which we call a retrieval scene.

It is also important that the user be able to satisfactorily grasp the relations among bookmarks for effective information retrieval and practical use of the obtained web documents. However, existing bookmarking methods present the titles of documents in a simple list style, making it difficult for users to understand the relations. When bookmarks are shared, understanding the relations is especially difficult since unknown bookmarks registered by other users are included.

To make bookmark sharing more useful, we therefore have to solve three problems: (1) the difficulty of referring to and using documents that are relevant to the retrieval purpose; (2) the difficulty of properly controlling the user scope for sharing; and (3) the difficulty of understanding the relations among bookmarks.

2.2 Support Policy

To solve problems 1 and 2, we developed a support system (Fig. 1) that extracts and presents useful bookmarks according to each retrieval scene, and also supports management of retrieval purposes and user groups. To solve problem 3, we implemented a method of visualizing the relations among bookmarks from various viewpoints as a system function. With this system, users can easily manage their retrieval purposes and share extracted bookmarks according to those purposes. Moreover, the visual presentation of the relations among bookmarks enables users to quickly grasp them.

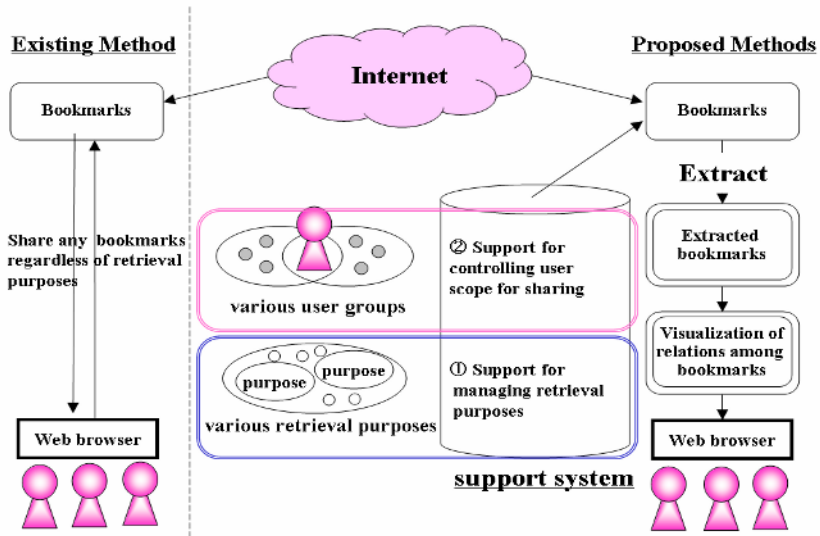


Fig. 1. Outline of support system

3 Support System

3.1 System Design

We designed the system to meet the following requirements.

a: Basic requirements

- (a1) Easy registration and editing of bookmarks
- (a2) Easy referral to web documents indicated by bookmarks

b: Bookmark sharing based on the retrieval scene

- (b1) Easy management of retrieval purposes
- (b2) Easy management of user groups
- (b3) Extraction and presentation of bookmarks according to the retrieval purposes and user groups

c: Easy understanding of the relations among bookmarks

- (c1) Visual presentation of relations among bookmarks
- (c2) Visualization supporting various viewpoints

d: Easy set up of the support system not requiring special knowledge or tasks

To satisfy these requirements, we implemented the following functions.

The bookmark management function supports the registration, editing, deletion, and preservation of bookmarks. It also enables the extraction of bookmarks according to the retrieval scene.

The retrieval purpose management function supports the creation of retrieval purposes and the corresponding settings. When users start bookmark sharing with this system, they begin by creating a retrieval purpose with a convenient name. The purpose thus created is used to extract bookmarks according to the retrieval scene.

The user group management function supports the creation of user groups and the corresponding settings. When users start sharing bookmarks or wish to change the

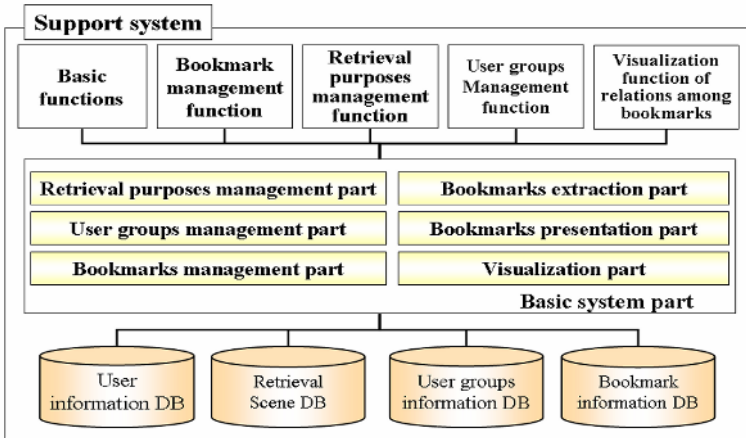


Fig. 2. Configuration of a support system

user scope, they create a user group with an appropriate name and decide on the group members according to the retrieval purpose. The created user group is used for the bookmark extraction according to the retrieval scene.

The bookmark relation visualization function shows the relations among bookmarks based on various visualization styles (described in Section 3.2). Users can observe the relations from different viewpoints according to each situation.

The basic functions support the management of user information for registering and editing bookmarks and for presenting the web document indicated by bookmarks.

To realize these functions, our support system consists of a retrieval purposes management part, a user groups management part, a bookmarks management part, a bookmarks extraction part, a bookmarks presentation part, and a visualization part (Fig. 2). In addition, the basic system part coordinates these parts and provides a system framework.

3.2 Visualization of Bookmark Relations

User Requirements

Properly grasping the relations among bookmarks makes information retrieval and the sharing of results more efficient and effective. However, the number of bookmarks generally increases in proportion to the progress of information retrieval activities. This makes it difficult to keep track of the contents of web documents indicated by bookmarks when the bookmarks are conventionally presented in list style with only a bookmark title. This is particularly a problem when bookmarks are shared among users since the bookmarks include information created by other users.

Research on the visualization of relations among web documents has led to a visualization method focusing on link connections and an interactive system based on this method [5], a visualization method based on browsing histories [6], and a conic visualization approach based on the semantic relation degree among documents [7]. However, with these approaches the characteristics of bookmarks created by users and

various viewpoints important for sharing are not considered. In this research, the visualization method draws the relations among bookmarks as a graph where nodes and edges respectively represent the bookmarks and the links connecting them.

User requirements with regard to understanding the relation among bookmarks can be met in the following ways based on our experience. Users can benefit from observing the relation among bookmarks (1) from a viewpoint focusing on one factor, (2) from a viewpoint connecting two factors, (3) by centering on the marked document, (4) based on the relation degree among bookmarks, and (5) by focusing on link connections.

In our support system, we provide several visualization styles based on various viewpoints to help users understand the relations among bookmarks.

Relations among Bookmarks

The degree of relation between two bookmarks x and y is calculated as

$$relDeg(x, y) = \sum_{k=1}^n \alpha_k |Char(x, y, att_k)|, (\alpha_1 + \alpha_2 + \dots + \alpha_n = 1, \alpha_k > 0). \quad (1)$$

Here, att_k means the k_{th} attributes of the bookmark. $Char(x, y, att_k)$ shows the coincidence of att_k between bookmarks x and y . The relation degree is defined as a value to show the relations among bookmarks by connecting several attributes. Currently, the registration date, the date of the most recent referral, the number of referrals, and the total number of links are used as att_k . Although we plan to consider the contents-based relation degree, such as the value of *tf-idf* [8], in future research, such a contents-based relation degree is not used in our current system since the realization of various visualization styles and their application to the support system based on the definition given above is our main priority. In equation (1), α_k shows the weight assigned to each attribute which is decided by each user. The relation degree thus calculated lets us perceive the relation between bookmarks as a kind of semantic distance.

Visualization Styles

The support system provides four visualization styles for observing the relations among bookmarks from various viewpoints. Here, we describe the features of each style.

Visualization style 1: Figure 3 shows the visualization where the bookmark of interest is placed at the origin point and nodes are arranged around it on concentric circles; the stronger the relation is, the closer a node will be to the origin. This visualization enables users to intuitively grasp the bookmark relation degree as the distance from the origin point. This style satisfies user requirements 2, 3, and 4.

Visualization style 2: Figure 4 shows the visualization where nodes are placed according to the value of one of their attributes on the y axis. The bookmark with the highest value for that attribute is placed on the x axis. This enables users to observe the relations by focusing on one factor (e.g., the number of referrals). This style satisfies user requirement 1.

Visualization style 3: Figure 5 shows the visualization where nodes are placed using the values of two attributes on the respective x and y axes. Bookmarks with higher

values for these attributes are placed closer to the origin point. This lets users observe the relations among bookmarks by focusing on two factors (e.g., when trying to distinguish a node registered before a certain date but referred to frequently). This style satisfies user requirement 2.

Visualization style 4: Figure 6 shows the visualization where the node of interest is placed on the origin point and other nodes are placed around it based on the value of two of their attributes on the respective x and y coordinates. The two attributes are calculated as values relative to those of the bookmark of interest. This enables users to observe the relations among bookmarks by centering on the one of interest and considering two factors. This style satisfies user requirements 2 and 3.

Moreover, hyperlinks connecting the documents indicated by bookmarks are presented as edges of a graph in all four styles. This satisfies user requirement 5.

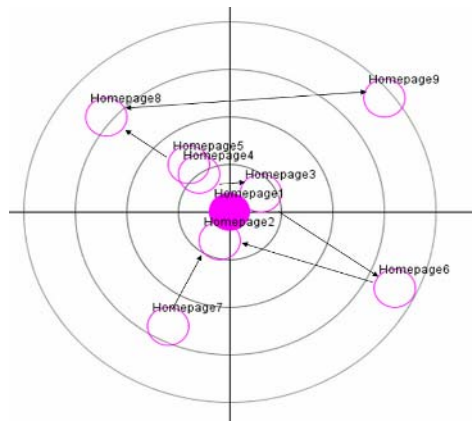


Fig. 3. Visualization style 1

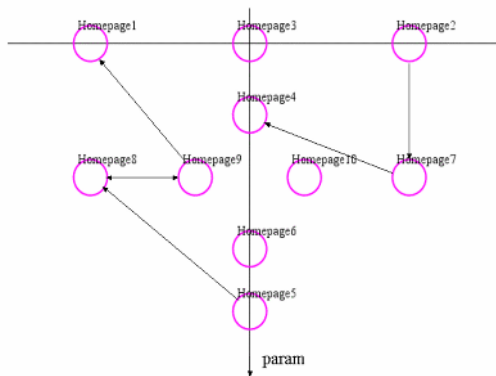


Fig. 4. Visualization style 2

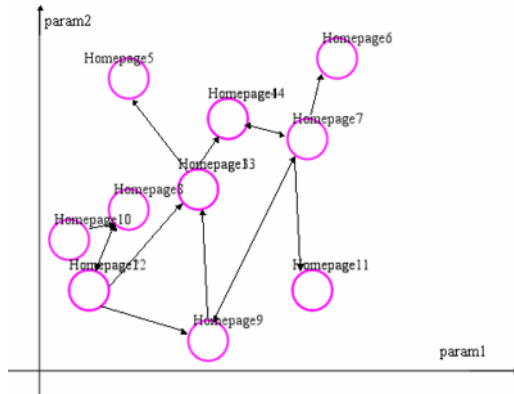


Fig. 5. Visualization style 3

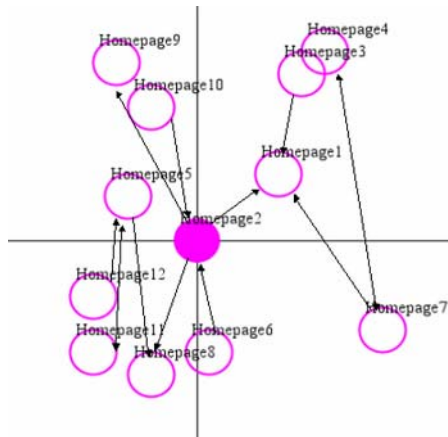


Fig. 6. Visualization style 4

3.3 System Implementation

We tested the main functions of the support system using typical execution examples. Figure 7 shows an example of managing the retrieval scene and sharing the extracted bookmarks based on the scene. This example shows that users can easily manage the retrieval scene by specifying the search purpose and the user group. Bookmarks appropriate for the retrieval scene are then extracted and presented.

Users can also visualize the relations among bookmarks by simply clicking a button provided by the support system on web browser (Fig. 8). The user selects the desired visualization style, the attributes used for it, and the bookmark of interest using a GUI provided by the support system. The user can then interactively observe the relations from various viewpoints by changing the visualization style. To use this system, users simply have to install a plug-in for the web browser.

Management of retrieval scene

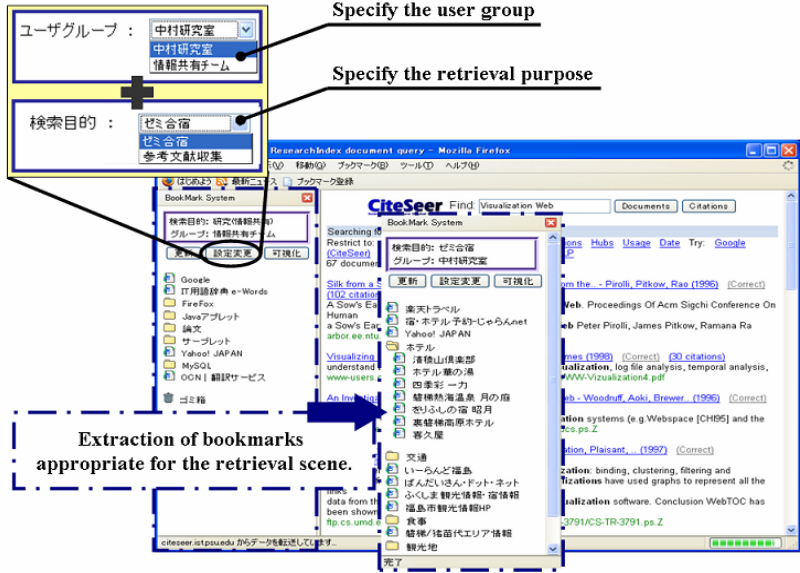


Fig. 7. User interface

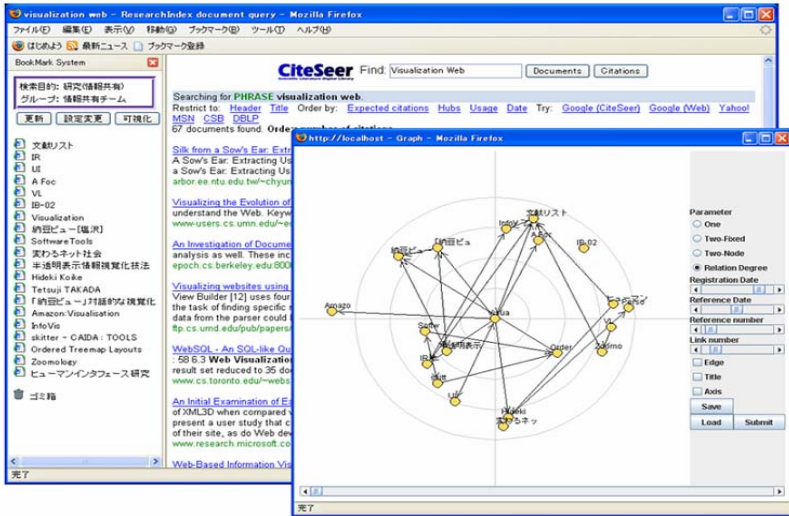


Fig. 8. An execution example of visualization

4 Comparison to Related Systems and Discussion

We compared the proposed system to existing approaches [1] [2] [4] with respect to the main features, management of retrieval purposes, management of user scope for sharing, and support for visualizing the relations among bookmarks (Table 1).

Table 1. Comparison of the proposed system to existing approaches

	Main features	Management of retrieval Purposes	Management of user scope for sharing	Supports visualization of bookmark relations
Proposed System	- Web browser used as user interface - Easy set up	Supports multiple retrieval purposes	Supports multiple user groups	Visualization considering various viewpoints
AniMAP [1]	- Specialized for travel planning - Requires exclusive Software	Only to design travel plans	---	Visualization using animal characters specialized for travel planning
Bookmark agent [2]	- automatic collection of bookmarks - Software agent explores similar bookmarks	---	---	---
Face [4]	- Web browser used as user interface - Assumes a common folder for sharing	Distinguish purposes by creating exclusive folders	Distinguish user groups by creating exclusive folders	Classification by icons prepared beforehand

AniMap [1] supports the sharing of web documents through browsing histories such as the number of pages referred to and the scope of browsing until a user decides on a recommended page. Visualization of bookmarks is also provided. Although the sharing of browsing histories has some merits regarding efficient retrieval, the application area is currently limited because this system has been implemented specifically for designing travel plans. Bookmark agent [2] automatically collects bookmarks related to a user's interests. This system is similar to our approaches in that it focuses on bookmarks, but this system differs from our approach since it mainly aims at finding bookmarks related to a user's favorite ones while our system is targeted towards the management of retrieval scenes and the sharing of bookmarks based on these scenes. Face, a commercial bookmark sharing service [4], is useful for checking for updates of web documents indicated by bookmarks and internal search of registered bookmarks. However, the retrieval scene is not fully considered since bookmarks are managed in exclusive folders for sharing.

Thus, the main features of our system – the extraction of bookmarks according to retrieval purposes, proper control of the user scope for sharing, and visualization of the relations among bookmarks – are unique.

5 Conclusion

We have developed a support system for extracting useful bookmarks according to a retrieval scene based on a search purpose and user groups. This system incorporates a visualization method to show the relations among bookmarks. Here, we have briefly described the underlying method of this system and its main functions. The

effectiveness of these functions has been confirmed through typical execution examples, and comparison to existing systems shows their uniqueness.

Acknowledgements

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan under Grant-in-Aid for Young Scientists (B) (No.17700599) and by Japan Society for the Promotion of Science under Grant-in-Aid for Scientific Research (B) (No.17300262).

References

1. Nakajima, S. et al.: Group-Based Collaborative Web Exploration by Sharing Bookmarks and Annotations, IPSJ Technical Report, Vol.2003, No.71, (2003) 177-184.
2. Mori, M. et al.: Bookmark-Agent: Sharing Bookmarks for Search Assists, Trans. IEICE, Vol.J83-D-I, No.5 (2000) 487-494.
3. Sano, K.: BisNet: An Information Sharing System Using Bookmarks of Web Browsers, JSAI Journal, Vol.20, No.4 (2005) 281-288.
4. FACE (<http://face.marsflag.com/>)
5. Shiozawa, H. et al.: The Natto View: An Architecture for Interactive Information Visualization, IPSJ Journal, Vol.38, No.11 (1997) 2331-2342.
6. Alan Wexelblat et al.: History-Based Tools for Navigation, Proc. of the 32nd Hawaii International Conference on System Sciences, IEEE Computer Society Press (1999).
7. Teraoka, T.: Adaptive Information Visualization Based on the User's Multiple Viewpoints, IPSJ Journal, Vol.39, No.5 (1998) 1365-1372.
8. Salton, G. et al.: Automatic Structuring and Retrieval of Large Text Files, Comm. ACM, Vol.37, No.2 (1994) 97-108.
9. Miyadera, Y., Hayashi, N., Nakamura, S., Yokoyama, S.: A Visualization System for Organizing and Sharing Research Information, Proc. 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, in LNAI 3683 (2005) 1288-1295.

Development of a Planisphere Type Astronomy Education Web System Based on a Constellation Database Using Ajax

Katsuhiko Mouri¹, Mamoru Endo², Kumiko Iwazaki³, Manabu Noda¹,
Takami Yasuda⁴, and Shigeki Yokoi⁴

¹Astronomy Section, Nagoya City Science Museum
17-1 2chome sakae naka-ku 460-0008 NAGOYA Japan
mouri@ncsm.city.nagoya.jp

²School of Information Science and Technology, Chukyo University

³School of Information Culture, Kinjo Gakuin University

⁴Graduate School of Information Science, Nagoya University

Abstract. In this study, we build a knowledge database that consists of topics planetarium curators have about astronomy and the stars. We suggest that the database should have a structure based on constellations and that mirrors the basic way in which humans perceive the night sky. Furthermore, we developed a web system for astronomy education using Ajax. The web system's interface is based on the planisphere, which is most basic tool for finding stars.

1 Introduction

There are a lot of sites on the Internet about stars and astronomy. Most of these sites are made available to the public by astronomical observatories or universities. Many of their observations, including beautiful images of celestial objects seen with big telescopes are freely accessible by the public[1,2]. At the same time, for general citizens who look at the night sky and then attempt to find out them, there is little useful information. This is because information about observed results or celestial objects are not tied to fundamental information about finding them at the night sky.

The authors have conducted research about astronomical education through the visualization of astronomical phenomena since 1993. Some examples include "The collision of Comet Shoemaker-Levy 9 with Jupiter", "The Vanishing Rings of Saturn", "Comet Hale-Bopp", and "Leonids Meteor Shower"[3]. Furthermore, we developed the "Powers of Ten" teaching materials[4].

In the "Powers of Ten", various astronomical phenomena and knowledge were ordered from a spatial viewpoint. We also created an exhibition with many panels and a computer graphics movie that illustrated this viewpoint. This showed where interesting celestial objects were in the universe in a hierarchical format. The national astronomical observatory has a similar scheme called the 4D2U project[5]. Their main purpose in that project is not astronomical education, but the visualization of their research data. Therefore, that project isn't comprehensive from the angle of astronomical education. Neither the "Powers of Ten" nor 4D2U can indicate where interesting celestial objects are in the night sky. No comprehensive database and web

system that developed on the point of view to look up at the sky actually have existed by now. There is great potential for the development of these resources for astronomical education.

On the other hand, a wealth of interesting information about the stars and astronomy is shown and explained in an easy-to-understand format at planetariums. The curator, acting as a commentator, selects information that is appropriate for the general public. Such content is accumulated as part of the curator's own knowledge. However, only planetarium visitors can currently access such information. Thus, a structure that can disseminate the contents of each curator's knowledge database to open public is desirable. Until now, no systematic database that builds information about such information has existed. A project focusing on astronomical observation data and astronomical news is being conducted by the National Aeronautics and Space Administration [6]. Furthermore, it can obtain a result of a reference with an ADS system with a XML form-type[7]. However, the stored contents and their structure differs from this paper, as we discuss below.

We suggest building a database that consists of the comprehensive information planetarium curators possess on stars and astronomy. Beyond that, we think the database should have a structure based on constellations, which illustrates human cognition about the night sky. In this paper, we use the word "constellation database" to refer to that database.

We also develop a web system that uses this constellation database. The most important aspect of this astronomical education is that after a user visits the web system, they go outside and can find them at the actual night sky. Thus, we chose the metaphor of the planisphere for the web system interface. The planisphere is the most general and popular tool for finding a star. Using the planisphere, knowledge can more easily be tied to the night sky. Constellation database that has useful knowledge for them is indispensable in such an education system.

Additionally, we use Ajax technology to improve the webservice's usability. In this paper, we use the word "planisphere type web system" to refer to that web system.

To this end, we built a knowledge database from the perspective of astronomical education and developed a planisphere type astronomy education web system.

2 Constellation Database

2.1 Database Requirements

Before building a constellation database, from the perspective of the astronomical education, the following points are necessary;

Compatibility

We need a format for the new astronomical education the database that can be used for a long time. Also, the format should offer portability and universality as well. Future improvements and modifications should also be simple.

There are four types of educational projects connected to this database: the planisphere type webservice; a star guide type webservice; a star guide for cellular

phones; and the database knowledge accumulated system. Universities, planetariums, schools, and so on are all concerned in these projects.

Internationalization

The concept of a constellation is a general idea shared by the whole world that was standardized by International Astronomical Union in 1930[8]. Thus, if our database structure could be a multi-lingual, it is easy for this concept to be used throughout the world.

Databases using XML[9] are currently attracting a lot of attention. XML has the following features. Using a tag that can be defined freely, XML can structuralize a document. This is suitable for embodying our database structure, which is made on a constellation basis as we discuss below. Also, each element can be changed comparatively easily. These characteristics are suitable for many-sided uses in several projects. Because XML uses UTF-8 code, it is possible for data in various languages to be held collectively. Thus, it is easy to ensure internationality.

For these reasons, we chose a text format and applied XML as one of the schemes to meet the above requirement. Items described using XML are stored in the directory structure.

2.2 Database Structure and Contents

Figure 1 shows an example of a conventional catalogue of stars. Data are grouped in one table, and they aren't structuralized. The distance to each star can't be perceived with the naked eye or with a telescope in the night sky because it is too far away. There is a custom that shows the position of the star and the celestial object in the virtual two-dimensional polar coordinate space of the celestial sphere. The two axes are right-ascension and declination. Therefore, catalogs of stars and celestial objects are built from catalog's number or two axes.

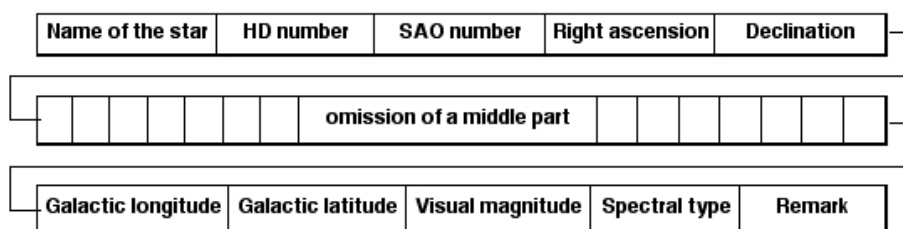


Fig. 1. Conventional star catalogue (BSC5 [10])

Such a catalogue is effective when only one celestial object is pin-pointed as a target. This format is easy for astronomy researchers to use. However, it isn't suitable when handling more than one celestial object, such as when working with a constellation.

Another method that shows position of the star and the celestial object on the celestial sphere is basing on constellation. The notion of constellations is much older than indications using coordinates; it originated in Mesopotamia more than 5000 years ago. The "Almagest", a treatise compiled by astronomer Ptolemy in Greece, is a

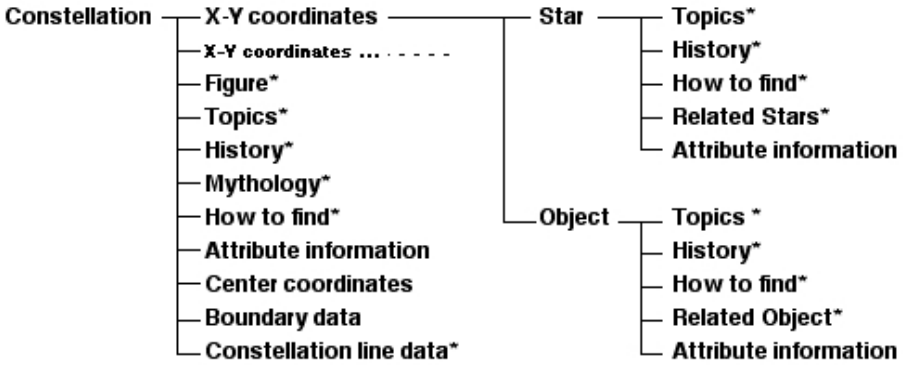


Fig. 2. XML based constellation database

classic astronomy text where the explanation of each star is based on the star’s position in relation to its constellation. Actually, when we look up at the night sky, it is easier to recognize specific constellations rather than specific coordinates. When expressing the position of a celestial object, constellations are easier for the general public to understand compared with coordinates. Even if a particular constellation can’t be found, people can still understand the concept behind it. Furthermore, no database that has structure of constellation based on has existed. So constellation database’s structure itself is a new idea. That structure can manage the knowledge that relates to the constellation integratively. Figure 2 shows an example of star information in our constellation database plan. Items are structuralized into an XML based database. Items that are marked with * are characteristic of this database. When a person looks up at the star, these items are important, or the item may add an interesting point. These items weren’t necessarily included in other databases, even though planetarium curators may have had this information. This is only part of a planetarium curator’s tacit knowledge. Still, it is crucial to include important knowledge that hasn’t been included in any database till now. The database should also be in a format that is easy to introduce to the public on the Internet.

Using the constellation database, it is possible to combine all these elements in one structure. Therefore, in view of a database for general-purpose use in the future, we propose a constellation database that uses XML as base as the standard in astronomical education.

3 A Planisphere Type Web System Using Ajax

We developed a web system that uses a constellation database. The most important aspect of this system from the perspective of the astronomical education is that after a user visits the web, they should look at the actual night sky. In other words, We would like to tie learning by web pages in the inside to the experience in the real nature. When a person looks up at the night sky, it is rare that they’d have access to their PC outside. Our web system is the same, too. Learning in advance is the purpose of this web system. However, many people will go out to find the stars with the planisphere in their hands.

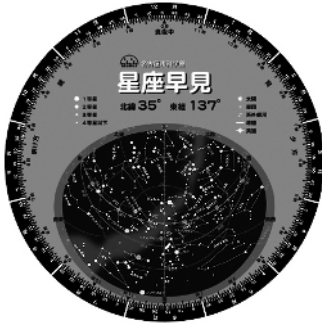


Fig. 3. Planisphere (Nagoya City Science Museum)

interface of our web system. After people learn with a planisphere type web system, they will go out and look up the real sky with real planisphere.

Additionally, we use Ajax technology to improve usability of the web-system. Ajax is a generic name for "Asynchronous JavaScript + XML"[11]. Ajax facilitates asynchronous communication between a server and a client. So, Ajax is suitable for scrolling a big image, offering good usability. In our web system, a user can look at the detail of a big planisphere image and can scroll through it. Additionally, the web system's interface is improved by applying a javascript function. We have already applied XML to our web system database as mentioning above. Thus, Ajax is the most suitable technology for our web system as it can show content dynamically.

Google earth, moon, and mars [12] are good examples of astronomy education systems. The Hubble site has a similar interface [13] that uses Flash technology. However, these examples offer only an actual view of the planet or nebula. Currently, there are no cases that use teaching tools, nor that add extra value using Ajax.

Planispheres have been an effective tool for astronomical education since the 17th century. A planisphere is tool that finds a star by finding the azimuth and altitude of the star from the observation date and time. Most elementary school students in Japan learn how to use it. In fact, the planisphere is the most popular and general tool for finding specific stars. It is good to choose the tool that the people actually use in the outside for the web system's interface. So we applied a planisphere as

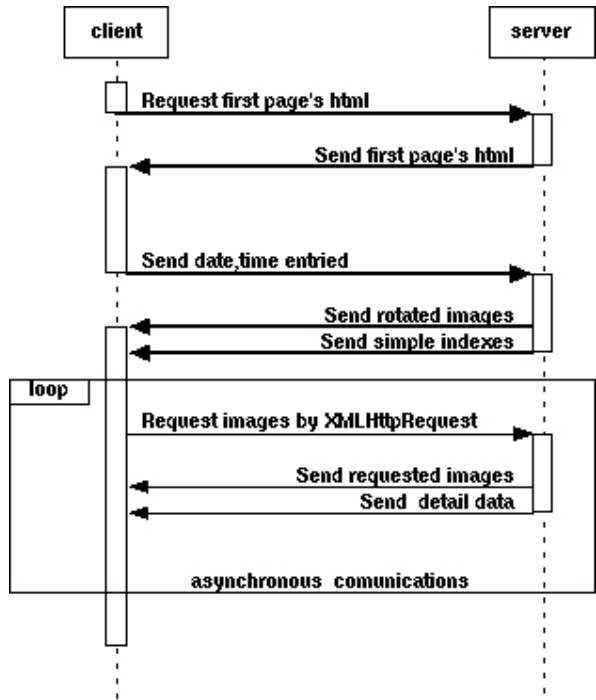


Fig. 4. Process flow

The communication between the database and the client using Ajax is outlined in figure 4. As a precondition, the server has an XML-based constellation database. We use the term "detail data" to mean the database records in this section. Also, the server has another simple database that consists of only the coordinates and classification of the attribute data. This means that "where the flag is marked on the planisphere". We use the term "simple index" to mean this another simple database.

First, the screen is loaded, and the user sets up the date and time. This date and time entry is sent to the server using asynchronous XMLHttpRequest. Servers generate images that are rotated to fit the date and time entry. Clients get the images as necessary using asynchronous XMLHttpRequest. At the same time, clients also get a simple index from the server.

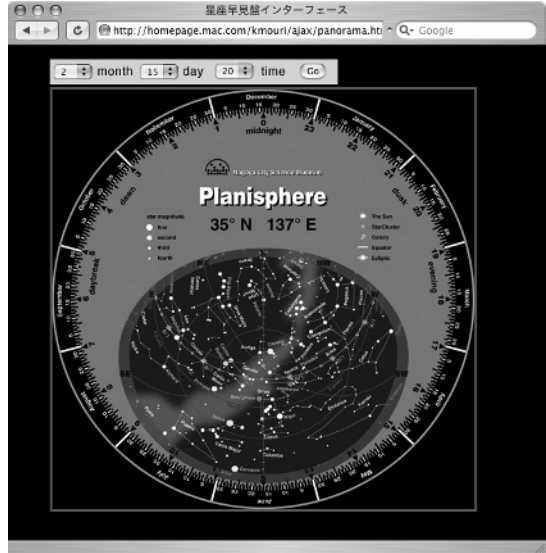


Fig. 5. The first page of the web system

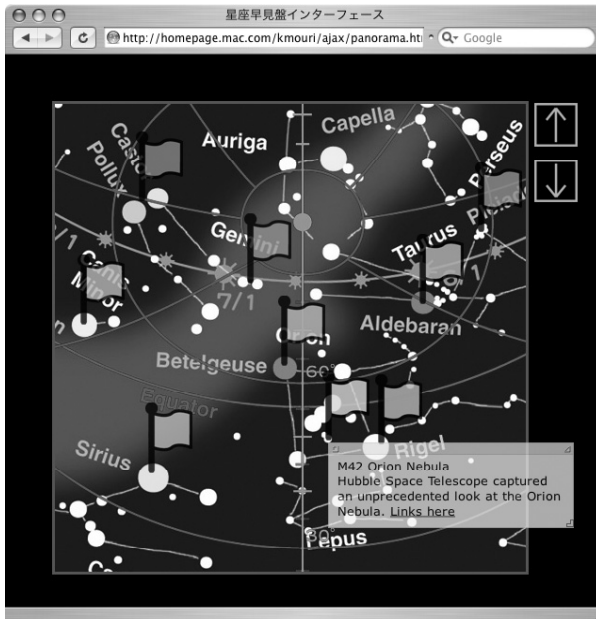


Fig. 6. Detailed images of the planisphere and flags

Clients send a request and receive images and detail data in advance, which allow clients to quickly indicate detail data's contents. Users can perform various operations including drag, scale-up, scale-down, and clicking flags on the planisphere interface. When users click the flags, the client indicates detail data copied with the flag on the planisphere in the most suitable format. A user can study stars, constellations, and celestial objects, based on the date and time input to the planisphere.

Users can recognize and memorize the contents of the constellation database by clicking on various places on the planisphere type web system. Also, the learning content on the system can be memorized and recalled when stars are seen outside with a real planisphere.

4 Conclusion

In this study, we built a knowledge database consisting of the knowledge held by planetarium curators on topics about the stars and astronomy. The structure of the database is based on constellations that mirrors the way humans perceive the night sky. It was very important to include important knowledge that had not been included in databases until now and in a format that was easy to introduce to the public on the Internet.

Furthermore, using Ajax, we developed a web system for astronomy education. The web system's interface was based on the planisphere, which is one of the most basic tools for locating stars. Using a planisphere-type interface, knowledge could be tied to the night sky more easily. Additionally, we used Ajax technology to improve usability of the web system.

We have not yet conducted strict evaluation verification of this system. Below are several user evaluations. According to evaluations by some planetarium curators, such a database was seen for the first time. And it is very interesting. They stated they would like to use it practically. Astronomy club members highly evaluated the planisphere-type interface. They also stated that was interesting to see the planisphere on the Internet. Furthermore, they were surprised that information was available there. They found it pleasant that the night sky could be seen with a planisphere. Another member said that it was good to recall the contents that listened to in the planetarium. It was very useful when looking up an actual starlit sky. There are about 900 members of astronomy club in our planetarium. We have plan to get more evaluations from them and improve our database and system.

In our future study, we would like to accumulate more knowledge for the constellation database and to further optimize the structure. Also, we would like to proceed with using this system with more projects, as we mentioned in the second section. The concept of the constellation is shared by people throughout the whole world. In terms of the planisphere web system, we would like to expand its applicable areas and make the system accessible to an international audience.

Acknowledgments

The authors would like to thank Mr. Takashi Yamada for giving guidance to this research and for making the planisphere. We would also like to thank Miss Masako

Kitahara for her continuing encouragement. This research was partially funded by 21st century COE program "Intelligent Media (Speech and Images) Integration for Social Information Infrastructure" from the Ministry of Education, Culture, Sports, Science and Technology, Japan. This research was also supported in part by the grant-in-aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Subaru Telescope, National Astronomical Observatory of Japan: <http://www.naoj.org/>
2. The European Southern Observatory: <http://www.eso.org/>
3. Katsuhiro Mouri, Masao Suzuki, Takami Yasuda, Shigeki Yokoi: Production and Practical Use of Teaching Materials based on 3-dimensional Computer-graphics Technology with Collaboration in Education of Astronomy. *The Journal of Information and Systems in Education* Vol. 1, No. 1(2002) 60-69
4. Katsuhiro Mouri, Masao Suzuki, Akihiro Yamamoto, Takami Yasuda: Production and Practical Use of Astronomical Teaching Materials: "The Powers of Ten" Computer Graphics. Vol.8 No.1(2001) 89-98 (in Japanese)
5. 4-dimensional digital universe project: <http://4d2u.nao.ac.jp/>
6. The NASA XML Project: <http://xml.nasa.gov>
7. The NASA Astrophysics Data System: <http://adswww.harvard.edu/>
8. International Astronomical Union: <http://www.iau.org/CONSTELLATIONS.241.0.html>
9. The World Wide Web Consortium: <http://www.w3.org/XML/>
10. Hoffleit D., Warren W.H. Jr, 1991, THE BRIGHT STAR CATALOGUE, Yale Univ. Obs., New Haven, Connecticut, V revised ed.: <http://cdsweb.u-strasbg.fr/cgi-bin/Cat?V/50>
11. Jesse James A New Approach to Web Applications: <http://adaptivepath.com/publications/essays/archives/000385.php>
12. <http://earth.google.com/>, <http://moon.google.com/>, <http://www.google.com/mars/>
13. <http://hubblesite.org/newscenter/newsdesk/archive/releases/2006/01/image/a+zoom>

Group Collaboration Support in Learning Mathematics

Tomoko Kojiri¹, Yosuke Murase², and Toyohide Watanabe²

¹ Information Technology Center, Nagoya University

² Graduate School of Information Science, Nagoya University

Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

{kojiri, murase, watanabe}@watanabe.ss.is.nagoya-u.ac.jp

Abstract. Our objective is to construct the effective collaborative learning support environment. Students acquire knowledge from their activities in a group. On the contrary, a group activity is composed of learning activities of individual students. So, in our approach two support mechanisms are introduced: the mechanism which assists a group to accomplish a group task by generating hints for deriving answers, and the mechanism which promotes individual students in the group activity by pointing out differences between group's opinions and student's opinion privately. In this paper, we focus on learning collaboratively mathematical exercises and implemented these support mechanisms in mathematics by using diagrams effectively. Our experimental results make it clear that our mechanisms could promote the group discussion so as to derive answers and urge students to join into the group discussion positively.

1 Introduction

The collaborative learning is one of learning styles in which students collaboratively solve common problems by exchanging their opinions [1]. When students in a group study exercises which have answers and answering paths such as mathematics, they often acquire knowledge by monitoring appropriately opinions which are exchanged in the discussion. If appropriate knowledge was not derived or impasse situation was occurred, students cannot get sufficient knowledge to solve exercises. In the collaborative learning, students not only behave passively to receive knowledge, but also provide their opinions actively to the group which may become effective knowledge to other students. To give their own knowledge to the group is also effective learning activity, since students have to arrange their opinions before uttering their opinions. Therefore, in the collaborative learning, to support the learning activity of group as well as to promote each student so as to enhance communication is important.

CSCL (Computer Supported Collaborative Learning) is a research topic which supports such a learning style [1], [2]. Researches or systems in this area can be classified globally based on their supporting targets: group activity and individual activity. Researches that focus on the group activity mainly support their coordination through communication [3], [4]. The objective of these researches was to solve the conflict among their opinions and then activate the discussion. Most of these researches did not analyze the contents of the utterances, but estimated the situation based on attributes of utterances such as types. In answering exercises that have an answer and

answering paths, whether students can derive the answers by themselves is important. Thus, utterances should be examined from a viewpoint of answering paths. Systems in CSCW (Computer Supported Cooperative Work) provided functions that support the group to accomplish the tasks easily [5], [6]. Most of them only supported students to work collaboratively, but did not promote students to accomplish their tasks successfully. On the other hand, most researches of CSCL that support individual activity of students focused on the differences between a group activity and a student's activity [7], [8]. These researches did not often evaluate the group activity totally. However, in the collaborative learning, students mainly acquire knowledge in the group activity, so students should not be urged to solve exercises individually but should be promoted to participate into the group activity. Therefore, to achieve an effective collaborative learning, a group should be supported so as to derive the learning goal and students should be promoted to join into the discussion.

In this paper, we propose an effective collaborative learning environment which embeds mechanisms that manage a group so as to achieve their learning goal and that urge students to participate into the group activity. In our approach, students are urged to derive an answer in a group. In order to enhance a group activity, the group is regarded as one supportable target. Then, a group discussion is monitored from a viewpoint of the progress along answering paths. Utterances are mapped to the target answering step, and appropriate advices that help a group to derive the answer are generated occasionally. On the other hand, since a group activity is composed of individual students' activities, to promote students to participate into the group activity is important. Therefore, in our approach the mechanism which indicates differences between their activities and the group's activity is also introduced. To note differences may make students have confidence in their opinions and urge them to exhibit their opinions in the group. Therefore, according to these two support mechanisms, a group is managed from a viewpoint of deriving the answer but also communicating others effectively. Thus, we focus on a collaborative learning in mathematical exercises of high school and introduce an effective collaborative learning environment within mathematical domain.

2 Collaborative Learning Support Environment in Mathematics

In mathematical exercises, one answer and several answering paths are generally existed. Students understand mathematical knowledge, such as formulas and their usages, by not only deriving the answer but also considering several answering paths that consist of different answering steps. If students know formulas and their usages well, they could apply appropriate formulas to various exercises. On the other hand, diagrams used to derive the answer help students understand exercises that are described only by characters conceptually. Ito, et al. [9] analyzed the importance of diagrams in solving exercises of trigonometrical functions. They classified diagrams drawn in solving exercises into 8 types on the basis of roles of the diagrams. In all types of diagrams, a diagram in which supplementary figures are added to figures of known equations is effective to derive new answering paths.

In the collaborative learning of exercises that have an answer and several answering paths, to support a group for deriving the answer along the specific answering

path is important. Since students acquire knowledge through discussion, they cannot acquire sufficient knowledge of the exercise if a group could not derive the answer. Thus, the role of monitoring group's learning activity and assisting a group so as to accomplish their effective learning is needed. As a method for assisting a group, to give right answer directly would not activate a group's discussion. The assistant function should give hints to encourage the group activity, if necessary. In mathematics, students attain to their answer by applying formulas or answering methods stepwisely. These formulas or answering methods are indicated by supplementary figures indirectly. Therefore, in our research, the mechanism which generates supplementary figures for deriving the next answering step automatically is introduced [10]. Since it is difficult to prepare supplementary figures for all answering steps, supplementary figures that correspond to formulas or answering methods are described as rules to apply figures to the current diagram. Then, the rules are selected on the basis of forward reasoning. Although all answering methods whose conditions satisfy the current diagram can be selected even if they were not appropriate for the target exercise, students can consider not only the correctness of indicated answering methods but also the effectiveness of the proposed supplementary figures.

On the other hand, students acquire knowledge from other students' utterances in discussion, so they need to participate into the group activity in order to activate the discussion. If students get new ideas and have the confidence with their ideas, they may utter their opinions to group easily. So, to make students note differences between their ideas and group's ideas urges them to propose their ideas to the group. In mathematics, to discuss various answering paths is effective. Differences of answering methods can be grasped according to figures in a diagram. Figures and their relations are different according to formulas so that all figures and their relations in a diagram are unique to each answering path. Therefore, in our environment a student's private canvas and a group's public canvas are prepared, and a mechanism which detects the differences between a diagram in a private canvas and that in the public canvas is introduced. For the purpose, an inner model which represents diagrams based on types of figures and remarkable relations between figures is introduced. Two diagrams are transformed into the corresponding representations in the inner model and the differences between diagrams are detected by comparing the representations.

Fig. 1 shows the conceptual imagination of our collaborative learning support environment in mathematics.

3 Mechanism for Indicating Difference Between Diagrams

The difference between diagrams can be defined from various viewpoints. If we look upon a diagram as a collection of figures, the number of existing figures and their types should be examined. However, if we regard a diagram as a mapping of equations into a two dimensional space, coordinates that figures take need to be compared. In the collaborative learning, to discuss various answering paths in a group is effective. When answering paths are different, the existing figures and their relations are different because figures correspond to equations. Moreover, since scales of diagrams are different among students, it is not appropriate to compare diagrams by their coordinates. Therefore, in our research figures and their meaningful relations in

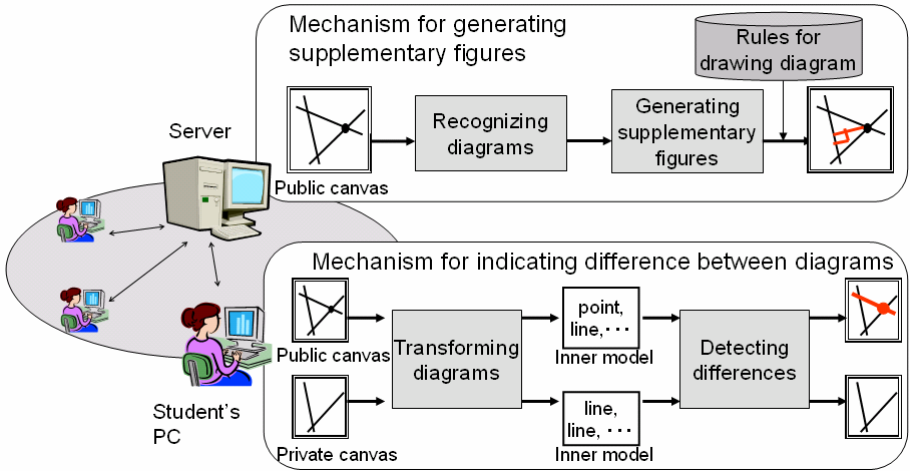


Fig. 1. Conceptual imagination of collaborative learning environment

diagrams of private canvas and a public canvas are compared. Then, detected differences are displayed to students so as to make notice of their original answering viewpoints.

The inner model of diagram that represents figures and their relations as a predicate form is introduced [10]. In order to represent meaningful figures, six predicates are introduced including x-axis and y-axis. In addition, 12 types of meaningful relations are also prepared. These relations are meaningful to discriminate the conditions of answering formulas.

Diagrams drawn in a public canvas and a private canvas are transformed into inner models, respectively. Then, by comparing inner models of diagrams, figures that are drawn on the basis of different viewpoints are able to be detected. In this process, first, predicates that represent figures are compared. If a predicate of a figure in one diagram does not match predicates in the other diagram, the corresponding figure is one of the differences between diagrams. If a predicate of a figure situates in the other diagram, relations related to them are examined. When the relations are different, these figures are regarded as different figures.

Fig. 2 shows our prototype system and an example of detecting the difference between diagrams. In the system, students push the type buttons of figures and point out the corresponding coordinates on the canvas in drawing figures. In addition, students can select buttons for specific relations. When a diagram is submitted on either public canvas or private canvas, the system transforms diagrams in both canvases to inner models, compares them, and detects different figures. In this example, since the diagram in the private canvas does not contain the line and the point, $line(a)$ and $point(c)$ in the public canvas are detected as different figures. Then, they were displayed in the public canvas with their color highlighted.

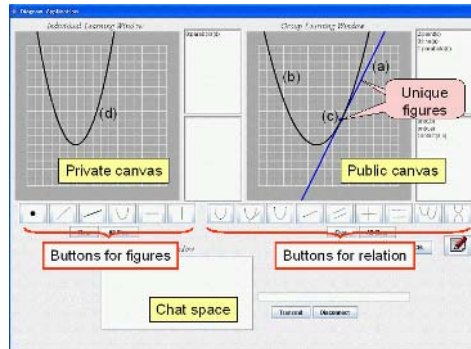


Fig. 2. Example of indicating differences between diagrams

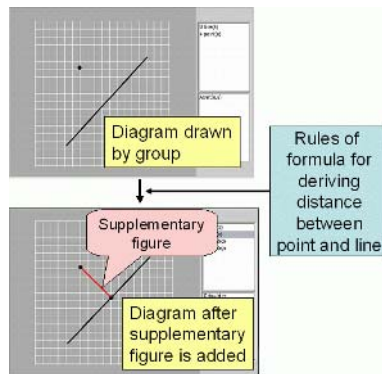


Fig. 3. Example of generating supplementary figure

4 Mechanism for Generating Supplementary Figures

The system generates advices when a group could not proceed the learning effectively. The advices should activate the discussion about various answering methods related to the exercise. In mathematics, the appropriate formula or answering method whose conditions satisfy the current diagram is selected and applied to the current diagram. Thus, to compare various formulas based on the current diagram and consider if they are appropriate for to the answering paths are important learning process.

Rules for drawing diagrams are defined as individual formulas or answering methods [10]. They indicate supplementary figures that are hints for deriving the formulas. Since conditions to apply formulas are defined according to the characteristics of derived equations, rules for drawing diagrams are described by using predicates defined in Section 3: namely figures and their relations. On the conditional part of rules, figures and their relations that indicate the condition for applying the corresponding rules are defined. On the action part, to add supplementary figures and their relations with the existing figures is described. Currently, two types of figures are prepared in

each formula. One is to emphasize figures in the conditional part, and another is to display figures that should be derived in applying the formula.

When an impasse situation is detected, one rule whose conditional part satisfies a current diagram is selected and applied to the current diagram. If conditional parts in plural rules are applicable to the current diagram, the rule whose conditional parts include the most predicates is chosen.

Currently, supplementary figures are generated when the button that requires supplementary figures is pushed. After the button has been pushed, rules for drawing diagrams are applied and supplementary figures are added to the public canvas automatically. Supplementarily added figures are different from other figures with the color in order to highlight them. Fig. 3 shows the example of supplementary figures. In this example, a line which connects to the existing point and line is generated in order to indicate the formula of deriving distance between point and line.

5 Experimental Result

Experiments were conducted based on our prototype system. In order to evaluate the effectiveness of each mechanism, two different experiments were executed.

5.1 Indicating Difference Between Diagrams

The effectiveness of the mechanism indicating differences between diagrams was evaluated. In this experiment, 6 examinees (“A” to “F”) in our laboratory were divided into 2 groups and asked to study two mathematical exercises collaboratively. For one exercise, examinees used the system in which the mechanism for indicating the difference between diagrams was embedded (*learning 1*), and for another exercise they studied using a chat system (*learning 2*).

The numbers of drawn figures in the public canvas and the private canvas were counted up for individual students in both learnings. The numbers were also counted according to the order of accessing each canvas. Table 1 shows the increase in the number of students’ drawings in *learning 1*, in comparison with that in *learning 2*. In this table, “A” to “F” corresponds to individual examinees. “Private” and “public” mean the canvas to which examinees draw. Namely, “public -> private” means that an examinee drew a diagram in a private canvas after he has drawn in the public canvas. Based on the result, the number of drawn figures in the private canvas increases for four examinees. So, our system is likely to promote examinees so as to derive the answer by themselves. Most students drew diagrams more frequently when they used the system which includes the mechanism for indicating differences between diagrams. Moreover, the numbers of “public -> private” and “private -> public” are also increased for more than three students, respectively. The purpose of indicating the difference between diagrams is to note the difference between answering viewpoints of a group and those of examinees. If examinees were aware of the differences, they may reconsider their answer in the private canvas or propose their opinions to the public canvas. Therefore, these results show that examinees were aware of the differences between their diagrams and group’s diagram.

Table 1. Increase in the number of students' drawings (the number drawings in *learning 1* - that in *learning 2*)

	A	B	C	D	E	F	Total
private -> private	2	6	-1	4	-4	1	8
Public -> public	-3	1	0	1	-3	-1	-5
Public -> private	1	5	1	2	-3	1	7
private -> public	1	0	1	2	-1	-1	2

5.2 Generating Supplementary Figures

Two experiments were conducted in order to evaluate the mechanism for generating supplementary figures. The objective of the first experiment is to evaluate the correctness of supplementarily generated figures. In this experiment, supplementary figures were requested at 39 steps in 25 exercises of two dimensional functions by using our prototype system, which includes 38 rules.

Table 2 shows the correctness of the selected rules for drawing diagrams. When the selected rules correspond to correct formulas to derive the next step, they are regarded as correct rules. If not so, it is counted as a wrong rule. Based on the result, our system could generate correct supplementary figures by 87% of the total steps. Reasons that wrong rules were selected are shown in Table 3. Since the number of predicates in the inner model is small, the combinations of predicates are limited. So, conditional parts of some rules are the same or include those of another. Although to discuss the correctness of supplementarily generated figures is effective to understand formulas deeply, to specify applicable rules for exercises may be necessary.

Table 2. Correctness of rules

Result	Number
Correct rules	34
Wrong rules	5

Table 3. Reason for selecting wrong rules

Reason	Number
Conditional part of selected rule was same as that of correct rule.	4
Conditional part of selected rule contains that of correct rule.	1

The objective in the second experiment is to examine the effectiveness of supplementarily generated figures. In this experiment, seven examinees in our laboratory were divided into two groups: *group 1* of four examinees and *group 2* of three examinees. They were asked to study an exercise of two dimensional functions in groups by using our prototype system that generates supplementary figures.

During the learning, supplementary figures were generated twice for *group 1*, while only once for *group 2*. Based on advices, *group 1* could select one appropriate answering path from derived two different paths. Also, *group 2* could found out the method for deriving the next step successfully. Fig. 4 and Fig. 5 are the number of utterances along the time sequence. The numbers of utterances were increased after the supplementary figures have been derived in both groups. Therefore, supplementary figures are estimated to be triggers of activating discussion.

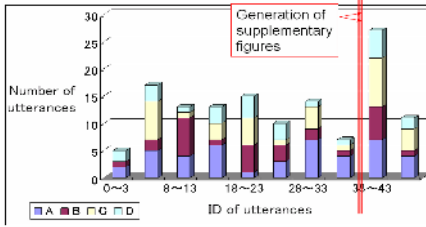


Fig. 4. Utterances of group 1

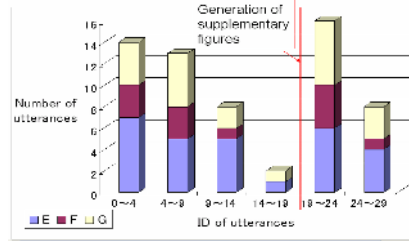


Fig. 5. Utterances of group 2

6 Conclusion

In this paper, the collaborative learning support environment which includes two support mechanisms, such as the mechanism which promotes group discussion of deriving answer and the mechanism which indicates differences between group’s activity and student’s activity, was introduced. Experimental results indicated these mechanisms were effective to active discussion for deriving answer. However, further evaluation is needed for proving effectiveness by analyzing performances of examinees in detail.

Currently, the system generates supplementary figures when students require. If the system detects impasse situation and generates supplementary figures automatically, students are not conscious of an existence of the system and the discussion among students may become more natural. We have already proposed the mechanism for detecting impasse situation of a group from text-based interaction [11]. So, in our next step, the mechanism that grasps the learning situation from diagrams and text-based interaction, and generates appropriate supplementary figures based on the detected impasse situation, should be constructed. Moreover, to construct the collaborative learning environment in the different domain is also necessary.

References

1. Dillenbourg, P. (eds.): “Collaborative Learning – Cognitive and Computational Approaches”, Elsevier Science Ltd. (1999).
2. Koschmann, T. (eds.): “CSCL: Theory and Practice of an Emerging Paradigm”, Lawrence Erlbaum Associates Publishers (1996).
3. Hesse, F. W., and Hron, A.: “Dialogue Structuring in Computer Supported Synchronous Discussion Groups”, Proc. of ICCE’99, Vol.1 (1999) 350-355.
4. Inaba, A., Ohkubo, R., Ikeda, M., and Mizoguchi, R.: “Models and Vocabulary to Represent Learner-to-Learner Interaction Process in Collaborative Learning”, Proc. of ICCE’03 (2003) 1088-1096.
5. Adelsberger, H. H., Collis, B., and Pawlowski, J. M. (eds.): Handbook on Information Technologies for Education and Training, Springer-Verlag (2002).
6. Carroll, J. M., Neale, D. C., Isenhour, P. L., Rosson, M. B., and McCrickard, D. S.: “Notification and Awareness: Synchronizing Task-oriented Collaborative Activity”, Trans. on Human-Computer Studies, Vol.58 (2003) 605-632.

7. Nakamura, M., and Otsuki, S.: "Group Learning Environment Based on Hypothesis Generation and Inference Externalization", Proc. of ICCE'98, Vol.2 (1998) 535-538.
8. Constantino-Gonzalez, M. A., and Suthers, D. D.: "A Coached Collaborative Learning Environment for Entity Relationship Modeling", Proc. of ITS 2000 (2000) 324-333.
9. Ito, T., Ohnishi, N., and Sugie, N.: "Problem Solving SCRIPT and Clarification of Drawings for Explaining Drawing Process", Trans. on IEICE, D-II, Vol.J77-D-II, No.4 (1994) 811-822 (in Japanese).
10. Murase, Y., Kojiri, T., and Watanabe, T.: "Dynamic Generation of Diagrams for Supporting Solving Process of Mathematical Exercises", Proc. of KES'05, Vol. 3 (2005) 673-680.
11. Kojiri, T. and Watanabe, T.: "Agent-oriented Support Environment in Web-based Collaborative Learning", Journal of Universal Computer Science, Vol. 7, No.3 (2001) 226-329.

Annotation Interpretation of Collaborative Learning History for Self-learning

Masahide Kakehi, Tomoko Kojiri, and Toyohide Watanabe

Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
Phone: +81-52-789-2735; FAX: +81-52-789-3808
{kakehi, kojiri, watanabe}@watanabe.ss.is.nagoya-u.ac.jp

Abstract. In collaborative learning, learners exchange opinions through communication. When learners study the same type of exercises after the collaborative learning, they recall that communication, find effective utterances, and derive an answer by themselves. Therefore, in an collaborative learning support environment, it is useful for learners to monitor communication history after the learning has been finished. In the collaborative learning of exercises that contain an answer and answering paths, grasping answering path that learners derive during the learning and detecting effective utterances based on their answering paths are important. In our approach, the function to add annotations to utterances during learning is introduced in order to grasp effective utterances that students think during the learning. Then, based on the added annotations, learning situation of other learners and effective utterances for learners are derived.

1 Introduction

People often coordinate with others learners through the network. In a learning field, to support such collaborative learning through the web is one of the hottest subjects[1],[2],[3]. During communication, learners utter what they understand and what they think. Also, they acquire knowledge from utterances of others. When learners study the same type of exercises after collaborative learning, they recall that communication, find effective utterances, and derive the answer by themselves. Therefore, in the collaborative learning support environment, it is useful for learners to monitor the communication history after the learning has been finished. However, there are a huge amount of utterances in the collaborative learning, so it is difficult for learners to find utterances that help them to derive the answer.

Effective utterances for learners are classified into two types: utterances that they think effective during learning and that they can notice after learning has been finished. When learners try to derive the answer by the same answering path than what they solved during the learning, they recall the former type of utterances. On the other hand, when they want to discover new ideas, the

latter utterances take an important role. Therefore, detecting effective utterances from derived or underived answering paths and providing them to learners are important.

There are many researches that analyze meaningful scenes from the collaborative learning history[4],[5]. Learners cannot notice effective utterances of collaborative learning.

Currently, the collaborative learning of exercises that contain several right answering paths, such as programming, is focused on. Namely, utterances concerned with the answering path that a learner derived are useful for detecting answers as the same answering path to know whether utterances are for the same answering path is important. On the other hand, utterances that are not related to his answering path may be effective to derive new answering paths.

The effective utterances for each answering path are different among learners. Therefore, in our research, a function to add annotations is embedded to our collaborative learning support environment. Then, the mechanism for extracting effective utterances for learners from each learner’s annotations is introduced. Utterances which learners who have the same viewpoint think important are effective utterances. On the other hand, utterances which learners who have different viewpoints think important are key utterances of different answering paths. In our approach, annotated utterances are classified into the same/different answering paths based on relations among learner and other learners who attached annotations. In this paper, we introduce the function for extracting effective utterances of the same/different answering paths and evaluate its effectiveness based on the experiment using a prototype system.

2 Effective Utterances in Collaborative Learning

Effective utterances are classified into two types; effective utterances that follow the answering path that a learner derived during learning and that belong to answering paths which a learner did not derive. Utterances that each learner thinks effective during the collaborative learning are the key utterances of answering path that he derived. Meanings of utterances that other learners’ key utterances are different for each learner. For example, let us assume the situation

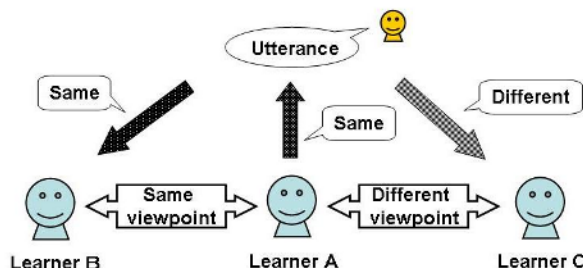


Fig. 1. Effective utterances of learners based on their viewpoints

as shown in Figure 1. In this example, the learner A and the learner B have the same answering viewpoints, while the learner C is different. When A think an utterance is a key utterance for the same answering path, the utterance may also be effective with the same viewpoint for B. However, for C, the utterance is a key utterance for the different viewpoint. Therefore, in order to classify effective utterances according to answering viewpoints, it is necessary to grasp answering paths of learners and learners' intentions for utterances.

The meanings of utterances are different among learners. Therefore, in our approach, the function by which learners can attach annotations to utterances is introduced. By this function, learners' intention to their utterances can be acquired. Since annotations indicate the current answering path of a learner, firstly, viewpoints of other learners are determined by comparing with their annotations. Then, based on the viewpoints of learners, annotated utterances are classified into the same or different viewpoints and provided to the learner automatically. Figure 2 shows a conceptual imagination of the system.

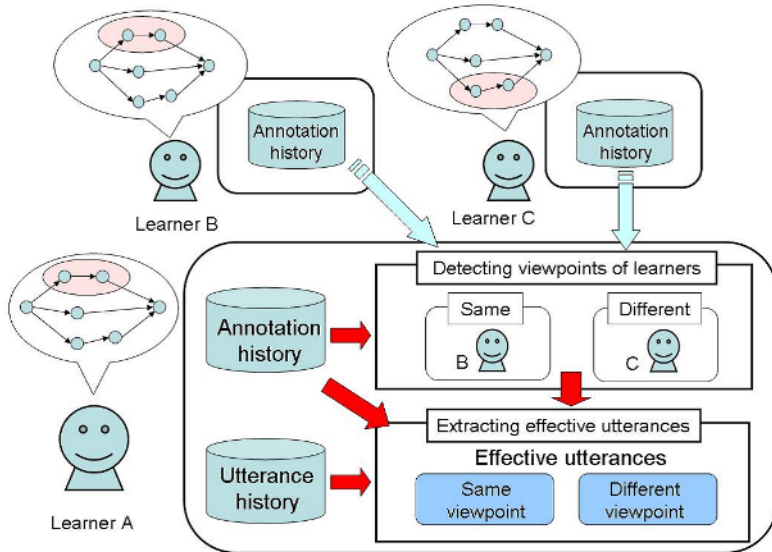


Fig. 2. Conceptual imagination of our system

3 Mechanism for Extracting Effective Utterance

3.1 Annotation

Two annotations are prepared so as to represent learners' intentions toward utterances:

1. same viewpoint, and
2. different viewpoint.

The annotations of the same viewpoint indicate that utterances are useful for a learner to derive an answer during the collaborative learning. The different viewpoint shows that utterances are not effective for deriving the answer by his answering path, but may be key utterances to be different answering paths.

3.2 Learners' Viewpoints

Table 1 shows viewpoints between learners who add annotations. When both learners attach annotations of the same viewpoint, their answering viewpoints are the same. If one learner adds an annotation of the same viewpoint and another attaches that of different viewpoint, they are regarded to have the different viewpoints. However, if both learners add annotations of different viewpoints, whether their answering viewpoints are the same or not is not specified. This is because two answering viewpoints that are different from one answering viewpoint do not always have the same. Table 2 shows a relation between a learner

Table 1. Viewpoints of learners who add annotations to the same utterance

Type of annotations	Same	Different
Same	Same viewpoint	Different viewpoint
Different	Different viewpoint	Unspecified

and an utterer whose utterances are annotated by the learner. If a learner attaches an annotation of the same viewpoint, the learner and the utterer may have the same viewpoint. If a learner attaches an annotation of the different viewpoint, the learner and utterer probably have the different viewpoints.

If a learner does not utter his opinions nor attach an annotation, his answering viewpoint is seemed to be not changing. Therefore, the answering viewpoint of the learner is determined as the same as the previous situation.

Table 2. Viewpoint of utterer whose utterance is annotated

Type of annotation	Viewpoint of utterer
Same	Same viewpoint
Different	Different viewpoint

3.3 Effective Utterances

All utterances which are annotated to learners are candidates of effective utterances. They are classified into the same and different viewpoints, in order for the learner to refer when they tackle with the same type of exercises. Table 3 shows the meanings of utterances to a learner according to the viewpoints of learners who attach annotations.

Table 3. Viewpoints of candidate key utterances

	Annotation : Same	Annotation : Different
Learner’s viewpoint : Same	Same viewpoint	Different viewpoint
Learner’s viewpoint : Different	Different viewpoint	Unspecified

The annotations that learners who have the same answering viewpoint attached mean the same to the learner. On the other hand, if learners have different answering viewpoints, their annotations of the same viewpoint are different viewpoint for the learner. However, their annotations of different viewpoints cannot be evaluated whether they have the same answering viewpoint of different answering viewpoint for the learner.

By discriminating utterances, candidate utterances that are selected based on their answering viewpoints are provided to learners. Learners use utterances of the same answering viewpoint to derive the same answering path, and also do those of different answering viewpoints to induce different ideas.

4 Prototype System

The function to add annotations to utterances is introduced to our collaborative learning support environment called HARMONY[6],[7]. Moreover, the function that displays a history of communication held with key utterances extracted is implemented. In HARMONY, a chat system is prepared as a communication tool. Figure 3 shows the interface of our system.

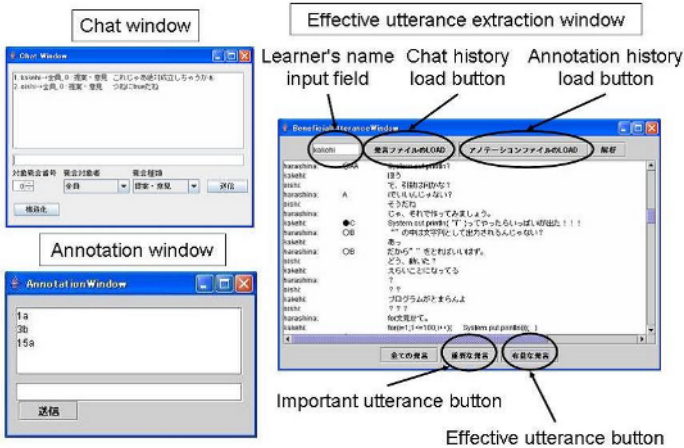


Fig. 3. Interface of the system

Learners use a chat window to communicate with other learners. When utterances are input to the chat window, IDs attached to the utterances are displayed to the text area of the chat window. Learners attach annotations through the

annotation window. On the annotation window, an ID of the target utterance and a type of annotation need to be specified. Types of annotations are indicated by alphabets: *a* represents the annotation of the same viewpoint, and *b* shows that of different viewpoint. For example, "3a" means the annotation of the same viewpoint for the utterance 3.

Effective utterance extraction window provides the history of a chat window with effective utterances marked by different symbols according to the answering viewpoint. On this window, learners select a file of target history from histories of collaborative learning by pushing the chat history load button. Learners also select an annotation file from an annotation database by pushing the annotation history load button. When the learner's name is input to user's name input field and the important utterance button is pushed, utterances to which the learner attached annotations are displayed with symbols that indicate types of annotations. Then, when effective utterance button is pushed, effective utterances are determined from all the annotation files selected by the learner and their symbols that indicate effective utterances are attached. Using these two buttons selectively, learners can review their collaborative learning history effectively by considering their selected viewpoints.

5 Experiment

To evaluate the mechanism for extracting effective utterances, two experiments were performed. First, the correctness of determining answering viewpoints of other learners was evaluated. In this experiment, one group of three examinees in our laboratory was asked to study an exercise concerned with the programming collaboratively. During the learning, examinees communicated with others through our chat system and made their program using Eclipse. In Eclipse, they were asked not to erase what they derived, but make the wrong sentences commented out when they wrote inappropriate program. Moreover, if they derived the answer triggered by specific utterances, IDs of the utterances were written down at the corresponding program. After the learning had been finished, the answering paths that individual examinees derived were compared, and correctness of answering viewpoints detected by the system was evaluated.

In this experiment, each answering viewpoint corresponds to each answering path. Since answering paths consist of several steps, answering viewpoints need to be compared for each step. The objective of the exercise was to construct the program that searches character strings from a file. There were four steps in a correct program. Examinees discussed each step in order. Here we focus on examinees A and B to evaluate the correctness of detected answering viewpoint. During the discussion, A and B constructed programs by the same viewpoint in steps 1 and 2, while by different viewpoints in steps 3 and 4. Figure 4 shows a transition of answering viewpoints of two examinees that are detected by the system. In this figure, x axis indicates IDs of utterances and y axis shows whether answering viewpoints are the same or different. In this figure, circles and triangles correspond to utterances of A and B respectively with their IDs attached.

Based on the result, our system could detect that examinees had the same answering viewpoint in the first two steps and different answering viewpoints in the latter two steps almost correctly. However, the system could not detect that the other examinees have correct answering viewpoints until the 13th and the 20th utterances, since they did not make utterances nor add annotations until those utterances. Similarly, the system could not detect that B has different viewpoint in step 3 until 109th utterance. The objective of our research is to detect the answering viewpoint of utterances to which annotation is attached, so it is not a problem that the system could not detect answering viewpoints when annotations or utterances are not put in. At the 128th utterance, the answering viewpoint of A from viewpoint of B changed to the same. At this time, B derived his answer differently to A, but made utterance along the viewpoint of A. Therefore, it is proved our system can detect the answering viewpoints of others correctly when learners utter their opinions with their own answering viewpoint and express their answering viewpoints to a group. Second, the

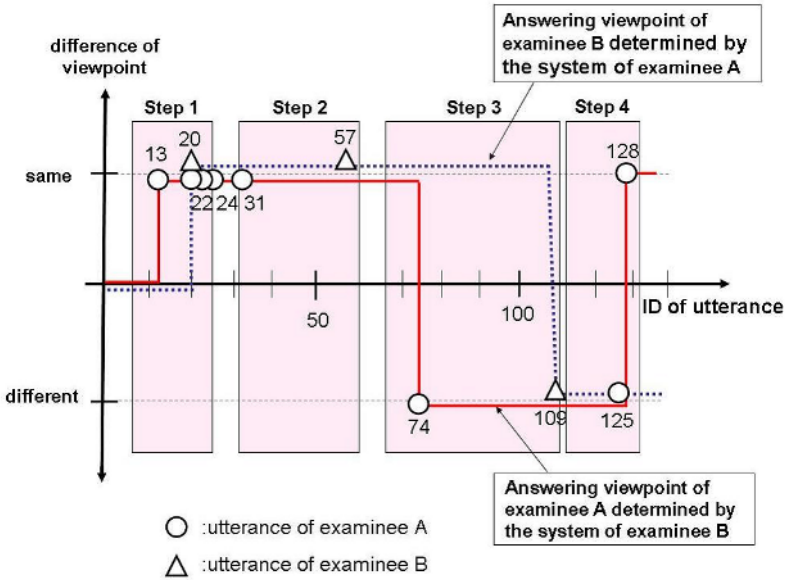


Fig. 4. Transition of answering viewpoints detected by the system

experiment that evaluates effectiveness of key utterances was tried. In this experiment, twelve examinees in our laboratory were asked to form 4 groups and study exercises of programming in groups. Then, after some weeks, they were divided into two groups and were asked to study the same exercise alone using our system of extracting effective utterances. In one group, examinees were shown chat history with marks of the same answering viewpoints. In the other

group, examinees were shown chat history with marks of the different answering viewpoints. Then, the programs derived by these examinees were compared.

Table 4 is the result of experiment. In the first group, five examinees were able to code the same programs as what they derived during the collaborative learning. Moreover, four examinees in the second group were able to derive the different programs from what they composed during the collaborative learning. Three examinees who could not make programs had not composed correct programs during the collaborative learning. Based on this result, to provide utterances with marks of the same or different answering viewpoints support examinees to derive answers according to the specific answering viewpoints.

Table 4. Results of experiment

	Group 1 with utterances of same viewpoints	Group 2 with utterances of different viewpoints
Same program	5	0
Different program	0	4
Not constructed	1	2

6 Conclusion

In this paper, we proposed a mechanism which detects viewpoints of learners and extracts effective utterances arranged by answering viewpoints from collaborative learning history. The experiment based on the prototype system shows that our mechanism could detect answering viewpoints of other learners correctly. Moreover, to provide utterances based on answering viewpoints was proved to support learners who derive answers in specific answering viewpoints.

Currently, the system only focuses on the answering viewpoints and types of annotations when detecting effective utterances. Therefore, learners' own utterances are also extracted as key utterances and provided to the learner himself. These utterances are not useful for the learner if he does not consider them important, so the system should be revised in order not to detect the learner's own utterances as key utterances.

Moreover, we should consider the mechanism for displaying extracted key utterances more effectively. Current our system provides all chat history with specific marks attached. However, if learners could see abstract of the conversation that contains only key utterances, they could acquire effective knowledge from chat histories more easily.

Acknowledgements

The authors would like to thank the 21st Century COE(Center of Excellence) Program for 2002, a project titled Intelligent Media (Speech and Images) Integration for Social Information Infrastructure, proposed by Nagoya University.

References

1. J. Zhao and H.U. Hoppe: Supporting Flexible Communication in Heterogeneous Multi-user Environments, Proc. 14th IEEE International Conference on Distributed Computing Systems, (1994), 422-429.
2. R. Plotzner, H.U. Hoppe, E. Fehse, C. Nolte, and F. Tewissen: Model-based Design of Activity Spaces for Collaborative Problem Solving and Learning, Proc. European Conference on Artificial Intelligence in Education, (1996), 372-378.
3. B. Goodman, M. Geier, L. Haverty, F. Linton, and R. McCready: A Framework for Asynchronous Collaborative Learning and Problem, Artificial Intelligence in Education, J. D. Moore et al.(Eds.), IOS Press, (2001).
4. L. Jackson: Concurrent Engineering in Construction Challenges for the New Millennium, Proc. 2nd International Conference on Concurrent Engineering in Construction, (1999), 37-46.
5. C. Bereiter: Situated Cognition and How to Overcome it, Social, Semiotic, and Psychological Perspectives, ed. D. Kirshner and J.A. Whitson, pp. 281-300, Hillsdale, NJ, (1997).
6. T. Kojiri, Y. Ogawa and T. Watanabe: Agent-oriented Support Environment in Web-based Collaborative Learning, Journal of Universal Computer Science, Vol. 7, Issue 3, (2001), 226-239.
7. T. Kojiri, K. Yamaguchi, and T. Watanabe: Visualization of Temporal Topic-flow in Collaborative Learning, Proc. of HCII2005, Vol. 5, No. 132, (2005).

A System Assisting Acquisition of Japanese Expressions Through Read-Write-Hear-Speaking and Comparing Between Use Cases of Relevant Expressions

Kohji Itoh¹, Hiroshi Nakamura¹, Shunsuke Unno¹, and Jun'ichi Kakegawa²

¹ Tokyo University of Science, 2641 Yamazaki, Noda, 278-8510 Japan

² Hyogo University of Teacher Education,
942-1 Shimokume, Yashiro, Kanto-gun, Hyogo, 673-1494 Japan
itoh@te.noda.tus.ac.jp

Abstract. We are developing a system aiming at supporting learners acquiring the ability of using expressions of Japanese as 2nd language. The system assists the author to develop a collection of expression notes edited and annotated to their locations in structured texts. The authored system assists the learners to retrieve and compare usage of the expressions. Based on the experience of an evaluation experiment, a collection of prototype expression notes and an expression relational map were introduced in order to facilitate authoring and learning by comparison of usage of the same or related expressions. We then describe the subsystem for diagnosing sentence-wise blank-filling composition and its authoring system, constructed using a lexicalized grammar with feedback of giving use cases of the erroneously used expressions. Finally we report on the phonetic assistance with visualized prosody in pitch and pause being synchronized with progression of the spoken moras of the voice of the authentic speaker as well as of the learner using a speech recognition engine.

1 Introduction

The systems of early days for assisting learning Japanese as second language were featured by incorporating electronic dictionary and letting translation in mother tongue to pop up when the learner clicks on a word in reading texts[1]. A recent example of more advanced assistance of learning second language makes use of a concordancer to retrieve sentences to show use cases of collocations[2].

We are developing a system to assist learning Japanese, as second language, based on the well known learning model according to which comparison of usage, with context, of the same or related expressions makes the learners to acquire ability of using expressions. The comparison is assisted by the system retrieving the expression notes edited by the authors with the text locations to which they are annotated[3][4].

The acquisition of using expressions are also phonetically assisted by making the learners to listen to and imitate the authentic speakers' narration reading the text, visualizing prosody of the authentic speaker and of the learner[4].

Also developed is assisting learning by production through diagnosis of sentence-wise blank-filling composition problems set up in the material text[5], with feedback of providing such sample textual locations as making aware of their mistakes.

In this paper we first review the prototype system constructed. And we next report on the preliminary evaluation experiment conducted using the prototype. We then describe incorporation of a collection of “prototype expression notes” and “expression relational map” we have decided to introduce based on the results of the experiment.

We then describe the subsystem for diagnosing sentence-wise blank-filling composition and its authoring. Finally we report on the phonetic assistance with visualized prosody in pitch and pause being synchronized with progression of the spoken moras of the authentic speaker and of the learner using speech recognition engine.

2 Prototype System for Comparison of Use Cases of Expressions

A prototype sub-system for assisting comparison of use cases of expressions was constructed. The teacher selects texts appropriate to her students and the system parses the texts to XML files being tagged according to the textual and syntactic structures. We made use of “KNP”[6] to parse the material texts to obtain XML files with syntactic tagging which are converted to DOM trees at run time.

She can author and annotate notes, for assisting the students, to those parts of the texts where such expressions as she thinks important appear. A note carries the (1) “semantic category” and the (2) “expression format” of the selected expression along with a (3) “memo” to which the teachers or the learners can freely add comments in the mother/mediating language of the learners, and also the (4) “syntactic feature” as well as the (5) “text location”.

The assistance for the author comes from a menu of semantic categories and expression formats for editing (1),(2), and also from a mechanism for matching the syntactic feature (4) of the specified expression with the partial structures of the texts, which assists the author to find the text locations where the expressions in concern appear [3]. (5) is automatically recorded by the system.

Now the system helps the learner making sense of the texts by way of the word notes that are embedded behind the words in the text display and pop up when the learner brings the cursor on the words. When the learner mouse-drag on a part of the texts, the notes annotated to that part appear in a menu. On selecting an expression from the menu, the learner is allowed to retrieve and select notes of the expression with the same semantic category (1) and/or expression format (2) as that of the expression note the learner has selected. The system highlights those textual locations to which the selected notes are annotated. The learner can compare cases of using those expressions naturally to acquire ability of discriminating the usage.

3 Preliminary Evaluation Experiment

We conducted a preliminary evaluation experiment of the prototype system explained in 2. with 14 students from abroad, of Tokyo University of Science, as subjects. They had finished learning basic Japanese as well as basic English.

The flow of the evaluation experiment was as follows:

- (1) pre-test (“basics problems” and “expression usage problems”),
- (2) study using the system (40 minutes), and after 1 week has passed:
- (3) post-test (“expression usage problems” and “extended problems”).

The “expression usage problems” of the post-test were the same as those of the pre-test and consisted of problems with blanks to be filled with expressions selected from a menu of expressions the learners were to learn in the study using the system.

After the pre-test, 14 students were divided into 2 groups of 7 members each, the experimental group and the controlled group, so as the average and the standard deviation of the pre-test results for each group were approximately the same.

In stage 2, the members of the experimental group studied using the full functionality of the prototype system, and for the members of the controlled group no service of retrieving sample text locations was given and only the comprehension support service was given.

The effectiveness of the expression usage explorative learning was measured by the difference, between the experimental group and the controlled group, of the score transitions from the pre-test to the post-test for the “expression usage problems” whose full score was 60 points.

As for subjects whose pre-test score was equal to or higher than 48 points, acquisition seemed almost saturated for the both groups. Small was the difference of the score transition for the subjects whose pretest score was less than 30 points.

In contrast, as for the subjects with the medium score, the difference of the down-from-full-score point transitions got visible. For instance, 2 subjects of the experimental group were seen to have made a noticeable progress by 7 and 10 points decrease compared with 2 of the controlled group by 5 and 3 points decrease.

After the experiment, a questionnaire was carried out. We list in the following the answers obtained from especially many subjects.

From the controlled group : “I think it is difficult to develop the ability in Japanese only by reading the given text”. “I feel it will enhance learning if a multiplicity of example sentences are provided in which the expressions in concern are used.”

From the experimental group: “It serves to understanding the expression in concern that we can refer to many example sentences in which the expression is used.” “It is confusing when we are provided too many examples.”

In summary, our system is likely to be useful in providing a multiplicity of use cases of the expressions, enabling learners to construct general rules by themselves.

The suggested problem of being confused by being provided too many usage examples may have been caused by the menu of the retrieved expression notes which indicated nothing identifiable as for the textual contents, so that once one moved to different text location it was difficult to return to the original location.

There became also apparent a problem of authoring to the effect that the authors had to edit exactly or almost the same contents to many different text locations.

Comparison between the usage of expressions of different but related categories would serve for the learners to differentiate the expressions, any assistance for which, however, was not implemented in the prototype.

4 Prototype Expression Notes and Expression Relational Map

In order to solve the problems stated at the end of section 3, we have decided to introduce a related collection of “prototype expression notes” as well as an

“expression relational map”, the latter being GUI visualization of the former. The prototype notes in the collection are related via a collection of “related pairs of semantic categories” and a collection of “resembling expression formats.”

An expression note object is generated from a prototype expression note object having a semantic category and an expression format, embedded with the textual location where such an expression featured by the prototype is used. And it can be edited of its memo field and its object reference is added to the list of the expression note objects the prototype note object holds.

When the learner selected an expression note from the menu shown on dragging on the text and has studied the content, she/he may simply exit the note for either going to the other expression notes or to the other text locations. Or she/he may as well click on “Map” button to go to the expression relational map. In the latter case the corresponding prototype node appears on the map. Now refer to Fig2.

A prototype node consists of concatenation of the semantic category-titled pane (we call S-pane) and the expression-format-titled pane (we call E-pane).

*Now on the prototype node is a “texts” button for popping up the menu of the text-annotated expression notes generated from the prototype note. On selecting from the menu with one line text tool tips display service, the textual location where such an expression as specified by the prototype is used is shown highlighted.

The S-pane carries a “sem” button for unfolding and folding the menu of the semantic categories for each of which there exists a prototype note with the same expression format as that of this node. The E-pane carries a “form” button for unfolding and folding the menu of the expression formats for each of which a prototype note exists with the same semantic category as that of this node.

On selecting either a semantic category or an expression format from the unfolded menu by the mouse left button, the corresponding prototype node with the other pane title being unchanged is placed on the Map.

On selecting a semantic category or an expression format (from the menu) by the mouse right button, a menu of the semantic categories or the expression formats in relation with the selected semantic category or the selected expression format appears. When one selects one from the latter menu, a prototype node appears carrying a default expression format or a default semantic category.

Once a prototype node appears on the map, a similar procedure beginning from the explanation marked by “*” in the above can follow.

Suppose the learner is led on the map to finding different use cases of a prototype expression with the semantic category “allowing”, and the use cases of different expression formats with the same semantic category. Then she/he may be guided to the related semantic category “forbidding” on the map, and again to finding different expression formats with the latter semantic category and different use cases.

5 Assisting Learning by Production by Diagnosing Compositions

The system also provides the learners with an environment to enable verifying acquisition of expressions by production.

For preparing the environment, the author is requested to edit, in the texts, “blank filling composition problems”, specifying the sentences with the list of the autonomous words (verbs, adjectives and nouns) to be used in composition. The system analyses the sentences being blanked to generate their semantic relational trees

making use of a Japanese LTAG (Lexicalized Tree-Adjoining Grammar) [5], which is a typical lexicalized grammar whose lexicons of functional words and “yogen” (Japanese verb or adjective) words will be given by the system providers and lexicons of “taigen” (Japanese noun) words will be given by the authors.

An LTAG lexicon consists of a surface lexical item and a tree structure consisting of *yp* and/or *tp* or *lp* with the surface item as a leaf, where *yp*, *tp*, *lp* represents, respectively, “yogen” phrase node, “taigen” phrase node and “locutional” phrase node. By way of substitution and adjoining of the nodes with sub-trees, the phrasal syntactic sub-trees are constructed along with the semantic relational trees through unification between the arguments provided in the nodes.

Using the LTAG-based generator program, the system then generates, from the resulted semantic relational trees, candidate variations of sentences among which the authors are requested to select those they think appropriate considering the context. The selected variations are recorded as local surface variants in the semantic relational trees (we call such a tree SRT-LSV).

The system in the near future will give a list of pairs of the specified autonomous words with high corpus-based probability of unintended modifying-modified relationship along with linking words or inflections. For the present, the authors are requested to give such a list (we call such a list UI-MM).

Given a blank to be filled in a context with the list of the autonomous words to be used, the learner composes sentences selecting word orders, inflections and functional words. A diagnostic system, making use of the SRT-LSVs and UI-MM, produces analytical diagnostic messages. The system, however, first retrieves and gives the learner, as feedback, such use cases of the expressions in the collection of texts as making aware of the error she/he committed, or of the un-reminded expressions. It is only after such feedback has been found useless that the diagnostic messages are given to her/him.

6 Phonetic Assistance of Acquiring Expressions

Based on the hypothesis that self-controlled voice prosody of the narration in reading use cases of expressions plays an important role in acquiring expressions, we are

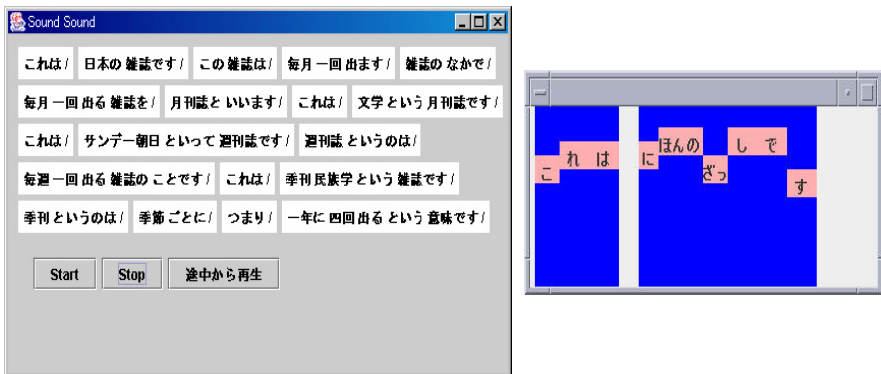


Fig. 1. Phonetic assistance with pause and pitch display synchronized with reproduction

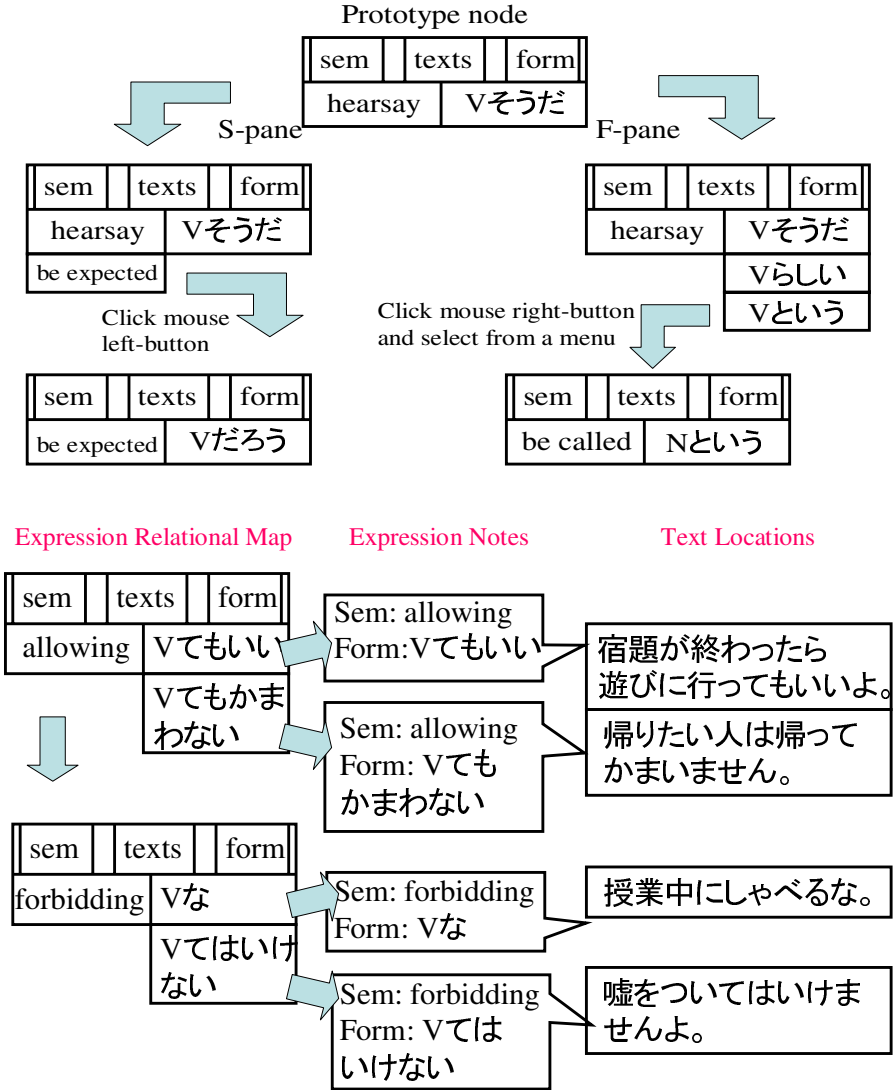


Fig. 2. How to use Expression Relational Map for navigating to related expressions and retrieving corresponding Expression Notes to compare use cases in text locations

developing a system to provide the learners with prosody learning assistance in the process of acquiring expressions by comparison of use cases.

In contrast to the prosody of English featured by rhythm and stress, the prosody of Japanese is characterized by pause and mora-wise pitch, the latter varying in time according to syntax-dependent “phrase components” plus word-dependent “accent components” owing to the Fujisaki’s model[7] of voiced Japanese. And we are developing visualization of pauses and pitches synchronized with visualization, in

characters, of syntax and moras in reproduction of the voice recorded by authentic speakers as well as of the voice in real time of the learner for learning by comparison.

For the purpose of synchronization we introduced a voice recognition engine with capability of learning parameters of the Hidden Markov Model and we made use of the time stamps of the moras the engine recognized using the sequence of morphemes, words and sentences given beforehand as constraint.

The pith of the voice is measured every 10 ms by a function incorporated in the engine using the autocorrelation method with dynamic time window processing. The pitch of a mora is determined by interpolation and averaging of the pitch data in the mora interval detected by the recognition function. Pause and pitch information is recorded in the tags of the XML files as well as in the DOM trees representing syntax-morphological structure of the sentences from which the voice-synchronized pause and pitch display is produced. Refer to Fig.1.

Also incorporated is the function with which the learner can command reproduction of the recorded voice of the authentic speakers or of himself from any of the chunks each of which we define as a phrase from a pause to the next pause of the authentic speakers' narration.

7 Conclusion

We described our system assisting learners acquiring the ability of discriminatory usage of Japanese expressions in understanding as well as producing sentences in contexts.

The system assists the author to develop a library of syntactically structured texts with a collection of Expression Notes each of which is created referring to the one selected from the collection of Prototype Expression Notes, edited and annotated to such textual locations where usage of the expression is found, and the author is also helped finding such locations by way of syntactic matching.

When the learner is interested in an expression, while being assisted by the Expression Notes as well as word notes to understand reading texts, they are guided to the Expression Relational Map on which the Prototype Expression Note appears as a node, and they can either directly retrieve and compare different use cases of the same expression or move to nodes of related expressions different in semantics or in form and retrieve their use cases in contexts to be compared.

In a preliminary system evaluation experiment with students from abroad of our university as subjects, we observed the subjects with the average score to have made a noticeable progress using the system.

Also described was the subsystem for diagnosing sentence-wise blank-filling composition and its authoring system, constructed using a lexicalized grammar called LTAG (Lexicalized Tree Adjoining Grammar) with feedback of giving correct use cases of the erroneously used expressions, or of un-reminded expressions.

Finally we reported on the phonetic assistance for acquiring expressions with visualized prosody in pitch and pause being synchronized with progression of moras with respect to the voice of the authentic speaker and of the learner based on a speech recognition engine.

We plan to evaluate effectiveness of the revised and function-augmented system.

References

1. Tera,S., Kitamura,T., Ochimizu,K.: DL: A System for Assisting Comprehension of Japanese, [http:// www.jaist.ac.jp/tera/](http://www.jaist.ac.jp/tera/)
2. St.John,E.: A Case for Using a Parallel Corpus and Concordancer for Beginners of a Foreign Language, *Language Learning &Technology* Vol.5, No.3, pp.185-203 (2001).
3. Kakegawa,J., Nakamura,H., Sekiya,M., Itami,M., Itoh,K.: Retrieving Sample Sentences by Using Natural Language Processing for Assisting Learners and Teachers of Japanese in Second Language Learning, *Japan Journal of Educational Technology*, Vol.25, No.2, pp.85-94 (2001)
4. Nakamura,H., Unno,S., et al: An Integrated Environment for Assisting Learning Usage of Japanese as Second Language, JSAL, Research Report, SIG-ALST-A503-09, pp.49-54 (2006)
5. Kakegawa,J. ,Kanda,H., Fujioka,E., Itami,M., Itoh,K.: Diagnosing Processing of Japanese for Computer-Assisted Second Language Learning, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong pp.537-546 (2000).
6. Kurohashi,S.: KNP version2.0 b6: Syntax Parser of Japanese, (1998)
7. Fujisaki,H., Narusawa,S.: Automatic Extraction of Model Parameters from Fundamental Frequency Contours of Speech, *Proc. 2001, 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pp.133-138(2002)

A Web-Based System for Gathering and Sharing Experience and Knowledge Information in Local Crime Prevention

Masato Goto¹, Akira Hattori², Takami Yasuda³, and Shigeki Yokoi³

^{1,3} Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
masato@nagoya-u.jp, {yasuda, yokoi}@is.nagoya-u.ac.jp

² Faculty of Information Technology, Kanagawa Institute of Technology,
1030, Shimo-ogino, Atsugi, Kanagawa, 243-0292, Japan
ahattori@ic.kanagawa-it.ac.jp

Abstract. In this paper, we propose a Web-based system to promote public safety and security that uses grassroots information to prevent and solve crime. The feature of our system is that it gathers and formalizes information related to crime or suspicious activities. The system provides useful information by arranging it with time and geographical attributes. As a result, people can find out about crime and suspicious happenings in their local area. Using this system enables people to share their experience and knowledge before crime actually occurs. This can lead to strategies for self-defense against ever-evolving criminal techniques.

Keywords: Crime Prevention, Local Community, CMS, Information Sharing.

1 Introduction

In Japan, it is becoming increasingly important to protect oneself against familiar crime or trouble in daily life [5]. To secure safety and a peace of mind, it is essential to obtain helpful information for the crisis management. In recent years, many people have been requesting the practical use of ICT (Information and Communication Technology) in the field of crime prevention. Consequently, the development of information-sharing systems that use the Internet and cellular phones is progressing rapidly [1],[2],[4]. There is also growth in the incidence of events that might lead to crime, such as phishing scams and fictitious claim. Other crimes stem from poor judgment by the victims, often related to money. It is therefore important to recognize criminal means and ways to prevent and solve crimes. An example is in noticing suspicious persons or vehicles, where each local resident may carry out the important role of controlling and preventing the crime.

Websites of public organizations offer solutions to or examples of crime cases in which consultation or notification have played a part. This information is very useful because it acknowledges details provided by victims of crime. However, suspicious events occur far more often than they are reported. When people see something suspicious and investigate, the suspect will generally deny they were about to commit a

crime. Therefore, it is difficult to know the real state of affairs, what kind of experiences people had, or how they coped with them before anything bad happened. However, intelligence and knowledge from individuals are as important as the information that official organizations give when performing crisis management. It is hoped that by putting this intelligence to work and by sharing experiences with people in one's local area, awareness of crime prevention will increase, thereby building the local community into a safe place [3],[5]. We need to discuss safety support based on the real needs of citizens.

The purpose of this study is to use the Web to support people's safety and security in their daily lives, aiming at gathering and distributing grassroots information to prevent and solve crime. First of all, we research the requirements of users by investigating what happens when a suspicious event takes place, what action they take in response, and what information they need. Second, we examine the Websites of certain public organizations to find tendencies about what kind of information is currently on offer in the present condition, and what kind of information is not being offered that is related to crime and trouble. An information gathering and sharing system is then designed based on those results for individual crisis management.

Features of this work are that we formally gather grassroots wisdom and stories of experiences, since it is difficult to grasp the real state of affairs at the moment, and appropriately manage and provide data intelligibly with characteristics about the information collected. Moreover, our system has a feature for clearly obtaining methods of action to take and deciding information to share at the regional level.

2 A Present State and Problems of Safe and Secure Information

In this section, we discuss needs investigation and grasping problems of the information which public organization offers by Web.

2.1 Demand of Safe and Secure Information

Through this research we are also aiming to offer information that gives people some sense of security in their daily lives when a suspicious event occurs that leads to a crime. As the first step in our investigation, we asked two questions to ten people from one's twenties to one's fifties. The following results were obtained:

- (1) What do you do when you feel uncomfortable about crime in your daily life?
 - Talk to or ask for advice from public organizations such as the police or a consumer center. Talk to family or a friend.
 - Search for information whether something similar has happened before. Want to hear the opinions of people in the same circumstances as myself.
 - Pay closer attention to related newspaper articles for a few days.
- (2) What kind of information do you want that relates to crime and suspicious events in your daily life? How do you want to obtain it?
 - Want to know whether an incident was isolated or whether there have been similar cases.
 - Want information about what to do from the specialists in public positions, such as administrators, lawyers, etc.

- Want information about precedents that happened in the past in the local area.
- Want to know public safety information such as the crime statistics of an area when one moves to a new address.
- Think that some kind of information about suspicious people and risky places should be shown on a map
- Would like a Website that provides detailed information about recent crime cases, which makes it easy to gather data quickly.

Two things are clear from question (1):

- Interviewees want to receive opinions from people in various positions.
- They want to obtain intelligence about similar events.

Moreover, from the results of question (2), we recognized that it was important to offer the information given below.

Before a suspicious event happens: Attention information / Statistical data / Crime map (crime-prevention map)

After a suspicious event happens: Information on similar cases / Solution

To respond to the replies to question (1), it is important to collect the desired information from not only official organizations but also from people who have actually encountered crime or similar trouble. The answers to question (2) reconfirm the importance of using information available on the Internet or via cellular phones.

2.2 Safe and Secure Information on the Web from Public Organizations

We found through the needs investigation that people need information offer by public organizations. Therefore, we investigated the Website of each prefectural police force and of the Consumer Affairs Center to find data dealing with familiar crimes and checked which types of information are offered. As a result, we found that concrete information on crimes and whether they were solved does indeed exist on Websites of those public organizations. In particular, there is plenty of information about actual cases of bag-snatching, burglary and pick-pocketing, false claims, and other criminal techniques. The contents include general notes and warnings, solutions for preventing crime and so on, which can be found from crime situation data that are clearly presented by month and location. The information is mostly provided in the form of maps, tables, graphs, and text.

The Shimane prefectural police is one organization that has decided to use a modern method of collecting information: they accept tips on suspicious-looking people and incidents before a crime actually occurs, such as potential theft, a suspicious person's appearance, reckless driving, suspicious sales, etc. However, it is very difficult to find systems for applying and sharing information about a local area on the Web. Therefore, it is difficult to grasp details from most public organizations about minor cases in local areas that might lead to more serious crimes in future.

2.3 Present Problems and Solutions

Consequently, regarding both the information gathered from residents and that from public organizations, we need to solve the following problems by analyzing data in terms of satisfying people's needs.

- There is no place to accumulate data on suspicious events that happen in our daily lives.
- It is difficult to know the facts of crime, suspicious incidents, or other types of trouble in one's surroundings.
- The relevant information about crimes and suspicious events are scattered widely.

To solve these problems, we consider that the followings points are important:

- Gather facts about one's surroundings. This includes information on the what, when, and where of suspicious events. This information should be gathered formally, by taking suitable offerings into account.
- Experience, wisdom, knowledge, etc. should be included in the information gathered from residents. This means local residents can effectively act as advisors when information is being accumulated.
- The information held by public organizations needs to properly sorted and offered as general-measure information as well as information specialized for local areas.
- It is necessary to design an offering structure according to the type of suspicious event and to the attributes of the information. Particularly, time and geography are the critical factors to arrange information.
- Observed information should specify the place. Information on incidents like phishing scams and fraudulent renovations should be arranged at the town level to consider the privacy.
- To find the most recent crimes, all information should be arranged by time.

In the following section we propose concrete system to meet these demands.

3 Local Crime Prevention Information Sharing System

In this section, we propose a system that realizes the solution plan given in Section 2.3.

3.1 System Overview

The feature of our system is that it gathers formalized information related to crimes and suspicious events, and offers coherent information by arranging time attributes and geography attributes. This system can easily search the latest criminal techniques, crime scenes, and the nature of crimes committed. Therefore, it is possible to gather and offer information that is different from that on mailing lists or BBS (Bulletin Board Systems). Besides these basic functions, we can easily add other functions as modules, since different users have different needs. Figure 1 shows a concept chart of the system.

This system collects on the server information about what happened and where from people with experience and advice on the events. This information is input by users via special online forms. This means we gather not only information from general users, but also that from public organizations that can provide reliable data. Such information is intelligibly displayed as a map combining a time category table by effectively using RSS (RDF Site Summary) and Google Maps. In this research, we constructed the system by using XOOPS, which is typical open-source CMS (Contents Management

System) software, taking into consideration factors such as account management and function management. XOOPS is suitable for this system because it contains many functions, including user management, a message-sending function, a search function, and site link construction, etc. The following subsections provide further details about gathering formalized information and offer ways to arrange time and geographic attributes.

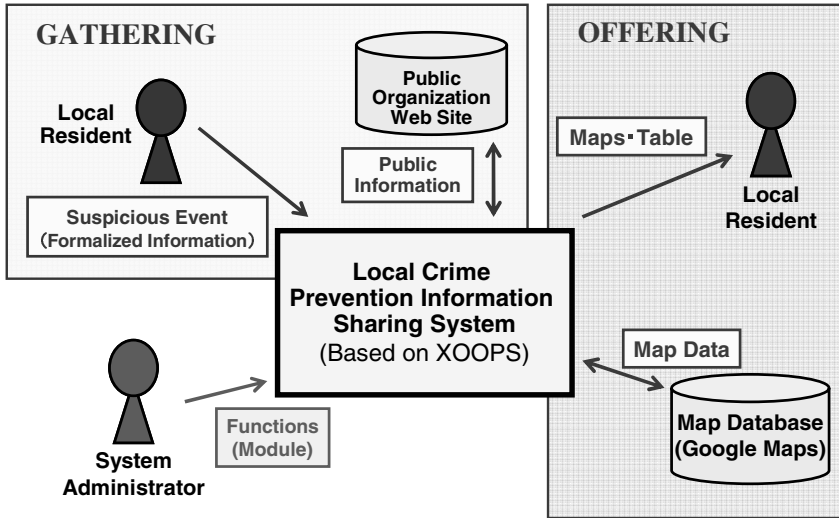


Fig. 1. Concept of the system

3.2 Information Gathering Phase

Target Information. Table 1 shows an example of information gathered by the system. To manage this information easily, the following four items are used to classify data for all cases: (a) time information; (b) place information [Range that individuals cannot specify]; (c) detailed information about incidents; and (d) comments. A formalized package is created by the addition of a few miscellaneous items to the above four classes.

Table 1. Gathered Information

a means, a start	suspicious event
Telephone	phishing scams, persistent persuasion, others
postal matters	false claims, financial fraud, others
Visitors	suspicious door-to-door sales, suspicious persuasion, others
Observations	suspicious persons, suspicious vehicles, suspicious objects, others
others	bad or smoky smells, an accident around the house, others

Management Method and Interface. To manage updated information and search for every element, one RSS file is generated in each one of the event packages collected. Moreover, since the number of senior-citizen victims has been increasing, our system provides a simple interface for people who are not accustomed to using computers (Fig. 2).

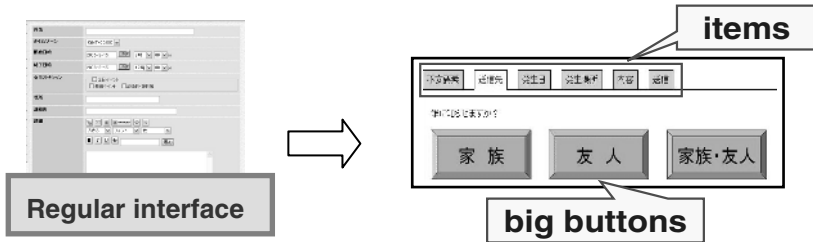


Fig. 2. Simple interface

3.3 Information Offering Phase

Our system mainly arranges and offers information by time and geographical attributes. It offers information by deciding which attribute should be dealt with independently based on the data requested.

日	月	日	2008年 3月	日	欄	欄	欄
1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	

Fig. 3. Time category table



Fig. 4. Local incidents map

Arrangement by time attribute (Fig. 3): Show the time of a suspicious event by using the table makes searching easy if contributed information is relevant to other cases. It is effective for finding the latest information on suspicious persons, etc.

Arrangement by geographical attribute (Fig. 4): This attribute deals with information for which location is important, and is offered via maps (Google Maps). Maps make it easier for users to grasp what happened and where in their neighborhood.

3.4 Other Functions

The functions we mentioned above are basic functions, and are mainly realized by XOOPS. However, additional functions can be added as modules according to

individual needs and ideas, and communities can be formed through interactions among members and visitors. The following two functions have also been developed.

- A function to support family and close friends when a suspicious or criminal event has occurred.
- A function for senior citizens that can change the interface to a simple one or show a software keyboard.

In another study we are working on an information promotion project for seniors in Nagoya City, called the “e-namo project.” The aim is to provide useful software for senior citizens such as a Web directory search system, a Web mail systems, and a software keyboard system. In Japan, due to rapidly aging population, crime committed against senior citizens is increasing. Consequently, there is a growing demand for software that includes practical safety and security functions, which has led us to examine the issue of crime-prevention information in an e-mail magazine by using RSS. This system will encourage cooperation with public organizations that offer crime-related information.

4 Evaluation and Discussion

We evaluated our system by analyzing comments from five users who cooperated in this study. The evaluation method was based on analysis of their opinions about the system. Most of them replied affirmatively to the question of “Do you want to try to use this system?” Examples of extracted comments include: I want to try it if it is easy to operate. / I want to use it if I am likely to find what I am looking for. / I think that such a system is necessary. Extracted comments from the question of “What do you think about the functions with which information can be shared by time and geographical attributes, and the other functions?”: I think it is good to grasp information about one’s local area on a map regarding suspicious persons, etc. / I want to try to use the mechanism that enables me to share information with my family.

From this evaluation, we consider that our system, developed based on the concept for this research, is effective. We also understand that it is important for the system to be easy to use when a crime or suspicious incident has occurred.

5 Conclusion and Future Works

In this research, we proposed a Web-based system that uses grassroots information to promote public safety and security in response to Japan’s recent rise in crime. We designed a system to acquire information about crimes and suspicious incidents in local area, which is in high demand but difficult to find. People are also able to use the system to access relevant data offered by public organizations such as the police. Moreover, this system can gather regular and reliable information via the user management function, and it can share requested information with time and space attributes. We have also developed a simple interface for people such as senior citizens, who are not accustomed to using computers. The usefulness of the system was

confirmed, although the evaluation was only a preliminary one. We predict that use of this system in practice will make it possible to share intelligence and experiences among people before a crime happens. In sharing the latest information, defenses can be built against ever-evolving criminal techniques. Moreover, this should produce a deterrence effect by raising local residents' awareness of crime.

Future work will involve developing a more convenient and useful system that includes a unique interface, information filtering techniques for selecting and notifying useful information to a user, deduction functions for predicting future crimes, and test it in a large-scale experiment.

Acknowledgement

This research was partially funded by Grants-in-Aid for Scientific Research, 21st century COE program "Intelligent Media (Speech and Images) Integration for Social Information Infrastructure" from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Koji Asano and Kiyoshi Nakano: Development of a Secure and Peaceful Community with Information and Communication Technology. IPSJ SIG Technical Report, 2005-EIP-27, pp. 9-16 (2005)
2. Mariho Kataoka, Kazuhiro Takeda, and Nobuhiko Nishio: WOWnet:Community-Based Security Network. IPSJ SIG Technical Report, 2005-UBI-7, pp. 67-73(2005)
3. Kimihiro Hino, Rikutarō Manabe and Osamu Koide: Achievement of Regional Safety Classes in Partnership with Various Subjects: Making Regional Safety Maps with 'Kakiko Map'. Reports of the City Planning Institute of Japan, No. 3, pp. 59-62(2004)
4. Kingston, R., Carver, S., Evans, A. and Turton, I.: Web-based public participation geographical information system: an aid to local environmental decision-making. Computers, Environment and Urban Systems, Vol.24, No.2, pp.109-125 (2000)
5. <http://www.kantei.go.jp/jp/singi/hanzai/dai3/3siryou2-3.pdf>

The Scheme Design for Active Information System

Ping Zong¹ and Jun Qin²

¹ Computer of College, Nanjing University of Posts and Telecommunications
210003 Nanjing, P.R. China
zong@njupt.edu.cn

² College of Communications and Media, Nanjing University of
Posts and Telecommunications
210003 Nanjing, P.R. China
qjun@njupt.edu.cn

Abstract. The scheme design of active information system can provide the high capability and flexibility of developing application systems. We propose a framework for the design and realization of the active information system, to realize the active requirements from database or application in a uniform and convenient mechanism. According to the framework of an active information system (AIS), some important techniques are discussed. The definition and processing of the event and rule are described. The execution model and realization strategy are explained. This approach is very effective and practical to develop the active information application system.

1 Introduction

Traditionally DBMSs have been passive: that is queries or transactions are executed only when explicitly requested. Data values of some database instances are related through certain dependencies and restricted through certain constraints [1, 2]. As the scale and complexity of data management increased, interest has grown in bringing active behavior into database, allowing them to respond independently to data related events. There are many demands for active functions in database applications, i.e., function to real time inspection and control, reaction to the instantaneous state in application system, automatic service from information system and so on [3, 4].

The community of information system researchers addressed issues of managing information within a large dynamic system. The general model of an information system is in form of a collection of semantic services [5]. Active database system (ADB) detects events and does related trigger actions as a result of this detection, according to the condition. Most of the active capabilities are provided by a set of ECA (event condition action) rules [6]. The distinction between a database and an information system is best appreciated when we consider their function. The task of a database is to store data and answer queries. The task of an information system is to provide a service [5]. We argue for a change in viewpoint, so AIS is an interactive service-providing systems rather than mere data transformation engines. This change offers a promising approach for addressing ADB shortcomings, and reveals a roadmap of additional features for ADBs [5]. However, the practical technical approach is not given.

In this paper, we propose the scheme design of active information system (AIS) that integrates the passive functions in traditional database systems and active functions in application environment together. The active requirements can be implemented by unified mechanism. For example, some applications require automatic monitoring of conditions defined over the database state and a capability to take actions when the state of the underlying database changes. The performance of database system application can be improved. According to the internal status or the external event of database system, the active data processing and application service are afforded. Active views provide active caching of materialized views to support subscription services and notification services, and to update user profiles specially [5]. Furthermore, we can receive the high performance and flexibility in AIS.

This paper is organized as follows. The framework of AIS is given in Section 2. Main design techniques are introduced in Section 3. Section 4 expresses the realization Strategy of AIS.

2 Framework of AIS

Currently, RDBMS is widely used. Most users are familiar with RDBMS, and the advantages of RDBMS are well known. Rule capability is provided in many commercial systems, but it is not sufficient as it only provides basic triggering capabilities [7]. In the general database system the data integrity and consistency can be executed automatically by DBMS. All aspects of the data integrity and consistency are defined beforehand during the system realization. Users can not set any especial event that they need, so that their many active requirements cannot be satisfied.

Information system is founded on a database system. If we can extend its capability for active processing, we provide the functions of AIS. Synchronously combining with the active requirements from applications, the active events from database or application may be applied in uniform and convenient methods. So software developer can define the events and rules in especial application environment. During the running of application system, AIS detects the occurring event automatically. If the event that the related rule needs rises, AIS accomplishes the correlative database operation, control operation and so on.

The framework of AIS is composed of passive service, active service, data definition language, data manipulation language, active data manipulation language, event processing, rule interpretation, active function interface, database management system, rule database and database. The framework shows in Figure 1.

The passive service is the basic user interface and deals with database queries and operations. The active service is the additional user interface and arranges active functions using events and reports especially, so that it enhances active capability of general information system. Data definition language and data manipulation language are the basic components that preside over the essential data processing. The active data manipulation language was designed for the description of active requirement and active processing from the application field. The event processing is for the event definition and the event detection. It is noticeable that event processing includes two parts. The first part is supported from database system. The second part is added from AIS. The rule interpretation activates the rule which is triggered by related events and

fulfills the related actions. The active function interface is a middleware which transfers messages between the DBMS and other components. The Rule database stores and manages rules of AIS.

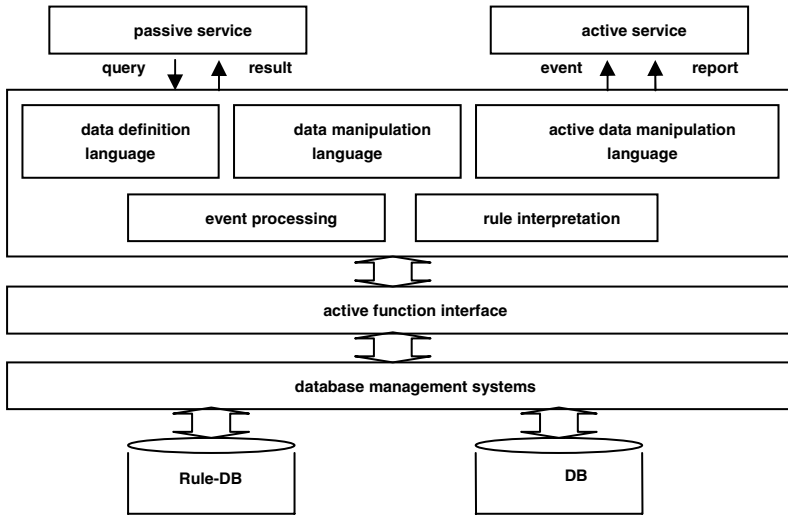


Fig. 1. Framework of AIS

According to this framework of AIS, we designed and realized the decision aided support system for preventing or controlling flood (DSS-PCF). The general requirements in the application field are arranged by the passive service. On the other hand, the active requirements from application field are arranged by the active service. Active function interface is a virtual machine that manages the active function which is not supported by DBMS.

3 Main Design Techniques

The idea to design AIS is defining the given events and related rules in conformity to the application condition in the information system. When any event which is needed by the rule occurs, AIS executes the database operation, system control and user exit routine in the agreement with the action in the related rule. Therefore the capability of active information processing is offered and satisfied for active requirements in application fields.

3.1 Event Definition and Processing

Events in DSS-PCF are separated into two kinds: basic events and complex events. The basic event is the elementary unit to define an event. It describes an event which can occur in any time from internal or external database system. The complex event is assembled with some basic events by means of algebra.

The kinds of the basic events are as follows:

1. A data processing event is the event which occurs after or before a database processing acts on the database item in database system. Its expression is:
[BEFORE | AFTER] □ INSERT | UPDATE | DELETE | QUERY □ database_item
2. A transaction processing event is the event which occurs at beginning, end or cancel of a database transaction processing. Its expression is:
event □ BOT | EOT | ABORT | [transaction_name] □
3. A time event is the event which occurs at appointed time or during a time period. Its expression is in two kinds:
event □ year. month. day □ hour : minute □
event (EVERY frequency [YEAR | MONTH | DAY | HOUR | MINUTE]
time WITH interval
4. A random event is that m events occur in an event list. A random event is independent of the time sequence. Its expression is:
ANY (m, e_1, e_2, \dots, e_n) $m \leq n$
5. A count operator describes that event occurs fixed times. Its expression is:
COUNT (event, times)
6. A user-defined event provides users to define some especial events which come from the application fields.
DEFINE EVENT event_name
BEGIN
event_specification
END

There are five arithmetic operators to construct the complex events. They are NOT, AND, OR, COUNT and SEQUENCE. The arithmetic operators of NOT, AND and OR are the same as relation algebra. The COUNT operator describes a complex event that event occurs fixed times. The SEQUENCE operator describes that several events occur in a determinate sequence.

An event appearance is identified by event processing. All occurring events insert the event recorder. After the action related with the event has been processed, this event is deleted from the list of the event recorder. The data processing event and transaction processing event utilize the event trigger in database system to realize the event processing. The time event is defined in application system and inspected by the subsystem for the time inspection.

A user-defined event is defined and activated by oneself. Mostly, the main problem addressed here is the description of services. It is embedded in the application with user_exit. A service description specifies how task in the service ought to be processed. So a user-defined event is to resolve the extra active services, which are from the business rules and can not be afforded by DBMS itself.

The complex event definition depends on the time. We design first matching algorithm as the distinguishing method for the complex event. That means the searches in the event recorder is in the time order. The firstly researched event which accords with the condition is matched firstly. An example complex event of Flood_Alarm is defined as follows:

```
DEFINE EVENT Flood_Alarm
BEGIN
  AFTER UPDATE Flood_Flux AND Flood_Flux ≥ 5000
END
```

3.2 Definition and Processing of ECA Rule

An active database system must provide a knowledge model (ECA rule model) and an execution model for supporting the reactive behavior [8]. This is in contrast with the execution which determines how a set of rules behaves presented in AIS. The knowledge model essentially supports the description of the active functionality. ECA rule is a production rule. It is composed of the event, condition and action. The language for active data processing is realized with the mechanism of ECA rule. The mechanism of ECA rule includes the rule specification, rule execution and rule management. The rule specification describes the event characteristic, relevant condition, action execution and relationship among them. The syntax of ECA rule in DSS-PCF is as follows.

```
rule_definition ::= RULE rule_id
                  ON event
                  WHERE condition
                  DO action
                  [ PRIORITY number]
                  END RULE
```

The defined ECA is stored in the rule database after the processing of the rule interpreter. ECA rule is a data item and can be stored in DB. It is managed with DBMS, but we design a special program module in the component of the active function interface to search and identify the appointed rule. According to the concept of the transaction processing in DBMS, a transaction processing defined by user is actually the process structured with some data processing in the certain sequence. The action of ECA rule explains how the data is managed. In order that DSS-PCF keeps in active status at all times, there is a stipulation that the last command of the command sequence must be the commit command. After the action in ECA rule finished, the system control is transferred to the event processing, in order to keep the activity and timeliness of DSS-PCF. Because the action of ECA rule is not changed anymore in the execution, ECA rule is not allowed in the form of the self recursion. A given example for the ECA rule expression is as follows:

```
RULE Flood_Schedule
  ON Flood_Alarm
  WHERE occur in 2 times in 1 day
  DO start flood prevention process
  PRIORITY 20
END RULE
```

3.3 Component Technique

Based on the component design method, we use directly the DBMS to fulfill the elementary data definition and processing function, which satisfies the passive function. The event or action of the non business logic can be defined as the component in DBMS [9]. In DSS-PCF there is an additional middleware, which is the active function interface, among communication. This middleware is in charge of the all communication between the active operation and DBMS. We encapsulate the active

processing function in the middleware as a component, in order to adapt the change of the active processing function and provide the reuse of active data object. The message queue in DSS-PCF may easily be transferred between the database requirement and the component with the active function interface, so that it is easy to realize active application system. Because there is the commonness among the different active business logic, it is very effortless to maintain the active application system and realize the reuse technique.

3.4 Execution Model

The execution model expresses the execution method of ECA rule. It includes the kinds of coupling modes, semantic description, action of ECA rule and relationship of user transaction among the various parts of active rules. With the processing of the triggered actions related to user transaction, the system attains the serialization and consistency for the transaction processing. The execution model is dependent on DBMS tightly and extends the normal transaction model. In the execution process of the ECA rule, the coupling mode between the event and condition or the condition and action can be separated into the immediate mode, delay model and independent model [6].

In DSS-PCF we use the immediate mode. It means that the condition is immediately evaluated, after the event was triggered in the coupling mode between the event and condition and that the action is immediately executed after the condition was evaluated successfully. So the complexity of the transaction processing is depressed. If an event triggers the multiple ECA rules, the collision problem will appear. We take the predefined priority of ECA rule to resolve the execution sequence of ECA rules, when an event triggers the multiple ECA rules. After multiple ECA rules are triggered, the rule interpretation executes the related rule by the predefined priority.

The rule scheduler offers the dynamic management of the ECA rule. In DSS-PCF the rule scheduler is the primary part to ensure the execution of every rule correctly. Performance comparison of the different active rule execution semantics should be done only in the context of equivalent rule programs. Hence, it should be guaranteed that, given the same initial database state and user or application transaction, the rule program yields the same final state independently of the adopted rule semantics, i.e., rule programs with different semantics are equivalent [10]. Using immediate execution semantics, action execution takes place after the event of the rule containing the action is triggered and its condition is evaluated [11]. We can refer to the concept of triggered transaction [12]. In the processing, every rule has its own process identifier after the rule triggered. And then this rule is put by the priority in the same as the process. If all triggered rules execute, the current transaction will be suspended. This transaction will be renewed until the all rules in the immediate coupling mode have been performed.

4 Realization Strategy of AIS

At the present software environment, there are two approaches of the design scheme. One approach is that AIS is developed in the environment of the active database

system. Another approach is that the toolkit for designing AIS will be developed beforehand. Then the active application system is developed on this toolkit.

As a result of the constant development of business database systems, for example ORACLE, the event definition and the stored procedure are realized in DBMS or in the developing tools. A developer controls the database via low layer interface, uses the SQL, database trigger or stored procedure to develop the interface of the database management. Certainly the developer can embed the programming language and the control mechanism from the operating system into the database applications with pro*C, user_exit, C++, etc.

In fact we utilize the approach that the toolkits will be developed for designing DSS-PCF. This approach is an ideal method to develop an AIS. It combines with the integration technology and development technology. So it is possible that the application systems can be developed with the high efficiency and the good quality. Based on the framework of AIS in Figure 1, data definition language and data manipulation language adopt the extended SQL supplied by DBMS. The other modules were developed by the toolkit implemented with C++. This toolkit can be considered as a virtual machine above DBMS. This virtual machine extends the functions of the traditional database system and offers the processing ability of the active service requirements. The rule database, the rule processing subsystem and AIS toolkit make up an AIS core. AIS toolkit is the foundation to sustain the ECA rule processing and time management. It solves the problems that traditional database systems can not directly support the event management, rule realization and action execution.

5 Conclusion

AIS can describe the function specifications in the application environment more availablely and make the reaction actively for the different events in the application environment. These actions are all kinds of the database processing and command operations related with the events occurring from the AIS. AIS advances the design quality and application functions of the information system. By the practice the design scheme is a very effective approach to develop the active information application system.

References

1. U. Dayal, B. Blaustains: The HiPAC Project: Combing Active Database and Timing Constraints. SIGMOD RECORD, Vol.17, No.1, March 1988
2. Shaoyuan Li, Xiaopei Luo: New Database Technology. Tsinghua Publishing Company, 1997
3. K. R. Dittrich: Active Database System. Institute for Information, University Zurich, 1992
4. Norman W. Paton, Oscar Diat: Active Database System. Uni. of Manchester, 1998
5. D. Goldin, S. Srinivasa, V. Srikanti: "Active Database as Information Systems" International Database Engineering and Applications Symposium (IDEAS'04). July 2004
6. David Botzer, Opher Etzion: Self-Tuning of the Relationships among Rules' Components in Active Databases Systems. IEEE Transactions on Knowledge and Data Engineering, March 2004

7. Lijuan Li, Sharma Chakravarth: An Agent-Based Approach to Extending the Native Active Capability of Relational Database Systems. 15th International Conference on Data Engineering (ICDE'99), March 1999
8. Chanho Ryu, Hyeok Han, Yongkeol Kim, Young-kuk Kim, Seongil Jin: Kernel Structuring Using Time-Triggered Message-Triggered Objects for Real-Time Active DBMS in Layered Architecture. Third IEEE International Symposium on Object-Oriented Real-Time Distributed Computing. March 2000
9. Hui Xu, Min Chen, Xiaoyu Zhang: Study of Component-Based Active Database Models. Computer Engineering & Science, Vol.23, No.6,2001
10. Elena Baralis, Andrea Bianco: Performance Evaluation of Rule Execution Semantics in Active Databases. 13th International Conference on Data Engineering (ICDE'97) . April 1997
11. Danilo Montesi, Elisa Bertino, Maria Bagnato, Peter Dearnley: Rules Termination Analysis Investigating the Interaction between Transactions and Triggers. International Database Engineering and Applications Symposium (IDEAS'02). July 2002
12. Kam-yiu Lam, Tony S.H. Lee: Approaches for Scheduling of Triggered Transactions in Real-Time Active Database Systems. 24 th. EUROMICRO Conference Volume 1 (EUROMICRO'98). August 1998

R-Tree Based Optimization Algorithm for Dynamic Transport Problem

Naoto Mukai¹, Toyohide Watanabe¹, and Jun Feng²

¹ Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
`naoto@watanabe.ss.is.nagoya-u.ac.jp`

`watanabe@is.nagoya-u.ac.jp`

² Hohai University, Nanjing, Jiangsu 210098, China
`fengjun-cn@vip.sina.com`

Abstract. The scheduling problem of transport vehicles is one of the important issues in our international society. Such a problem should be solved optimally and instantaneously to keep high profitability and conveniency. However, it is hard to find an optimal solution in such a transport problem because transport requests occur constantly and the optimal solution changes according to the requests. Therefore, the strategy to find a near-optimal solution is a practical way to solve the problem at low calculation cost. In this paper, we focus on the range constraints of the optimization problem. The assignment of requests is optimized by request exchange between transport vehicles in their belonging group (i.e., range constraint) to keep the calculation cost low. The group of transport vehicles should be dynamically changed on the basis of their positions. Hence, we introduce an indexing tree called R-Tree and reconstruct the tree in certain intervals to index moving transport vehicles dynamically. In the last of this paper, we report the influence of group size and the update interval on the performance of transport vehicles and then show the effectiveness of our optimization algorithm.

1 Introduction

The management of transport systems changes over the years with the development of global positioning system (GPS). In fact, the system called demand-bus (or demand-taxi) [1,2] provides the position information of buses to users in real time over the Internet. Moreover, the traveling paths (routes) of demand-buses can be changed according to the request of customers. The scheduling problem of transport vehicles (such as demand-bus or ambulance-cars) should be solved optimally and instantaneously. However, it is hard to find the optimal solution of the scheduling problem because transport requests occur constantly and the optimal solution changes according to the requests. The scheduling problem can be regarded as a combinational optimization problem between vehicles and requests, i.e., “which vehicle is assigned to the request?”. Basically, such combinational optimization problems are NP-hard. Hence, the strategies called

heuristic algorithms [3,4] (such as genetic algorithm or ant colony optimization) to find near-optimal solution are practical ways to solve the problem.

In this paper, we focus on the range constraint of problem space to reduce the calculation cost. We regard the range constraint as the group of transport vehicles. The assignment of requests is optimized by request exchange among vehicles in the group. However, the groups should be dynamically changed because transport vehicles change their positions according to the assigned requests. Therefore, we introduce a tree structure called R-Tree[5] to index transport vehicles on the basis of their positions. A leaf node of the tree represents a minimum unit of a group. A hierarchy of the tree represents an optimization level.

The remainder of this paper is as follows: Section 2 formalizes the transport problem we address. Section 3 describes how to form groups of transport vehicles. Section 4 describes how to optimize the assignment of requests and presents two cost measures: time and distance. Finally, Section 5 concludes and offers our future work.

2 Formalization

In this section, we formalize our target transport problem. A service area of transportation is given by a graph which consists of nodes N and edges E in Equation (1). A node represents a traffic cross-point, and an edge represents a traffic road between two nodes. We denote a position of node n on x - y plane by $n(x)$ and $n(y)$.

$$N = \{n_1, n_2, \dots\}, E = \{e(n, n') | n, n' \in N\} \quad (1)$$

There are two kinds of transport requests: one-to-many and many-to-many. One-to-many represents collecting or delivering, i.e., vehicles visit pick-up nodes in turn. Many-to-many represents transporting, i.e., vehicles transport customers (or packs) from pick-up nodes to drop-off nodes. In this paper, we focus on the former transport request. Thus, a set of requests is given by R in Equation (2). Each request r is specified by three parameters: n is a pick-up node, t is an occurrence time, and q is a quantity of request (i.e., the number of customers or packs).

$$R = \{r_1, r_2, \dots, r_i, \dots\}, r_i = (n_i, t_i, q_i) \quad (2)$$

A fleet of K vehicles is given by V in Equation (3). Each transport vehicle v is specified by three parameters: n is a position (node), c is a capacity, and p is a scheduled path (route) which is given by a sequence of pick-up nodes. The sum of quantities ($\sum q_i$) assigned to one vehicle does not exceed a capacity c . As mentioned above, transport requests occur constantly while vehicles are moving. Thus, scheduled paths of vehicles also change constantly according to pick-up nodes assigned to vehicles.

$$V = \{v_1, v_2, \dots, v_j, \dots, v_K\}, v_j = (n_j, c_j, p_j) \quad (3)$$

3 R-Tree Based Grouping

In this section, we consider “how to divide K vehicles into some groups?”. A set of groups $G(t)$ at time t is given by a subset of vehicles in Equation (4). Each group $g(t)$ includes m vehicles (v_1, v_2, \dots, v_m) . The maximum group size (the number of vehicles in a group) is set to M .

$$G(t) = \{g_1(t), g_2(t), \dots\}, g(t) = \{v_1, v_2, \dots, v_j, \dots, v_m | m \leq M\} \quad (4)$$

A shape of a group is given by a minimum bounding rectangle (MBR) of vehicles which is contained by the group in Equation (5) where (x, y) is the position of upper left vertex, w is the width, and h is the height.

$$\begin{aligned} MBR(g(t)) &= (x, y, w, h) \\ &= (\min(n_j(x)), \min(n_j(y)), \max(n_j(x)) - x, \max(n_j(y)) - y) \end{aligned} \quad (5)$$

As mentioned above, the assignment of requests is optimized (i.e., request exchange) in units of groups. It is obvious that vehicles in each group should be located close each other. Hence, the combination of rectangles, which minimizes the sum of area of groups as in Equation (6), is selected as an ideal group of vehicles.

$$\min \left(\sum_{g(t) \in G(t)} Area(MBR(g(t))) \right) \quad (6)$$

Here, we introduce an indexing structure called R-Tree. R-tree is a height balanced tree to index spatial objects. Leaf nodes of the tree represent minimum bounding rectangles of spatial objects, and a parent node of the leaf nodes represents a minimum bounding rectangle of the leaf node. Such parent-child relations are fulfilled in the whole of the tree. Hence, we apply R-Tree to index the minimum bounding rectangles of vehicles. For example, as shown in Figure 1(a), there are 8 vehicles and 4 groups, and the groups are bounded by minimum bounding rectangles recursively. However, R-Tree is only adopted for static spatial objects, but not for moving spatial objects like vehicles in our target problem. Therefore, R-Tree is built repeatedly in the update interval I to keep groups of moving vehicles shown in Figure 1(b). We denote R-Tree which is built at time t as $tree(t)$. In particular, vehicles periodically send their positions to an index server (which maintains R-Tree), and the index server builds R-Tree (which represents groups of vehicles) in the update interval I by using collected positions of vehicles. If update interval I is short, the adequacy of groups (i.e., the accuracy of positions) is high, but the costs of sending positions and building tree is also high.

4 R-Tree Based Optimization

4.1 Request Assignment

If a request occurs at the present time t , the request is sent to the index server. The index server searches one vehicle for the request from $tree(t_u)$ (t_u is the

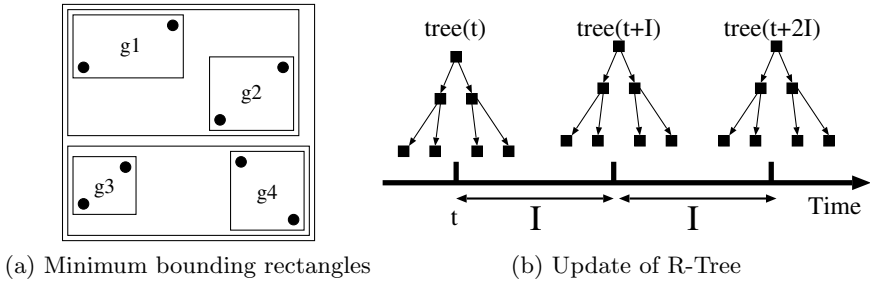


Fig. 1. Structure of R-Tree

latest updated time) by the rule in Equation (7). The rule represents that the tree node, whose distance between the center of tree node and the pick-up node of request is minimum, is adequate for the request. The search step of starts from the root node of $tree(t_u)$, and descends to a child node selected by the rule. At the end of the search step, the request is assigned to the vehicle which is contained in the reached leaf node. For example, Figure 2(a) illustrates an indexing tree ($K = 8, M = 2$), and vehicle v_6 is selected in the search step. In terms of the calculation cost, it is obvious that this search step is less than the linear search because the calculation cost of this search is $M \times \log_K M$.

$$\min \left(\sqrt{\left(\left(x + \frac{w}{2} \right) - n_i(x) \right)^2 + \left(\left(y + \frac{h}{2} \right) - n_i(y) \right)^2} \right) \quad (7)$$

4.2 Request Exchange

After the request assignment step, the assignment is optimized by the request exchange. First, we define an optimization level L which controls the balance between local and global optimizations. The balance depends on the group size of

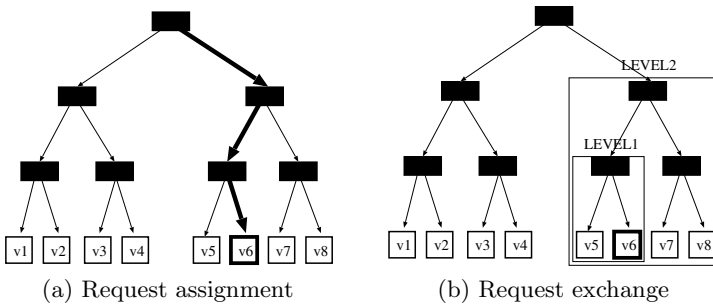


Fig. 2. R-Tree based optimization

vehicles (i.e., the problem space for request exchange). Hence, we regard the height of tree layer as the optimization level L , and the vehicles contained in the optimization level as a group for optimization (request exchange). For example, in Figure 2(b), the lowest layer of tree (v_5 and v_6) is the optimization level 1, the upper layer (v_5, v_6, v_7 , and v_8) is the optimization level 2.

Next, the request exchange is applied to all combinations of two vehicles (v and v') contained in the optimization level. In terms of the calculation cost, it is obvious that the number of exchange increases at a faster rate because the calculation cost of request exchange is ${}_2C_{ML}$. Here, we define two cost measures for the request exchange: time and distance. The distances of paths for v and v' are given by $|p|$ and $|p'|$. The time measure is longer distance from scheduled path of vehicles in Equation (8). Thus, the time measure tends to equalize the distribution of traveling load among vehicles. The distance measure is the sum of route distance of vehicles in Equation (9). Thus, the distance measure tends to centralize the traveling load to a few vehicles. The assignment of requests which minimizes the measure is selected from all combinations of two vehicles in the request exchange step.

$$time(v, v') = \max(|p|, |p'|) \tag{8}$$

$$distance(v, v') = |p| + |p'| \tag{9}$$

5 Experiment

5.1 Experimental Setting

We set three evaluation values of this problem: traveling distance, elapsed time, and optimization cost. The traveling distance is the total route distance of vehicles. The elapsed time is the interval time between occurrence and pick-up times of requests. The optimization cost is the calculation cost of request exchange step. Our experiment investigates the effects on the evaluation values by three parameters: group size M , update interval I , and optimization level L . The default parameter setting is described in Table 1. The service area is set to a grid network (40×40) which contains 1600 nodes. The number of vehicles is set to 30, and vehicles move between nodes in a unit time $1t$. The request is occurred at the rate of $1.0/t$ in the service area. One process is set to $1000t$, and the average of 10 processes is graphed as the result.

Table 1. Parameter setting

parameter	value	parameter	value
number of vehicles K	30	group size M	3
capacity of vehicles c	5	update interval I	1
quantity of requests q	1	optimization level L	2

5.2 Experimental Results

Figure 3 illustrates the results related to group size M . The group size M is set from 2 to 5. The size of problem space depends on the group size M . Hence, larger group size decreases both traveling distance and elapsed time, but increases optimization cost. The results also indicate that distance measure requires higher computational capacity than time measure. The reason is that the traveling load is centralized to a few vehicles, it brings about increase in the number of requests for exchange.

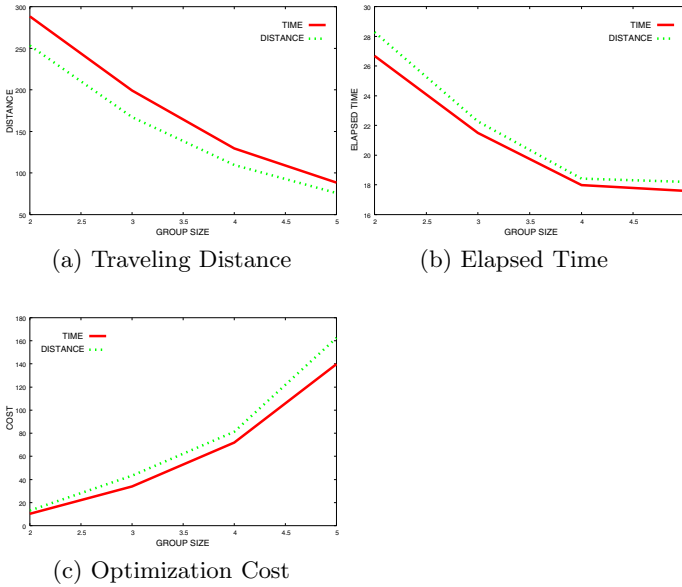
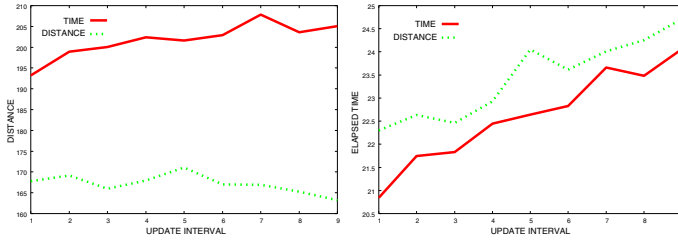


Fig. 3. Experimental Results on Group Size M

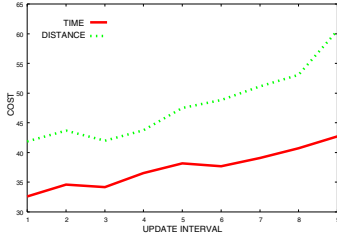
Figure 4 illustrates the results related to the update interval I . The update interval I is set from 1 to 9. The adequacy of groups (i.e., the accuracy of positions) depends on the update interval I . Hence, the update interval has no direct effect on traveling distance. However, longer update interval increases both elapsed time and optimization cost. The reason is that a request is not always assigned to the nearest vehicle.

Figure 5 illustrates the results related to optimization level L . The optimization level L is set from 0 to 3 (level 0 is no optimization). It is obvious that higher optimization level produces good effects on both traveling distance and elapsed time in return for the increase of optimization cost. Here, it should be noted that the difference of traveling distance between time and distance measures shrinks with increasing of optimization level, but the difference of elapsed time between time and distance measures spreads with increasing of optimization level. It means that time measure is more rational for both customers and vehicles at high optimization level.



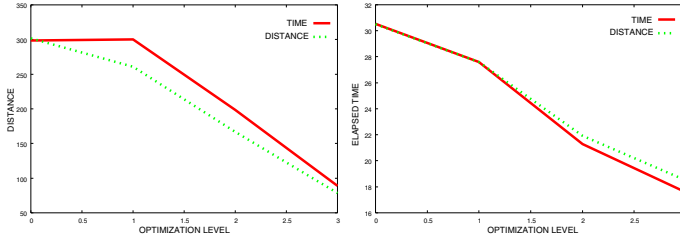
(a) Traveling Distance

(b) Elapsed Time



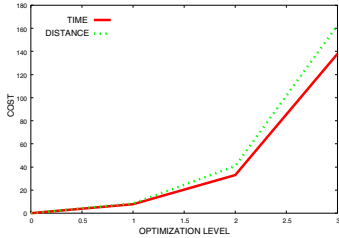
(c) Optimization Cost

Fig. 4. Experimental Results on Update Interval I



(a) Traveling Distance

(b) Elapsed Time



(c) Optimization Cost

Fig. 5. Experimental Results on Optimization Level L

6 Conclusion

In this paper, we focused on the scheduling problem of transport vehicles. In order to restrict problem space of combinational optimization for the scheduling problem, we introduced an indexing structure called R-Tree to form groups of vehicles, dynamically. Our optimization algorithm consisted of two steps: request assignment and request exchange. In the request assignment step, requests are assigned to a group of vehicles on the basis of the indexing tree. In the request exchange step, the assignment of requests is optimized by request exchange based on time or distance measures. Our results indicated that our algorithm leads to near-optimal solution for the transport scheduling problem effectively, but the performance of our algorithm depends on three capital parameters: group size M , update interval I , and optimization level L . In our future work, we would like to extend our algorithm to deal with more specific constraints such as the topology of road networks.

References

1. Noda, I., Ohta, M., Shinoda, K., Kumada, Y., Nakashima, H.: Is demand bus reasonable in learge scale towns? Technical Report 2003-ICS-131, IPSJ SIG Technical Report (2003) in Japanese.
2. Harano, T., Ishikawa, T.: On the validity of cooperated demand bus. Technical Report 2004-ITS-19, IPSJ SIG Technical Report (2004) in Japanese.
3. Qili, K., Ong, K.: A reactive method for real time dynamic vehicle routing problem. In: Proc. of ICTAI 2000. (2000) 176–181
4. Tian, Y., Song, J., Yao, D., Hu, J.: Dynamic vehicle routing problem using hybrid ant system. In: Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems. (2003) 970–974
5. Guttman, A.: R-trees: a dynamic index structure for spatial searching. In: Proc. of ACM SIGMOD 1984. (1984) 47–57

Knowledge-Based System for Die Configuration Design in Cold Forging

Osamu Takata¹, Tsubasa Mitani¹, Yuji Mure², Masanobu Umeda¹,
and Isao Nagasawa¹

¹ Kyusyu Institute of Technology, Japan

² Kagoshima Prefectural Institute of Industrial Technology, Japan

Abstract. We propose a knowledge-based system for process planning in cold forging. In our study, we had developed FORTEK-L (FORging Technology of process planning using Expert Knowledge for Layout), which is widely applicable to various shapes and types of equipment. In the next step, we have developed a knowledge-based die configuration design system, FORTEK-D (Die configuration design), which can generate suitable die configurations from layouts obtained from FORTEK-L. In the development of FORTEK-D, we have analyzed and reproduced the expert engineer's thinking process during die configuration design. Moreover, we have formulated the data representation, inference method, and knowledge-base for use with this system. In accordance with the formulation, we have developed a prototype system. We applied FORTEK-D to actual forged products and obtained appropriate results. The knowledge using the basic principle is widely applicable to various shapes and on a variety of equipment. Therefore, the abilities to maintain and operate the system have been improved.

1 Introduction

Forging is the process of forming steel or non-ferrous metal billet into the required product shapes through some intermediate products by hammering or pressing, as shown in **Fig. 1**. This process is applicable to mass production and is, therefore widely used in manufacturing automobile parts.

Forging process planning is very important because it has great influence on the cost and quality of final products. However, it is time-consuming and requires considerable experience and knowledge about forming and equipment. Because the number of expert engineers is decreasing, it is essential to create a way for companies to inherit this knowledge by developing systems using a knowledge-based approach that incorporates a thinking process of forging process planning and formulates utilization of that knowledge.

For this reason, many kinds of knowledge-based systems have been developed [1,2,3,4,5,6,7]. These can be classified into two types. Systems of the first type formulate a new process plan by retrieving previous cases and revising them; however, such a system cannot generate a plan if a similar case is not stored in its database. Systems of the second type generate new process plans by using

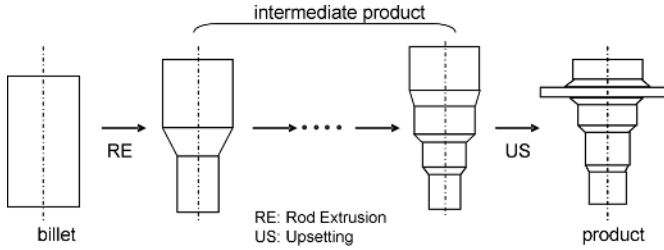


Fig. 1. An example of multi-stage cold forging

knowledge of the forming method of one process, its working limits, etc. For this type of system to be applicable to various kinds of products, it is necessary to analyze and formalize a large number of cases in order to build a cumulative knowledge-base. Therefore, neither of these two types of systems is applicable to a wide range of forged products and equipment.

The basic concept of the present study was to faithfully reproduce the thinking processes of expert engineers in forging process planning. In our approach, expert knowledge of process planning is decomposed into working limits, die configurations, and metal flow. As a result, we developed FORTEK-L (FORging Technology of process planning using Expert Knowledge for Layout), which is widely applicable to various shapes and types of equipment [8].

In the next step, we propose a knowledge-based die configuration design system, FORTEK-D (Die configuration design), which can generate suitable die configurations from layouts obtained from FORTEK-L.

This paper describes the outline of the prototype system and explains the formulation of the data representation, the inference method, and the knowledge-base. We show its usefulness by applying the system to various practical layouts obtained by FORTEK-L.

2 Outline of FORTEK-D

As shown in **Fig. 2**, for a one-stage layout composed of a pre-formed product and a forged product, and the given equipment conditions, FORTEK-D generates components of die configuration such as inners, rings, a knockout-pin, a back-up ring, and a punch. To generate die configuration, the process for die design knowledge which is elicited from experts' designers is used.

We analyzed the thinking processes of expert engineers [9,10] and formulated the die configuration design using the five-step operation shown in **Fig. 3**.

First, the die is decomposed into the upper and lower dies roughly. In this process, FORTEK-D determines the parting-line¹ by considering the pre-formed and forged products, and also the forming method.

Next, FORTEK-D determines the details of the upper and lower dies. In the process of determining the upper die, the structures and dimension for the

¹ the division line of the upper and lower dies.

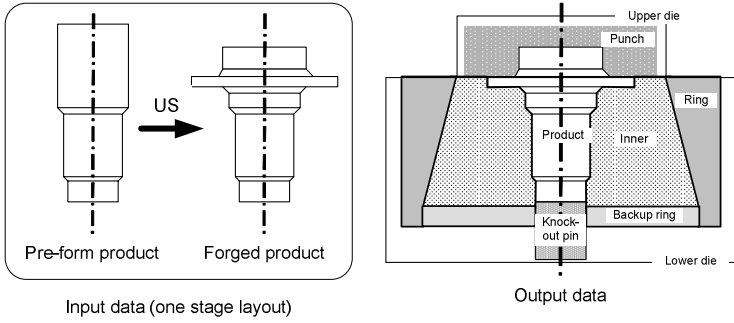


Fig. 2. Input and Output data

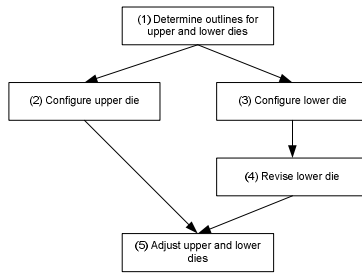


Fig. 3. Operation of die configuration design

punch are determined. Also, in the process of determining the lower die, the structures and dimension for inners, rings, the knockout-pin, and backup-rings are determined. In addition, it is necessary to revise (divide) inners and rings with respect to absorbing forming loads.

Finally, the upper and lower dies are adjusted with respect to the connectivity.

3 Implementation of FORTEK-D

3.1 Data Representation

The input, intermediate, and output data are uniformly represented by semantic networks [11], to maintain the knowledge-base easily. In the semantic networks, the node indicates objects such as the punch, inners, rings, the knockout-pin, the backup-ring, products, and basic elements. Each object has attribute values. On the other hand, the link indicates the relation with each object such as the **part-of** and the **neighboring**. The neighboring relation represents a pointer between objects.

In the connection between two objects for the neighboring relation, the shape of an object is same as the mirrored shape of another. For an example, a shape of a product is same as the mirrored shape of the die because the product is

formed by the die. In the neighboring relation, the mirrored shape of an object is not defined the shape explicitly, but defined as the reference for another. Using the neighboring relation, if the shape of an object changes, the shape of the neighboring object does not have to be changed.

Fig. 4 shows the semantic networks for the die configuration shown in Fig. 2.

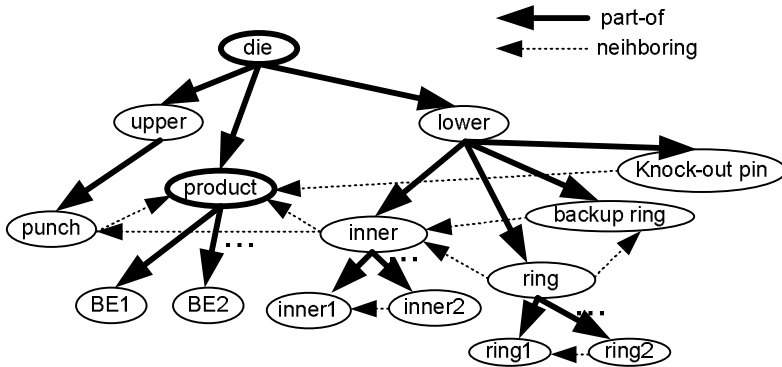


Fig. 4. Data representation (semantic networks)

3.2 Inference Method

We analyzed and classified knowledge for the five-step operation and found that the contents of knowledge for each operation are different, but the usage of it is similar irrespective of the operation. As a result, all of the knowledge-base can be described in the same manner.

In FORTEK-D, all data are managed uniformly as semantic networks with objects, as mentioned in Section 3.1.

To execute the five-step operation, we use the inference method to apply data in the knowledge-base to manipulate instance attributes using object manipulation functions, as shown in Table 1. Moreover, inferences can proceed simultaneously if the previous operation has been completed. For example, in Fig. 3, process (2) and process (3) could proceed simultaneously, if process (1) has been completed.

Table 1. Object manipulation functions

function	explanation	usage
get	refer to an attribute value of an instance	get(class, instance, attribute)
put	store an attribute value of an instance	put(class, instance, attribute, value)
generate	generate an instance object	generate(class)

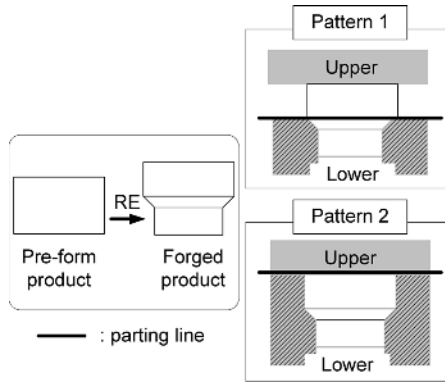


Fig. 5. Example of knowledge

requirements	get parameters	calculate	parameters	condition part
最大径力key				$\text{when}(\text{member}(\text{BE_keys}))$ $\text{when}(\text{member}(\text{BE_Targets}))$ $\text{when}(\text{member}('RE', \text{FM1}))$ $\text{when}(\text{member}('US', \text{FM2}))$
最大径力target				
				$\text{fortek_get}(\text{product}, \text{Product}, \text{beList})$ $\text{fortek_get}(\text{product}, \text{Product}, \text{keys})$ $\text{fortek_get}(\text{product}, \text{Product}, \text{targets})$ $\text{fortek_get}(\text{product}, \text{Product}, \text{formingRules})$ $\text{fortek_get}(\text{formingRule}, \text{FR}, \text{formingRule})$
				$\text{dcall}(\text{UTIL}, \text{getDataPart}, [\text{beList}, \text{be}, \text{du}], [\text{DUUS}])$ $\text{dcall}(\text{UTIL}, \text{splitList}, [\text{DuS}, [\text{MAX}], [], []], [\text{LEF}, \text{RIGHT}])$ $\text{max}(\text{DuS})$ $\text{length}(\text{LEF})$ $\text{nth}((\text{N}+1), \text{beList})$
				$\text{fortek_generate}(\text{upper})$ $\text{fortek_generate}(\text{lower})$
				$\text{fortek_putA}(\text{upper}, \text{Upper}, \text{upperBe}, \text{LEFT})$ $\text{fortek_putA}(\text{upper}, \text{Upper}, \text{lowerBe}, \text{RIGHT})$

Fig. 6. Example of knowledge representation

3.3 Knowledge-Base

Knowledge is composed of the **if-then** rule, which has the condition and conclusion parts. If the condition part is satisfied, then the conclusion part is executed.

Fig. 5 shows an example for the knowledge of decomposition of upper and lower dies. In the example, there are two decomposition patterns. Fig. 6 shows the knowledge representation for pattern 1 in Fig. 5. First, the system checks

the knowledge-base for data about the condition of the forming-method and the connectivity of the maximum diameters. Next, it calculates values of instances attributes and then generates new instances such as upper and lower dies. Finally it applies the calculated values to the instances.

Using this knowledge representation, users can maintain independent knowledge-bases that do not have to be consistent with the content of other knowledge-bases, because knowledge-bases are classified under procedural flows.

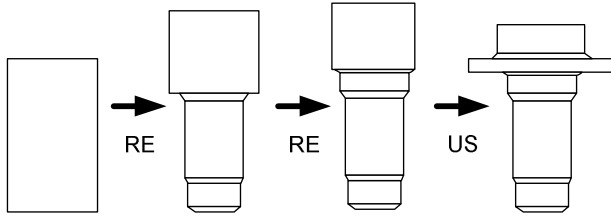


Fig. 7. An example of input data

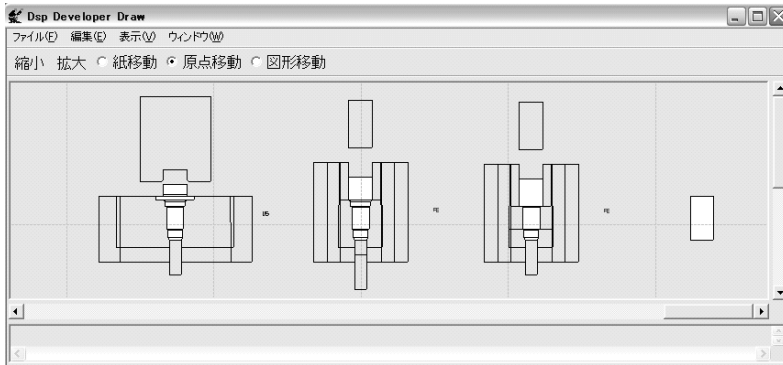


Fig. 8. An example of series of die configurations

4 Execution Results

We built the proposed implementation mechanism and constructed knowledge-bases. Moreover, we applied FORTEK-D to actual forged products² to evaluate the usefulness and effectiveness of the proposed system. The prototype was implemented in the knowledge-based tool DSP [12], which is developed over Inside Prolog [13].

For Fig. 7, process planning for the layout generated by FORTEK-L, FORTEK-D generates the series of multi-stage process of die configurations, as shown in Fig. 8. As shown in these figures, a desirable die configuration was successfully obtained.

² The number of applied products is five.

The total execution time was less than 1.0 seconds. In this case, 57 kinds of knowledge were stored.

5 Conclusions

We analyzed the thinking processes of expert engineers and formulated die configuration design as a five-step operation: determination of the outline for the upper and lower dies, configuration of the upper die, configuration of the lower die, revision of the lower die, and adjustment of the upper and lower dies.

Moreover, we implemented the data representation, inference method and, knowledge-base. In data representation, all data are managed uniformly as semantic networks with objects. Knowledge-bases are defined as if-then rules that are managed with respect to the above five-step operation. In order to execute the five-step operation, the inference method applied to the knowledge-base manipulates instance attributes using object manipulation functions.

Finally, we applied FORTEK-D to actual forged products to evaluate the usefulness and effectiveness of the proposed system. A desirable die configuration design was successfully obtained.

The knowledge using basic principles is widely applicable to various shapes and types of equipment. Therefore, the abilities to maintain and operate the system have been improved.

Our next goal will be formulate and incorporate experts' knowledge of metal flow in conjunction with the physical/numerical simulation in designing a detailed die configuration.

References

1. Bariani, P., Benuzzi, E., Knight, W.A.: Computer aided design of multi-stage cold forging process: Load peaks and strain distribution evaluation. *Annals of the CIRP* **36**(1) (1987) 145–148
2. Sevenler, K., Raghupathi, P.S., Altan, T.: Forming-sequence design for multistage cold forging. *J. of Mechanical Working Technology* **14** (1987) 121–135
3. Lange, K., Guohui, D.: A formal approach to designing forming sequences for cold forging. *Trans. of the NAMRI/SME* (1989) 17–22
4. Mahmood, T., Lengyel, B., Husband, T.M.: Expert system for process planning in the cold forging of steel. *Expert Planning Systems* **322** (1990) 141–146
5. Takata, O., Nakanishi, K., Yamazaki, T.: Forming-sequence design expert system for multistage cold forging: Forest-d. In: *Proc. of Pacific Rim International Conference on Artificial Intelligence '90*. (1990) 101–113
6. Yang, G., Osakada, K.: A review of expert system for process planning of cold forging. *Manufacturing Review* **6**(2) (1993) 101–113
7. Kim, H.S., Im, Y.T.: An expert system for cold forging process design based on a depth-first search. *J. of Materials Processing Technology* **95** (1999) 262–274
8. Takata, O., Mure, Y., Nakashima, Y., Ogawa, M., Umeda, M., Nagasawa: Knowledge-based system for process planning in cold forging using adjustment of stepped cylinder method. In: *INAP-2005*. (2005) 117–126

9. Verson, M.D.: Impact Machining. Verson Allsteel Press Company (1969)
10. JSTP: Forging Technology (in Japanese). Corona Publishing Co. (1995)
11. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach (2nd Edition). Prentice Hall (2002)
12. Umeda, M., Nagasawa, I., Higuchi, T.: The elements of programming style in design calculations. In: Proceedings of the Ninth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. (1996) 77–86
13. Katamine, K., Umeda, M., Nagasawa, I., Hashimoto, M.: Integrated development environment for knowledge-based systems and its practical application. IEICE Transactions on Information and Systems **E87-D**(4) (2004) 877–885

An Automatic Indexing Approach for Private Photo Searching Based on E-mail Archive

Taketoshi Ushiana¹ and Toyohide Watanabe²

¹ Faculty of Design, Kyushu University,
4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
ushiana@design.kyushu-u.ac.jp

² Graduate School of Information Science, Nagoya University,
Furo-cho, Chikuka-ku, Nagoya 464-8603, Japan
watanabe@is.nagoya-u.ac.jp

Abstract. Recently, much attention has been paid to manage amount of personal contents effectively. Private photo is one of the most important types of personal contents. In many of conventional approaches for private photo searching, a user is required to assign annotations manually to the private photos of the user by oneself. However, assigning annotations is tedious for a user. This paper introduces an automatic indexing approach for private photo searching based on e-mail message archive. Many of e-mail messages of a user may describe some experiences of the user. In this approach, the temporal expressions are extracted from an e-mail message, and they are used for estimate the relativity between a term in the e-mail message and a temporal interval in the personal time space of a user. The weight of a term to a private photo is calculated on the basis of the relativity and the granularity of the temporal interval.

1 Introduction

Today, there are a lot of digitalized personal contents such as digital photograph and digital video around us. The recent remarkable progress of HDD technology allows a user to buy a large capability HDD in low price. A large capability HDD enables a user to store whole of personal contents in a PC. Traditionally, folder hierarchy is the most popular approach for managing their contents. However, it is hard for a user to recognize a large scale of hierarchy. Recently much attention is paid to desktop search to solve this problem. Some web search engine providers such as Google, Yahoo, and MSN have offered tools for desktop search. Such desktop search tools are designed for document retrievals, and they have the feature in which an individual file and web page retrieval are integrated seamlessly. One of the defects of the conventional desktop search tools is that they do not provide enough functions to search multimedia files such as music files and video files. These tools uses file names and metadata to search multimedia files. In order to obtain the search results that satisfy user's requirement, it is necessary to assign suitable metadata to multimedia files.

Files that are stored in a personal computer can be classified into two types: public content and private content. Public contents are assumed that two or more people use it through the Internet. Music contents and movies are typical examples of public

contents. When a user searches multimedia files, metadata plays an important role. It is relatively easy to acquire the metadata of public contents. For instance, some metadata of a CD track can be automatically obtained from the Internet by a player software on a personal computer. On the other hand, it is difficult to obtain metadata of private contents. It is necessary to assign metadata for private contents manually. However, assigning metadata for all private contents is very tedious for a user. It is expected to develop a technique for search private contents without manual annotation.

Photo is one of the most comprehensive type in private contents. Today, many persons store a lot of their private photos in their own personal computers. We focus on private photos as the target objects of search of this paper. This paper introduces a new technique to search private photos without any manual annotation using keyword query.

A private photo of a user is representation of personal experience of the user in the real world. Our technique uses estimated experiences of a user for searching private photos of the user. To assume personal experiences of a user, this technique uses sent and received e-mail messages of the user. Estimated experiences are able to be searched by a keyword, so the system enables the user to search a private photo by keyword by assigning a private photo to an experience.

2 Approach

The searching target of this paper is a private photo archive that are taken by a digital camera. A user who wants to search in this photo search system gives one or more keywords to the system as a query. Search results are scored and ranked to the query. The main theme of this paper is how to assign weight to each relation between keyword and photo.

In this research, we focused on the photos that are taken with a digital camera. Metadata are embedded in a photo taken by a digital camera using the EXIF format. The date and time when a photo was taken is one of the metadata of EXIF. If the snapping date and time is recorded in a photo, the photo can be related to experiences of a user. Many of e-mail messages of a user may describe some experiences of the user. In our approach, the temporal expressions are extracted from an e-mail message, and they are used for estimate the relativity between a term in the e-mail message and a temporal interval in the personal time space of a user. The weight of a term to a private photo is calculated on the basis of the relativity and the granularity of the temporal interval. Fig. 1 shows an overview of this indexing approach.

3 Weight Assignment

This section describes how to decide the weight of an index term to a private photo.

This paper uses following notations: t_i denotes a term, m_i denotes a e-mail message, e_i denotes a temporal expression, I_i denotes a time interval, and p_i denotes a private photo.

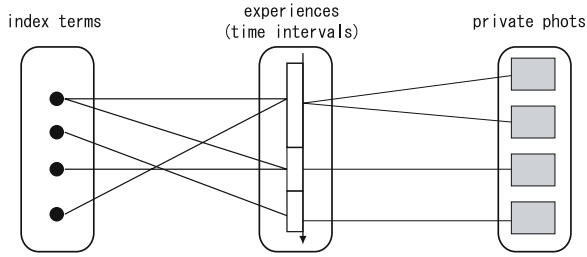


Fig. 1. Indexing based on time intervals

3.1 Three Types of Weighting

Document search is well known as one of the most important fields in information retrieval, and many techniques about term weighting in a document have been proposed. There are three types of term weighting are used in many of document search techniques.

1. Local weight l_{ij} : Local weights are used for taking recall rate higher. l_{ij} represents the local weight for term i in document j . Many document search techniques apply TF values for local weights. A TF value represents how many times a term occurs in a document.
2. Global weight g_i : Global weights are used for taking precision rate higher. g_i represents the global weight for term i in the entire document collection. Many document search techniques apply IDF values for global weights. An IDF value represents how many times documents containing a term appears in the document collection.
3. Normalization factor n_j : In document search techniques, normalization factors are used for cancelling the effects of document length. Many document search techniques apply values based on the number of terms in a document.

The weight of a term i to a document j can be described by the following form.

$$a_{ij} = \frac{l_{ij}g_i}{n_j} \tag{1}$$

This weighting schema is adaptable besides document search techniques. We apply this schema for term weighting for a private photo.

3.2 Indirectly Term Weighting to a Private Photo

E-mail messages have following features about the index terms and temporal expressions that they contain.

- An e-mail message contains one or more index terms.
- An e-mail message contains one or more temporal expressions.
- A temporal expression appeared in different e-mail messages may has semantic relation about the same experience of a user.

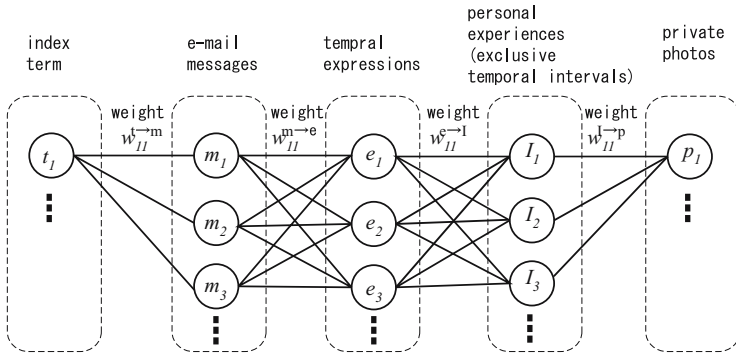


Fig. 2. An term weighting structure of private photos using e-mail messages and temporal expressions, and personal experiences

We treat an e-mail message is a meaningful unit that represents an experience of a user. Our technique weights a term to a private photo based on co-occurrence of a term and a temporal expression in an e-mail message. We assume that an index term t_i represents a user experience that was performed in a temporal expression e_j if t_i and e_j are contained by the same e-mail message.

A term is contained by multiple e-mail messages. An e-mail message contains multiple temporal expressions. Some temporal expressions concern the same user experience. A user experience may be represented by a private photo. Using our interconnected relationships our technique calculates a weight of term to a private photo.

Fig. 2 shows the weighting schema of our technique. In the figure, $w^{t \to m}$ represents a weight of a term to an e-mail message, $w^{m \to e}$ represents a weight of an e-mail message to a temporal expression, $w^{e \to I}$ represents a weight of an temporal expression to a user experience and $w^{I \to p}$ represents a weight of a user experience to a user’s private photo. Multiple paths exist between from the index term t_1 to the private photo p_1 . The weight of t_1 to p_1 is calculated as the sum of weights of all the paths from t_1 to p_1 . In general, a weight of a term i to a private photo m can be represented as the following formula.

$$w_{in} = \sum_j \sum_k \sum_l (w_{ij}^{t \to m} \times w_{jk}^{m \to e} \times w_{kl}^{e \to I} \times w_{ln}^{I \to p}) \tag{2}$$

The rest of this section describes how to calculate the above four types of weights. Each type of the weights consists of global weight, local weight and normalize factor.

Weighting a Term to an E-mail Message. A term weight to an e-mail message is estimated based on the same approach of term weighting of document search.

An index term may appear multiple times within the same e-mail message. The larger the number of appearance of a term in an e-mail message, the term is more important for representing content of the e-mail message. The local weight of a term t_i to an e-mail message m_j is represented as $l_{ij}^{t \to m}$. $l_{ij}^{t \to m}$ is defined as the following formula where TF_{ij} is the number of occurrence of a term t_i within an e-mail message m_j .

$$l_{ij}^{t \rightarrow m} = \text{TF}_{ij} \tag{3}$$

The smaller the number of documents that contain a term, the term is more important. The global weight of a term t_i is represented as $g_i^{t \rightarrow m}$. $g_i^{t \rightarrow m}$ is defined as the following formula where MF_i is the number of e-mail messages that contain a term t_i .

$$g_i^{t \rightarrow m} = \frac{1}{\text{MF}_i} \tag{4}$$

The larger the number of terms in an e-mail message, the average number of term occurrences in an e-mail message is higher. The normalize factor of a message m_j is represented as $n_j^{t \rightarrow m}$. $n_j^{t \rightarrow m}$ is defined as the following formula where LEN_j is the number of terms within a message m_j .

$$n_j^{w \rightarrow m} = \text{LEN}_j \tag{5}$$

The weight of a term t_i to a message m_j is defined as following based on the formula 2 and the above weights and factor.

$$\begin{aligned} w_{ij}^{t \rightarrow m} &= \frac{l_{ij}^{t \rightarrow m} g_i^{t \rightarrow m}}{n_j^{t \rightarrow m}} \\ &= \frac{\text{WF}_{ij}}{\text{MF}_i \times \text{LEN}_j} \end{aligned} \tag{6}$$

Weighting an E-mail Message to a Temporal Expression. Multiple temporal expressions may appear in one e-mail message. An e-mail message is weighted to a temporal expression according to its occurrence. The local weight of an e-mail message m_i to a temporal expression e_j is represented as $l_{jk}^{m \rightarrow e}$. $l_{jk}^{m \rightarrow e}$ is defined as the following formula where EF_{jk} is the number of occurrence of e_k in m_j .

$$l_{jk}^{m \rightarrow e} = \text{EF}_{jk} \tag{7}$$

The same temporal expression hardly appears multiple times in the same e-mail message, So in many cases $l_{jk}^{m \rightarrow e}$ may take either 1 or 0. We don't consider temporal relationship between temporal expressions. Such relationships can be considered when estimating the weight of a temporal expression to a personal experience.

The smaller the number of temporal expressions in an e-mail message, a temporal expression is higher related to the contents of the message. The global weight of an e-mail message m_j is represented as $g_j^{m \rightarrow e}$. $g_j^{m \rightarrow e}$ is defined as the following formula where NE_j is the number of occurrence of temporal expressions in m_j .

$$g_j^{m \rightarrow e} = \frac{1}{\text{NE}_j} \tag{8}$$

The normalize factor of m_j to e_k is represented as $n_k^{m \rightarrow e}$. In this paper, we set the factor 1 as follows.

$$n_k^{m \rightarrow e} = 1 \tag{9}$$

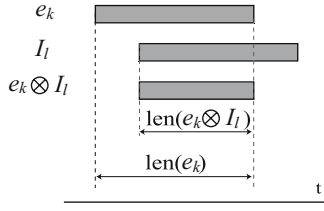


Fig. 3. Relation between time intervals

The weight of an e-mail message m_j to a temporal expression e_k is defined as following based on the formula (2) and the above weights and factor.

$$\begin{aligned}
 w_{jk}^{m \rightarrow e} &= \frac{l_{jk}^{m \rightarrow e} g_j^{m \rightarrow e}}{n_k^{m \rightarrow e}} \\
 &= \frac{EF_{jk}}{NF_k}
 \end{aligned}
 \tag{10}$$

Weighting a Temporal Expression to a Personal Experience. In our technique, a personal experience is represented as one or more exclusive time intervals. We suppose that the granularity of an exclusive time interval is specified by a user when the user executes a query.

In order to define the weight of temporal expression to a personal experience (an exclusive time interval) we introduce an operation on a temporal expression and exclusive time interval. The operation to calculate the overlapping part of a time expression e_k and an exclusive time interval I_l is represented as $e_k \otimes I_l$. The function that give the length of a time interval I_l is represented as $len(I_l)$. Fig.3 shows an example of the operation.

The larger the common part of a time expression and an exclusive time interval is, the higher they may relate each other. The local weight $l_{kl}^{e \rightarrow I}$ of a temporal expression e_k to an exclusive time interval I_l is estimated based on the size of common part between them. $l_{kl}^{e \rightarrow I}$ is defined as following fomula.

$$l_{kl}^{e \rightarrow I} = len(e_k \otimes I_l)
 \tag{11}$$

If the granularity of a temporal expression is large, it has highly possibility to overlap many exclusive temporal intervals. Otherwise, if the granularity of a temporal expression is small, it may be important for a specific experience. The global weight of a time expression e_k is represented as $g_k^{e \rightarrow I}$, and estimated based on inverse of the length of e_k . The global weight $g_k^{e \rightarrow I}$ is defined as following formula.

$$g_k^{e \rightarrow I} = \frac{1}{len(e_k)}
 \tag{12}$$

If the granularity of an exclusive time interval is large, it may overlap many temporal expressions. The normalize factor to an exclusive temporal interval I_l is represented as $n_l^{e \rightarrow I}$ and defined as the following formula.

$$n_l^{e \rightarrow I} = \text{len}(I_l) \quad (13)$$

The weight of a temporal expression e_k to an exclusive temporal interval I_l is defined as the following formula.

$$\begin{aligned} w_{kl}^{e \rightarrow I} &= \frac{l_{kl}^{e \rightarrow I} g_k^{e \rightarrow I}}{n_l^{e \rightarrow I}} \\ &= \frac{\text{len}(e_k \otimes I_l)}{\text{len}(e_k) \times \text{len}(I_l)} \end{aligned} \quad (14)$$

Weighting a Personal Experience to a Private Photo. In our approach, a personal experience is represented as one or more exclusive temporal intervals. If the creation time of a private photo can be taken, the photo can be related to one exclusive time interval. When a digital camera is used, the creation time information is embedded to the photo image in the EXIF format. The weight of an exclusive temporal interval I_l to a private photo p_n is represented as $a_{ln}^{I \rightarrow p}$, and is defined as the following formula.

$$a_{ln}^{I \rightarrow p} = \begin{cases} 1 & \text{the photo } p_n \text{ was taken in the interval } I_l \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

4 Experimental Results

To evaluate the introduced approach we performed an experiment. In this experiment, we ask a subject to execute some search queries for private photos, and evaluate the relevance of the results. The e-mail messages that were stored in the mail folder for memorial were used for the experiment. The number of the e-mail messages was 50, and they were received by the subject from August 11, 2004 to May 5, 2005. The extraction of the temporal expressions in the e-mail messages were performed manually. The experiment consists of 5 queries.

The experimental results are as following: the average of precision values is 0.9, the average of recall values is 0.4 and the average of F-measure values is 0.6. The results show a high expectation for the adaptability of proposed approach. However, this experiment was performed with small number of e-mail messages and queries, so we cannot examine precise evaluation. We plan to perform further experiments with more large volume of e-mail messages.

5 Related Works

Lifestreams[1] and Time-Machine Computing[2] provide a user to manage personal files with time. These methods organize personal files based on the creation time and access times of each file, but they do not provide keyword-based access to the personal files. On the other hand, our technique provides a user with a kind of keyword-based access to a private photo based on temporal organization.

Ohmoto et al. had developed a method for automatically deciding attribute values based on an IS-A hierarchy of keywords and temporal inclusion relationship between video objects[3]. This method does not consider weights of a keyword. Moreover, this method assumes that one or more keywords are assigned to some video objects manually. On the other hand, our technique does not require manual keyword assignment of a user.

Hori et al. introduced a method for searching video scenes that are captured with a wearable camera[4]. In this method, the e-mail messages and web pages that are browsed with a wearable computer. A term that is appeared in the e-mail messages or web pages is assigned video scenes based on the access time of the e-mail messages and web pages. This method does not consider temporal expressions in an e-mail message or web page. And this method does not provide ranking score for results.

6 Conclusion

This paper introduces a technique for searching private photos of a user without any manually added annotation. In this technique, a term in an e-mail message of a user is assigned to the personal time space of the user on the basis of the temporal expressions appeared in the same e-mail message. The granularity of a temporal expression is considered for weighting a term to a user's private photo.

Our technique treats an e-mail message as one relevant unit and uses the frequency information for term weighting in an e-mail. The validity of weighting in this technique is not high, because many e-mail messages contain less number of terms than a scientific paper or article in a newspaper. In order to defeat this difficulty, we plan to unify relate e-mail messages on the basis of reference association and content similarity. Such message groups are expected to enlarge the validity of statistical information such as term frequency.

The introduced method uses e-mail messages for indexing private photos of a user. The e-mail messages that a user sent or received usually contain some junk mails. Moreover, some e-mail messages can contain the temporal expressions that are not concerned with activities of the user. In order to defeat this difficulty, some filtering techniques to increase the validity of temporal expression are expected to be developed.

This approach is adaptable to other text-based media such as web page and office documents. The performance of the search is expected to be increased by considering such text-based media.

References

1. Freeman, E., Gelernter, D.: Lifestreams: A Storage Model for Personal Data. *SIGMOD Record* **25**(1) (1996) 80–86
2. Rekimoto, J.: Time-Machine Computing: A Time-centric Approach for the Information Environment. In: *Proc. of UIST'99*. (1999) 45–54
3. Oomoto, E., Tanaka, K.: OVID: Design and Implementation of a Video-Object Database System. *IEEE Trans. on Knowledge and Data Engineering* **5**(4) (1993) 626–643
4. Hori, T., Aizawa, K.: Context-based Video Retrieval System for the Life-Log Applications. In: *Proc. of MIR'03*. (2003) 31–38

Analysis for Usage of Routing Panels in Evacuation

Toyohide Watanabe and Naoki Harashina

Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
watanabe@is.nagoya-u.ac.jp

Abstract. This paper proposes an evacuation support method, based on routing panels, and evaluates analytically this method by means of computer simulation. For this simulation, the multi-agent paradigm is adopted as our basic modeling view. On this modeling, we focus on the behaviors of agents promoted in-structively by routing panels. As a result, two characteristic suggestions are derived: one is that routing panels set on critical sections, in which were estimated by Monte Carlo method, can support the rapid and safe evacuation; and another is that the route finding way to be applicable by the global disaster information is better than that of the local disaster information.

1 Introduction

The counter measure for natural disasters is one of most important social problems, and recently various types or large scales of disasters have been often reported from every place in the world. As a counter measure for disasters, it is important to lead sufferers instantly into their safety situation from a viewpoint of evacuation. Sugiman et al. experimented the evacuation phenomena analytically with respect to a scenario that indicators enforce sufferers to escape from disasters occurred in underground towns. In this experiment, they proposed two different evacuation support methods: a macro-based method which informs many sufferers of their exit and direction; and a micro-based method which provides only a few sufferers with the routing information about this exit and direction. The micro-based method was superior to the macro-based method. Additionally, Sugiman et al. made it clear that the ratio of indicators to sufferers may change drastically the result between two methods.

In general, it is not easy to do experiments practically because disasters are irregularly and suddenly happened. From this viewpoint, the computer simulation is very effective to investigate the counter measure. Ishida et al. validated the correctness of Sugiman's method by means of computer simulation [1]. Their experimental environment is applied to the station area in subway. Of course, we cannot absolutely apply the evacuation plan in the subway station to human actions as the practical tasks. Kagami et al. analyzed first the characteristics of human behaviors by requesting questionnaires to various generations, and modeled individual sufferers as agents with respect to the questionnaire results [2].

It is successful to characterize individual agents as simulation targets when we look upon computer simulation as a powerful means. In this paper, we address an evacuation support method, based on the routing panel, and evaluate the effectiveness and safety from viewpoints of the allocation of routing panels, routing procedures, and so on. We typically look upon an earthquake as one of disasters to be discussed in this paper, and look for emergence procedures to be performed soon after the occurrence. In our computer simulation, we assume that the routing panels are allocated into intersections and also the routing information about confusion states, traffic situation and so on is at all collected from the pre-assigned base server.

2 Framework

We address an evacuation support method based on the routing panels with a view to solving complex task of route selection and avoiding confusion status on the earthquake occurrence. Our approach adopts the strategy which determines the locations of routing panels by using Monte Carlo method, and which changes the display contents of routing panels by sensing disaster situations.

2.1 Allocation of Routing Panels

The feature which determines the locations of routing panels depends on how many persons can pass by individual panels or routes. In the practical experiment, the estimation is very difficult because the problems such as where the disasters are occurred, which directions and routes should be selected, which routes are prohibited to pass through, and so on could not be always solved successfully. Thus, we construct virtually an urban model and analyze the panel locations appropriately by the computer simulation. In this case, we compute the approximate solution by making use of Monte Carlo method. As Monte Carlo method is a kind of simulation means based on the random number, the more the experimental trials is the higher the accuracy is. As well as the positioning problem of road indicators, the intersections, in which the number of agents who passed through is too many, are regarded as panel places.

2.2 Confusion Information

The safety and effectiveness of evacuation are not always kept at better level, because the moving velocities of crowds are decreased if many people gather at once in the same places. When the evacuator select only by themselves their own roads, the confusion state becomes worse. In order to avoid the undesirable phenomena, we change dynamically the routing information on the panel places by monitoring and analyzing the correct traffic data globally. The confusion information is derived basically from the number of agents who pass through the preset points. Our confusion information classifies into two types: local information and global information.

Local Information:

This information depends on the nodes which the routing panels are set and the links among their candidate nodes, which indicate evacuating agents according to routing panels. We assume that two different candidate nodes are $c1$ and $c2$, and also the numbers of corresponding agents on these candidate links are $n1$ and $n2$, respectively. Under this assumption, the local information which the panels provide for agents is:

- $c1$ when $n1 < n2$,
- $c2$ when $n1 > n2$,
- $c1$ or $c2$ when $n1 = n2$.

Figure 1 shows this situation: the confusion situation for links may be avoided if the traffic data about links could be changed appropriately.

Global Information

This information is related to the traffic data on the panel locations with passable possibilities, concerned to two preset candidate nodes. The locations are computed from candidate nodes and neighboring nodes of preset target nodes.

We assume that the coordinate values of two candidate nodes $c1$ and $c2$ are $(x1, y1)$ and $(x2, y2)$, and the coordinate values of target nodes are $(xg1, yg1)$ and $(xg2, yg2)$, respectively. Under this assumption, we define two rectangles of width $|x1 - xg1|$ and height $|y1 - yg1|$, and of width $|x2 - xg2|$ and height $|y2 - yg2|$. They are sampling locations, and are used to count up the number of agents who exist in the sampling locations for some period. When the numbers of agents who are included in individual locations are $n1'$ and $n2'$, and the sizes of two rectangles are $A1$ and $A2$, the information with which our panels provide agents is:

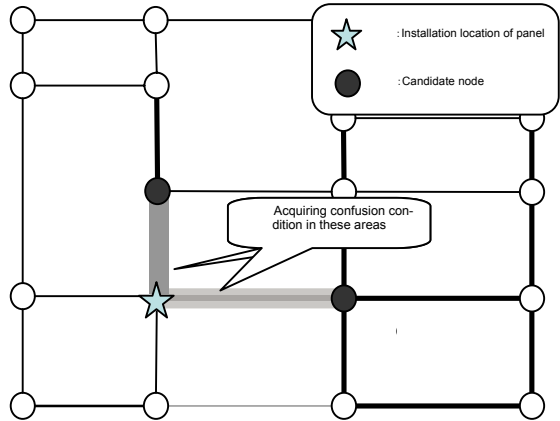


Fig. 1. Evaluation based on local information

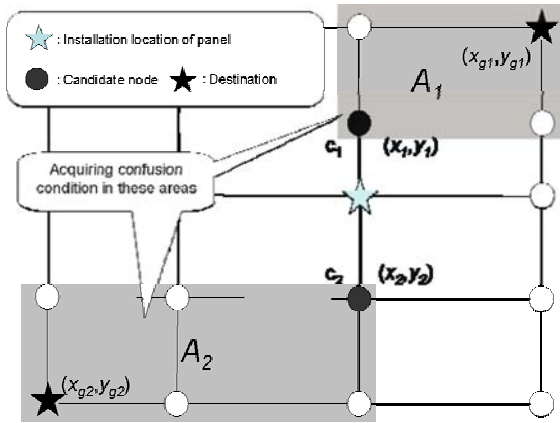


Fig. 2. Evaluation based on global information

- $c1$ when $n1/A1 < n2/A2$,
- $c2$ when $n1/A1 > n2/A2$,
- $c1$ or $c2$ when $n1/A1 = n2/A2$.

Figure 2 shows this situation. We can avoid the confusion state when the number of agents who reach to destination places are estimated according to the density of agents, and the number of agents to be navigated to the destination places is controlled.

3 Simulation Environment

Road Network

We make use of two simulation data as road networks: artificially composed grid-mesh data and map data extracted from the central area in Nagoya city, Japan. Figure 3 is an example of grid-mesh road network. Figure 4 is an example of map-based road network, whose pixel corresponds to 2 m and which is 800 *600 pixels. Each node is attended with the following attributes:

(Node attribute)

- x and y coordinate values,
- number of connective nodes,
- connective node number.
- Also, the links have the following attributes:

(Link attribute)

- x and y coordinate values of two terminal ends,
- link length,
- link width,
- information about whether road is passable.

In our computer simulation, we generate regularly events which changes the routing states whether some roads are passable or not on the basis of the un-passable probability of roads. Moreover, we assume that the road which became un-passable once cannot be reset to be passable again. Table 1 arranges link features in our simulation. The link width is classified into 3 types. The number of passable agents corresponds to the maximum limitation of agents who can pass through particular roads at once, and is propositional to the link width. It is assumed that the un-passable probability is reverse-propositional to the link width on the basis of the manual of closed roads, established as the national evacuation standard.

Table 1. Link features in simulation

Link width	Number of passable agents	Un-passable probability of road(%)
10	2	2
20	4	1
30	6	0

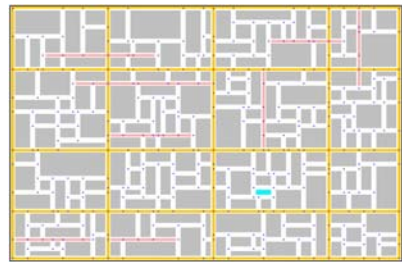


Fig. 3. Grid-mesh road network

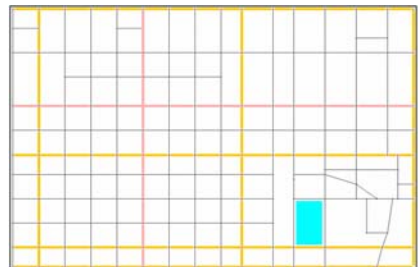


Fig. 4. Map-based road network

Agent

Agents are allocated at random into roads at first, and controlled repeatedly in accordance with our evacuation-specific algorithm on the node conflict. Of course, since agents do not know the routing nodes to the destination but knows only the location of destination, all agents are fully controlled by the algorithm. Also, the moving speed takes

two pixels for one unit time (i.e., four seconds) because in our practical observation it is said that the evacuation speed of a whole group in the disaster is about 1 m/second. In case that there are many destinations on one road network, individual agents select their nearest destination from currently displayed panels. Our evacuation-specific algorithm is specified as the processing flow in Figure 5.

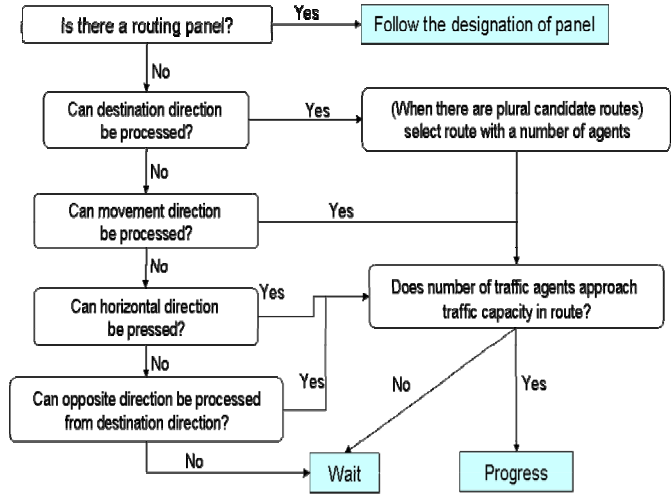


Fig. 5. Evacuation-specific algorithm

Routing Panel

Individual routing panels attend the following attributes:

(Panel attribute)

- located node number,
- candidate node 1,
- candidate node 2,
- information acquisition method.

The candidate node indicates whether agents can go ahead along appropriate roads. These candidate nodes are set in advance before our simulation, using the following conditions from the link between located node and candidate nodes.

(Condition)

- 1) link whose road width is large and which is close to destination,
- 2) link whose road width is large,
- 3) link which is close to destination.

This criteria are in the order from the upper (1) to the lower (3): 1 is attached with high priority; and 3 is done with low priority.

4 Experiment

4.1 Evaluation Criteria

Our simulation criteria are the effectiveness and safety. They are defined as follows:

(Definition)

- Effectiveness: This criterion is evaluated by the number of evacuated persons.
- Safety: This is attained by evacuation actions which make it possible to avoid confusion routing or un-passable road selection. This typical action may be to select roads whose width is large as possible.

The estimation expression for safety is defined as:

$$S = \sum_{t=ts}^{tf} r_w(t) / r_a(t)$$

Where,

- ts*: start time of simulation,
- tf*: end time of simulation,
- ra(t)*: number of agents who cannot evacuate at time *t*,
- rw(t)*: number of agents who pass through road of width 30.

Agents can select a road whose width is proportional to *S*: it is possible to attain the evacuation activity with the concept of safety.

4.2 Experimental Environment

Our experimental environment is shown in Table 2, common to all experiments.

Experimental-1

Figure 6 shows the panel locations, applied our method to Figure 3. The symbols “★” indicate desirable panel places. On the basis of Monte Carlo method 100 simulation experiments are repeated.

Experiment-2

Figure 7 is an experimental result: a vertical axis represents the safety and a horizontal axis is the time. Table 3 summaries the safety evaluation by the simulation .

Experiment-3

Figure 8 is another result, applied to Figure 3. The safety is evaluated and the result is arranged in Figure 9. Our experimental result in Figure 7 is superior to human estimated allocation in Figure 9. Table 4 summarizes the result.

Table 2. Parameter setting

Number of agents	100, 200
Maximum limitation time	300 (UT)
Event occurrence time	40 (UT)
Road width	30

Table 3. Safety evaluation by simulation

Number of agents	S (without panels)	S (with panels)
100	76.2	111.2
200	57.4	122.2

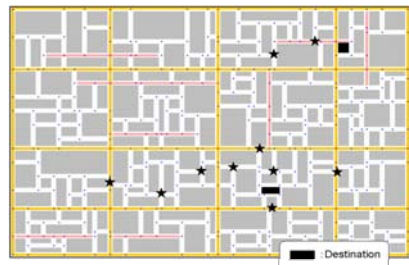


Fig. 6. Panel locations by simulation

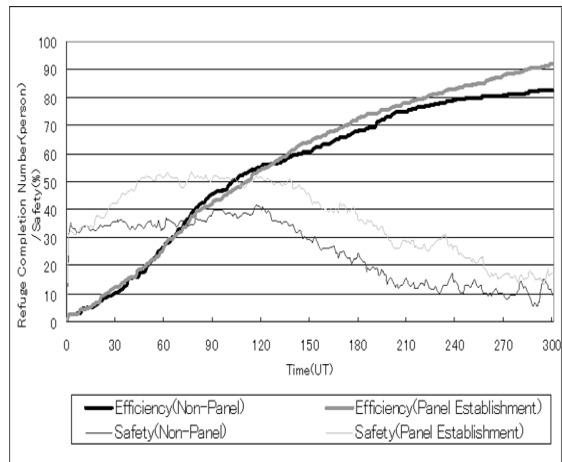


Fig. 7. Evaluation of panel location by simulation

Experiment-4

Figure 10 is a result, and Table 5 summaries the evaluation for safety. The more the number of agents is, the better the result of dynamic routing control is.

Experiment-5

Figure 11 is a result, and Table 6 shows the safety evaluation. The usage of global information is more useful to support the evacuation than that of local information.

Table 7 shows all of simulation results. Additionally, we experimented using the practical road maps, and arrange the results in Table 8. Approximately, the result is the same, except the case that the corresponding experiments are not well performable.

Table 4. Safety evaluation by manual

Number of agents	S (by simulation)	S (by human beings)
100	111.2	97.5
200	122.2	97.2

Table 5. Dynamic and Static setting

Number of agents	S (dynamic setting)	S (static setting)
100	111.2	115.8
200	122.2	102.8

Table 6. Usage of global and local data

Number of agents	S (local information)	S (global information)
100	111.2	143.1
200	122.2	129.1

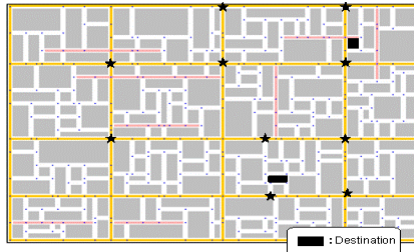


Fig. 8. Panel location by manual

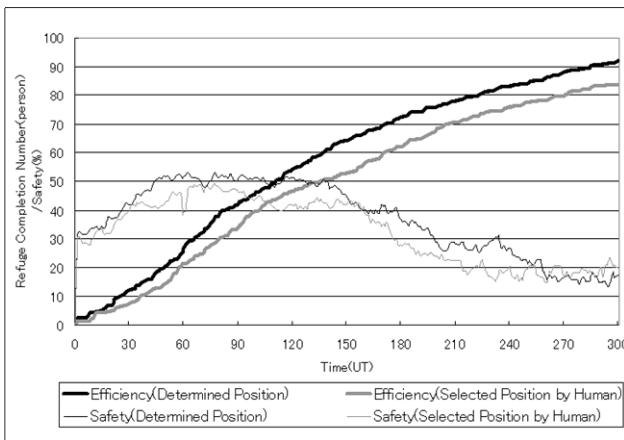


Fig. 9. Evaluation of panel location by manual

Table 7. Summary in grid-mesh road network

Experiment	effectiveness	safety
Experiment-2 (with/without panels)	With panel	With panel
Experiment-3 (by simulation/manual)	Simulation	Simulation
Experiment-4 (dynamic/static setting)	Dynamic setting	--
Experiment-5 (local/global information)	Global information	Global information

Table 8. Summary in map-based road network

Experiment	effectiveness	safety
Experiment-2 (with/without panels)	With panel	With panel
Experiment-3 (by simulation/manual)	Simulation	Simulation
Experiment-4 (dynamic/static setting)	Dynamic setting	Dynamic setting
Experiment-5 (local/global information)	Global information	--

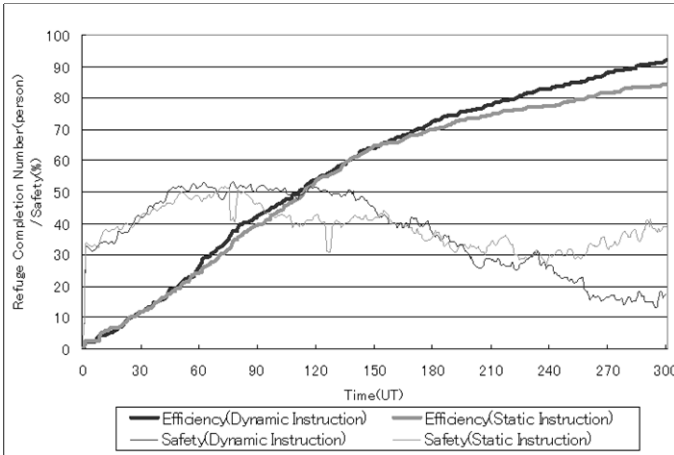


Fig. 10. Dynamic and Static setting

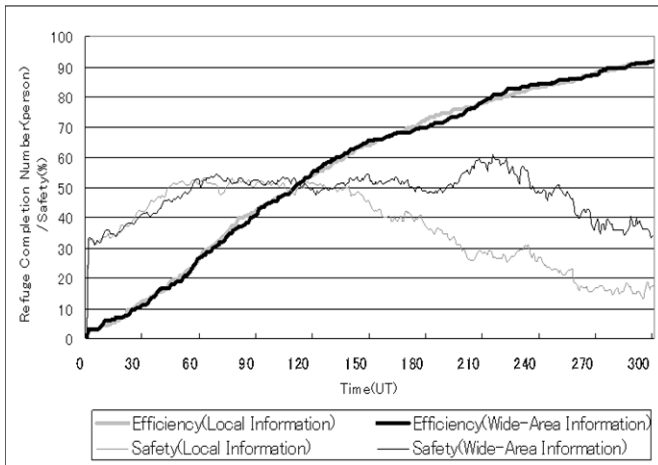


Fig. 11. Comparison between global and local data

5 Conclusion

In this paper, we proposed an evacuation support method on the basis of routing panels for destinations, and evaluated the effectiveness and safety through computer simulation. These results are useful to support the evacuation planning, because the routing panels are distributed commonly and the utilization is kept in the practical step.

References

1. Minami,K., Murakami,Y, Kawasoe,T., and Ishida,T.: Evacuation Simulation by Using Multi-agent System, Annual Report.of JSAI2002, 2B1-04(2002)
2. Kagami,H.: Relationship Between Mortality and Building Damage, Proc.of Texas Disaster Symposium, Houston, USA (2003).

Facial Expression Classification: Specifying Requirements for an Automated System*

Ioanna-Ourania Stathopoulou and George A. Tsihrintzis

Department of Informatics
University of Piraeus
Piraeus 185 34
Greece
{iostath, geoatsi}@unipi.gr

Abstract. Automated facial expression classification arises as a difficult, yet crucial, pattern recognition problem in the design of human-computer interaction and multimedia interactive service systems. In the process of building NEU-FACES, a novel system of ours for processing multiple camera images of computer user faces to determine their affective state, we conducted and present here an empirical study in which we specify related design requirements, study statistically the expression recognition performance of humans, and identify quantitative facial features of high expression discrimination and classification power.

Keywords: Facial expression analysis, human-computer interaction, multimedia interactive services, affective computing.

1 Introduction

Facial expressions are particularly significant in communicating information in human-to-human interaction and interpersonal relations, as they reveal information about the affective state, cognitive activity, personality, intention and psychological state of a person and this information is, in fact, difficult to mask. Similarly, images that contain user faces are instrumental in the development of more effective and friendlier methods in human-computer interaction. Indeed, the facial expressions “neutral”, “smiling”, “sad”, “surprised”, “angry”, “disgusted” and “bored-sleepy” arise very commonly during a typical human-computer interaction session and, thus, vision-based human-computer interaction systems that recognize them could guide the computer to “react” accordingly and attempt to better satisfy its user needs.

It is common experience that the variety in facial expressions of humans is large and, furthermore, the mapping from psychological state to facial expression varies significantly from human to human and is complicated further by the problem of *pretence*, i.e. the case of someone’s facial expression not corresponding to his/her true psychological state. These two facts make the analysis of the facial expressions of

* Support for this work was provided by the General Secretariat of Research and Technology, Greece, under the auspices of the PENED-2003 program.

another person difficult and often ambiguous. This problem is even more severe in *automated facial expression classification*, as face images are non-rigid, have a high degree of variability in size, shape, color and texture and variations in pose, facial expression, image orientation and conditions add to the level of difficulty of the problem.

To address this problem, researchers have developed either *face feature*-based methods or methods which rely on *image-based representations of the face* (see, for example, [1] for a presentation of the highlights of some of these previously published methods). The former approaches rely on basic, image-extracted facial features and their location or computed relations between them, while the latter approaches utilize the entire face as input to artificial neural network-based classifiers. Recently, we have been developing a novel automated facial expression classification system [2-5], called NEU-FACES, in which certain features extracted as variations between the neutral and other common expressions are fed into neural network-based classifiers.

For use in the development, training, and testing of facial expression classifiers, appropriate extensive facial databases are required. These databases are non-trivial to create, as they need to be sufficiently rich in both facial expression variety and representative samples of each expression. Moreover, the creators of the database need to make sure that the human models form their true facial expressions when posing. In the past years, only a relatively small number of relevant face databases have been presented in the literature. These include : (1) The AR Face Database [6], which contains over 4,000 color images of 126 persons' faces in front view, forming different facial expressions under various illumination conditions and occlusion (eg., sun glasses and scarf). The main disadvantage of this database is its limitation to containing only four facial expressions, namely "neutral", "smiling", "anger", and "scream". (2) The Japanese Female Facial Expression (JAFPE) Database [7] contains 213 images of the neutral and 6 additional basic facial expressions, as formed by 10 Japanese female models. (3) The Yale Face Database [8], which contains 165 gray-scale GIF-formatted images of 15 individuals. These correspond to 11 images per subject of different facial expression or configuration, namely, center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink. (4) The Cohn-Kanade AU-Coded Facial Expression Database [9], which includes approximately 2000 image sequences from over 200 subjects and is based on the Facial Action Coding System (FACS), first proposed by Paul Ekman [10]. (5) The MMI Facial Expression Database [11], which includes more than 1500 samples of both static images and image sequences of faces in front and side view, displaying various expressions of emotion and single and multiple facial muscle activation.

Although many of the aforementioned face databases were considered for the development of our system (NEU – FACES), either the number of different facial expressions or the number of representative samples of them were found insufficient for developing NEU-FACES up to a fully operational form and, thus, we decided to create our own facial expression database. In this paper, we present an empirical study of the facial expression classification problem in images, as well as details of the process followed in creating our facial expression database. More specifically, the paper is organized as follows: in Section 2, we analyze the performance of expression

classification by humans and describe our data acquisition technique. In Section 3, we identify differences between facial expressions, which we quantify and analyze in terms of discrimination power in Section 4. Finally, in Sections 5, we draw conclusions and point to future work.

2 Empirical Study of Facial Expression Classification by Humans

Our study of expression classification by humans consisted of three steps:

1. *Observation of the user's reactions during a typical human-computer interaction session:* From this step, we concluded that the facial expressions corresponding to the “neutral”, “smiling”, “sad”, “surprised”, “angry”, “disgusted” and “bored-sleepy” psychological states arose very commonly in human-computer interaction sessions and, thus, form the corresponding classes for our classification task

2. *Data Acquisition:* To acquire image data, we built a three-camera system, as in Fig. 1. Specifically, three identical cameras of 800*600 pixel resolution were placed with their optical axes on the same horizontal plane and successively separated by 30-degree angles. Subjects were asked to form facial expressions, which were photographed by the three cameras simultaneously.

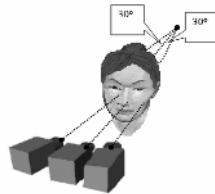


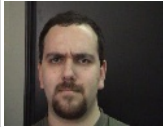

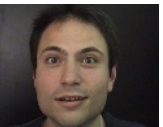






Fig. 1. The geometry of the data acquisition setup

To ensure spontaneity, the subject was presented with pictures on a screen behind the central camera. These pictures were expected to generate such emotional states that mapped on the subject's face as the desired facial expression. For example, to have a subject assume a “smiling” expression, we showed him/her a picture of funny content. We photographed the resulting facial expression and only then asked him/her to classify this expression. If the image shown to him/her had resulted in the desired facial expression, the corresponding photographs were saved and labeled; otherwise, the procedure was repeated with other pictures. The final dataset consisted of 250 different persons, each forming the seven expressions: “neutral”, “smiling”, “sad”, “surprised”, “angry”, “disgusted” and “bored-sleepy”.

3. *Questionnaires -- Classification of expressions by humans:* To understand how humans classify facial expressions and estimate the corresponding error rate, we developed a questionnaire in which each we asked 300 participants to classify the facial expressions in 36 images. Each participant could choose from 11 of the most common

facial expressions, such as: “angry”, “smiling”, “neutral”, “surprised”, etc., or specify some other expression he/she thought appropriate.

Table 1. Typical face image subsets in our questionnaire

<u>The three sets of our questionnaire</u>			
1st Set			
2nd Set			
3rd Set			

Specifically, our dataset was consisted of 3 subsets of images, typical examples of which are shown in Table 1:

- various images of individuals placed in a background and mimicking an expression,
- a sequence of facial expressions of the same person without a background, and
- facial images of various individuals without a background.

We found that the “surprised” expression was the one recognized with the lowest error rate of 22%. Next, follow the “smiling” and “neutral” expressions, with error rates of 30% and 35%, respectively. The highest error rate corresponded to the “sad” expression and rose to 88%, while the “angry” and “disappointment” expressions had an error rate of 80% and 76%, respectively. These are summarized in Fig. 2. From these findings, we conclude that the facial image classification task is quite challenging and the fact, that expressions such as “angry”, “sad” and “disappointed” seem to differ significantly from person to person or some people may be too shy to form them clearly, may result in high classification error rates.

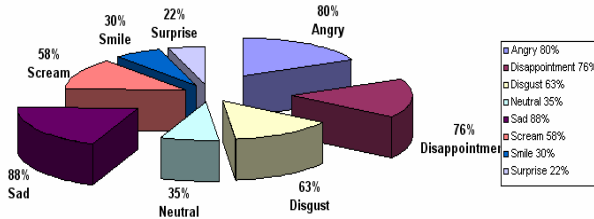


Fig. 2. Error rates in recognizing the expressions in our questionnaire

3 Feature Selection

From the collected dataset, we tried to identify differences between the “neutral” expression of a model and its deformation into other expressions, as typically highlighted in Table 2. To convert pixel data into a higher-level representation of shape, motion, color, texture and spatial configuration of the face and its components, we locate and extract the corner points of specific regions of the face, such as the eyes, the mouth and the brows, and compute their variations in size, orientation or texture between the neutral and some other expression. This constitutes the *feature extraction process* and reduces the dimensionality of the input space significantly, while retaining essential information of high discrimination power and stability. The extracted features, the face regions, and the dimension ratios used to classify the expressions are summarized in Fig. 3.

Table 2. Differences among facial expressions



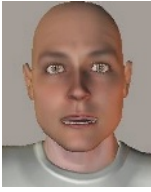



<i>Variations between Facial Expressions:</i>	
Smiling	Bored-Sleepy
 <ul style="list-style-type: none"> • Bigger-broader mouth • Slightly narrower eyes • Changes in the texture of the cheeks • Occasionally, changes in the orientation of brows 	 <ul style="list-style-type: none"> • Head slightly turned downwards • Eyes slightly closed • Occasionally, wrinkles formed in the forehead and different direction of the brows
Surprised	Sad
 <ul style="list-style-type: none"> • Longer head • Bigger-wider eyes • Open mouth • Wrinkles in the forehead (changes in the texture) • Changes in the orientation of eyebrows (the eyebrows are raised) 	 <ul style="list-style-type: none"> • Changes in the direction of the mouth • Wrinkles formed on the chin (different texture) • Occasionally, wrinkles formed in the forehead and different direction of the brows
Angry	Disgusted-Disapproving
 <ul style="list-style-type: none"> • Wrinkles between the eyebrows (different textures) • Smaller eyes • Wrinkles in the chin • The mouth is tight • Occasionally, wrinkles over the eyebrows, in the forehead 	 <ul style="list-style-type: none"> • The distance between the nostrils and the eyes is shortened • Wrinkles between the eyebrows and on the nose • Wrinkles formed on the chin and the cheeks



Fig. 3. The extracted features (gray points), the measured dimensions (gray lines) and the regions (orthogonals) of the face

4 Quantification of Feature Discrimination Power

We have found that a number of high discrimination power features may correspond to the location and shape of face components, as these vary significantly among the “neutral”, “smiling” and “surprised” facial expressions. For example, the distribution of the mouth and face dimension ratios over a broad range of face images are shown in Fig. 4 for these three different facial expressions. Similarly, the orientation and texture of specific portions of the face also vary significantly between expressions. Typical results of these two measures are demonstrated in Table 2.

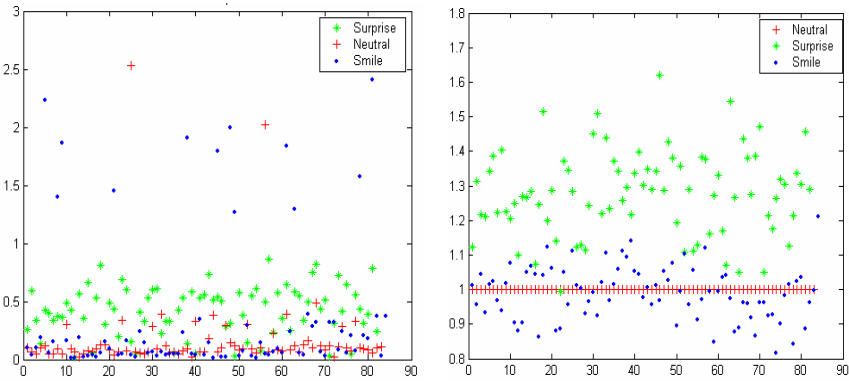






















Fig. 4. Distribution of mouth (left) and face (right) dimension ratio

As far as the collected *side view* images are concerned, our study showed that formation of some expressions involves deformation of a person’s head sides and, thus, additional classification features may be derived from side view face images. In fact, features may be more evident in side view rather than front view images for certain expressions. For example, better discrimination between the “neutral” and “smiling” expression seems to be achieved in front view images, whereas the “surprised” expression seems to be better identified in side view images. Similarly, the “sad” and “angry” are better discriminated in front rather than in side view images as forehead texture, one of the corresponding classification features, is better computed in front rather than side view images. Thus, we conclude that better facial expression classification results can be achieved by using images of several views of a person’s face.

Table 3. Different measures of the facial expressions

<i>Measures of texture</i>				
Region between the brows:				
<i>Expressions</i>	<i>Input Image</i>	<i>Difference between the relevant 'neutral expression'</i>	<i>Texture Measure</i>	<i>Possible facial expression class</i>
Neutral			0	'neutral'
Smiling			16	'neutral', 'smiling', 'surprised'
Surprised			6	'neutral', 'smiling', 'surprised'
Angry			44	'angry', 'disgusted'
Disgusted			175	'angry', 'disgusted'
Forehead:				
Neutral			0	'neutral', 'smiling', 'angry', 'disgusted'
Smiling			0	'neutral', 'smiling', 'angry', 'disgusted'
Surprised			8	'surprised', 'angry'
Angry			0	'neutral', 'smiling', 'angry', 'disgusted'
Disgusted			0	'neutral', 'smiling', 'angry', 'disgusted'

5 Conclusions – Future Work

Automated face detection and expression classification in images is a prerequisite to the development of intelligent human-computer interaction and multimedia interactive service systems. However, the development of integrated, fully operational such automated systems is non-trivial, a fact that was further corroborated by statistical results for human classifiers in our paper. Towards building such systems, we have been developing a novel automated facial expression classification system [1-5], called NEU-FACES, in which features extracted as variations between the neutral and other common expressions are fed into neural network-based expression classifiers.

Although at first we considered several existing face databases which contained a number of representative samples of various facial expressions, either the number of different facial expressions or the number of representative samples of them were found insufficient for developing NEU-FACES up to a fully operational form and, thus, we decided to create our own facial expression database. In this paper, we

presented an empirical study of the facial expression classification problem in images, specified related system design requirements, studied statistically the expression recognition performance of humans, and identified quantitative facial features of high expression discrimination and classification power, and presented details of the process followed in creating our facial expression database.

In the future, we will extend this work in the following directions: (1) we will enhance our database with additional facial expressions and corresponding samples, (2) we will investigate the application of quality enhancement techniques to our image dataset and seek to extract additional classification features from them, and (3) we will extend our database so as to contain *sequences of images of facial expression formation* rather than simple static images of formed expressions and seek in them additional features of high classification power. This and related work is currently in progress and will be presented on a future occasion.

References

- [1] I.-O. Stathopoulou and G.A. Tsihrintzis, "Detection and Expression Classification Systems for Face Images (FADECS)," *Proc. 2005 IEEE Workshop on Signal Processing Systems (SiPS05)*, Athens, Greece, November 2 – 4, 2005
- [2] I.-O. Stathopoulou and G.A. Tsihrintzis, "A neural network-based facial analysis system," *Proc. 5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisboa, Portugal, April 21-23, 2004
- [3] I.-O. Stathopoulou and G.A. Tsihrintzis, "An Improved Neural Network-Based Face Detection and Facial Expression Classification System," *Proc. IEEE International Conference on Systems, Man, and Cybernetics 2004*, The Hague, The Netherlands, October 10-13, 2004
- [4] I.-O. Stathopoulou and G.A. Tsihrintzis, "Pre-processing and expression classification in low quality face images", *Proc. 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Smolenice, Slovak Republic, June 29 – July 2, 2005
- [5] I.-O. Stathopoulou and G.A. Tsihrintzis, "Evaluation of the Discrimination Power of Features Extracted from 2-D and 3-D Facial Images for Facial Expression Analysis," *Proc. 13th European Signal Processing Conference*, Antalya, Turkey, September 4-8, 2005
- [6] A.M. Martinez and R. Benavente, "The AR face database", CVC Tech. Report #24, 1998.
- [7] M. J. Lyons, J. Budynek, & S. Akamatsu, "Automatic Classification of Single Facial Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21 (12): 1357-1362 (1999)
- [8] The Yale Database: <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>
- [9] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis", *Proc. of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, Grenoble, France, 2000
- [10] P. Ekman and W. Friesen W., *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*, Englewood Cliffs, NJ: Prentice Hall (1975)
- [11] M. Pantic, M.F. Valstar, R. Rademaker and L. Maat, "Web-based Database for Facial Expression Analysis", *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME'05)*, Amsterdam, The Netherlands, July 2005.

On the Software Engineering Aspects of Educational Intelligence

Thanasis Hadzilacos^{1,2} and Dimitris Kalles¹

¹ Hellenic Open University, Patras, Greece
{thh, kalles}@eap.gr

² Research Academic Computer Technology Institute, Patras, Greece

Abstract. Students who enroll in the undergraduate program on informatics at the Hellenic Open University (HOU) demonstrate significant difficulties in advancing beyond the introductory courses. We use decision trees and genetic algorithms to analyze their academic performance throughout an academic year. Based on the accuracy of the generated rules, we analyze the educational impact of specific tutoring practices and reflect on some software engineering issues involved in the development of organization-wide measurement systems.

1 Introduction

All measurements affect that which is being measured. However, measurement for performance evaluation greatly affects what is being measured and may distort the process being evaluated. This can lead to misleading policies in education. For example, the number of computers per 100 high school students is an indicator of educational ICT utilization (other things being equal). When however the EU Commissioner pronounces this indicator to be used as a funding goal (“*by 2004 all member states should have 1 computer per 10 students*”) the result is that schools keep old machines and that the total ICT budget is shifted toward hardware.

It is reasonable to suggest that student success is a natural success indicator of a University (of a course, of a class, or of a teacher). However, if that success is used as a criterion for tutor contract renewal, and if students must evaluate their own teachers, then tutors may tend to lax their standards. This paper is about dealing with this issue in the context of a large (over 25,000 students) Open-and-Distance-Learning (ODL) University, the Hellenic Open University (HOU) and in particular its undergraduate Informatics program. We ask how we can detect best distance tutoring practices in an ‘objective’ way and effectively disseminate them.

In this paper, we describe a measurement strategy that we have used in HOU, to analyze whether some specific tutoring practices have a significant effect on the performance of junior students. Appreciating the inherent caveats of measurement *per se* and of measurement for performance evaluation, we have developed aggregate statistics that rely on the methodologies of decision trees and genetic algorithms. To account for possible wide-scale adoption, we clearly acknowledge the implications and the potential for related processes at HOU, which brings into the front the software engineering aspects of devising a system to disseminate the findings.

This paper is structured in four subsequent sections. First, we briefly review the problem of predicting student performance at large, and the related techniques we have been using at HOU. We then single out two clearly different policies on dealing with students who have failed an exam and devise a set of experiments to observe whether these policies can be consistently compared. We then discuss the validity of the results and potential implications of our findings from an educational point of view and elaborate on the organizational and human-issues aspects of obtaining quality information for measurement and of communicating that information to all players in an organizational context. Finally, we reflect on the export potential of our approach.

2 Background

Drop-out is a significant issue in ODL universities [1, 2] and at HOU it mostly occurs very early in the studies, as a result of failure in a junior year module. For the Informatics program, three modules are heavily populated and high-risk: “Introduction to Informatics” (INF10), “Fundamental Software Engineering” (INF11) and “Mathematics” (INF12). Collectively, these modules cover fundamental topics on mathematics, software engineering, programming, databases, operating systems and data structures.

Key demographic characteristics of students (such as age, sex, residence etc), their marks in written assignments and their presence or absence in plenary meetings may constitute the training set for the task of learning (and explaining, and predicting) whether a student would eventually pass or fail a specific module.

Initial experimentation at HOU [3] consisted of using several machine learning techniques to predict student performance with reference to the final examination. The WEKA toolkit [4] was used and the key finding, also corroborated by tutoring experience, is that success in the initial written assignments is a strong indicator of success in the examination. A surprising finding was that demographics were not important.

We then followed-up with experimentation [5] using the GATREE system [6], which produced significantly more accurate and shorter decision trees (when compared to other conventional approaches; this is also the reason why we still use that approach). That stage confirmed the qualitative validity of the original findings (also serving as result replication) and set the context for experimenting with accuracy-size trade offs.

Our technical approach is based on the generation of decision trees via genetic algorithms [6]. A decision tree like the one in Fig. 1 (similar to the ones actually produced by GATREE) tells us that a mediocre grade at an assignment, turned in at about the middle (in the time-line) of the module, is an indicator of possible failure at the exams, whereas a non-mediocre grade refers the alert to the last assignment. An excerpt of a training set that could have produced such a tree is shown in Table 1.

A student who fails a module examination can sit the exam on the following academic year. Such “virtual” students are only assigned to student groups for examination purposes and the group tutor is responsible for marking their papers only. Virtual students are not entitled to attending plenary sessions or submitting assignments, but tutors may decide to relax this regulation.

Subsequently, we focus on the basic informatics INF10 and INF11 modules since they also demonstrate only one clear difference in the tutoring practice.

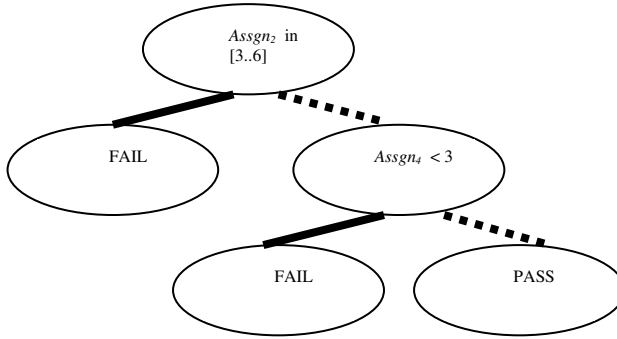


Fig. 1. A sample decision tree

Table 1. A sample decision tree training set

Assgn ₁	Assgn ₂	Assgn ₃	Assgn ₄	Exam
9.1	5.1	4.6	3.8	FAIL
7.6	7.1	5.8	6.1	PASS
...

One step taken by tutors of the INF10 and INF11 modules is to hold a plenary marking session for each module after an examination, to discuss variations in individual marking styles based on a predefined assignment of points to exam questions. This is especially important for problems that involve design or prose argumentation. We note that this practice is not widespread within HOU.

A further ad hoc step taken (during the 2003-4 academic year) by the INF11 tutors was to assign one experienced tutor to all virtual students, as opposed to distributing virtual students across tutors. These students were fully (and only) supported by an asynchronous discussion forum and by a synchronous virtual classroom. This also served as a constraint on the “degrees of freedom” of our educational experiment.

A summary description of these policies is shown in Table 2.

Table 2. Summary description of approaches

Approach	INF10	INF11
Post-exam plenary marking session	√	√
Grouping of virtual students	...	√

3 The Experimental Environment

We use GATREE for all experiments, with the default settings for the genetic algorithm operations (cross-over probability at 0.99, mutation probability at 0.01, error rate at 0.95 and replacement rate at 0.25). All experiments were carried out using 10-fold cross-validation, on which all averages are based. The experiments were made with a configuration of 150 generations with a population of 150 trees per generation.

Our methodology is the following: we attempt to use the student data sets to develop success/failure models represented as decision trees. We then use the differences between the models derived when we omit some attributes to reflect on the importance of these attributes. The results are then used to comment on alternative educational policies for dealing with virtual students.

The measurement is based on partitioned data sets. As opposed to treating each student population as an individual data set, we have partitioned them into module groups, according to how tutors are assigned to groups. This partitioning allows us to examine an important question: do the grading practices of one tutor apply predictably well to a group supervised by another tutor at the same module?

To answer the question, we devised a very detailed study based on inducing models for one group and testing them at another. An example is shown in Table 3.

Table 3. A template for tabulating cross-testing results

Data Set	D_1	D_2	...	D_n
D_1	CV_1	$V_{1,2}$
D_2	...	CV_2
...
D_n

A few words on notation are in order. D_i refers to the student group supervised by tutor i . CV_i refers to the cross-validation accuracy reported on data set D_i . $V_{i,j}$ refers to the validation accuracy reported on data set D_j , using the model of data set D_i .

All data refer to the 2003-4 academic year and they do not differentiate between typical and virtual students. To appreciate the scale, we note that for modules INF10 and INF11, n takes the value of 31 and 16 respectively (with student populations at about 1000 and 500 students respectively).

For each module, we first computed the above table at its basic version, with the assignment grades as attributes and the *pass/fail* flag as class attribute. We then computed the above table at its extended version, which included two additional attributes: the tutor and the year of first sitting the exam for that module. We then subtracted the two tables (*basic* – *extended*) and computed the average of all cells. Table 4 shows the results, also including calculations for standard deviations and experiments with some accuracy/size trade-offs.

Table 4. Accuracy results for decision trees

Data Set	Smallest Trees	Allow Larger Trees if Accuracy Grows as well
INF10	- 4.92 (10.39)	- 3.65 (9.12)
INF11	0.28 (8.01)	0.24 (8.58)

We note that group models are clearly more aligned within the INF11 module which demonstrates a model “smoothing” across its training set partitions. This success indicator suggests that the failure explanation must be traced solely to academic performance (i.e. assignments) and that virtual students are not disadvantaged.

4 On User Requirements for an Educational Intelligence System

The wider context of our research is to investigate the building an “early warning and reaction system” for students with “weak” performance. As the cautious reader might imagine, user acceptance of such a system has delicate operational and political aspects that can easily transcend technical issues. To deal with this, we first need to think who the user might be.

Note that while we do not expect the individual or aggregate results to necessarily hold at other educational establishments, we believe that the measurement approach (with a relatively wide assortment of tools) should be readily applicable beyond our application case study. We also explore these directions below.

A sensitive point is that it would be unwise to simply consider the higher or lower overall *absolute* accuracy rate of (any) model in one module as an indicator of success of an approach, at least at this early stage of the research. It is for this reason that in the experiments described above we never pit one module’s accuracy against another module’s accuracy; besides referring to different student populations (including differences in population sizes), a module also refers to different tutors and to another scientific field.

An obvious plausible user is any student; therefore a key issue is how to communicate the model to the student. One option is to “publicize” the model in advance to the students. Even though the findings may not recur in the next year (though they might hold remarkably well [5]), an aura of “precaution” could be implicitly but effectively communicated, highlighting an otherwise obvious advice: unless one performs consistently well, a failure is quite likely. Another option suggests that the model should be only “publicized” to tutors, who then must make individual decisions how to use it. This can have some variances, however. For example, “using” the model can be both served by the tutor who ignores students who do not send in their first assignment and by the tutor who persistently goes with focused counseling after students who do badly in their third assignment. A middle-of-the-road course may be also possible.

A sensitive point is that it would be unwise to simply consider the higher or lower overall *absolute* accuracy rate of (any) model in one module as an indicator of success of an approach, at least at this early stage of the research. It is for this reason that in the experiments described above we never pit one module’s accuracy against another

module's accuracy; besides referring to different student populations (including differences in population sizes), a module also refers to different tutors and to another scientific field.

The above discussion suggests that, besides the students, tutors are also candidate users. A very sensitive point is the interpretation of the individual model differences within each module. It might be tempting to think that this singles out groups within modules (and, as a result, their tutors) which seem to be not integrating into a module-wide view. This could have far-reaching effects if not properly managed. Unless, we painstakingly understand all background that may have led a tutor to adopt a particular approach (and, in that course, analyze all false alarms), pitting tutors' performance against each other is bound to create strife that will endanger the legitimacy of the approach. A full background analysis can be, however, safely ruled out due to the excessive "detective" costs that it entails. It is far more instructive to communicate the findings without any impact on performance evaluation and allow each individual tutor to reflect on their tutoring approach. It is very interesting that this is remarkably similar to the issue of whether we can actually use the models derived to better target each individual student.

Note that, in both cases, the issue of how one publicizes the measurement results can have profound differences in how these results are received and interpreted. We believe that any approach which might lead us into taking micro-managing decisions would create a distraction. What is more important, we claim, is to detect and observe the trends within the module itself and try to understand what macro-managing actions need to be taken at the module level.

We now explore whether this suggests that the university at large is a primary user.

We cannot yet answer whether the approach of the INF11 tutors is an approach that would have had replicable educational results in the other modules. The most obvious reason is that exact replication of the above experiments is impossible. Had we wanted to experiment with INF11 approach in INF10, we cannot hope to ever again observe the given set of students and their assignment to groups within modules, as well as the given set of tutors and their assignment to groups. This is one of the reasons that we progressively narrowed down our experiments: we started at only one undergraduate program, then focused on the most junior and well-subscribed modules, then singled out the two ones that demonstrated one difference only at the policy level.

Note that the question of credibility of the results runs across all our experiments. In [5] we first set out an initial presentation of building classifiers for student performance using decision trees and genetic algorithms. In [7] we experimented with using the "virtual student" property as a discriminator of the classification – therein we demonstrated that when the discriminating property is suppressed, this is an indication of the fact that the second chance these students get is a substantial one. In [8] we argue that the results we are obtaining are consistent over several experimental rounds with relatively small data sets; this is essential to boost the credibility of the derived models.

By establishing credible statistics at this level, we are now progressively working in the opposite direction, that of expanding our experimental range. This must be done across modules of the same program, then across programs, and at the same time, we

must deal with the question of whether these models are consistent across consecutive academic years [7, 7].

Our measurement approach is self-contained, in the sense that all data required are available at the university registry. This eliminates a substantial risk factor, that of having to collect the data from the tutors and the consolidate it. At the same time, however, we have not yet dealt with the software and data architectural aspects of collecting the data. While it is straightforward for an investigating team of several researchers to collect the data directly from the “source” for pilot projects, establishing an organization-wide process also entails issues of dealing with privacy issues (we must not disclose any individual student record).

Such issues will probably be affected by the decision of whether the models are computed centrally or in a decentralized fashion (by devolving responsibility to the tutors, for example). In any case, deploying our measurement scheme in an organization-wide context would also lend support to our initial preference for short models [5]. At the same time, the possibility of a decentralized scheme also suggests that we should strive to use tools that do not demand a steep learning curve on the part of the tutors. As a result we intend to continue favoring GATREE compared to other software for the particular data analysis tasks. This will probably be a recurring theme in our attempts to diffuse our approach as classical approaches using statistics [9, 10] can be cumbersome to disseminate to people with a background on humanities or arts, and this could have an impact on the user acceptance of such systems.

The above all raise the fundamental question of whether one measures the performance of actors (students or tutors) or the performance of the system at large (the ODL system implemented in HOU). Earlier, we have also conjectured that it is the latter alternative that has the most potential from an educational point of view [5]. In the opening of this paper we have furnished a measurements-related discussion that supports this conjecture. We strongly believe that the experiments documented in this paper suggest that this conjecture is valid.

5 Conclusions

In this paper we have used an advanced AI method in order to obtain information necessary for the application of organizational (educational in this case) policies and we have explored the potential data flows from the data collection stage to the user presentation stage.

Quality control systems are fundamental in any large organization and measurement is a *sine-qua-non* method of quality control. We have argued that deciding what to measure and how to convey the measurement (and the underlying measurement message) in a university context is a very complex task, especially when the potential users can range from one (the university itself) to well over several hundreds (the tutors) and to tens of thousands (the students). To do this successfully, we always treat our experiments as an integral part of a software engineering project.

If we tread carefully and succeed, we expect that this will render our approach applicable to other educational settings, as well. Taking the sting out of individual performance evaluation but still being able to convey the full unabridged message should be a venerable target for every learning community.

Acknowledgements

This paper shares setting-the-context paragraphs with some references [5, 7, 8], as the accurate description of the educational environment is a significant factor in conveying our subsequent results concisely. However, it duplicates neither experiments nor results.

References

1. Open Learning (2004). Special issue on "Student retention in open and distance learning". 19:1.
2. Xenos, M., Pierrakeas, C., & Pintelas, P. (2002). A survey on student dropout rates and dropout causes concerning the students in the Course of Informatics of the Hellenic Open University. *Computers & Education*, 39, 361-377.
3. Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using Machine Learning techniques. *Applied Artificial Intelligence*, 18:5, 411-426.
4. Witten, I., & Frank, E. (2000). *Data mining: practical machine learning tools and techniques with Java implementations*. San Mateo, CA: Morgan Kaufmann.
5. Kalles, D., & C. Pierrakeas (2006). Analyzing student performance in distance learning with genetic algorithms and decision trees (to appear in: *Applied Artificial Intelligence*).
6. Papagelis, A., & Kalles, D. (2001). Breeding decision trees using evolutionary techniques. *Proceedings of the International Conference on Machine Learning*, Williamstown, Massachusetts, pp. 393-400, Morgan Kaufmann.
7. Kalles, D., & Pierrakeas, C. (2006). Using Genetic Algorithms and Decision Trees for a posteriori Analysis and Evaluation of Tutoring Practices based on Student Failure Models (to appear in: 3rd IFIP conference on Artificial Intelligence Applications and Innovations, Athens, Greece).
8. Hadzilacos, Th., Kalles, D. Pierrakeas, C. & M. Xenos (2006). On Small Data Sets Revealing Big Differences. *Proceedings of the 4th Panhellenic conference on Artificial Intelligence*, Heraklion, Greece, Springer LNCS 3955, pp. 512-515, 2006.
9. Werth, L.H. (1986). Predicting student performance in a beginning computer science class. *Proceedings of the 17th SIGCSE technical symposium on Computer science education*, Cincinnati, OH, pp. 138-143.
10. Minaei-Bidogli, B., Kashy., D.A., Kortemeyer, G., & Punch, W.F. (2003). Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education conference*, Boulder, CO.

Feature Model Based on Description Logics^{*}

Shaofeng Fan and Naixiao Zhang

LMAM, Department of Information Science,
School of Mathematical Sciences,
Peking University, Beijing 100871, China
{fsf, znx}@is.pku.edu.cn

Abstract. Intelligent interactive systems have begun to adopt knowledge and software engineering technologies in an attempt to effective development. Feature models have been widely used in knowledge and software engineering for the reuse purpose. However, due to the lack of a formal semantics of feature models, it is rather difficult to perform rigorous consistency reasoning on them. Without guaranteed consistency of feature models, the quality of interactive systems based on them, can not be guaranteed. In this paper, how to formalize feature models with Description Logics is investigated. Following the proposed translation principles, each feature model is formalized into an *ALCQI* knowledge base. Hence the consistency reasoning on the feature model turns into the consistency reasoning on the corresponding *ALCQI* knowledge base. Especially, the latter reasoning can be automatically performed via the description logic reasoner RACER.

1 Introduction

There is growing awareness of the importance of the development processes for intelligent interactive systems through knowledge engineering and software engineering approaches [7,10]. Domain engineering is a software reuse approach, which aims to develop reusable software assets focusing on a particular application domain. The most important result of domain engineering is the feature model [6]. The prominent and distinctive user requirements are denoted by common and variant features, which are in turn captured into a graphical feature model. So far quite a number of feature-centered domain engineering methods have been proposed, such as FODA [8], ODM [11] and KAPTUR [2].

By capturing user requirements of a particular application domain, feature models act as the start-point of software development. The quality of the interactive systems is strongly affected by the consistency of the feature models. Any system, based on an inconsistent feature model, is quite prone to be inconsistent. Therefore, the consistency reasoning becomes a critical problem. However, due to the lack of a formal semantics, there is no automated tool to perform consistency checking of a feature model. It is neither efficient nor reliable to validate

^{*} This paper was supported by the National Natural Science Foundation of China under Grant. 60473056.

it by hand. Especially it becomes infeasible to accomplish the reasoning when there are a great deal of features and constraints between them.

Description Logics (DLs) is a formalism for representing knowledge and reasoning about it [1]. DL languages have formal model-theoretic semantics, and their main strength lies in the support of powerful reasoning mechanisms. Much attention has been paid on the application of DLs [1].

We believe that there is a strong similarity between description logics systems and feature models, both of which represent concepts in a particular domain and define how various properties relate among them. Hence, in this paper we propose feature models based on description logics. We present how to formalize a feature model with the DL *ALCQI*. Thus the feature model is provided with the DLs formal semantics, and the consistency of feature models is translated to the consistency of the knowledge base consequently. Then the automated reasoning on the DL representation of the feature model can be performed using the DL reasoner–RACER. Thus the consistency of the feature model is checked automatically with high efficiency and reliability.

The remainder of the paper is organized as follows. Section 2 gives a brief overview of feature models and DLs. Section 3 presents our proposal of feature models based on DLs. A case study is given in section 4 to demonstrate our approach. In section 5, related works are compared and distinguished. Section 6 concludes the paper and indicates the future work.

2 Background

2.1 Graphic Feature Model

A feature model consists of a feature diagram and some additional information, such as rationale, constraints and dependency rules. A feature diagram provides a graphical tree-like notation that shows the hierarchical organization of features. It consists of a set of nodes, a set of directed edges, and a set of edge decorations. The root of the tree represents a concept node. All other nodes represent features.

Here we take the graphical notation introduced in [6]. Assuming a feature is selected, we have the following definitions on its child features:

- *Mandatory* feature: The feature must be included into the description of a concept instance, pointed to by a simple edge ending with a filled circle.
- *Optional* feature: The feature may or may not be included into a concept instance, pointed to by a simple edge ending with an empty circle.
- *Alternative* feature: Exactly one feature from a set of features can be included into a concept instance. The nodes of a set of alternative features are pointed to by edges connected by an arc.
- *Or* feature: One or more features from a set of features can be included into a concept instance. The nodes of a set of or features are pointed to by edges connected by a filled arc.

However not all arbitrary feature configurations have practical meaning. We identify three kinds of inter-dependencies among features, i.e. feature constraints:

- *Require* constraint: The presence of some feature in a concept instance requires the presence of some other feature.
- *Exclude* constraint: The presence of some feature excludes the presence of some other feature.
- *Cardinality* constraint: The cardinality constraint represents the quantity relation between features.

An instance of a feature model consists of an actual choice of features matching the constraints imposed by the diagram.

Definition 1. *Given a feature model, if its extension is not empty, i.e. there exists an instance satisfying the feature diagram and feature constraints, the feature model is said to be consistent.*

2.2 Description Logic \mathcal{ALCQI}

In \mathcal{ALCQI} , concepts and roles are built inductively from atomic concepts and atomic roles with constructors. The syntax and semantics of \mathcal{ALCQI} is summarized in Table 1, where A and P denote atomic concepts and atomic roles, C and D denote concepts, R denotes roles, n denotes a strict positive integer. Then we can define \perp as $\neg\top$, $\forall R.C$ as $\neg(\exists R.\neg C)$, and $(\exists^{\leq n} R.C)$ as $\neg(\exists^{\geq n+1} R.C)$. The semantics is specified through the notion of *interpretation* \mathcal{I} that consists of a non-empty set $\Delta^{\mathcal{I}}$ and an interpretation function $\cdot^{\mathcal{I}}$.

Table 1. The syntax and semantics of \mathcal{ALCQI}

Constructor	Syntax	Semantics
universal concept	\top	$\Delta^{\mathcal{I}}$
atomic concept	A	$A^{\mathcal{I}}$
concept negation	$\neg C$	$\Delta^{\mathcal{I}} - C^{\mathcal{I}}$
intersection	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
existential restriction	$\exists R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}}, (x, y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$
qualified number restriction	$\exists^{\geq n} R.C$	$\{x \in \Delta^{\mathcal{I}} \mid \#\{y \mid (x, y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\} \geq n\}$
reverse role	P^-	$\{(o, o') \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid (o', o) \in P^{\mathcal{I}}\}$

A DLs *knowledge base*(KB) \mathcal{K} comprises two components, the TBox and the ABox. TBox (denoted as \mathcal{T}) is a finite set of terminological axioms. A concept C is satisfiable with respect to \mathcal{T} if there is an interpretation \mathcal{I} such that $C^{\mathcal{I}}$ is nonempty. A TBox \mathcal{T} is consistent if there is an interpretation \mathcal{I} such that all concepts are satisfiable. Then the interpretation \mathcal{I} is named a *model* of \mathcal{T} .

3 Formalization of Feature Models with \mathcal{ALCQI}

In order to automatically perform consistency checking on feature models, here we present the formalization of feature models into the \mathcal{ALCQI} KB. Then the consistency checking of feature models will be translated into the consistency reasoning about the \mathcal{ALCQI} KB.

3.1 Translation Rules

Given a feature model FM including a feature diagram FD and feature constraints, the corresponding KB $\mathcal{K}=\varphi(FM)$ can be gained by the following rules:

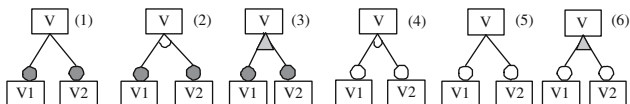


Fig. 1. Basic structures in feature diagrams

- Step 1.** Every node V in FD is formalized into an \mathcal{ALCQI} concept C ;
- Step 2.** Suppose node V and its child nodes $V_i(i \geq 1)$ are formalized into \mathcal{ALCQI} concept C and C_i respectively. Then each edge connecting the node V and its child nodes V_i is translated into an \mathcal{ALCQI} role R_i , which represents the relation between the concept C and C_i ;
- Step 3.** The edge decorations can be translated into \mathcal{ALCQI} terminological axioms. For each basic structure¹, as shown in Fig.1, we have:
 - (1) For structure (1), i.e. $mandatory(V_1, V_2)$, denoting V_1 and V_2 are *mandatory* features of V , we use the following terminology axiom to model it:

$$C \sqsubseteq \forall R_1.C_1 \sqcap \forall R_2.C_2$$

- (2) Structure (2), i.e. $alternative(V_1, V_2)$, represents that V_1 and V_2 are *alternative* features of V . We introduce the following terminology axioms:

$$C \sqsubseteq C_1 \sqcup C_2, C_1 \sqsubseteq C, C_2 \sqsubseteq C \sqcap \neg C_1$$

The first axiom expresses the covering of subconcepts, and the latter two axioms ensure that the concept C_1 and C_2 are disjoint.

- (3) Structure (3), i.e. $or(V_1, V_2)$, means that V_1 and V_2 are the *or* features of V . It is obvious that this structure can be expressed by the the combination of structure (1) and (2). Then making use of the rules (1) and (2), the following terminology axioms are introduced for structure (3):

$$C \sqsubseteq C_1 \sqcup C_2 \sqcup (\forall R_1.C_1 \sqcap \forall R_2.C_2)$$

$$C_1 \sqsubseteq C, C_2 \sqsubseteq C \sqcap \neg C_1$$

$$(\forall R_1.C_1 \sqcap \forall R_2.C_2) \sqsubseteq C \sqcap \neg(C_1 \sqcup C_2)$$

¹ Feature diagrams can be normalized into the basic structures. Without losing the universality, we discuss the structures including two features in order to keep the presentation tersely.

- (4) Structure (4), i.e. *mandatory(optional(V_1), optional(V_2))*, denotes parent feature V has two *optional* feature V_1 and V_2 . It means that if V is included in a concept instance, V_1 , V_2 , neither, or both of them are included in the instance. This structure can also be expressed by the combination of structures (1) and (2), i.e.

$$\textit{alternative}(V_1, V_2, \textit{mandatory}(V_1, V_2), \textit{None})$$

Here *None* represents the situation that none of the subfeatures of a parent feature is included in a concept instance. For *None*, we introduce an *ALCQI* atomic concept *Null*. Note that *Null* is not equal with the *ALCQI* concept \perp , since $\perp^I = \emptyset$, while $\textit{Null}^I \neq \emptyset$. Then corresponding axioms can be introduced by making use of the rules (1) and (2).

- (5) For structure (5) and (6), the translation can be similarly accomplished as previous.

Step 4. For feature constraints, we have:

- (1) V_1 *require* V_2 , means that if feature V_1 is included in a concept instance, then V_2 must be included. To formalize this constraint, we need to find out the related information of edges between features. The edges from the nearest common ancestor node² of V_1 and V_2 to V_1 , denoted by $E_{11} \dots E_{1i}$, can be gained by the following algorithm:

```
// temp as variable and parent(temp) as parent node of temp
// CommonV as the nearest common ancestor node of V1 and V2
temp = V1
while(temp != CommonV)
{
  if it is mandatory structure between temp and its sibling node
    then note the directed edge  $E_{1i}$  from parent(temp) to temp
    else temp = parent(temp)
}
```

For feature V_2 , the effective directed edges $E_{21} \dots E_{2j}$ can be gained analogously. Then the following terminology axiom can be introduced:

$$(\forall R_{11} \dots (\forall R_{1i}. C_1)) \sqsubseteq (\forall R_{21} \dots (\forall R_{2j}. C_2)),$$

in which R_{11}, \dots, R_{1i} are atomic roles corresponding to E_{11}, \dots, E_{1i} , and R_{21}, \dots, R_{2j} corresponding to E_{21}, \dots, E_{2j} .

- (2) Constraint V_1 *exclude* V_2 means that if feature V_1 is included in a concept instance, then V_2 must not be included, and vice versa. Then the corresponding terminology axiom is introduced:

$$(\forall R_{11} \dots (\forall R_{1i}. C_1)) \sqcap (\forall R_{21} \dots (\forall R_{2j}. C_2)) \sqsubseteq \perp,$$

declaring that no concept instance includes both feature V_1 and V_2 .

² Obviously, there must be a common ancestor node of V_1 and V_2 , because the root is one of their common ancestor.

- (3) The *cardinality* constrains between features can be translated into terminology axioms with qualified number restriction constructors in \mathcal{ALCQI} .

The translation rules have covered all the elements of feature models. Hence any feature model can be translated into an \mathcal{ALCQI} KB.

The correctness of the translation function φ can be proven by building two mappings, one from the concept instance of FM to the model of $\varphi(FM)$ and the other from the model of $\varphi(FM)$ to the concept instance of FM . Since the translation rules proposed here are properly intuitionistic, the proof details are omitted here due to space limitation. Readers interested in the validity of φ are referred to [5]. Note that the ABox of the KB $\varphi(FM)$ is empty because no individuals are involved in a feature model FM . Therefore the following discussion on KB reasoning is limited to the TBox of $\varphi(FM)$.

3.2 Automatic Reasoning Using RACER

Given a feature model FM , each instance of FM corresponds to a model of KB $\varphi(FM)$. If FM is consistent, then the set of instances of FM is nonempty, i.e. there must be an instance satisfying FM . Thus there must be a corresponding model of $\varphi(FM)$, so $\varphi(FM)$ is consistent, and vice versa. Hence we have:

Theorem 1. *Given a feature model FM and its corresponding \mathcal{ALCQI} KB $\varphi(FM)$, FM is said to be consistent if and only if $\varphi(FM)$ is consistent.*

Now the consistency checking of a feature model is translated into the consistency reasoning of the corresponding knowledge base.

To accomplish the automatic consistency reasoning, we adopt the description logic reasoner RACER (*Renamed ABox and Concept Expression Reasoner*) [9]. From a KB $\varphi(FM)$, we can define its corresponding RACER script. By adopting RICE [9] as the reasoning client, we can load the script. Then we can use the following command to check the consistency of the knowledge base:

(check - tbox - coherence knowledgebasename (tbox(current - tbox)))

If the result is *NIL*, all the concepts are satisfiable, i.e. the knowledge base is consistent, which demonstrates that the corresponding feature model is consistent. Otherwise, the corresponding feature model is inconsistent.

4 Case Study

To clarify our approach, we present how to translate a feature model of the course concept into an \mathcal{ALCQI} KB. The example feature diagram is shown in Fig.2. The constraints between the features include:

- (1) “*Required*” *exclude* “*TA*” to ensure the teaching quality of the course.
- (2) “*Phd*” *require* “*Prof*” to ensure the profundity of the course.
- (3) Every course can only be taught by one teacher to avoid it to be reopened.
- (4) Every teacher teaches at most one course to ensure the fair assignment of teaching tasks.

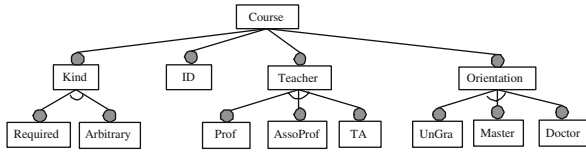


Fig. 2. Feature diagram of Course concept

According to the translation rules proposed in this paper, the above feature model is formalized into the following *ALCQI* KB, as shown in Table 2. Then we define the RACER script, and load it to the RACER knowledge base. Using the consistency checking command, we get the result ‘NIL’, which means that the current feature model is consistent.

Table 2. The fragment of \mathcal{K} corresponding to the feature model of Course

$CCourse \sqsubseteq \forall kind. CKind \sqcap \forall id. CId \sqcap \forall taughtby. CTeacher \sqcap \forall orient. COrient$
$CKind \sqsubseteq CRequired \sqcup CArbitrary$
$CRequired \sqsubseteq CKind \ CArbitrary \sqsubseteq CKind \sqcap \neg CRequired$
.....
$(\forall kind. CRequired) \sqcap (\forall taughtby. CTA) \sqsubseteq \perp$
$\forall orient. CPhd \sqsubseteq \forall taughtby. CProf$
$CCourse \sqsubseteq (\exists^{\leq 1} taughtby. CTeacher)$
$CTeacher \sqsubseteq (\exists^{\leq 1} taughtby^-. CCourse)$

If another constraint “every professor should teach two courses” is added, the corresponding RACER script should be updated by introducing another axiom: (implies $CProf$ (at-least 2 $taught\ CCourse$)). Then the knowledge base will be checked to be no longer consistent. Through modifying the feather constraints, it is feasible to maintain the consistency of the feature model.

5 Related Work

With feature models being more and more widely used in the development of intelligent interactive systems, the formalization of feature models and the consistency reasoning on them have become considerably worthwhile issues. An adapted form of OCL is proposed in [12] to formally describe feature relations in feature models. However, the automated consistency checking of feature models was not further explored. In [13], the first-order logic in Z is used to formalize and verify feature models. Yet the cardinality constraint between features was not identified, which is naturally formalized using DLs in this paper. Benavides et al. present an algorithm to transform an extended feature model into a CSP [3], and further propose to use constraint programming to reason on feature models [4]. However, their feature models still lack the support of feature constraints, which is pointed out in [4] to be one challenge they have to face in the future.

6 Conclusion

In this paper, we have proposed an approach to formalizing feature models with description logics. The consistency checking of feature models can be performed automatically. Moreover, the reasoning is extraordinarily reliable within rigorous logic framework. In fact, through this approach feature models are provided with formal description logics semantics, which will facilitate the development of feature-oriented interactive systems in practice. Now we are investigating how to introduce typical concrete domains, such as numbers and time intervals, into *ALCQI* to express more kinds of feature constraints and exploring the experimental evaluation results about the consistency checking.

References

1. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook*. Cambridge University Press, 2003.
2. S. Bailin. Domain Analysis with KAPTUR. In Tutorial of TRIAda'93, Vol. I, ACM, NewYork, NY, September 1993.
3. D. Benavides, A. Ruiz-Cortés, and P. Trinidad. Coping with automatic reasoning on software product lines. In *Groningen Workshop on Software Variability Management*, pages 1–14, 2004.
4. D. Benavides, P. Trinidad, and A. Ruiz-Cortés. Automated reasoning on feature models. In *CAiSE'05*, LNCS. Springer, pages 491–503, 2005.
5. D. Calvanese, M. Lenzerini, D. Nardi. Unifying class-based representation formalisms. *Journal of Artificial Intelligence Research*, 1999, 11(2):199–240.
6. K. Czarnecki, U. Eisenecker, *Generative Programming: Methods, Tools, And Applications*, Addison-Wesley, 2000.
7. Eui-Chul Jung, et.al. DIF Knowledge Management System: Bridging Viewpoints for Interactive System Design. *Proceedings of 11th Human Computer Interaction International Las Vegas, Nevada USA, July 22-27, 2005*
8. Kang KC, Cohen SG, Hess JA, Novak WE, Peterson AS. Feature-Oriented Domain Analysis (FODA) feasibility study. *Technique Report, CMU/SEI-90-TR-21*.
9. V. Haarslev and R. Mäöler. *RACER Users Guide and Reference Manual*, 2004.
10. Ralf Mäöler. Reasoning about domain knowledge and user actions for interactive systems development. In: *Proceedings IFIP Working Groups 8.1/13.2 Conference, Domain Knowledge for Interactive System Design*, May, 1996.
11. M. Simos, D. Creps, C. Klinger, L. Levine, and D. Allemang. *Organization Domain Modeling (ODM) Guidebook, Version 2.0*. Technical Report for STARS, 1996.
12. D. Streitferdt, M. Riebisch, and K. Philippow. Details of formalized relations in feature models using ocl. In *Proceedings of the 10th International Conference and Workshop on the Engineering of Computer-Based Systems*, pages 297–304, 2003.
13. J. Sun, H. Zhang, Y. Li, and H. Wang. Formal semantics and verification for feature modeling. In *Proceedings of the 10th IEEE International Conference on Engineering of Complex Computer Systems.(ICECCS 2005)*, pages 303–312, 2005.

PNS: Personalized Multi-source News Delivery

Georgios Paliouras¹, Mouzakidis Alexandros¹, Christos Ntoutsis²,
Angelos Alexopoulos³, and Christos Skourlas²

¹ Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece
{paliourg, alexm}@iit.demokritos.gr

² Department of Informatics, Technological Institute of Athens, Greece

³ Department of Informatics and Telecommunications, University of Athens, Greece

Abstract. This paper presents a system that integrates news from multiple sources on the Web and delivers in a personalized fashion to the reader. The presented service integrates automatic information extraction from various news sources and presentation of information according to the user's interests. The system consists of source-specific information extraction programs (wrappers) that extract highlights of news items from the various sources, organize them according to pre-defined news categories and present them to the user through a personal Web-based interface. Dynamic personalization is used based on the user's reading history, as well as the preferences of other similar users. User models are maintained by statistical analysis and machine learning algorithms. Results of an initial user study have confirmed the value of the service and indicated ways in which it should be improved.

Keywords: Personalization, Information Extraction, Machine Learning.

1 Introduction

The rapid increase of interest for the Internet stems from the fact that it is very easy to access and publish information on the Web. At the end of 1993 the community of the World Wide Web numbered about 300,000 users, the majority of which were university researchers and large IT companies. Today it is estimated that there are over 1 billion users on the Internet, most of which cannot be considered computer experts. The increase in the number of users does not only cause an increase of information, but also increases the variety of information that is available (music, news, products, movies, services, etc.) so that more and more users find it useful to surf the Web and contribute to its expansion. This phenomenon is responsible for the "explosion" of the Web, which leads to the "*information overload*" of Web users. Moreover, the large number of Web users and the globalization of the economy led businesses to offer e-services such as e-commerce, e-learning and e-papers. The increased competition and the need for successful services obliged the businesses to add value to e-services in order to create loyal visitors – customers.

To realize this added value and to face the reality of information overload personalization techniques have been developed. Web personalization is simply defined as the process of making Web-based information systems adaptive to the

needs and interests of individual users. Typically this concerns data collection about the users, analysis of these data, and retrieval of the suitable data for the specific user at the suitable time [1].

Users have their own preferences about news content and news sources. Databases of news sources change continuously, making it impossible for users to identify and follow every single news item that is published and could be of interest to them. Therefore, news delivery on the Web is one of the most typical services where personalization can add significant value. Nowadays all large news media (newspapers and news channels) offer in some grade personalization services.

However, the personalization of individual news sources is not a sufficient solution to the problem of information overload, as the users still have to visit many different sites, in order to keep up-to-date with current news. Therefore, an integrated service that aggregates information from various sources and presents it to the users according to their own preferences is very desirable. This is the kind of service by PNS, the system presented in this paper. PNS includes information extraction programs (wrappers), which retrieve continuously highlights of new items that appear at various news sources. This information is organized according to predefined news categories and presented to the users through a Web-based interface. Personalization is achieved with the use of a separate personalization server that provides a variety of services. PNS makes use of four types of adaptive personalization: (a) personal user statistics, (b) stereotype modeling, (c) community modeling, (d) news itemsets. Each of the four types requires the acquisition and maintenance of a different user model, which is achieved with the use of statistical analysis and machine learning methods.

The rest of this paper is structured as follows. Section 2, reviews the state-of-the-art systems for news personalization. Section 3 briefly describes the design and implementation of PNS. Section 4 presents the results of an initial user study, while in the last section conclusions and future directions are discussed.

2 Related Work

A wide variety of both research prototypes and commercial systems offer personalized news on the Web. The goals of these systems are [2]:

- *Personalized news presentation*: The system can provide personalized news, tailored to individual preferences.
- *Personalized advertisements*: The system publishes advertisements targeted to individual groups of people.
- *Effective search capability*: It is possible to search for news items related to a given topic by providing meaning to some keywords.

For successful Web personalization the following three steps are important [1,3]:

1. Gathering of useful information about the user and his interests. The collection of data is performed explicitly, through form-filling, and/or implicitly, through the logging of usage data, possibly combined with legacy data.
2. Creating user models. The collected data are processed and interesting patterns are discovered. Users are clustered and modeled according to their interests.

Non-adaptive user models are predefined and cannot change even if the user interests have changed. Machine learning methods are used to create adaptive user models that capture changes in the user's interests.

3. News filtering/ranking. News articles to be presented are chosen, together with the order of presentation. When the filtering is based on the content of the articles then it is called content-based filtering. Due to the high cost of data preprocessing and analysis, an alternative (or complementary) personalization technique that can be used is collaborative filtering, where the system groups the users into communities according to common characteristics and reading interests.

Table 1 summarizes some well-known personalized news systems with a small description of the algorithms that are used for personalization.

Table 1. Summary of Web news personalization systems, according to their features

<i>System</i>	<i>Data Collection</i>	<i>User modeling</i>	<i>Filtering</i>
Personal Wall Street Journal, San Francisco Chronicle, Fishwrap [4]	Explicit input, legacy data	Non-adaptive	Content-based
Krakatoa, Anatagonomy [5]	Explicit and implicit input	Adaptive	Content-based, Collaborative
SmartPush [6]	Explicit input	Non-adaptive	Content-based
Newsweeder [7]	Explicit and implicit input	Adaptive	Content-based, Collaborative
Aggrawal and Yu [8]	Explicit and implicit input	Adaptive	Collaborative
WebMate [9]	Explicit and implicit input	Adaptive	Content-based
NewsDude [10]	Explicit input	Adaptive	Content-based
Findory (http://www.findory.com)	Implicit input	Adaptive	Content-based
Google (http://news.google.com), Yahoo (http://dailynews.yahoo.com)	Explicit input	Non-adaptive	Content-based
Newsjunkie [11]	Explicit and implicit input	Adaptive	Content-based

3 Personalized News Service Description

The Personalized News Service (PNS) provides users with personalized access to news items from multiple Web sources. For user modeling, the system makes use of a generic Personalization Server (PServer)¹. In comparison to the existing systems this service has the following characteristics:

1. *Data collection*: The system collects both explicit (optional) user data, through a registration form, and implicit data through usage logging.

¹ PServer has been developed in the Institute of Informatics and Telecommunications of NCSR "Demokritos" and will soon be made available under a BSD-like license.

2. *User modeling*: The system is highly adaptive with the use of PServer.
3. *Filtering*: The system uses content-based and collaborative filtering.

The key feature of the proposed system that differs from many existing ones is that it aggregates news highlights from multiple Web sources. News integration is achieved through a combination of source-specific information extraction (wrappers) and RSS input. To our knowledge, PNS is the first system that provides highly adaptive personalization together with multi-source news integration.

3.1 Description of the PNS Content Server

The PNS Content Server is the main server of the system that scans news sources, at regular time intervals. The output of the server is a personalized electronic newspaper consisting of recent article titles that match the interests of the user.

Figure 1 shows the overall architecture of the system, which is made up of three main units: a) Content Scanner, b) Content Selector, c) Content Presenter, as well as the Content Index Database where highlights about the news items and the wrappers are stored. Respecting the copyright of the sources, the server does not store the content of the articles, but simply indexes it, according to its own categorization.

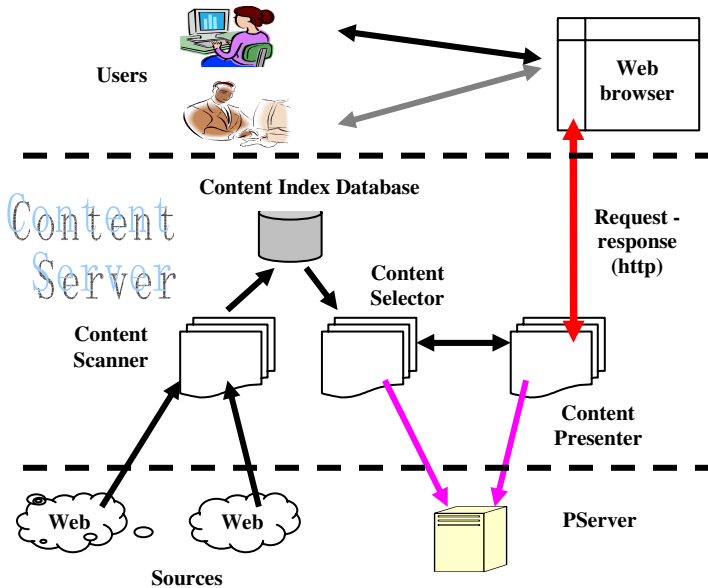


Fig. 1. Overall architecture of the personalized news delivery service

The system collects information about users in two ways:

1. A username and password are specified for logging and memorization purposes. During the registration, the user can also provide personal information, such as age, gender, occupation, etc. for improved personalization.
2. The browsing activity of users is tracked and stored for personalization purposes.

The user's input and choices are forwarded to the PServer from the Presenter. Additionally, the Presenter supplies the registration and identification information to the PServer. This information is used for maintaining the user models. The three component modules of the system are described in more detail below.

3.2 Content Scanner

Content Scanner is an autonomous module that retrieves information about news items from a range of content sources at specified time intervals, using information extraction techniques. This is a typical *web information integration* system that extracts and combines data from multiple web sources. Content Scanner follows the local-as-view approach where, for every information source a specific wrapper is used to extract the desired information. Following this approach, it is simple to add or delete sources and it is also easier to describe constraints on the contents of the sources. Furthermore, the Scanner obtains news through RSS feeds. The extracted information is stored in the Content Index Database. The input for the Content Scanner is a set of news sources associated with a set of wrappers that are used to identify and extract the relevant information.

Figure 2 shows the architecture of Content Scanner. Because of the hierarchical structure of the content sources, the extraction of the relevant news is performed in two levels, thus using two levels of wrappers. The wrapper of the first level (source wrapper) extracts from the main page of a news source (e.g. yahoo), the URL address of the most recent articles. This URL is used by the content wrapper, which extracts the news highlights and stores them in the database. The wrappers are source-specific HTML patterns. For example, in order to extract the titles of articles in a specific source the TITLE tag was used (start pattern: <TITLE>, end pattern: </TITLE>). The wrappers are stored in the Content Index Database and can thus easily be maintained.

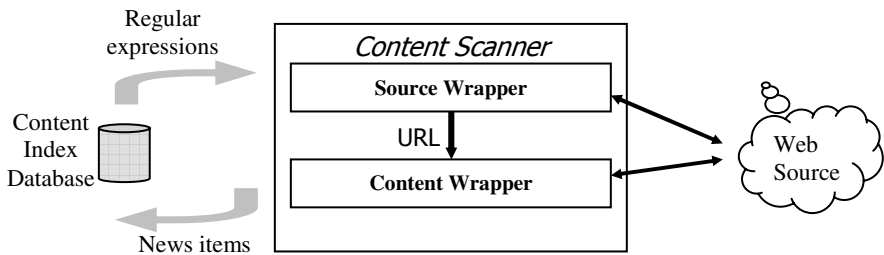


Fig. 2. Architecture of Content Scanner

The Content Scanner can easily be extended so to extract news highlights from additional Web sources. This can be done by defining new wrappers for the additional sources and inserting the wrappers into the Content Index Database.

3.3 Content Selector

The Content Selector chooses the recent content from the Content Index Database to be used for the composition of the personalized newspaper for a particular user. For

the selection both content-based filtering and collaborative filtering are used. In the first case, news is classified along two orthogonal dimensions, the category (e.g. Sports) and the source (e.g. yahoo). For example, a user may prefer to read financial and sports news, while another might be interested specifically in the world news of yahoo. PNS also supports various types of collaborative filtering: (a) according to personal characteristics, optionally provided by them, the users are assigned to a stereotype, the model of which is dynamically maintained based on usage data, (b) users are clustered into communities according to their common preferences alone, (c) news are clustered into news itemsets according to the usage data.

3.4 Content Presenter

The Content Presenter module is the user interface of PNS. All services are available through this unit, which is responsible for the following tasks:

- Registration of new users.
- Identification of registered users.
- Presentation of the daily news that exist in the Content Index Database according to the user's preferences (personal e-paper).
- Presentation of the daily news according to the preferences of users that belong to the same stereotype.
- Presentation of the daily news according to the preferences of users that belong to the same user community.
- Presentation of the daily news that belong to the same news itemset as the currently viewed item.
- Personalized presentation of the news of previous days.
- Text -search and presentation of news titles using keywords.
- Ability to retrieve news highlight in specific dates from the database.

3.5 Personalization Server

The Personalization Server (Pserver) is a general purpose personalization server, using a feature-based representation of user models. Pserver constructs and maintains models for individual users, stereotypes, user communities and feature groups. For PNS the features of users are the sources and categories of news articles. User models are used by the Content Selector to personalize the content that is presented to the users. Each personal user model may contain the following information: (a) personal information about the users, as provided during the registration and (b) sources and categories of news articles with a weight parameter which is based on the frequency at which the user chooses the particular source or category. Stereotypes are similar to personal user models, but they accumulate frequency statistics for all users with the same personal characteristics. User communities are also aggregate models, but they do not contain personal information about the users. Finally feature groups are orthogonal to the user communities in that they aggregate statistics about features, rather than users.

User communities and feature groups are not predefined, but are constructed with the use of machine learning algorithms. Pserver's architecture supports the usage of

alternative machine learning algorithms for that purpose. One such example is Cluster Mining [12], which discovers patterns of common behavior by looking for all fully connected sub-graphs (cliques) of a graph that represents the user's characteristic attributes. It starts by constructing a weighted graph $G(A,E,W_A,W_E)$. In order to construct feature groups, the set of vertices A corresponds to the features used in the user models, and the set of edges E corresponds to feature co-occurrence as observed in the models. For instance, in our application that we examine, if the user reads economy news from BBC and ANT1 an edge is added between the relevant vertices. The weights on the vertices W_A and the edges W_E are computed as aggregate usage statistics. An example graph is shown in Figure 3.

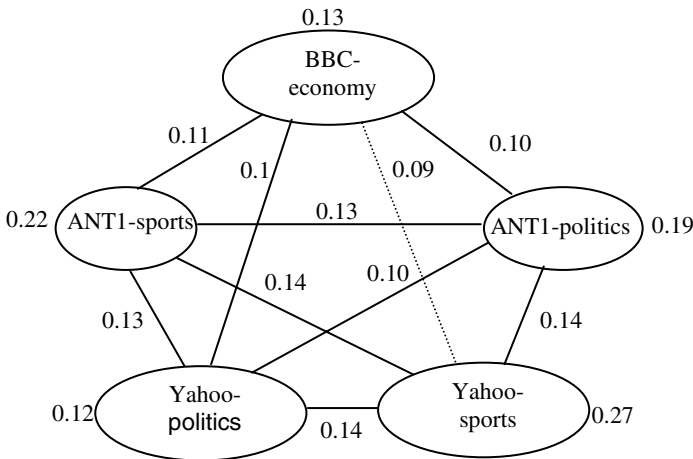


Fig. 3. Feature graph for cluster mining

The connectivity of the graph is usually very high. For this reason we make use of a *connectivity threshold* aiming to reduce the edges of the graph. In our example in Figure 2, if the threshold equals 0.1 the edge ("BBC-economy", "yahoo-sports") is dropped. Cliques are then found on the reduced graph. In addition to the construction of feature groups, this algorithm can be used to construct user communities, by placing users at the vertices of the graph. However, PServer supports the use of any other clustering algorithm for that purpose, too. The administrator of the system should specify the frequency at which communities and feature groups will be updated. The communication between PServer and the Content Server is done by simple HTTP requests and replies.

4 User Study

In order to evaluate PNS, users of different background were asked to test the system for a short period of time. On a daily basis, the system collected the most recent news, which were then presented to the users. The users were asked to fill an electronic

questionnaire with their observations. The role of the user study was to gather information in several different areas, with the general aims of:

- Validating the personalization services.
- Evaluating the functionality of the system.
- Providing input into the design of the system.

Table 2 presents the most important subjects that were evaluated, followed by the results according to the answers supplied by the users. The results obtained from this initial user study confirm the added value provided by the service and point out a number of interesting improvements. The most important subject is the enhancement of the system with more new sources and categories, which will make the system more interesting for the users and the added value of personalization more clear.

Table 2. Summary of the results of the user study

Subject tested	Positive	Partly	Negative
Satisfaction with the order that news highlights are presented	55%	35%	10%
Satisfaction with the news presentation	70%	25%	5%
Satisfaction with the number of news categories	20%	45%	35%
Satisfaction with the number of news sources	10%	50%	40%
Feeling of ease in searching for news	70%	20%	10%
Satisfaction from the interface	70%	30%	0%

5 Conclusions and Future Work

News personalization is an emerging technology that serves both users and businesses. Despite the wide-adoption of personalization by various news sources on the Web, there is still a need for an integrated service that will aggregate information from multiple sources and present the results to the user in a personalized manner. Our service (PNS) uses source-specific information extraction programs to retrieve highlights of news articles and organize the extracted information according to predefined news categories. Using a Personalization Server (PServer), the system provides a personalized view of the collected news items through a Web interface. Dynamic personalization techniques are used for that purpose, analyzing both the user's own interaction with the system, as well as the preferences of similar users. The results of an initial user study have confirmed the added value provided by the service and have pointed to a number of interesting extensions.

The most highly demanded extension is the increase of the coverage of the system with new sources. The extension of the news retrieval module (Content Scanner) is also very important. Frequent changes of the structure of the news sources require the manual update of the wrappers, which is a time-consuming process. Wrapper

induction and verification techniques [13, 14] can be used to automatically update the wrappers.

Another important issue is to create a module that detects sudden changes in news trends and create new topics and categories. For example, during the war in Iraq, CNN dedicated a special page with news from the war. The proposed system could not extract any information about the war, since “war” isn’t a predefined topic for extraction.

Concluding, news aggregation and dynamic personalization from various sources is a promising, highly requested application, which also leads to a variety of interesting research challenges. The service presented in this paper will serve as a base for the development of further innovative applications and services.

Acknowledgements

The presented work is part of a long-term project of the Software and Knowledge Engineering Laboratory at the Institute of Informatics and Telecommunications of NCSR “Demokritos”. Part of this work was done in collaboration with the Department of Informatics of the Technological Institute of Athens, in the context of the research project PA_CO_CLIR (Parallel, COntent Based Cross Language Information Retrieval) that is co-funded by the European Social Fund and National Resources (EPEAEK-II)-ARXIMHDHS.

References

- [1] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, 2003, Web Usage Mining as a Tool for Personalization: A Survey, *User Modeling and User-Adapted Interaction*, v. 13, n. 4, pp. 311-372.
- [2] L. Ardissono, L. Console, and I. Torre 2000, On the application of personalization techniques to news servers on the Web, *Lecture Notes in Computer Science*, Torino, Italy, pp. 1-12.
- [3] A. Kobsa, J. Koenemann, and W. Pohl. 2001, Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *The Knowledge Engineering Review* 16 (2), pp. 111-155.
- [4] P.R. Chesnais, M.J. Muckle, and J.A. Sheena. 1995, The fishwrap personalized news system. In *Proceedings IEEE 2nd Intl Workshop on Community Networking Integrating Multimedia Services to the Home*, Princeton, New Jersey, USA.
- [5] T. Kamba, K. Bharat and M.C. Albers. 1995, The Krakatoa Chronicle - an interactive personalized newspaper on the Web In *Proceedings 4th Intl WWW Conference*, p. 159-170.
- [6] T. Kurki, S. Jokela, R. Sulonen and M. Turpeinen, 1999, Agents in delivering personalized content based on semantic metadata In *Proceedings 1999 AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace*, p. 84-93.
- [7] K. Lang. 1994, Newsweeder: An adaptive multi-user text filter. *Technical Report*. School of Computer Science, Carnegie Mellon University.
- [8] C.C. Aggarwal and P.S. Yu, 2002, An Automated System for Web Portal Personalization, *Technical Report, IBM T. J. Watson Research Center Yorktown, USA*.

- [9] L. Chen and K. Sycara, 1998, WebMate: A personal agent for browsing and searching. *In Proceedings of the Second International Conference on Autonomous Agents, Minneapolis*, p. 132-139.
- [10] D. Billsus, and M. J. Pazzani, 1999, A Hybrid User Model for News Classification. In *Kay J. (ed.), UM99 User Modeling - Proceedings of the Seventh International Conference*, pp. 99-108. Springer-Verlag, Wien, New York, USA.
- [11] E. Gabrilovich, S. Dumais, and E. Horvitz, 2004, Newsjunkie:Providing Personalized Newsfeeds via Analysis of Information Novelty, *In Proceedings of the 13th international conference on World Wide Web*, New York,USA, pp.482-490
- [12] G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C.D. Spyropoulos, 2000, Clustering the Users of Large Web Sites into Communities, *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 719-726, Stanford, California.
- [13] N. Kushmerick, 2000, Wrapper Verification, *World Wide Web J.* **3**(2), pp.79-94, Special issue on Web Data Management.
- [14] N. Kushmerick, 1997, Wrapper induction for information extraction, *PhD Thesis*, University of Washington.

Relational Association Mining Based on Structural Analysis of Saturation Clauses

Nobuhiro Inuzuka, Jun-ichi Motoyama, and Tomofumi Nakano

Nagoya Institute of Technology,
Gokiso-cho Showa, Nagoya 466-8555, Japan
inuzuka@nitech.ac.jp,
ha8bu3@phaser.eclom.nitech.ac.jp,
nakano@center.nitech.ac.jp

Abstract. Restricting the form of rules is an important issue of relational association rule mining. The proposing method PIX extracts properties from given examples and to use them to form rules. An property of an instance consists of an addressing part which specifies objects related to the instance and description part which says something among the objects. Extracted properties are used like as an item in market basket database and an APRIORI-like algorithm calculates frequent item sets. The paper describes also an experiment in a sample application.

Keywords: relational data mining, association rules, saturation, property extraction.

1 Introduction

Association rule mining algorithm typically targets data in a single relation table, where each entity is represented in a single tuple. Market basket database is a typical target of association rule mining. It is a set of records of shopping done by super-market customers. A record is a set of items bought by a customer at a time, i.e. a content of a basket. A target of association rule miner is to enumerate frequent combination of items in the records. The frequency is called the support of the combination. Other measures, such as confidence, are also used for evaluation.

Structured data, such as in bio-chemistry, web mining, and natural language processing, however, are not appropriately represented there. Inductive logic programming (ILP) has been proven its effectiveness in such areas as knowledge discovery framework[3,8].

Relational association rule mining is one of promising method and has been intensively studied recently[2,3,5,7]. Warmr[2,3] successfully integrated the level-wise efficient mining algorithm[1] with ILP framework. A topic of Warmr is to restrict clauses to be explored. It gives an effective biases for restriction.

We propose a method, PIX (Property Item Extractor), to extract properties from examples. A property extracted has a restricted clausal form biased by the argument mode. Then we formulate the mining task by treating the property as

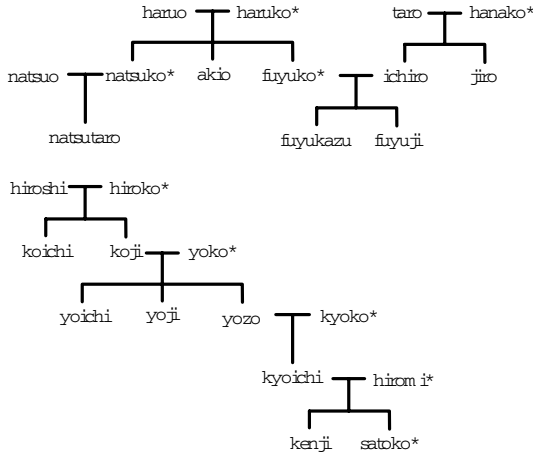


Fig. 1. A family example

items like in market basket databases. The extracted property from the examples can be used with an APRIORI-like algorithm to induce interesting rules.

2 Idea of the Method

Let us think of a simple example about family. It contains three relation **parent**, **male** and **female**. Fig.1 figures the relations, where the name with * means that the name is in female relation and the other names are in male. We consider another relation **grandfather**, which is a unary relation representing one’s grandfather, as a target relation. We target to induce rules describing the target.

When we focus on an instance **grandfather(koji)** we can see the facts,

$$\text{parent}(\text{koji}, \text{yoji}) \wedge \text{male}(\text{yoji}), \text{ and} \tag{1}$$

$$\text{parent}(\text{koji}, \text{yozo}) \wedge \text{parent}(\text{yozo}, \text{kyoichi}) \wedge \text{male}(\text{kyoichi}). \tag{2}$$

These say that **koji** has a son **yoji** and he has a grandson **kyoichi**.

For a person the thing that he/she has a son or has a grandson is his/her property. That is, these are properties which are possibly possessed by some instances of **grandfather**. There are many formulae like Formulae (1) and (2). To restrict formulae to avoid large calculation cost the idea is to treat only properties appeared in some chosen examples.

Formulating properties. Similarly to [3] we use the idea of mode of arguments, where a mode label $+/-$ denotes that the argument is used as input/output, respectively. In the example, we assume **parent**(+, -), **male**(+) and **female**(-). For a relation q we denote the mode of its i -th argument by $\text{mode}_q(i)$. For example, $\text{mode}_{\text{parent}}(1) = +$ and $\text{mode}_{\text{parent}}(2) = -$.

Then **parent** has a role like a function from parent to child, although it is not a deterministic mapping. We refer a relation which includes both $+$ and $-$ labels

a *path* relation[4], which makes paths from a term to others, like *koji* to *yoji*. A relation including only $+$ is called a *check* relation. A literal with a path/check relation is called a path/check literal.

In Formula (1) the part `parent(koji, yoji)` consists of a path literal and `male(yoji)` is a check literal. We can observe the similar thing also in (2). The path parts connects the term *koji* of the instance to another term *yoji* or *yozo* and to *kyoichi*. The check literals describe property for terms derived by the paths. In this example we can observe that a property consists of conjunction which consists of path literals which generate terms connected in some relation and a fact on the term described in a check predicate.

As a general framework we assume the following setting:

1. Several relation tables are given.
2. One of the tables are selected for a target. It means that tuples in the relation table are assumed to be real entities.
3. Primitive properties of the entities can be formulated in the conjunctive form of couple of path literals and a check literal.

We can consider association rules of the primitive properties as items.

3 Extracting Properties from an Example

Preliminaries. This section first gives some notations. E^+/E^- denote sets of all positive/negative examples of a target relation, where examples are ground atoms. B is the set of all tuples (or atomic formulae) in relation tables that we can use for the mining task. B can be regarded as background theory. For an example e and background theory B a clause F whose head is e and satisfies: (1) $e \wedge B \equiv e \wedge F$ and (2) if clause F' satisfies $e \wedge B \equiv e \wedge F'$ then it holds $F' \rightarrow F$, is called a saturation clause of e wrt B . Let S_e denote the set of body literals of the saturation of e wrt a background theory or related relation tables.

Properties. A property of an example e , e.g. (1) and (2) for `grandfather(koji)`, is a conjunction of literals in S_e , the conjunction which connects the term in the example to another terms and ends in a check literal.

For an example $e = p(t_1, \dots, t_m)$ and the set S_e of body literals of its saturation, we define the concept of property as follows:

- For a check literal $c = q(t_1^c, \dots, t_{m_c}^c) \in S_e$, the set $\mathbf{prop} = \{c\}$ is a property candidate with an unsolved term set $U = \{t_1^c, \dots, t_{m_c}^c\} - \{t_1, \dots, t_m\}$ and a solved term set $S = \{t_1, \dots, t_m\}$.
- Let \mathbf{prop} be a property candidate with an unsolved term set U and a solved term set S and $d \in S_e$ be a path literal. Let I_d (O_d) denote the set of terms in d 's input mode arguments (output mode arguments, respectively). When $U \cap O_d \neq \emptyset$, $\mathbf{prop} \cup \{d\}$ is a property candidate with an unsolved term set $(U - O_d) \cup (I_d - S)$ and a solved term set $S \cup O_d$.
- When \mathbf{prop} be a property candidate with an empty unsolved term set, \mathbf{prop} is a property of e .

Table 1. An algorithm PIX, extracting property items from an example

```

PIX(e):
input      e : a positive example  $e = p(t_1, \dots, t_k)$ ;
output    I : the set of property items extracted from e;
           C : the set of corresponding property clauses;

1.  I :=  $\emptyset$ ; C :=  $\emptyset$ ;
2.  Se := the set of body literals of the saturation of e;
3.  Thead :=  $\{t_1, \dots, t_k\}$ ;
4.  C :=  $\{\ell \in S_e \mid \ell \text{ is a check literal}\}$ ; P :=  $\{\ell \in S_e \mid \ell \text{ is a path literal}\}$ ;
5.  For each  $\ell \in C$  do
6.    prop :=  $\{\ell\}$ ;
7.    Unow :=  $\{t_i \mid \ell = q(t_1, \dots, t_k) \wedge m_q(i) = '+'\}$ ;
8.    P' := P;
9.    While  $U_{\text{now}} - T_{\text{head}} \neq \emptyset$  do
10.     ll =  $\{q(t_1, \dots, t_s) \in P' \mid$ 
11.        $\exists i \in \{1, \dots, s\}, \text{mode}_q(i) = '-' \wedge t_i \in U_{\text{now}} - T_{\text{head}}\}$ ;
12.     prop := prop  $\cup ll$ ;
13.     P' := P' - prop;
14.     Unow :=  $\{t_i \mid q(t_1, \dots, t_k) \in ll \wedge m_q(i) = '+'\}$ ;
15.     pclause := prop  $\cup \{e'\}$ , where e' is the atom obtained from e by
16.       replacing its relation name by a new name;
17.     Replace all different terms in pclause by different variables;
18.     item := the head of pclause;
19.     I := I  $\cup \{\text{item}\}$ ; C := C  $\cup \{\text{pclause}\}$ ;
20.  return (C, I);

```

An algorithm to extract all property of an example is shown in Table 1.

For an example grandfather(koji) its saturation is as follows.

```

grandfather(koji) : -male(kenji), female(satoko), parent(kyoichi, satoko),
parent(kyoichi, kenji), male(kyoichi), parent(yozo, kyoichi), male(yoichi),
male(yoji), male(yozo), parent(koji, yoichi), parent(koji, yoji),
parent(koji, yozo), male(koji).

```

From this saturation we can have the following seven properties.

```

{male(koji)}
{parent(koji, yoichi), male(yoichi)}
{parent(koji, yoji), male(yoji)}
{parent(koji, yozo), male(yozo)}
{parent(koji, yozo), parent(yozo, kyoichi), male(kyoichi)}
{parent(koji, yozo), parent(yozo, kyoichi), parent(kyoichi, satoko), female(satoko)}
{parent(koji, yozo), parent(yozo, kyoichi), parent(kyoichi, kenji), male(kenji)}

```

Property items and property item sets. We define a *property item*, a generalised form of a property. The name ‘property item’ is to remind ‘item’ of the setting of association rule mining.

For an example e and one of its property prop , consider a clause $c = ‘e’ \leftarrow \text{prop}$, where $e’$ is an atom with a new relation name and the arguments of e . We define a *property clause* and *property item*. A property clause is obtained from c by replacing all different terms in the clause by different variables. We call the head of the property clause a property item. For a property item item , $\text{item}(e)$ denotes a ground instance of it instantiated by the arguments of e .

Property clauses obtained from the properties of $e = \text{grandfather}(\text{koji})$ is,

$$\text{item1}(A) \leftarrow \text{male}(A). \quad \text{item2}(A) \leftarrow \text{parent}(A, B), \text{male}(B).$$

The relation names item1 and item2 is introduced as a new names. By the notation, $\text{item1}(e) = \text{item1}(\text{koji})$ and $\text{item2}(e) = \text{item2}(\text{koji})$.

When pclause is a property clause of an example e and item is the corresponding property item, it holds $B \cup \{\text{pclause}\} \models \text{item}(e)$. When we say just a *property item*, it is a property item of a certain example. A *property item set* is a set of property items.

For a set I of property items and the set C of corresponding property clauses, we can represent an example as a subset of property items,

$$t_e = \{\text{item} \in I \mid B \cup C \models \text{item}(e)\} \subseteq I.$$

It is the set of all property items that the example possesses. We call the set an *example transaction* by analogy to the market basket database,

The mining task. We regard a set $\{\text{item1}, \dots, \text{item}n\}$ of property items, an associate rule, $R = ‘\text{item1}, \dots, \text{item}n \implies \text{positive}’$. The interestingness is defined by the support and confidence as follows, where C is the set of property clauses that correspond to the property items.

$$\text{support}(R) = \frac{|\{e \in E^+ \cup E^- \mid B \cup C \models \text{item1}(e), \dots, \text{item}n(e)\}|}{|E^+ \cup E^-|} \tag{3}$$

$$\text{confidence}(R) = \frac{|\{e \in E^+ \mid B \cup C \models \text{item1}(e), \dots, \text{item}n(e)\}|}{|\{e \in E^+ \cup E^- \mid B \cup C \models \text{item1}(e), \dots, \text{item}n(e)\}|} \tag{4}$$

In the above $\text{support}(R)$ is a bit different from the standard definition. We used this for the situation where instances are decided to positive and negative. The method we proposed, however, can be used also with the standard support.

Then, we formulate the mining task as follows.

The task of relational rule mining:

- Given E^+ , positive examples and E^- , negative examples,
- B , background theory or related relation tables,
- sup_{\min} , minimum support, and conf_{\min} , minimum confidence,

Enumerate all rules ‘ $I' \implies \text{positive}$ ’

s.t. $\text{support}(I' \implies \text{positive}) \geq \text{sup}_{\min}$ and
 $\text{confidence}(I' \implies \text{positive}) \geq \text{conf}_{\min}$,
 where $I' \subseteq I$ and I is a set of property items.

4 The Mining Algorithm

As we have already seen, property items can be extracted from the saturation of examples. Interesting rules consists of frequent property items in positive examples, and then such property items are likely to be appeared in randomly chosen positive examples.

The algorithm that we propose is outlined as follows.

1. Choose a certain number of positive examples from given example set.
2. Extract property items from the chosen examples, using PIX algorithm.
3. Enumerate all interesting rules consisting of the extracted property items.

From the randomly chosen examples PIX algorithm extracts property items. Then the third step we use an APRIORI-like level wise algorithm.

Only the difference of the algorithm of the third step from APRIORI algorithm is the first level generation of frequent item sets. APRIORI generates all frequent singleton item sets. Our version does not make it contain items taking 100% support. The 100% support means that the property item or clause is likely tautology. The original task of association rule mining is hard to come across such item, which is bought by all customers all transactions, although the logical setting some conjunction is easier to be had and is likely tautology (we can not say it definitely because of it is not derived deductively). The algorithm deletes such items because tautological literals give no information.

An explanatory example. Table 2 shows an example run of the algorithm with the family example. We used background knowledge of Fig. 1. The target is *grandfather* but we assume that the user does not understand the meaning of it. The four positive and four negative examples are given as shown in Table 2.

First it calculate saturation and extracts property items for each of positive examples. The table shows only property items whose supports exceed the minimum support (here, 30%). *item3* has its support 100% and then was deleted. After that an APRIORI-like algorithm calculates all of frequent property item sets in the level-wise way. Five sets of frequent item sets are discovered in this examples. Finally the confidence is calculated and four rules are found.

5 Experiments

This section reports a preliminary experiments with an implementation of the algorithm. The algorithm is implemented with SWI-prolog with a Pentium-4 1.8GHz CPU Unix machine with 512 MB main memory.

Table 2. An example of deriving rules for the family example

Positive examples	
$E^+ = \{\text{grandfather}(\text{haruo}), \text{grandfather}(\text{hiroshi}), \text{grandfather}(\text{yozo}), \text{grandfather}(\text{koji})\}$	
Negative examples	
$E^- = \{\text{grandfather}(\text{haruko}), \text{grandfather}(\text{taro}), \text{grandfather}(\text{yoko}), \text{grandfather}(\text{ichiro})\}$	
All property items whose support is more than 30%:	
item1(A) \leftarrow parent(A, B), parent(B, C), parent(C, D), male(D).	
item2(A) \leftarrow parent(A, B), parent(B, C), male(C).	
item3(A) \leftarrow parent(A, B), male(B). % always true	
item4(A) \leftarrow male(A).	
Frequent item set (minimum support 30%)	
item1(A).	item2(A).
item1(A), item2(A).	item2(A), item4(A).
Found interesting rules (minimum confidence 60%)	
item1(A) \implies positive.	item4(A) \implies positive.
item1(A), item2(A) \implies positive.	item2(A), item4(A) \implies positive.

We conducted an experiment using the data of East-West challenge[6]. Data for 120 trains divided into 57 positive and 63 negative examples is used.

The proposed algorithm induced 32, 223, 807 and 5136 frequent property item sets when the minimum support is 10%, 30%, 50%, and 70%, respectively, when it used all examples to extract property items. Our experiments observed the number frequent item sets induced using a limited number of examples for property extraction. Even when we used limited number of examples for extraction the ratio of the numbers to the number when we used all examples reached 100%. The numbers of examples at which the ratio becomes 100% was 20, 9, 4 and 2 for 10%, 30%, 50% and 70% minimum support, respectively.

6 Conclusions

This paper proposed a method to extract properties from examples to construct association rules in a multi-relational data mining task. The proposed algorithm uses exacted properties from positive examples and construct frequent patterns by APRIORI-like way. By a preliminary experiment, only with relatively small portion of examples gives whole number of properties.

Property items treated are restricted. Each property is extracted and treated individually although two or more property are possibly connected essentially. For example it may extract a property for a person that he/she has a daughter and another one that he/she has a grand-son independently. However the mother of the grand-son may or may not the daughter. We need further study need or needless of an enlargement for such property.

References

1. R. Agrawal, T. Imielinski N. A. Swami : “Mining association rules between sets of items in large database”, Proc. SIGMOD, pp. 207–216. ACM, 1993.
2. L. Dehaspe, L. De Raedt. “Mining association rules with multiple relations”, In: Proceedings of the 7th International Workshop on Inductive Logic Programming, pp.125–132, 1997.
3. L. Dehaspe, H. Toivonen : “Discovery of Relational Association Rules”, in Relational Data Mining, pp. 189–212, Springer-Verlag, 2001.
4. M. Furusawa, N. Inuzuka, H. Seki, H. Itoh : “Induction of Logic Programs with More Than One Recursive Clause by Analysing Saturations”, Proc. 7th International Workshop, ILP-97, pp. 165–172, LNAI 1297, Springer, 1997.
5. B. Goethals and J. Van den Bussche: “Relational association rules: getting warmer”, Proc. ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining, LNCS, 2447, pp.125–139. Springer, 2002.
6. R. S. Michalski, J. B. Larson : “Inductive Inference of VL decision rules”, Workshop in pattern-Directed Inference Systems, Hawaii, SIGART Newsletter, ACM, 63, 38–44, 1977.
7. S. Nijssen, J. N. Kok : “Efficient Frequent Query Discovery in Farmer”, Proc. Principles of Knowledge Discovery and Data Mining 2003 (PKDD2003), 2003.
8. L. De Raedt, H. Blockeel, L. Dehaspe, and W. Van Laer : “Three Companions for Data Mining in First Order Logic”, in Relational Data Mining, pp. 105–139, Springer-Verlag, 2001.

Examination of Effects of Character Size on Accuracy of Writer Recognition by New Local Arc Method

Masahiro Ozaki¹, Yoshinori Adachi¹, and Naohiro Ishii²

¹ Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
ozaki@isc.chubu.ac.jp, adachiy@isc.chubu.ac.jp

² Aichi Institute of Technology, Yakusa-Cho, Toyota, Aichi, Japan 470-0392
ishii@aitech.ac.jp

Abstract. In the previous studies, authors proposed a new local arc method with new similarity evaluation function to the off-line writer recognition, and obtained high recognition ratios. However, the calculated similarity values are comparatively close each other. Therefore, it is very difficult to improve the accuracy any more. In this study, it is assumed that the amount of the characteristic features appeared to the curvature distribution may be different as the difference of the character size. Then, similarity values of five character sizes were compared, and the optimum character size for the new local arc method was examined. As a result, it turned out that the similarity values are greatly influenced by the character size, and the best character size and arc chord length are obtained for the writer recognition by the new local arc method.

1 Introduction

We have studied the off-line writer recognition [1-7] and have proposed the new local arc method for Japanese "hiragana" characters [4-5] and Chinese characters [6, 7]. In the beginning [2-4], two-dimensional (2D) Fuzzy membership functions were used for simplicities, and over 98% recognition ratio was obtained after omitting inadequate characters for recognition. Therefore, the recognition process happen to be abandoned from lack of suitable characters.

To overcome these problems, a characteristic of each stroke was tested in the previous papers [4-7]. Then, curvature distributions of the strokes found to be able to use as a feature of each writer. The new local arc method with the new similarity evaluation function and with the new searching method was proposed in the previous paper [7]. However, the calculated similarities are comparatively close each other. Therefore, it is very difficult to improve the accuracy any more.

In this study, it is assumed that the amount of the characteristic features appeared to the curvature distribution is different by the difference of the character size. Then, five character sizes were examined to propose the optimum character size for the new local arc method.

2 Effect of Character Size in Writer Recognition

2.1 Five Character Sizes

To examine the effect of character size in the writer recognition, five character sizes shown in **Figure 1** were used.

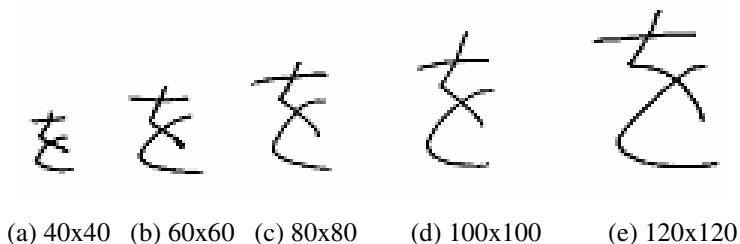


Fig. 1. Five types of character sizes

The distinctions of these character sizes are as follows:

- (a) 40x40: Smaller than the usual character size
- (b) 60x60: Slightly smaller
- (c) 80x80 and (d) 100x100: Ordinary size and used to write in letters of reports
- (e) 120x120: Bigger than the usual

Japanese people usually write in sizes (c) or (d), then his writing habit may appear in these sizes.

2.2 Collection of Sample Characters

Total 1200 characters, 5 (writers) x 8 (types of character) x 5 (character sizes) x 6 (times), were collected. The types of character are Japanese “hiragana” characters, “す(su)”, “と(to)”, “こ(ni)”, “の(no)”, “は(ha)”, “ま(ma)”, “る(ru)”, and “を(wo)”. Five character sizes are shown in the above, 40x40, 60x60, 80x80, 100x100, and 120x120, respectively. Five writers wrote those characters for 6 times for each character.

2.3 Calculation of Similarity Value

As same as the previous works [4-7], similarity values were obtained as follows:

- Step 1: Local arc distributions as 132 dimension vector, 12 (directions from 0 to 180 degrees) x 11 (curvatures from -5 to 5), were calculated for each 1200 characters.
- Step 2: The dictionaries were made from five characters out of 6 characters in each type and size by Principal component analysis. Therefore, 240 dictionaries, 8 (types of characters) x 5 (character sizes) x 6 (times), were created for each writer.

Step 3: Similarity values were obtained from the cosine value of the dictionary vector and the test character vector.

Step 4: The similarity value was obtained as the combination of two or three characters similarity values by Equation (1).

$$\eta = 1 - (1 - \eta_1)(1 - \eta_2)(1 - \eta_3) \quad (1)$$

According as the change of character size, suitable arc chord length must change. Therefore, the similarity values of different chord lengths were calculated through these 4 steps for each chord length. In this study, we tested 6 chord lengths, 5, 9, 13, 17, 21, and 25 (dots).

3 Result and Discussion

In the previous studies, we didn't care about character size. However, there must exist suitable character size that contains enough writing habit for recognition.

3.1 Writer Recognition by One Character

To examine the relation between character size and chord length, writer recognition was carried out by using the similarity values obtained from Step 1 to Step 3 without taking combination. The result is listed in **Table 1**.

Table 1. Writer recognition ratio at 5 character sizes and 6 chord lengths

Chord length	Character size				
	40x40	60x60	80x80	100x100	120x120
5	0.613	0.675	0.713	0.738	0.704
9	0.563	0.633	0.633	0.642	0.592
13	0.596	0.629	0.650	0.683	0.583
17	0.588	0.629	0.654	0.688	0.588
21	0.558	0.583	0.642	0.671	0.542
25	0.533	0.571	0.588	0.663	0.508

In case of 40x40, character size is too small and writing habit cannot appear sufficiently. As the character size increases up to 100x100, the recognition ratio increases. But character size 120x120 is too large to get writing habit properly. Therefore, character sizes 80x80 or 100x100 are recommended.

About chord length dependence, chord length 5 gives best results indicating that we do not need long chord even applying it to a large character. This phenomenon can be explained as follows: If chord length increases, it is possible to explain curvature more precisely. But in this work, we use integer value to express curvature, i.e. -5 to 5. Then, chord length 5 (dots) is enough to express curvature. Furthermore, short chord can be applied more often and can fit every part of strokes. Then, it leads correct classifications.

3.2 Effect of Type of Character on Chord Length

In this study, 8 types of Japanese “hiragana” characters were used, i.e. “す(su)”, “と(to)”, “に(ni)”, “の(no)”, “は(ha)”, “ま(ma)”, “る(ru)”, and “を(wo)”. The shapes of characters are depicted in **Figure 2**.

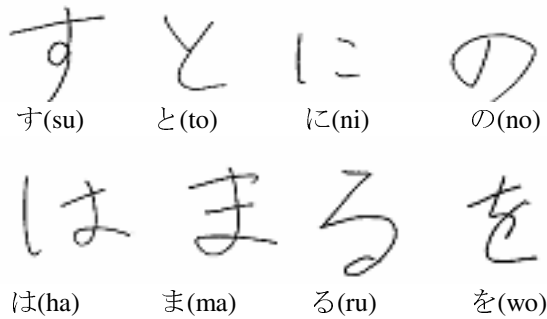


Fig. 2. Eight types of Japanese “hiragana” characters

The curvatures of these characters are totally different as shown in the figure. It is possible to give different recognition ratio. In **Table 2**, the character type dependence of recognition ratio is tabulated.

Table 2. Character type dependence of writer recognition ratio that is the average of those obtained from character sizes 80x80 and 100x100

Character type	Chord length					
	5	9	13	17	21	25
す(su)	0.900	0.750	0.767	0.800	0.817	0.817
と(to)	0.917	0.667	0.783	0.783	0.800	0.817
に(ni)	0.783	0.683	0.583	0.600	0.500	0.433
の(no)	0.550	0.517	0.467	0.433	0.450	0.400
は(ha)	0.817	0.583	0.583	0.567	0.550	0.483
ま(ma)	0.783	0.733	0.783	0.767	0.750	0.733
る(ru)	0.383	0.517	0.667	0.667	0.717	0.750
を(wo)	0.667	0.650	0.700	0.750	0.667	0.567

In cases of す(su), と(to), and は(ha), chord length 5 gives best results, but in case of る(ru), chord length 25 gives best result. In cases of の(no) and ま(ma), roughly same results are obtained for all lengths. It is indicating that there exists suitable chord length for each type of character, however overall speaking; chord length 5 gives reasonable results. Therefore, we recommend chord length 5.

3.3 Writer Dependence of Character Size

As mentioned above, 80x80 or 100x100 is suitable character size to extract writing habit on the whole. However, the sizes of written character are different from one writer to the other. Then, we examined the suitable character size for each writer, and the result is shown in **Table 3**.

Table 3. Writer dependence of character size

Writer No.	Character size				
	40x40	60x60	80x80	100x100	120x120
1	0.646	0.688	0.750	0.667	0.563
2	0.646	0.479	0.542	0.708	0.750
3	0.771	0.854	0.833	0.813	0.750
4	0.625	0.792	0.792	0.896	0.854
5	0.375	0.563	0.646	0.604	0.604

Writer no.2 and 4 usually write fairly large character. Therefore writing habit exists in the larger character. On the other hand, writer no.1 and no.3 write slightly smaller character than the average. Therefore, the recognition ratio is large at the small character region. Overall speaking, character size 80x80 gives good results.

3.4 Combination of Different Characters

As same as the previous work, similarity value is obtained by Equation (1) in Step 4. The results of writer recognition are depicted in **Figure 3**, where character size is 80x80 and arc chord length is 5. The average writer recognition ratio obtained from one character, two characters combination, and three characters combination are 68.8%, 90.6%, and 97.8%, respectively.

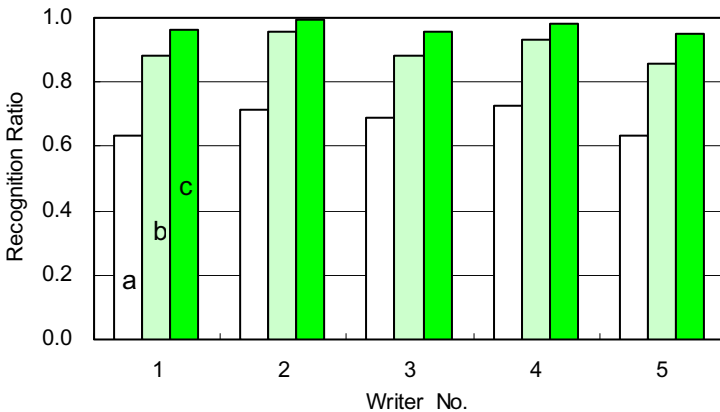


Fig. 3. Recognition ratio obtained from the new similarity evaluation function, Eq. (1). a: Only one character used in Eq. (1). b,c: Combination of two or three characters in Eq. (1), respectively.

4 Conclusion

In this work, the effects of character size and chord length on the writer recognition ratio are examined. Then, the followings were obtained:

- (1) Character size 80x80 is most suitable for writer recognition. Because writer used to write this size of characters and writing habit appear most strongly.
- (2) Chord length 5 gives most accurate results. Because short chord can express a feature of stroke properly.
- (3) Each writer has his own habit and each character has different feature in writing.
- (4) The new local arc method and the new similarity evaluation function with character size 80x80 and arc chord length 5 give 97.8% writer recognition ratio.

In this study, the number of writers and the numbers of characters are too small. Therefore, we have to extend these in the future.

References

- [1] M. Ozaki, Y. Adachi, N. Ishii, and T. Koyazu: Fuzzy CAI System to Improve Hand Writing Skills by Using Sensuous (1996) Trans. of IEICE Vol.J79-D- II NO.9 pp.1554-1561.
- [2] M. Ozaki, Y. Adachi, and N. Ishii: Writer Recognition by means of Fuzzy Similarity Evaluation Function (2000) Proc. KES 2000, pp.287-291.
- [3] M. Ozaki, Y. Adachi, and N. Ishii: Study of Accuracy Dependence of Writer Recognition on Number of Character (2000) Proc. KES 2000, pp.292-296.
- [4] M. Ozaki, Y. Adachi, N. Ishii and M. Yoshimura: Writer Recognition by means of Fuzzy Membership Function and Local Arcs (2001) Proc. KES 2001, pp.414-418.
- [5] M. Ozaki, Y. Adachi and N. Ishii: Development of Hybrid Type Writer Recognition System (2002) Proc. KES 2002, pp.765-769.
- [6] Y. Adachi, M. Liu and M. Ozaki: A New Similarity Evaluation Function for Writer Recognition of Chinese Character (2004) Proc. KES2004, pp.71-76.
- [7] M. Ozaki, Y. Adachi, and N. Ishii: Writer Recognition by Using New Searching Algorithm in New Local Arc Method (2005) Proc. KES2005, pp.775-780.

Study of Features of Problem Group and Prediction of Understanding Level

Yoshinori Adachi, Masahiro Ozaki, and Yuji Iwahori

Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
adachiy@isc.chubu.ac.jp, ozaki@isc.chubu.ac.jp,
iwahori@cs.chubu.ac.jp

Abstract. In the previous works, we have developed Web education system using the dynamically changing three layered learning materials and estimation method of understanding level to maintain the eagerness for self learning by changing learning material based on the understanding level. In those studies, the ratio of correct answers was used directly to estimate understanding level. In this study, we developed an understanding level judging system based on the two parameter logistic function obtained from a cumulative frequency distribution and Fuzzy membership functions based on response time and number of explanation referred.

1 Introduction

Authors have developed Web based education systems [1-6] and also the dynamically changing learning materials according to the intelligibility of the learner, and reported some research results [3,4]. However, in the Web education system, it is necessary to attempt to overcome the difficulty of the continuance of the learning volition in the individual learning. Therefore, it is important to change difficulty level according to the understanding level, and to make the sense of fulfillment by the learning. From the experiments, changing level of learning materials according as learners understanding level was found to be efficient for self learning [4].

In this study, the cumulative frequency distribution of the number of correct answers for each problem group is approximated by the two parameter logistic function, and from the two parameters, easiness of measuring understanding level is judged. Moreover, the specificity of the answer by the type of the problem is measured by C.P values in the S-P table, and the effect of the elimination of the problem that has extremely different tendency is considered. And, after the understanding level was estimated from the 2 parameter logistic function, it is corrected by fuzzy membership functions taking the influence of the terminology reference frequency and the answer time into consideration.

We proposed the equations that estimate the whole understanding level from the understanding level of each problem group. Then, it was confirmed to be able to calculate an approximately correct understanding level by checking with the result of the questionnaire.

2 New Judging Algorithm

a. Elimination by S-P Tables

To develop judging algorithm of understanding level, three kinds of record such as number of correct answers, number of referring explanations, and learning response time were collected. The 100 problems from Fundamental Information Technology Engineer Examinations were prepared. These are classified into 5 problem groups as listed in **Table 1**. Twelve subjects solved these problems and answers were arranged by S-P tables [7]. From the caution index (C.S and C.P), abnormal problems and subjects whose caution index were larger than 0.9 were omitted from further considerations as shown in **Table 2**.

Table 1. Five problem groups

Group	1	2	3	4	5
Area	Hardware	Operating System	System Architecture	System Application	System Development
No. of problems	18	14	14	26	28

Table 2. Number of omitted problems and subjects

Group	1	2	3	4	5
No. of problems	2	1	2	7	10
No. of subjects	1	0	2	1	1

b. 2 parameter logistic function

From each problem group, number of correct answers was arranged in a cumulative frequency distribution function and fitted by the following 2 parameter logistic function [8].

$$U(\theta) = \frac{1}{1 + \exp(-Da(\theta - b))} \quad (1)$$

where a is a discrimination parameter and b is a difficulty parameter like as Item Response Theory [8]. D is constant 1.7 and θ is ratio of correct answer.

In **Figure 1**, the cumulative frequency distributions and fitted 2 parameters logistic functions are shown. The values of two parameters are also listed in it.

From the figure, features of each group can be analyzed as follow:

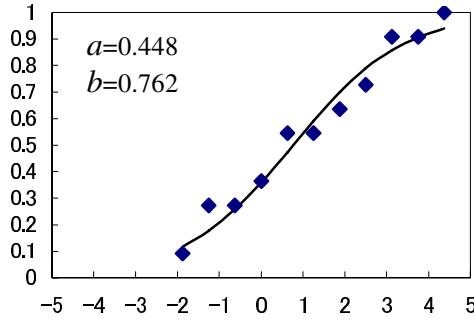
Group 1: Parameter a is smallest in these 5 problem groups indicate that it is not suitable to discriminate ones ability. Parameter b is also smallest indicating that this problem is easiest.

Group 2: Parameter a is largest and parameter b is also large. This problem group is difficult and gives sensitive understanding level.

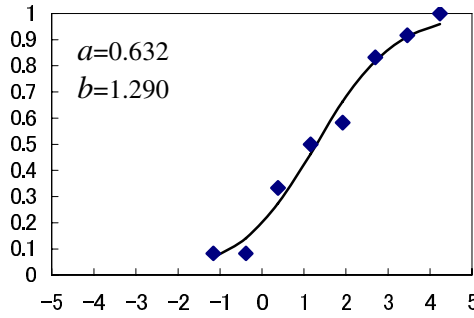
Group 3 and Group 4: These are similar to each other. These are not so difficult and give better discrimination result than Groups 1 and 5.

Group 5: This group is the most difficult and is not suitable to discrimination.

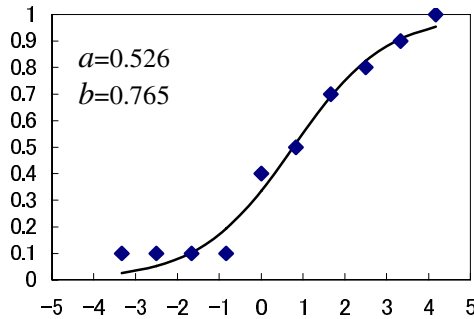
In the figure, x axis is ratio of correct answer. 0 indicates 50% probability. y axis indicates understanding level. 0.5 indicates 50% subjects can get correct answers and his understanding level is middle of the group.



(a) Group 1

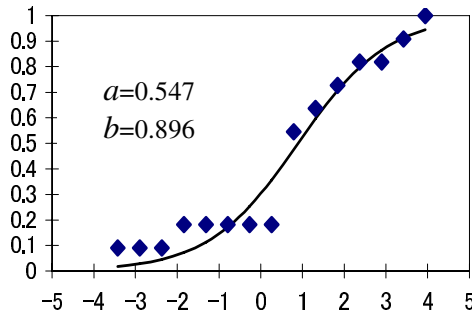


(b) Group 2

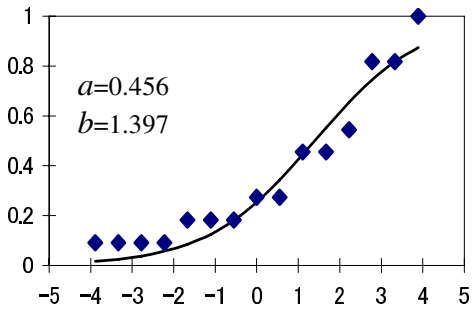


(c) Group 3

Fig. 1. Cumulative frequency distribution and fitted 2 parameter logistic functions



(d) Group 4



(e) Group 5

Fig. 1. (continued)

c. Combination of problem groups

The understanding level of each problem group can be easily obtained from the 2 parameter logistic functions. However, the understanding level of whole area cannot be obtained. Therefore we introduce the following equation to combine these five understanding levels.

$$U = \min \left[\frac{1 - \left\{ (1-u_1)(1-u_2)(1-u_3)(1-u_4)(1-u_5) \right\}^{\frac{1}{5}}}{\left(u_1 u_2 u_3 u_4 u_5 \right)^{\frac{1}{5}}} \right] \tag{2}$$

where u_i is the understanding level of problem group i .

d. Correction by terminology reference frequency and response time

From the terminology reference frequency, the following correction is obtained for each problem.

$$C_R = \begin{cases} (3-R)/3 & \text{if } R < 3 \\ 0 & \text{others} \end{cases} \tag{3}$$

where R is reference frequency.

From the response time relation between correct answer and wrong answer shown in **Figure 2**, we make a fuzzy membership function for the correction of each problem as follows:

$$C_T = \begin{cases} T/T_{AVE} & \text{if } T_{AVE} \leq 30, T \leq T_{AVE} \\ 1 - (T - T_{AVE})/2\sigma & \text{if } 30 < T_{AVE} \leq 50, T \geq T_{AVE} \\ 1 & \text{others} \end{cases} \quad (4)$$

where T is answer time and T_{AVE} is average response time.

These correction values are averaged in whole problems and simply multiply to U in equation 2 and it becomes total understanding level.

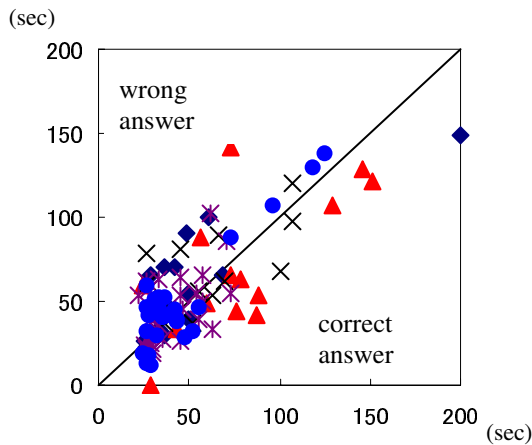


Fig. 2. Response time relation between correct and wrong answers

3 Experimental Results

We applied the above algorithm to 3 extra subjects and estimate understanding level. Furthermore their carrier and results of questionnaire are compared with the estimated

Table 3. Correction of estimated understanding level

Problem group		1	2	3	4	5
No. of correct answers	Subject A	13	8	10	11	13
	Subject B	12	7	7	11	14
	Subject C	10	10	7	10	15
Estimated understanding level (corrected)	Subject A	0.858 (0.811)	0.463 (0.463)	0.909 (0.909)	0.475 (0.465)	0.655 (0.615)
	Subject B	0.790 (0.705)	0.274 (0.274)	0.515 (0.504)	0.475 (0.442)	0.745 (0.661)
	Subject C	0.592 (0.527)	0.819 (0.762)	0.515 (0.511)	0.357 (0.346)	0.818 (0.702)

results in **Table 3**. The total corrected understanding levels of subject A, B, and C are 0.628, 0.491, and 0.549, respectively. From the questionnaire, Subject A is studying Information Technology and Subject B studied it only at school. These phenomena are well agreed with the estimated understanding level as seen in the table.

4 Conclusion

In this work, we proposed a new algorithm to judge understanding level by using 2 parameter logistic functions and fuzzy membership functions. This algorithm is based on three theories, i.e. Item Response Theory, S-P Table Analysis, and Fuzzy theory. By the proposed algorithm, judged understanding level is adequately estimated and well agreed with the questionnaire to the learner.

References

- [1] Adachi Y., Kawasumi K., Ozaki M., and Ishii N.: Development accounting education CAI system (2000) Proc. Intl. Conf. on Knowledge-Based Intelligent Eng. Sys. & Allied Tech., pp.389-392
- [2] Kawada H., Ozaki M., Ejima T., Adachi Y.: Development of the Client/Server System for CAI Education (2002) J. of Nagoya Women's University, No.48, pp.113-120 (in Japanese)
- [3] Takeoka S., Ozaki M., Kawada H., Iwashita K., Ejima T., Adachi Y.: An Experimental CAI System Based on Learner's Understanding (2002) J. of Nagoya Women's University, NO.48, pp.177-186 (in Japanese)
- [4] Ozaki M., Koyama K., Adachi Y. and Ishii N.: Web Type CAI System with Dynamic Text Change from Database by Understanding (2003) Proc. Intl. Conf. on Knowledge-Based Intelligent Eng. Sys., pp.567-572
- [5] Koyama H., Takeoka S., Ozaki M., Adachi Y.: The Development of Authoring System for Teaching Materials – The Dynamically Personalized Hyper Text based on XML – (2003) IEICE Tech. Rep. IEICE Educ. Tech. ET2003-55, pp.23-27 (in Japanese)
- [6] Adachi Y., Takahashi K., Ozaki M., and Iwahori Y.: Development of Judging Method of Understanding Level in Web Learning(2005) Proc. Intl. Conf. on Knowledge-Based Intelligent Eng. Sys., pp.781-786
- [7] Sato T.: “Introduction to Educational Information Technology” (1989) Corona Publishing Co. (in Japanese)
- [8] Toyoda H.: “Item Response Theory –Introduction–” (2002) Asakura Publishing Co.(in Japanese)

Graph-Based Data Model for the Content Representation of Multimedia Data

Teruhisa Hochin¹

Kyoto Institute of Technology, Mastugasaki Goshokaidocho, Sakyo-ku, Kyoto-shi,
Kyoto 606-8585, Japan

Abstract. The contents of multimedia data has complex relationships including deeply nested whole-part and the many-to-many relationships. This paper proposes a data model incorporating the concepts of directed graphs, recursive graphs, and hypergraphs in order to represent the contents of multimedia data. In the proposed data model, an instance is represented with a directed recursive hypergraph called an *instance graph*. The logic-based operation called *rewrite* is introduced. It gives us powerful querying capability because regular expressions on paths can be specified in retrieving instance graphs from collection graphs.

1 Introduction

In recent years, handling multimedia data stored in databases has extensively been investigated. Content retrieval of multimedia data is included in the topics on handling multimedia data. An approach uses graphs representing the contents of multimedia data. Petrakis *et al.*[12] have proposed the representation of the contents of medical images by using directed labeled graphs. Uehara *et al.*[15] have used the semantic network in order to represent the contents of a scene of a video clip. Directed labeled graphs are frequently used in these researches.

The characteristic of recursive graphs, whose nodes may recursively be graphs, is required in representing the contents of multimedia data. Let us consider a picture including a butterfly on a flower. The positional relationship between the butterfly and the flower is described with a graph. One node is for the butterfly, and the other is for the flower. The edge represents the positional relationship between them. Here, the butterfly has six legs, one body, four wings, and so on. The content representation of these parts and their relationships forms a graph. This graph should be included in the node representing the butterfly itself because this graph is the content representation of the butterfly. Therefore, recursiveness is required in representing the contents of multimedia data.

Moreover, the concept of hypergraphs, where a set of nodes is treated as an edge, is also useful in representing the contents of multimedia data. Let us consider a picture where three butterflies are on two flowers. Hypergraphs are very convenient in representing the contents of this picture. If a hypergraph is a directed one, three butterflies are represented as the initial nodes of a directed hypergraph, and two flowers are represented as its terminal ones because a hypergraph could have one or more initial nodes and one or more terminal ones.

If the content of this picture is represented with an ordinary directed graph, the node representing a butterfly is connected to two nodes representing the flowers by two edges. In this case, at least six edges are required to represent these relationships. On the other hand, only one hypergraph can represent these relationships. As the representation of the contents of multimedia data through hypergraphs can be natural as well as simple, introducing hypergraphs to the framework for the multimedia content representation is preferable.

As we have seen, the characteristics of recursive graphs and hypergraphs as well as directed graphs will be preferred in the representation of the contents of multimedia data. The graphs having the characteristics of recursive graphs, hypergraphs, and directed graphs are called *directed recursive hypergraphs* in this paper. If graphs are simply directed ones, they are naturally represented in object-oriented, or graph-based data models[5,11,14,2,13,9]. However, graphs having the characteristics of recursive graphs or hypergraphs cannot naturally be represented in these data models. Hypernode model[8] and Hy⁺[3] incorporate the concept of recursive graphs. In these models, regular expressions on paths can be specified in the retrieval. Recursive queries can also be written. These query facilities give these data models sufficient querying power. However, graphs having the characteristic of hypergraphs cannot naturally be represented. The directed recursive labelnode hypergraph model[1] and the self-structured semantic relationship model[4] incorporate the concepts of hypergraphs as well as recursive graphs. These models may have more expressive power in representing the contents of multimedia data. However, the operations in querying are not sufficient. The operations are mainly for updating the structures of graphs. Recursive queries are not considered. Regular expressions on paths cannot be specified.

In this paper, a graph-based data model is proposed. The proposed data model is called the *Directed Recursive Hypergraph data Model* (DRHM). This model incorporates the concepts of directed graphs, recursive graphs, and hypergraphs. An *instance graph* is the fundamental unit in representing data. A *collection graph* is a graph having instance graphs as its components. A *shape graph* of a collection graph represents the structure of the collection graph. The *rewrite* operation is a logic-based one. Recursive queries can be specified through this operation. It enables users to specify regular expressions on paths. It is used in updating instance graphs as well as querying on a database.

This paper is organized as follows: In Section 2, the proposed data model is informally described by using examples. Some considerations are made in Section 3. Lastly, Section 4 concludes this paper.

2 Directed Recursive Hypergraph Data Model

2.1 Instance Graph

In DRHM, the fundamental unit in representing data is an *instance graph*. An instance graph is a directed recursive hypergraph. An instance graph has a label composed of its identifier, its name, and its data value.

Here, DRHM is described by using a simple example.

Example 1. Consider the representation of the picture shown in Figure 1(a). In this picture, a butterfly is on flowers. Two fore-legs, one middle-leg, and two hind-legs of the butterfly as well as a head, a fore-wing, and a hind-wing appear in the picture. Two fore-legs are on a flower, and two hind-legs are on another flower. Figure 1(b) represents the contents of this picture in DRHM. In Figure 1(b), an instance graph is represented with a round rectangle. For example, $n111$, $n112$, $g1$, $g11$, and $g12$ are instance graphs. An edge is represented with a curve which is consisted of a broken curve and a dotted one. A broken curve surrounds a set of initial elements of an edge. A dotted one surrounds a set of terminal elements of an edge. For example, $n111$ and $n112$ are connected to $n113$ by the edge $e11$. When a set of initial elements of an edge contains only one element, and that of terminal elements also contains only one element, the edge may be represented with an arrow for simplicity. The edge $e16$ in the instance graph $g12$ is an example of this representation. An instance graph may contain instance graphs, and edges. For example, $g1$ contains $g11$, $g12$, $e13$, and $e14$.

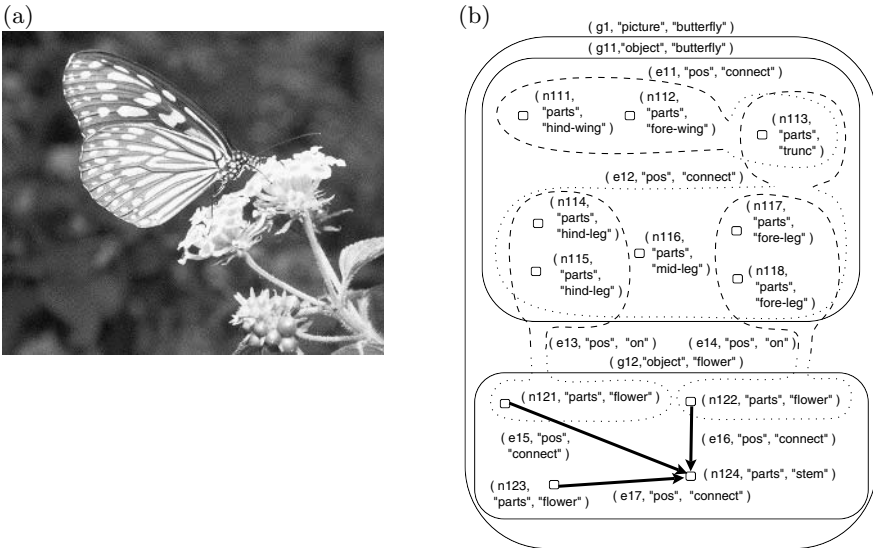


Fig. 1. (a) a picture and (b) an instance graph representing its contents

2.2 Collection Graph

A set of the instance graphs having the similar structure is captured as a *collection graph*. A *collection graph* is a graph whose components are instance graphs.

Example 2. An example of a collection graph is shown in Fig. 2. In this figure, a collection graph is represented with a dashed dotted line. A collection graph has a unique name in a database. The name of the collection graph shown in Fig. 2 is *Picture*. The instance graph $g1$ is the one shown in Fig. 1. The instance

graph *g2* is for another picture. These instance graphs are called *representative instance graphs*.

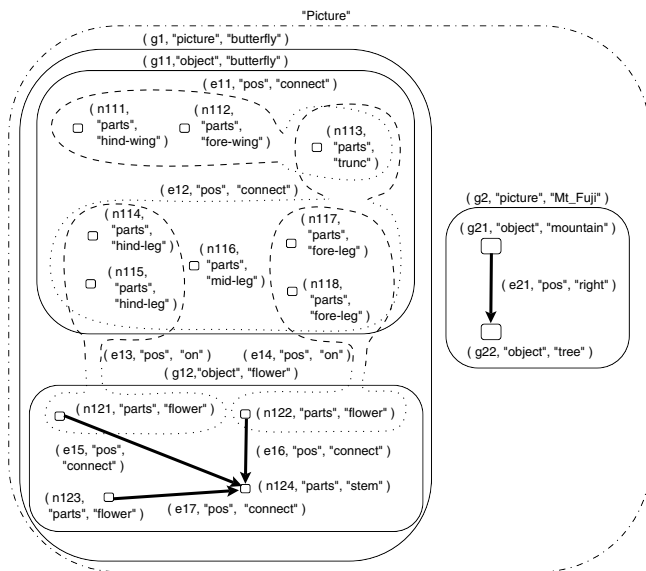


Fig. 2. An example of a collection graph

2.3 Shape Graph

The structure of a collection graph is represented by the graph called a *shape graph*.

Example 3. Figure 3 shows the shape graph for the collection graph **Picture** shown in Fig. 2. This shape graph represents the following structures. An instance graph **picture** includes an instance graph **object**. An instance graph **object** is connected to an instance graph **object** by an edge **pos**. An instance graph **object** contains an instance graph **parts**. An instance graph **parts** is connected to an instance graph **parts** by an edge **pos** inside of an instance graph **object** or outside it.

The shape graph has the nature of the *hard shape*[10]. That is, a shape graph does not have to exist prior to the creation of a collection graph. It may, of course, exist prior to the collection graph creation. A shape graph must exist while a collection graph exists. Inserting an instance graph results in the creation of a shape graph if the shape graph describing the definition of the instance graph does not exist yet. Updating an instance graph may also result in the creation of a shape graph. However, once shape graphs are created, they are not deleted by deleting instance graphs. Shape graphs can be deleted only by the operation deleting shape graphs.

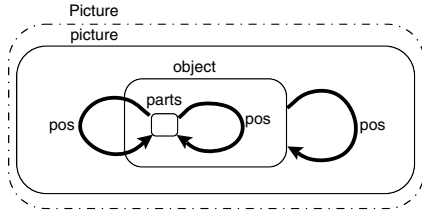


Fig. 3. A shape graph

2.4 Rewrite Operation

Next, we describe how to retrieve instance graphs from a database. For retrieving and modifying instance graphs, the operation *rewrite* is introduced. This operation is a logic-based one, and rewrites collection graphs. In this operation, two kinds of information are specified. One is on the target of the operation. The other is on the destination of the operation. The target and the destination of the operation are collection graphs. In specifying the structure of the target and the destination of the operation, a kind of graph, which is called a *query graph*, is used. The structure of a query graph is similar to that of a shape graph, which represents the structure of a collection graph. The major difference between the query graph and the shape graph is the label of a graph. The label of a query graph is a triple of variables for an identifier, a name, and a value in order to specify the desired instance graphs. The result of the rewrite operation is also a collection graph.

Example 4. An example of the *rewrite* operation is shown in Fig. 4. This is for obtaining the instance graphs, whose names are **picture**, including instance graphs, whose names are **object** and **parts**, and edges connecting them, where an instance graph **parts**, which is included in an instance graph **object**, is connected to an instance graph **parts**, which is also included in an instance graph **object**, by an edge whose name is **on**. The first argument of the rewrite operation is a destination query graph. The second one is a query graph specifying a retrieval condition. The instance graphs, which are in the collection graph **Picture**, satisfying the retrieval condition represented by the query graph become the instance graphs in the collection graph **My-picture**. The labels of query graphs, e.g. **X**, **Z1**, may be used in a retrieval condition. Variables for an identifier, a name, and a value of a label *X* are represented with X_{id} , X_{name} , and X_{val} , respectively, in Fig. 4.

In a query graph, regular expressions can be specified.

Example 5. An example of the *rewrite* operation including a regular expression is shown in Fig. 5. This is for obtaining the instance graphs, whose names are **picture**, including instance graphs, whose names are **object** and **parts**, and edges connecting them, where an instance graph **parts**, which is included in an instance graph **object**, is connected to an instance graph **parts**, which is also

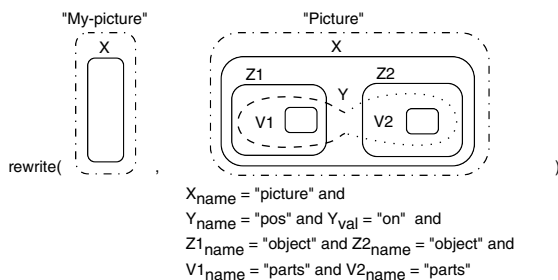


Fig. 4. An example of the rewrite operation

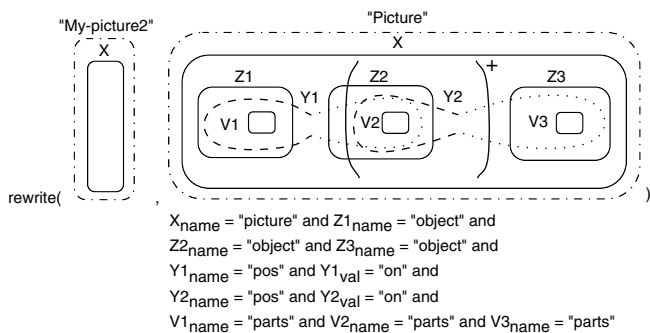


Fig. 5. Specification of the regular expression

included in an instance graph *object*, by two or more edges whose names are *on*. Brackets are used in order to represent regular expressions. The plus mark (+) denotes that the part in brackets may appear one or more times.

As the specification includes diagrams, the most convenient way in specifying query graphs is to follow the graphical query specification[5,11,2,6]. Implementation of this kind of language is one of the interesting issues. Using a generation environment for a visual language[16] may be a way to implement this kind of visual language. Implementing the visual environment for specifying query graphs is out of the scope of this paper.

The rewrite operation can also be used for inserting instance graphs into a collection graph. The rewrite operation only having a destination query graph means insertion. When an instance graph is inserted into a collection graph, identifiers of the instance graphs and the edges are assigned by the system.

If a variable for an identifier is used in a destination query graph in order to specify a specific instance graph or edge, this means modification of the instance graph or edge.

Formal definition of DRHM is omitted in this paper because of the limitation of pages. Please refer the manuscript[7] for the formal definition.

3 Consideration

Directed recursive hypergraphs can be simulated with recursive graphs and ordinary directed edges by using the fact that nodes can be included in another node in recursive graphs. That is, a directed recursive hypergraph, whose initial and terminal elements of a hyperedge are sets of nodes, can be simulated by connecting a node including the nodes in an initial set of the hypergraph to a node including the nodes in a terminal set of the hypergraph through an ordinary directed edge. In this case, users always have to decide whether an initial or terminal element of an edge is a node or a set of nodes.

Let us consider the head of a person. As a person ordinarily has only one head, a person is modeled to have a head. For example, a node for a person and a node for a head are defined, and these nodes are connected by an edge whose name is **have**. When a person having two heads is found, this definition must be changed in order that a person could have two heads. To this end, a node for heads, which is referred to as a **heads** node, will be introduced to contain **head** nodes. An edge whose name is **include** will also be introduced to connect these two kinds of nodes. The **have** edge, which connected the **person** node to the **head** node, will be changed to connect the **person** node to the **heads** node. In this case, the instances of the **have** edge must be changed according to the change of the definition. Another way is to introduce another **have** edge, which connects a **person** node to a **heads** node. Anyway, it is not preferred that the change of the number of instances causes the change of the definition of data because the change of the definition will result in the change of code of multimedia application program. Another method of using recursive graphs and ordinary directed edges is to catch everything as a set of elements. Following this method, two kinds of nodes and one kind of edge must always be defined. For example, a node for a head and a node for heads are defined, and an edge whose name is **include** is defined to connect these kinds of nodes. It is cumbersome that two kinds of nodes and one kind of edge are always defined in order to treat one thing.

The number of elements is treated in the framework of cardinality constraints. A cardinality constraint restricts the mapping from an element to one or more elements. It is carefully designed that a mapping is one-to-one, one-to-many, or many-to-many in the traditional information system, e.g. management of company information. The cardinality may not be so important in the modeling of the contents of multimedia data. Multimedia application system may not pay attention to the cardinality. For multimedia application system, one element merely means that the number of elements happens to be one. One or more elements are preferred to be captured as one thing in multimedia application system. As the result, adopting the concept of hypergraphs will bring us natural and simple representation of the contents of multimedia data. Therefore, the characteristics of recursive graphs and hypergraphs as well as directed graphs will be required in the representation of the contents of multimedia data.

4 Concluding Remarks

This paper proposed a graph-based data model. The concepts of directed graphs, recursive graphs, and hypergraphs are introduced to the proposed data model. An instance graph is the fundamental unit in representing data. A collection graph is a graph whose components are instance graphs. A shape graph of a collection graph represents the structure of the collection graph. The *rewrite* operation used for retrieving and updating instance graphs is a logic-based one. Introducing regular query graphs enables us to specify recursive queries. The depth of an edge is introduced to show whether an instance graph can be decomposed or not. Shape graphs can be managed by representing them with instance graphs.

The storage structure for a database based on the proposed data model and the efficient query processing are other subjects of future work.

References

1. Boley, H., Directed Recursive Labelnode Hypergraphs: A New Representation-Language, *Artificial Intelligence*, **9** (1977) 49–85.
2. Consens, M. P., and Mendelzon, A. O., GraphLog: a Visual Formalism for Real Life Recursion, *Proc. of 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'90)* (1990) 404–416.
3. Consens, M. P., and Mendelzon, A. O., Hy⁺: A Hypergraph-based Query and Visualization System, *Proc. of 1993 ACM SIGMOD Int'l Conference on Management of Data* (1993) 511–516.
4. Fujiwara, Y. and Gotoda, H., Representation Model for relativity of Concepts, *Int'l Forum on Information and Documentation*, **20**(1) (1995) 22–30.
5. Gyssens, M., *et.al.*, A Graph-Oriented Object Database Model, *IEEE Trans. on Know. and Data Eng.*, **6** (1994) 572–586.
6. Hochin, T., DUO: Graph-based Database Graphical Query Expression, *Proc. of 2nd Far-East Workshop on Future Database Systems* (1992) 286–295.
7. Hochin, T., *et. al.*, A Directed Recursive Hypergraph Data Model for Representing the Contents of Multimedia Data, *Mem. Fac. Eng. Fukui Univ.*, **48** (2000) 343–360.
8. Levene, M. and Loizou, G., A Graph-Based Data Model and its Ramification, *IEEE Trans. on Know. and Data Eng.*, **7** (1995) 809–823.
9. Lucarella, D. and Zanzi, A., A Graph-Oriented Data Model, *Proc. of 7th Int'l Workshop on Database and Expert Syst. Applications (DEXA'96)* (1996) 197–206.
10. Nakata, M., Hochin, T., and Tsuji, T., Bottom-up Scientific Databases Based on Sets and Their Top-down Usage, *Proc. of Int'l Database Engineering & Applications Symposium* (1997) 171–179.
11. Paredaens, J., Peelman, P. and Tanca, L., G-Log: A Graph-Based Query Language, *IEEE Trans. on Know. and Data Eng.*, **7** (1995) 436–453.
12. Petrakis, E. G. M., and Faloutsos, C., Similarity Searching in Medical Image Databases, *IEEE Trans. on Know. and Data Eng.*, **9** (1997) 435–447.
13. Rosenberg, A. L., Addressable Data Graphs, *J. ACM*, **19** (1972) 309–340.
14. Su, S. Y. W., Guo, M., and Lam, H., Association Algebra: A Mathematical Foundation for Object-Oriented Databases, *IEEE Trans. on Know. and Data Engineering*, **5** (1994) 775–798.

15. Uehara, K., Oe, M., and Maehara, K., Knowledge Representation, Concept Acquisition and Retrieval of Video Data, Proc. of Int'l Symposium on Cooperative Database Systems for Advanced Applications (1996) 218–225.
16. Zhang, K., Zhang, D.-Q., and Cao, J., Design, Construction, and Application of a Generic Visual Language Generation Environment, IEEE Trans. on Software Engineering, **27**(4) (2001) 289–307.

NCO-Tree: A Spatio-temporal Access Method for Segment-Based Tracking of Moving Objects

Yuelong Zhu, Xiang Ren, and Jun Feng

College of Computer & Information Engineering, Hohai University
No.1 Xikang Road, Nanjing, Jiangsu 210098 China
ylzhu@hhu.edu.cn, rxdf007@hotmail.com, fengjun@hhu.edu.cn

Abstract. With the continued advances in wireless communications and geo-positioning, an infrastructure is emerging that enables location-based services which rely on the tracking of the continuously changing positions of entire populations of service users, termed moving objects. The main interest of these services is to efficiently store and query the positions of moving objects. To achieve this goal, index structures are required. In this paper we propose a new index structure for moving objects in networks: NCO-Tree. It efficiently supports the Segment-Based tracking approaches and its optimization. We give the structure description, insertion and search algorithms, then evaluate it with experiment.

1 Introduction

Location-based service (LBS) is a service that provides location-based information to mobile users. The main idea is to provide the user with a service that is dependent on positional information associated with the user [1] (such as tell the user his current location, traffic jams on his way home, or even the best way to some place according to the current traffic instance). These services need tracking technique, which continuously monitor the current position and anticipate future position of a population of moving objects. A typical scenario is given in [2]. It assumes that moving objects are constrained by a road network and they are capable of obtaining their positions from an associated GPS receiver. Moving objects, also termed clients, send their location information to a central database, also termed the server, via a wireless communication network. After each update from a moving object, the database informs the moving object of the representation it will use for the object's position. The moving object is then always aware of where the server thinks it is located (It is decided by tracking approach). The moving object issues an update when the predicted position deviates by some threshold from the real position obtained from the GPS receiver.

There are three existing tracking approaches discussed in [2], They are point-based tracking, vector-based Tracking, and segment-based tracking. They are the same in monitoring but differ in how they predict the future positions of a moving object. They are compared in [3], and as a result the segment-based tracking prove to be the best choice for tracking technique.

How to represent and index moving objects' positions in database, so as to reduce update frequency is a challenge, because less update means less communications. This paper starts from the existing tracking approaches, discusses access method for each approach, and proposes a new index structure for the segment-based tracking approach: Network Constrained Object Tree (NCO-Tree). NCO-Tree efficiently stores and retrieves objects moving in networks. It manages the predicted positions of the moving objects with segment-based tracking approach and is capable of answering queries about the current distribution of the moving objects. We give the update and query algorithms for this structure.

The remainder of the paper is structured as follows: Section 2 introduces three existing tracking approaches, and covers improvements of using road-network modifications described in [3] for segment-based tracking. Section 3 proposes the NCO-Tree Structure. Section 4 gives the update and query algorithm of the NCO-Tree. Section 5 gives the experiment and evaluation. Section 6 concludes the paper and proposes some future work.

2 Tracking Approaches

There are three tracking approaches according to [2]. They are point-based tracking, vector-based tracking, and segment-based tracking.

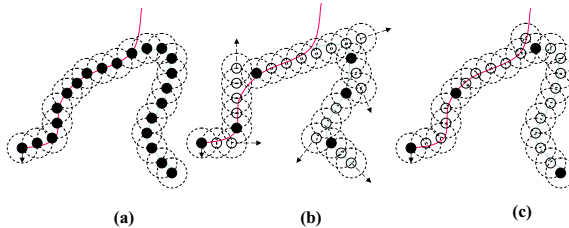


Fig. 1. Tracking approach: (a) point-based (b) vector-based (c) segment-based policy [2]

2.1 Point-Based Tracking

Point-based tracking approach predicts all the objects to be stock-still. Using this approach, the server represents a moving object's future positions as the most recently reported position. A moving object issues an update when its distance to the previously reported position deviates from its current GPS position by the specified threshold. An example of point tracking is presented in Fig.1(a). Here, the circles indicate the threshold and (solid) points indicate predicted positions that result from an update being issued by the object. The two bold lines (distinguished by color) indicate connected segments of the road network.

This tracking approach only stores moving objects' most recently reported positions without their velocities. For effectively access data describing using this approach, a static Spatio-Temporal index structure would be ok. For example:

the R-Tree [4] or the R*-Tree [5]. They are the same in structure but different in update algorithm.

2.2 Vector-Based Tracking

In vector-based tracking, the future position of a moving object is given by a linear function of time (i.e., by a start position and a velocity vector). A moving object issues an update when its distance to the computed position (the GPS receiver computes the position of moving object with the linear function) deviates from its current GPS position by the specified threshold. Fig.1(b) shows the velocity vectors that are used for prediction. Solid points indicate predicted positions that result from updates, while the remaining positions are simply predicted.

Vector-based tracking approach predicts objects moving along tracks described with linear functions of time. This character adapts to the Time-Parameterized R-Tree (in short: TPR-Tree) which proposed in [6], or its optimized structure [7] [8].

The TPR-Tree is a R*-tree based indexing structure that supports efficient querying of the current and future positions of moving objects. It describes that an object's position at some time t is given by $x(t) = (x_1(t), x_2(t), \dots, x_d(t)) = x(t_{ref}) + v(t - t_{ref})$. Here, d is the dimension and $x_1(t), x_2(t), \dots, x_d(t)$ describe the object's position projection on them. t_{ref} is a reference time. It is the update time. $x(t_{ref})$ and v stand for the position and velocity of the moving object at t_{ref} respectively.

2.3 Segment-Based Tracking

The segment-based tracking considers that objects' movement should be constrained by road network. In segment-based tracking, the future positions of a client are given by a movement at constant speed along the identified segment, which is represented as a ployline. The speed used is that of most recently reported by the client. When the predicted position reaches the end of its segment, the predicted position remains at the end from then on. An example of segment-based tracking is shown in Fig.1(c).

Compared with the above tracking approaches, segment-based tracking is sensitive to the fidelity of the road network representation used. In [3], they improve the segment-based approach with modification of the road network: connect segments with General Segment Connection (in short: GSG) algorithm and call them routes, then anticipate objects move along the routes. Such modification greatly reduces the update frequency and proved to be doable.

3 NCO-Tree

Fig. 2 depicts the segment model and the route model of road networks. In segment model, there are seven segments, signed: $E_{12}, E_{23}, E_{24}, E_{46}, E_{67}, E_{56}, E_{68}$, and the same road network could be described as four routes in route model,

signed: R_1, R_2, R_3, R_4 . Paper [3] describes segments connection algorithm seriously. Here, we consider objects moving along the routes instead of the segments and call this approach as route-based tracking.

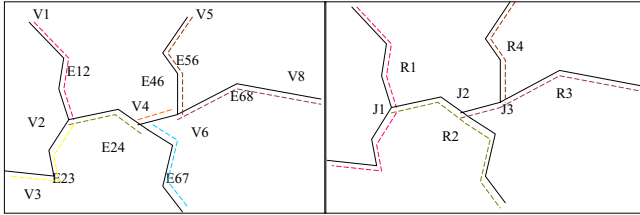


Fig. 2. Segment model and route model

In order to index objects tracked by using route-based tracking approach, we propose a new index structure: the Network Constrained Object Tree (in short: NCO-Tree) to efficiently store and retrieve objects moving in networks. The index structure consists of two parts: one is a R-Tree at the top, which indexes road network modelled by routes, another is a set of 1D TPR-Trees at the bottom which index objects moving along the routes. A hash structure is proposed in the top level which contains the top R-Tree’s routes identification and their pointer to the bottom TPR -Trees. Fig.3 depicts this structure.

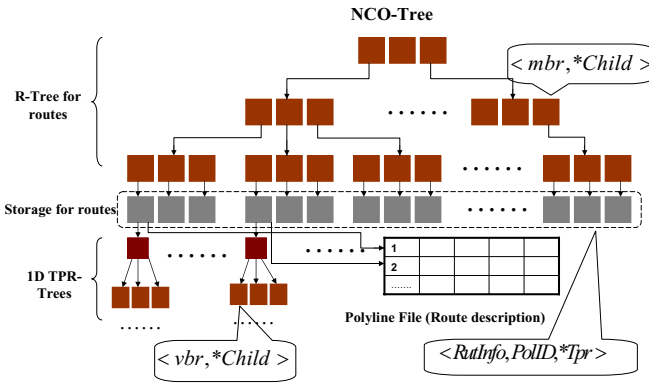


Fig. 3. Structure of NCO-Tree

In the top R-Tree, the routes are indexed by using MBR approximation. The node of this tree contains the pair $\langle mbr, *Child \rangle$, where mbr is the MBR of the node (we use $\langle x_0, y_0, x_1, y_1 \rangle$ to describe the MBR where $\langle x_0, y_0 \rangle$, $\langle x_1, y_1 \rangle$ is the bottom-left and the top-right point respectively), and $*Child$ is a pointer to the child node or the mid-cell (leaf node).

In order to connect the top R-Tree with real routes and the bottom 1D TPR-Trees, we introduce the mid-cell called: storage for routes. It contains the triple

$\langle RoutInfo, PolID, *Tpr \rangle$ where *RoutInfo* describe the information of the route (such as the length, the start point and the end point), *PolID* is ID of the representation of the route (a ployline), and **Tpr* points to the corresponding bottom 1D TPR-Tree.

The hash structure contain the pair $\langle PolID, *Tpr \rangle$, where **Tpr* is a pointer to the corresponding bottom 1D TPR-Tree. The hash structure is organized by *PolID*. Hence, we have two top level index structures: an R-Tree and a hash structure pointing to bottom level TPR-Trees. These two top level index structures are used as follows: the insertion algorithm for moving objects takes *PolID* as an argument, and uses the top level hash structure to find the bottom 1D TPR-Tree (we believe that the moving object tell the tracking system its position with a calculated *PolID* by a matching algorithm), and the insertion algorithm for route just inserts *PolID* and a pointer with Null value to the hash structure.

A set of 1D TPR-Trees (the bottom TPR-Trees) index moving objects which move along the routes. Here, one 1D TPR-Tree index objects moving on one and the same route. In order to describe this relationship, we need to modify the 1D TPR-Tree structure.

We know that in TPR-Tree, moving objects are described by its coordinates and velocities. For example, $\langle X, V \rangle$ describe an object moving in x-coordinate, *X* is x-coordinate value, and *V* is its velocity vector. Here we believe that objects moving along the route could be described by similar expression: $\langle Pos, V_{pos} \rangle$ ($Pos \in [0, 1]$), where *Pos* describe the position of the the object (here, 0 is the begin and 1 stands for the end of the route), and $V_{pos} = V/Len$ describe the velocity of the object which moving on the route (here, *Len* is the length of the route). Following the representation of moving objects, we let $t_{ref} = t_l$ and capture a one-dimensional time-parameterized bounding interval:

$$[Pos^+(t), Pos^-(t)] = [Pos^+(t_l) + \vec{V}_{pos}^+(t - t_l), Pos^-(t_l) + \vec{V}_{pos}^-(t - t_l)]$$

as $(Pos^+, Pos^-, \vec{V}_{pos}^+, \vec{V}_{pos}^-)$, where:

$$Pos^+ = Pos^+(t_l) = \min_i(o_i \cdot Pos^+(t_l)); \vec{V}_{pos}^+ = \min_i(o_i \cdot \vec{V}_{pos}^+)$$

$$Pos^- = Pos^-(t_l) = \min_i(o_i \cdot Pos^-(t_l)); \vec{V}_{pos}^- = \min_i(o_i \cdot \vec{V}_{pos}^-)$$

Actually, we call it: Velocity Bounding Rectangle(VBR), and we store $\langle vbr, *Child \rangle$ in the modified 1D TPR-Tree, where **Child* points to the child node or the moving object (leaf node).

4 Insertion and Query Algorithms

4.1 Insertion

There are two kinds of insertions for this index structure: route insertion and moving object insertion. The route insertion is needed to construct the road network. The moving object insertion is needed when an object is created or its motion is out of threshold. It is also necessary to perform a moving object insertion when an object changes from one route to another.

Route Insertion. The algorithm for route insertion is very simple: just insert the route identification with a null pointer in the hash structure. The insertion of the route in the top R-Tree is postponed to the insertion of the first moving object traversing it. In this way, we keep the top R-Tree as small as possible.

Moving Object Insertion. The algorithm for moving object insertion takes as argument the client moving object’s position with a *PolID* calculated by a matching algorithm, the hash structure $\langle PolID, *Tpr \rangle$, the description of the top R-Tree node with $\langle mbr, PolID, *Tpr \rangle$, the description of the moving object with $\langle vbr, *Child \rangle$. The algorithm is described as below:

```

1.  $*Tpr_{obj} \leftarrow \{ *Tpr \mid PolID = PolID_{obj} \wedge *Tpr \in \langle PolID, *Tpr \rangle \}$ 
2. If  $*Tpr = NULL$  then
   {
3.  $R - Tree \leftarrow \langle mbr_{obj}, *Child_{obj} \rangle$ 
4.  $mid - cell \leftarrow \langle RoutInfo_{obj}, PolID_{obj}, *Tpr_{obj} \rangle$ 
5.  $\langle PolID, *Tpr \rangle \leftarrow \langle PolID_{obj}, *Tpr_{obj} \rangle$ 
   }
6.  $TPR - Tree_{obj} \leftarrow \langle vbr_{obj}, *Child_{obj} \rangle$ 

```

4.2 Query

In segment-based tracking scenario, we need to know the distribution of the moving objects. Given a spatial-temporal query window $w = (x_1, x_2, y_1, y_2, t_{now})$, the query is to find objects within the area $r = (x_1, x_2, y_1, y_2)$ for now.

For this window query, the algorithm receives a spatial-temporal query window w and proceeds in three steps. In the first step, a search in the top R-Tree is performed to find the routes’ MBRs that intersect the spatial query window r . Then, in the second step, the route intersects are searched using the real route representation (i.e: polylines), and the result is a set of windows:

$$w' = ((PolID_1, Pos_{11}, Pos_{12}, t_{now}), \dots, (PolID_n, Pos_{n1}, Pos_{n2}, t_{now}))$$

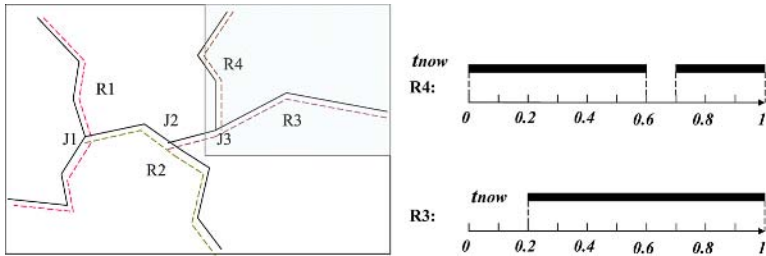


Fig. 4. Query window and result set

where n is the set size. The windows are disjoint and ordered. An example result of this procedure can be seen in Fig.4.

In the last step, we just search in bottom 1D TPR-Trees with $w_i = (PolID_i, Pos_{i1}, Pos_{i2}, t_{now})$ using TPR-Tree search algorithm, and get the result set. The algorithm is described as below:

```

1.  $w = ((PolID_1, Pos_{11}, Pos_{12}, t_{now}), \dots, (PolID_n, Pos_{n1}, Pos_{n2}, t_{now})) \leftarrow R-Tree.Search(w = (x_1, x_2, y_1, y_2))$ 
2.  $Result = \{\}$ 
3.  $For\ i = 1\ to\ do$ 
{
4.  $Result_i \leftarrow TPR_i.Search(w_i = (PolID_1, Pos_{11}, Pos_{12}, t_{now}))$ 
5.  $Result = Result \cup Result_i$ 
}
6. Return  $Result$ 
    
```

5 Experiment and Evaluation

We know that for a tracking system, less update means less communications and better performance. A moving object issues an update (a re-insert actually) when its distance to the computer position deviates from its current GPS position by the specified threshold. Here we compare the TPR-Tree with the NCO-Tree in their update frequency. For TPR-Tree, $x(t)_{computer} = x(t_{ref}) + v(t - t_{ref})$, and update arises when $x(t)_{computer} - x(t)_{real} > D_{lim}$. For NCO-Tree, $Pos(t)_{computer} = Pos(t_{ref}) + V_{pos}(t - t_{ref})$, and update arises when $Pos(t)_{computer} - Pos(t)_{real} > D_{lim}$.

In order to examine the performance of the NCO-Tree, we did an experimental evaluation. We track moving objects with two approaches. In Vector-Based Tracking, we index objects with TPR-Tree. In Route-Based Tracking, we index identical objects with one NCO-Tree. We control these objects moving in the same map. Update arises when the computer position deviates from its current GPS position by the specified threshold. We got the result in Fig.5. It shows that index objects with NCO-Tree could efficiency reduce the update frequency.

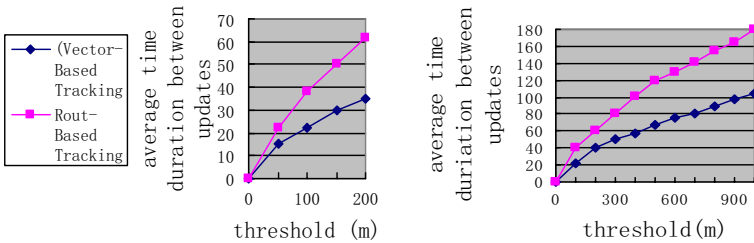


Fig. 5. Update frequency compare between TPR-Tree and NCO-Tree

6 Conclusion

In this paper, we proposed NCO-Tree for segment-tracking of moving objects. It is a mixture of R-tree and TPR-Tree by adding road information to the moving objects. As compared with TPR-Tree, it effectively reduces the frequency of the tracking update. In our future work, predicted search algorithm would be researched so as to sustain predicted query, with proper experiment to prove it.

References

1. J.D. Chung, O.H. Paek, J.W. Lee, and K.H. Ryu. Temporal Pattern Mining of Moving Objects for Location-Based Services. *Proc. Int'Conf. Database and Expert Systems Applications*, pages 331-340, 2002.
2. A. CCivilis, C. S. Jensen, J. Nenortaite, and S. Pakalnis. Efficient Tracking of Moving Objects with Precision Guarantees. *Proc. Int'Conf. Mobile and Ubiquitous Systems: Networking and Services*, pages 164-173,2004.
3. Alminas CCivilis, Christian S. Jensen, Senior Member, and Stardas Pakalnis. Techniques for Efficient Road-Network-Based Tracking of Moving Objects. *IEEE*, VOL.17,NO5,MAY 2005
4. A.Guttman. R-Trees: A Dynamic Structure for Spatial Searching. *Proc. of ACM SIGMOD84*, pages 47-57, 1984.
5. N. Beckmann, H. -P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. *Proc. of ACM SIGMOD90*, pages 322C331, 1990.
6. S. Saltenis, C. S. Jensen, S. T. Leutenegger, and M. A. Lopez. Indexing the Positions of Continuously Moving Objects. *Proc. of ACM SIGMOD'2000*, pages 46-53, 2000.
7. S. Saltenis,C. Jensen. Indexing of Moving Objects for Location-Based Services. *ICDE*, pages 802-813, 2002.
8. Y. F. Tao, D. Papadias, and J. M. Sun. The TPR*-Tree: An Optimized Spatio-temporal Access Method for Predictive Queries. *Proc. of VLDB03*, pages 790C801, 2003.

The Reasoning and Analysis of Spatial Direction Relation Based on Voronoi Diagram

Yongqing Yang, Jun Feng, and Zhijian Wang

Hohai University, Nanjing, Jiangsu 210098 P.R. China
yq_yang2002@hhu.edu.cn

Abstract. This paper studies the direction relations between two spatial objects and presents the reasoning model of spatial direction relation based on Voronoi diagram. The algorithm of the reasoning is described detailedly and an example is analyzed. This paper also discusses the six typical circumstances about spatial direction relation based on Voronoi diagram. These include the direction relations of point-point, point-line, point-region, line-line, line-region and region-region. At last, the advantages and disadvantages about the model are summarized.

1 Introduction

Spatial relations mainly include topology, direction and distance relations. Spatial topological relation is a kind of topological invariant by topological transformation such as adjacency and connectivity relations. Direction relation is also a kind of order relations. It describes a certain order of spatial objects such as forward and back, up and down, left and right, east, south, west, north and so on. Distance relation means the distance between two spatial objects. There are a lot of the research which relate to distance and topology. But the research of direction relation lags behind oppositely. Up to now, the models of describing spatial direction relation mainly have seven categories: Cone-Based model [1], Projection-Based model [1], Double-Cross model [2], 2D String model [3], MBR (Minimal Bounding Rectangle) model [4, 5], Direction Relation Matrix model [6], Voronoi-Based MBR model [7]. Thereinto, Cone-Based model and Projection-Based model belong to experience models. Double-Cross model is on the basis of Projection-Based model. The three models are used for point objects and can't be applied to region objects. The basic idea of 2D String model is based on the method of sign projection. The borders of different 2D spatial objects are projected on X axis and Y axis respectively. Spatial relations are expressed and judged by the character strings which come from the projection and have order relations. Therefore, the higher dimension problems are solved by the model which uses the method with one dimension. It is difficult to assure reliability and maturity and it doesn't express topology relations. It is also difficult to extend 3D space. MBR model and Direction Relation Matrix model describe spatial direction relations intuitively. But for two spatial objects which the diameter ratio is very large, it is not exact to use MBR model and Direction Relation Matrix model to describe spatial direction relations. The basic idea of

Voronoi-Based MBR model is that the four sides of a spatial entity's MBR are regarded as four growth cells, the four growth cells create four sides' Voronoi diagram and then the direction relations are formally described by two entities' Voronoi region and Voronoi boundary. But it also exists the same deficiency as MBR model. Furthermore, it exists very bigger errors for the slope linear objects.

This paper presents a new approach to the problem about spatial direction relation, this approach describes the direction relations between two spatial objects very well. The presented method about spatial direction relation is based on Voronoi diagram.

The rest of the article is structured as follows. In Section 2, the reasoning model of spatial direction relation based on Voronoi diagram is presented and the algorithm about spatial direction relation based on Voronoi diagram is also explained detailedly. Section 3 analyzes the six typical circumstances about spatial direction relation based on Voronoi diagram such as point-point, point-line, point-region, line-line, line-region and region-region. Section 4 is an analysis and evaluation of an example about the algorithm. Finally, conclusions are given in the last section.

2 Algorithm Description of Spatial Direction Relation Based on Voronoi Diagram

The Voronoi region of an entity has a special meaning—the *influence – region* of itself and is defined as the area containing all locations closer to itself than to any other.

[Definition] Voronoi Diagram: suppose having a set of spatial objects, $P = \{P_1, P_2, P_3, \dots, P_n\} \subseteq IR^2$, $\{1 \leq n \leq \infty\}$, P_i may be a point object P_P , or a line object P_L or a region object P_R . A region object is not necessarily convex, and may have holes in which another region may exist. The object Voronoi region of P_i can be defined as $V(P_i) = \{y | Dist(y, P_i) \leq Dist(y, P_j), i \neq j\}$, where $Dist(y, P_i)$ means the minimal distance between point y and P_i , $V(P_i)$ is the Voronoi polygon of spatial object P_i , $V(P) = \{V(P_1), V(P_2), \dots, V(P_n)\}$ is called Voronoi Diagram.

The Voronoi Diagram of point, line and region objects is shown in Figure 1. Indeed, a Voronoi Diagram (Voronoi-Based tessellation) is closer to human perceptions.

Because spatial direction between two objects takes on complexity and variety and there have a lot of directional line segments. A directional line segment means a direction vector which starts from a reference point to a target point. It is impossible to find out each directional line segment one by one. If considering each normal corresponding to each directional line segment, these normal can form Voronoi boundary between two objects (Figure 2), i.e. the curve which E0 denotes. So if knowing the Voronoi diagram between two objects, Voronoi boundary is also known, the direction relation is judged by the relation which direction and Voronoi sides are vertical with each other. Consequently, the key problem

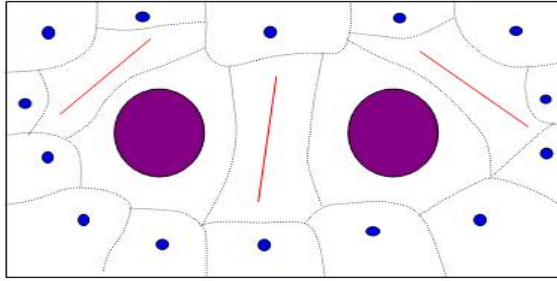


Fig. 1. Voronoi Diagram of Point, Line and Region Objects

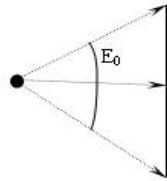


Fig. 2. Relationships Between Directional Line and Voronoi Boundary

consists in the calculation of Voronoi sides' direction. The article presents the following method to calculate the direction of Voronoi sides.

Algorithm: In R^2 , there is a XY plane in the Cartesian system of coordinates, for common Voronoi boundary between two objects A and B, $f(x, y)$ is a curve equation about the Voronoi boundary. x is a gather of horizontal coordinates which corresponding to the common Voronoi boundary between the two objects and y is a gather of vertical coordinates which corresponding to the common Voronoi boundary between the two objects. The curve is divided into n parts, each part is regarded as a linear line segment. One end is selected for start-point, the boundary is represented as a sequence of points $(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. It is shown in Figure 3. Each linear line segment's slope is decided by the following equation:

$$k_i = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}, (x_i \neq x_{i-1}, i = 1, 2, \dots, n) \tag{1}$$

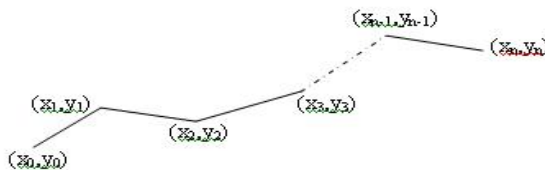


Fig. 3. Partitions of Voronoi Boundary

The obliquity corresponding to the slope is:

$$\alpha_i = \arctan k_i, (i = 1, 2, \dots, n) \tag{2}$$

If $x_i = x_{i-1}$, $\alpha_i = \pm \frac{\pi}{2}$, the average obliquity is:

$$\alpha = \frac{\sum_{i=0}^{n-1} \alpha_i}{n} \tag{3}$$

The average obliquity of Voronoi boundary is:

$$\bar{\alpha} = \begin{cases} \alpha, \pi + \alpha & 0 \leq \alpha < \pi/2 \\ \pi + \alpha, 2\pi + \alpha & -\pi/2 < \alpha < 0 \end{cases}, 0 \leq \bar{\alpha} < 2\pi \tag{4}$$

So the obliquity of direction between two spatial objects A and B is:

$$\beta = \bar{\alpha} \pm \pi/2 \tag{5}$$

If A is a conference object, the obliquity of direction between A and B is $\beta_{AB} = \bar{\alpha} + \pi/2$; On the contrary, if B is a conference object, the obliquity of direction between B and A is $\beta_{BA} = \bar{\alpha} - \pi/2$. In this way, the direction between two spatial objects is judged by the value of β , it is not necessary to make choice of conference object firstly.

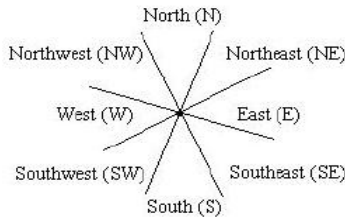


Fig. 4. Eight-Direction Model

According to the azimuth angle which is calculated by the above algorithm, the spatial direction relation between two objects is determined by the comparison between the azimuth angle and the eight-direction model [1](Figure 4).

3 The Reasoning and Analysis of Spatial Direction Relation Based on Voronoi Diagram

In GIS application and research domain, the research of spatial direction relation lags behind oppositely. Spatial direction is defined as using azimuth to describe the position relation between two objects. Spatial objects mainly include point,

line and region on map. The direction relations are discussed by the following six typical circumstances such as point-point, point-line, point-region, line-line, line-region and region-region based on Voronoi diagram.

(1) Point-Point

For the direction relation of two points, the Voronoi border E_1 is a perpendicular bisector which is created by the connecting line between two points. E_1 is a perpendicular bisector in Figure 5. Point x is an arbitrary point within E_1 . The obliquity of direction is decided by the vertical line of the perpendicular bisector. It is also decided by the connecting line of the two points.

(2) Point-Line

E_2 is a Voronoi border between point and line segment in Figure 6. It is also a parabolic curve. Point x is an arbitrary point within E_2 . The obliquity of the direction between point and line segment is determined by the equation (5) in above-mentioned algorithm.

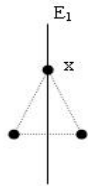


Fig. 5. Voronoi Border Between Two Points

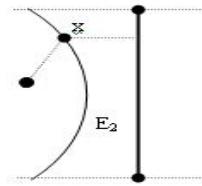
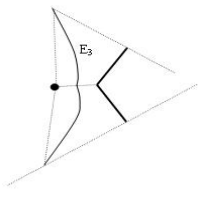


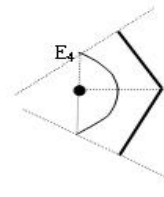
Fig. 6. Voronoi Border Between Point and Line

(3) Point-Region

A region can be regarded as a polygon. The polygon can be divided into two categories such as convex polygon and concave polygon. It is shown in Figure 7a and Figure 7b respectively. E_3 is a Voronoi border between point and convex polygon in Figure 7a and E_4 is also a Voronoi border between point and concave polygon. The obliquity of the direction between point and region is determined by the equation (5) in above-mentioned algorithm.



(a)



(b)

Fig. 7. Voronoi Border Between Point and Region

(4) Line-Line

The relations between two lines consist of three parts in Figure 8. It is displayed in Figure 8a, Figure 8b and Figure 8c respectively. E_5 , E_6 and E_7 are

Voronoi borders corresponding to the relations. Thereinto, E5 is a angle bisector which is formed by the intersection of the extending lines of the two line segments. The beeline sector is E5. E6 and E7 are all angle bisectors. In Figure 8a and Figure 8b, the obliquity of the direction between the two line segments is decided by the equation (5) in above-mentioned algorithm too. But to Figure 8c, the direction between the two line segments is expressed by the two Voronoi borders each other.

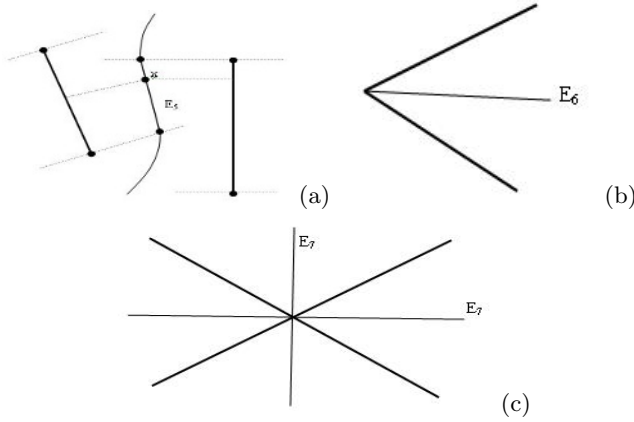


Fig. 8. Voronoi Border Between Two Lines



Fig. 9. Voronoi Border Between line segment and region

(5) Line-Region

It is the same as that in Point-Region, the region is considered as a polygon. There are two circumstances which are illustrated in Figure 9a and Figure 9b respectively. One is the instance of line segment and concave polygon, the other is the instance of line segment and convex polygon. E8 and E9 are Voronoi borders. The spatial direction relation between line segment and region is also determined by the equation (5) in above-mentioned algorithm.

(6) Region-Region

The two regions are treated as two polygons. There are three instances displayed in Figure 10a, Figure 10b and Figure 10c respectively. The two polygons are convex in Figure 10a but concave in Figure 10b. In Figure 10c, one is a convex polygon and the other is a concave polygon. E10, E11 and E12 are Voronoi

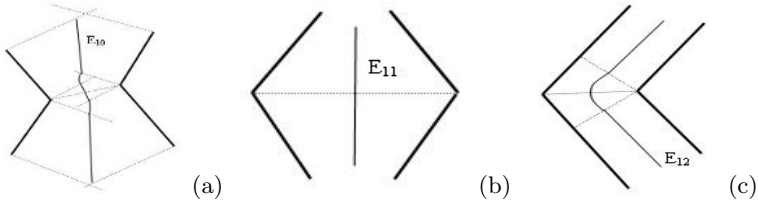


Fig. 10. Voronoi Border Between Two Regions

borders. The direction relation between two regions is all the same determined by the equation (5) in above-mentioned algorithm.

As a result, the key problem is the calculation of the direction of Voronoi border. However, the above-mentioned algorithm is apt to solve the problem.

4 Analysis of an Example

If knowing a equation of Voronoi boundary in R^2 is $y = x^2$, and $0 \leq x \leq 100$. It is a parabola segment. The focus is $(0, 1/4)$. The equation of directrix is $y = -1/4$. A sequence of points will be acquired as follows:

$$(0, 0), (1, 1), (2, 4), \dots, (i, i^2), (i + 1, (i + 1)^2), \dots, (100, 100^2).$$

The focus can be regarded as a point and the directrix can be regarded as a line segment. The line segment is $y = -1/4$ and $0 \leq x \leq 100$. It is shown in Figure 11. Then

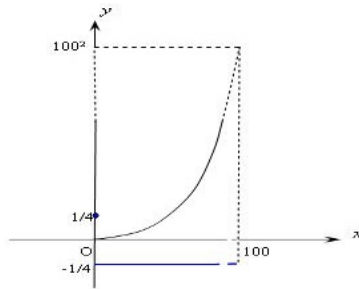


Fig. 11. An Example of A Point and A Line Segment

$$k_i = \frac{i^2 - (i - 1)^2}{i - (i - 1)} = 2i - 1, \quad i = 1, 2, \dots, 100$$

$$\alpha_i = \arctan(2i - 1), \quad i = 1, 2, \dots, 100$$

$$\alpha = \frac{\sum_{i=1}^{100} \alpha_i}{100} = 1.54$$

Therefore, the obliquity of direction between two spatial objects(i.e. a point and a line segment) is

$$\beta = \alpha + \pi/2 = 3.11$$

The value really reflects the relations between the two spatial objects. Therefore, it is feasible to describe spatial direction relation based on Voronoi diagram. At the same time, the example also demonstrates the correctness and accuracy of the algorithm. The time complexity of the algorithm is $O(n)$.

5 Conclusion

Spatial direction relation is a kind of important spatial relations. It is feasible to describe spatial direction relation with Voronoi diagram. The presented method in this paper mainly has four advantages as follows:

- (1) It doesn't need to distinguish source object and reference object. It is reflexive.
- (2) It is not influenced by the shape, location, distance of object and there is a right result.
- (3) Multi-profiles of two objects are considered. The direction of two objects is described by an aggregation of multi-direction. The deficiency of simplification which using other models such as MBR model, 2D String model and so on is avoided.
- (4) It is easy to combine with the achievement of the topology model based on Voronoi diagram.

But if the map is very complex, this method can also bring some deviations because of the decision of the range of Voronoi border.

The spatial relations are synthetical and complicated. This paper doesn't take into account distance and topology relations. The future work will concentrate on the combination of topology, direction and distance and search a kind of uniform reasoning model based on Voronoi diagram.

References

- [1] Andrew UF. Qualitative Spatial Reasoning About Cardinal Directions. In: Mark D, White D, eds. *Proc. of the 7th Austrian Conf. on Artificial Intelligence*. Baltimore: Morgan Kaufmann, 157–167, 1991.
- [2] Christian F. Using Orientation Information for Qualitative Spatial Reasoning. In: Frank AU, Campari I, Formentini U, eds. *Proc. of the Int'l Conf. on GIS*. Berlin: Springer-Verlag, 162–178, 1992.
- [3] Chang S K, Shi Q Y, and Yan C W. Iconic Indexing by 2-D String, *IEEE Trans. Pattern Anal. Machine Intelligence*, 1987(9):413–428.
- [4] Huang P W, Lee C H. Image Database Design Based on 9D-SPA Representation for Spatial Relations. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL.16,NO.12,DECEMBER 2004.

- [5] Goyal R, Egenhofer M J. Similarity of Cardinal Directions. In: Jensen C S, Schneider M, Seeger B, Tsotras V J, eds. *Advances in Spatial and Temporal Databases*. Redondo Beach: Springer-Verlag. 2001, pp.36–55.
- [6] Yang Y Q, Feng J, and Wang Z J. The Reasoning Models of Direction and Topology Relations in Road Network. *Journal of Computational Information Systems*. Vol.2, No.2. 2006, p. 713–718
- [7] Li C M, Zhu Y H, and Chen J. Direction Relation Description and Determination Based on Voronoi Diagram in GIS. *Journal of Liberate Army Mapping Institute*. VOL.15, No.2 Jun. 1998.

Some Experiments of Face Annotation Based on Latent Semantic Indexing in FIARS

Hideaki Ito and Hiroyasu Koshimizu

School of Information Science and Technology, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393 Japan
{itoh, hiroyasu}@sist.chukyo-u.ac.jp

Abstract. This paper describes annotation of face images in keywords based on latent semantic indexing, and experimental results in FIARS. Two latent semantic spaces are constructed from visual and symbolic features. These features are corresponding to lengths of some places of a face and keywords. One latent semantic space is constructed from visual features, the other space is constructed from both features. The former space is used for retrieving similar face images, and the latter for seeking keywords to a given face image. Moreover, the two types of visual features are utilized. One is specified in terms of the lengths of face parts, and the other in terms of points on the outlines of a face and its parts. As an experiment, recall and precision ratios of assigned keywords are measured using the two types of the visual features.

1 Introduction

To develop an annotation system of images is progressed[2,5,7,12,14], in recent. Moreover, a mechanism to annotate images in words which represent conceptual and emotional characters, and impressions of images is required based on human sensibility. In order to achieve face image annotation, such a mechanism is required, moreover, it is necessary to integrate visual and symbolic features. Keywords are corresponding to the symbolic features, and lengths of some places of a face are corresponding to visual features, respectively. The keywords represent the features of a face, visual impression, etc.

We have been developing a face image annotation system based on latent semantic indexing[1,6], called FIARS (the face image annotation and retrieval system)[4]. The beginning of developing FIARS, only one latent semantic space is utilized, which is made from symbolic and visual features. By expanding this system, we have developed the mechanism that two latent semantic spaces are combined. These latent semantic spaces are called the numerical latent semantic space and the combined latent semantic space. The former is constructed from the visual features only, and the latter is constructed from both visual and symbolic features. Moreover, two types of visual features are provided. One is lengths of face parts, the other points on the outlines of face parts and a face.

Some automatic image annotation systems have been developing. Such annotation systems treat to annotate regions of an image in keywords[3,8,14]. Symbolic and visual features are treated separately in traditional image and text

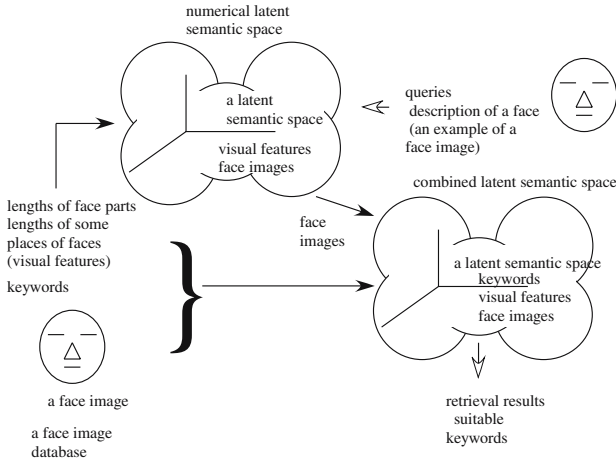


Fig. 1. FIARS: An overview of annotation of face images using two latent semantic spaces

retrieval systems. However, to achieve face annotation, it is required that visual and symbolic features have to be integrated. These two types of features are treated in the same way, in latent semantic indexing[9,13]. Some systems are proposed to reduce semantic gap using the latent semantic indexing[8,13]. On the other hand, it is required to specify faces in terms of visual features. Description of face images using singular values are proposed in[11].

This paper is organized as follows. Section 2 presents an overview of FIARS. Section 3 shows construction of two latent semantic spaces and their utilization. The implementation of the system is shown in Sec. 4. Some experimental results are presented in Sec. 5. Finally, concluding remarks are described in Sec. 6.

2 An Overview

A face image is specified in terms of some keywords and some visual features. The face image database is a collection of the face image descriptions. The description consists of two types of data. One is symbolic data which are keywords. The other is numeric data for visual features. The description of a face image is represented in a vector as a result, called a face image description vector.

Figure 1 shows an overview of annotation mechanism using two latent semantic spaces based on the face description vector. Two latent semantic spaces are constructed, which are the numerical latent semantic space and the combined latent semantic space. The former is constructed from only visual features, and the latter from all of the elements of face image description vectors, respectively.

At first a face image is given to be annotated, as a query. Its similar face images are sought in the numeric latent semantic space. As this result, some face images are retrieved. The positions of the retrieved face images in the combined latent semantic space are computed, which are represented in forms of vectors. The

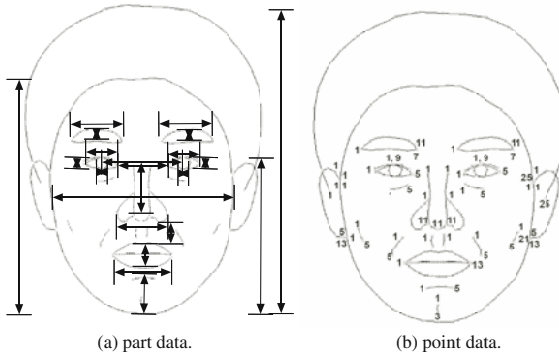


Fig. 2. Two types of visual features of a face image

barycenter of such vectors is calculated. For seeking keywords this barycenter vector is treated as a query. The keywords which satisfy a defined similarity to the query are retrieved as an answer set. The retrieved keywords are seemed as keywords for assigning to the given face image.

The face image database consists of 300 face images.

3 Construction and Utilization of Latent Semantic Spaces

A face image is specified in a face description vector. The attributes of the vector are set according to symbolic features and visual features. If an attribute is a keyword and the keyword is assigned to the face image, its value is 1, otherwise 0. There are around 70 keywords, in current.

FIARS deals with two types of visual features which are shown in Figure 2. They are called part data and point data, respectively. (a) shows part data. The lengths of 24 face parts are measured, for example, the length of an eye, the length of a face, etc. Some of part data in a face description vector are a sequence of the lengths of the face parts. On the other hand, when the point data are used as a visual feature, the outlines of a face and face parts are captured from a face image, which are shown in Figure 2 (b). About 300 points on the outlines are captured. Among them, about 30 points are selected every 10 points. The distances between all two points are computed. These distances are the values of the visual features. About 500 distances are computed for one face image.

The visual features are numeric values, and the values of keywords are binary values, respectively. The mean values and the variances of these features have large difference, when their distributions are compared. For balancing them, an actual observed value v'_i of the visual feature is normalized as $v_i = (v'_i - \overline{v'_i}) / \sigma_{v'_i} + 1/2$. $\overline{v'_i}$ and $\sigma_{v'_i}$ are the mean value and the standard deviation of v'_i .

A face description vector F_{Ci} is represented in the form of $(k_1, k_2, \dots, k_a, v_1, v_2, \dots, v_b)^T$. k_s and v_t are keyword s and visual feature t , respectively. A face description matrix, A_C ($n \times m$), is a collection of the face description vectors, (F_{C1}, \dots, F_{Cm}) . This matrix is decomposed into three

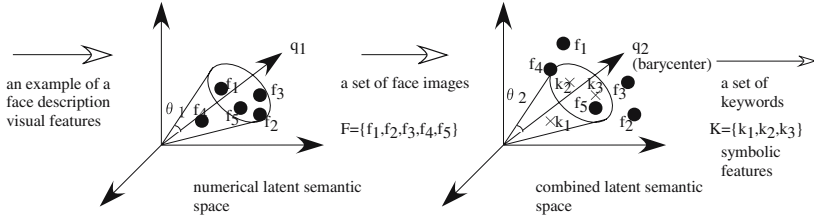


Fig. 3. An overview of a procedure for retrieving keywords based on a barycenter

matrices by the singular value decomposition. They are an attribute matrix W_C , a singular matrix Σ_C , and a face object matrix D_C . Face objects are vectors for representing face image vectors in the latent semantic space. Let the singular values, $\sigma_{C1} \geq \dots \geq \sigma_{Cl}$, be obtained. When a certain threshold is set as σ_{Ct} , the reduced singular matrix Σ'_C is obtained, which consists of $\sigma_{Ci} (1 \leq i \leq k)$, such that $\sigma_{Ci} \geq \sigma_{Ct}$. A_C is approximated as $A_C = W_C \Sigma_C D_C^T \approx W'_C \Sigma'_C D'^T_C$. The sizes of W'_C , Σ'_C and D'_C are $n \times k$, $k \times k$ and $k \times m$. As this result, a certain space is created, which consists of the attributes and the face objects. The position of keyword pw_i is $pw_i(x_1, \dots, x_k) = (w'_{i,1} \times \sigma'_1, \dots, w'_{i,2} \times \sigma'_k)$, the position of face object pd_j is $pd_j(x_1, \dots, x_k) = (\sigma'_1 \times d'_{1,j}, \dots, \sigma'_k \times d'_{k,j})$, respectively. On the other hand, the numerical latent semantic space is constructed using the vector, in which elements are corresponding to the visual features, only.

Figure 3 shows an overview of a procedure for assigning keywords to a given face image. At first, the numerical latent semantic space is used. Then a query is specified in the visual features of the given face image. This query is reformed into a query vector, q_1 . The similarity between the query vector q_1 and a face image vector f_i is defined as $\cos \theta_1 \leq \frac{(q_1, f_i)}{|q_1| \cdot |f_i|}$, that is, the cosine measurement. Here, θ_1 is a threshold. If a face image vector satisfies this similarity, such satisfied face images are obtained as similar face images to the given face image.

Next, the combined latent semantic space is used to find keywords. At this phase, an input is a set of face images F which are obtained at the first phase, and an output is a set of keywords. The elements of F are represented by means of vectors in the combined latent semantic space, also. The barycenter of the face image vectors of F is computed. This barycenter is a vector whose elements are mean values of the elements of the face image vectors. This barycenter is used as a query q_2 for seeking keywords. The similarity between the query vector q_2 and a keyword vector k_i is measured, like the previously defined similarity that retrieves face images in the numerical latent semantic space. As shown in Figure 3, the threshold θ_2 is determined. If a keyword vector satisfies the similarity, such keyword vectors are retrieved as an answer set.

4 Implementation of FIARS

FIARS consists of the following components:

- to define face images in terms of face description vectors into the face image database. Keywords and values of visual features are specified.

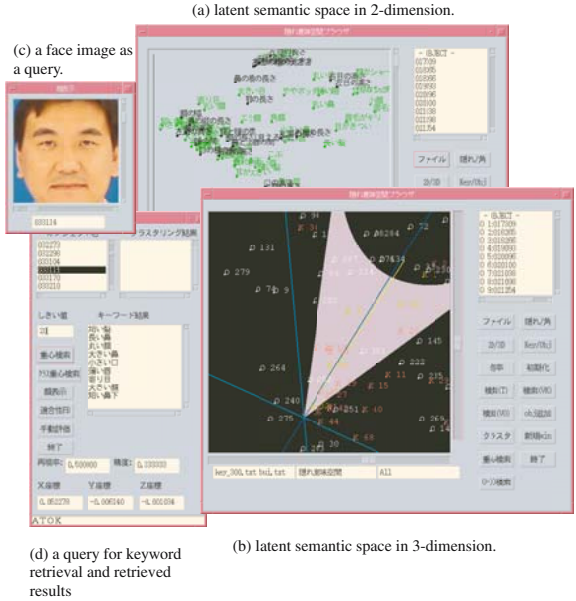


Fig. 4. Some screens in FIARS

- to transform measured data of a face image into a face description vector.
- to construct latent semantic spaces. Based on the definitions of face description vectors, the numerical latent semantic space and the combined latent semantic space are constructed.
- to retrieve keywords to a given face image. Since it is difficult to specify numeric data in exact, actual data are captured from face images. Moreover, a threshold for seeking each latent semantic space is able to be specified.
- to retrieve similar face images to the given face image. These face images are able to be shown with their visual features.
- to show some spaces at the same time. The system deals with several types of elements in the space. For example, face images, keywords, and elements of visual features. Since an inherent space is constructed depending on the visual futures, it is necessary to compare such spaces in graphical.

Figure 4 shows some windows. Latent semantic spaces are multi-dimensional spaces in original. The system provides the mechanism to show the spaces in two- or three-dimension. (a) and (b) show the spaces in the two- and three-dimensions, respectively. (c) shows the face image which is used as a query. The query is specified using the form shown in (d). A threshold to define similarity is specified also. Then the region to be retrieved is shown by means of a fan-shaped region in (a). Retrieved keywords are shown in window (c).

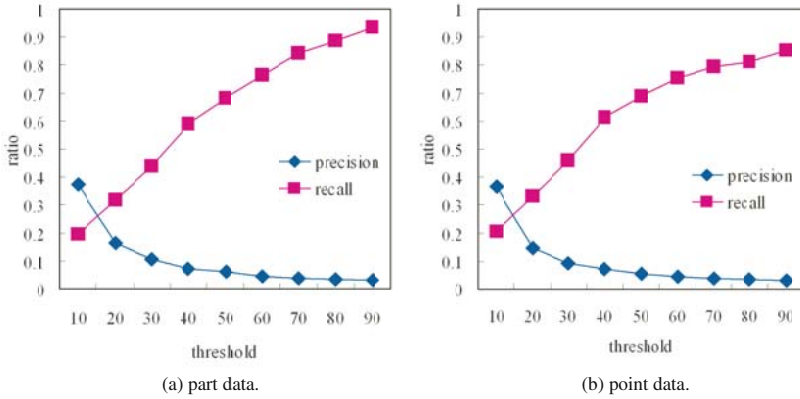


Fig. 5. Recall and precision ratios of face image retrieval using the numerical latent semantic space, when the part data and the point data are used

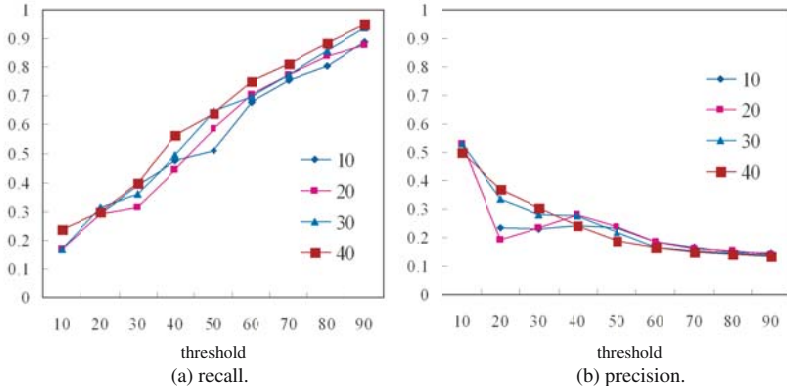


Fig. 6. Recall and precision ratios of retrieved keywords using the part data

5 Experimental Results

At first, similar face images are retrieved when one face image is given as a query. For retrieving similar face images to this query, the numerical latent semantic space is used. Figure 5 shows recall and precision ratios of the retrieval results. In this experiment, thresholds are changed. The angle for defining the similarity is changed 10° to 90° as the threshold θ_1 , see Fig. 3. The changes of the ratios are very marked, when the angle between the query and the face image vector is changed within 10° and 40° . Figure 5 (a) and (b) show two ratios when part data and point data are used as a visual feature, respectively. As shown in this figure, these ratios are similar to each other. Therefore, it seems that a face image is suitably represented in the part data as well as in the point data.

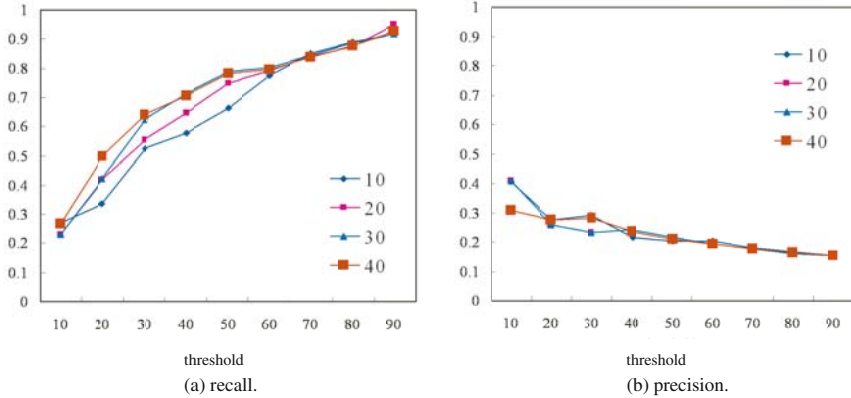


Fig. 7. Recall and precision ratios of retrieved keywords using the point data

To evaluate efficiency with respect to assignment of keywords, recall and precision ratios are measured. The recall and precision ratios are defined as $recall = (\text{the number of correct keywords which are retrieved}) / (\text{the number of keywords which are assigned in advance})$, and $precision = (\text{the number of correct keywords which are retrieved}) / (\text{the number of keywords which are retrieved})$.

Figure 6 and 7 show recall and precision ratios when the parts data and the point data are used as a visual feature, respectively. The threshold θ_1 and θ_2 is changed, simultaneously. θ_2 is a threshold for retrieving keywords at the second phase, see Fig.3. At the same time, the angle for retrieving similar face images θ_1 is changed between 10° to 40° .

When efficiencies of the retrieval results are compared, that the part data and the point data are applied, it seems that the efficiency using the point data is slightly better than the efficiency using the part data based on our experiments, intuitively. The number of elements of a face description vector based on the point data is greater than one based on the part data. When the point data is used, the main factors for constructing the combined latent semantic space are the visual features, and similar face images are located near in the space. While, when the part data are utilized, the number of the keywords is greater than the number of the visual features. Since composition of the space is reflected from keywords strongly when the combined latent semantic space is constructed, there is the case that keywords have harmful influence on their positions. It is considered that keywords will make the combined latent semantic space confusing.

6 Concluding Remarks

FIARS has been developing as an annotation system and an image retrieval system for face images based on the latent semantic indexing. Efficiency of keyword assignment is measured in the sense of recall and precision ratios. Now, we plan to develop two mechanisms for improving the efficiency. One mechanism

is to make interrelationship among the keywords clear. The other mechanism is relevance feedback.

Acknowledgement

We wish to thank Tsuda, T. and Sawada, S. for their cooperation to implement the system and for some experiments. The face images are used by permission of Softopia Japan, Research and Development Division, HOIP Laboratory.

References

1. Berry, M., Drmaç, Z., and Jessup, E.R.: Matrices, Vector Spaces, and Information Retrieval. SIAM Review, Vol.41, No.2, 1999.
2. Blie, D. and Jordan, N. J.: Modeling Annotated Data. Proc. SIGIR, 2003.
3. Carneiro, G., Vasconcelos, N.: A Database Centric View of Semantic Image Annotation and retrieval. Proc. SIGIR, 2005.
4. Ito, H. and Koshimizu, H.: Keyword and Face Image Retrieval Based on Latent Semantic Indexing. Proc. IEEE International Conference on SMC, 2004.
5. Jeon, J., Lavrenko, V., and Manmatha, R.: Automatic Image Annotation and retrieval using Cross-Media Relevance Models. Proc. SIGIR. 2003.
6. Kontostathis, A. and Pottenger, W. M.: A Framework for Understanding Latent Semantic Indexing (LSI) Performance. Information Processing and Management, Vol. 42, 2006.
7. Li, J. and Wang, J. Z.: Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. IEEE Trans. PAMI, Vol.25, 2003.
8. Monay, F., Gatica-Perez D.: PLSA-based Image Auto-Annotation: Constructing the Latent Space. Proc. ACM MM, 2004.
9. Pečenović, Z., Ayer, S., and Vetterli, M.: Joint Textual and Visual Cues for Retrieving Image Using Latent Semantic Indexing. Proc. International Workshop on Content-based Multimedia Indexing, 2001.
10. Softopia Japan Foundation: Face Image database. <http://www.hoip.jp/web=catalog/top.html>.
11. Tian, Y., Tan, T. Wang, Y., and Fang, Y.: Do Singular Values Contain Adequate Information for Face Recognition?. Pattern Recognition 36, 2003.
12. Tsai, C.-F., McGarry, K., and Tait, J.: Qualitative Evaluation of Automatic Assignment of Keywords to Images. Information Processing and Management, Vol. 42, 2006.
13. Zhao, R. and Grosky, W.I.: Narrowing the Semantic Gap? Improved Text-Based Web Document Retrieval Using Visual Features. IEEE Trans. on Multimedia, 4(2), 2002.
14. Zhou, X., Chen, L., Ye, J., Zhang, Q., and Shi, B.: Automatic Image Semantic Annotation Based on Image-Keyword Document Model. CIVR 2005, Leow, W.-K. et al. (Eds.), LNCS 3568, 2005.

Extended Virtual Type for a Multiple-Type Object with Repeating Types

Hideki Sato¹ and Masayoshi Aritsugi²

¹ Daido Institute of Technology, 10-3 Takiharu-cho, Minami-ku,
Nagoya 457-8530, Japan

² Gunma University, 1-5-1 Tenjin-cho, Kiryu 376-8515, Japan

Abstract. INADA is an enhanced C++ persistent programming language and compliant with the ODMG standard. INADA provides *multiple-type object* which enables any persistent objects to be extended by obtaining any type and by losing any unnecessary types. Furthermore, INADA provides *virtual type* which enables any persistent objects to be accessed through a virtual type derivable from other base/virtual types. However, it does not allow the type which a virtual type is derived from to be a repeating one. To overcome this constraint, this paper proposes *extended virtual type* which allows a virtual type to be derived from a repeating one and shows a method of implementing it in INADA.

1 Introduction

Object databases [1] possess an excellent ability to model real-world entities. However, it is widely recognized that intelligent databases need support for (1) objects that may evolve over time, and may exhibit a different behavior in different contexts (*object extension* facility [2,3,4,5,6,7,8]), (2) views mechanism to adapt database schema to the changing needs of users (*object viewing* facility [9,10,11,12,13,7,14,15]).

We have proposed *multiple-type object* [16] as *object extension* facility, which enables any persistent objects to be extended by obtaining any type and by losing any unnecessary types. It has been implemented in an enhanced C++ persistent programming language called INADA, which is compliant with the ODMG standard. In addition to the ordinary type selection method, we have proposed *AEE* method for realizing flexible and intelligent objects [17], in which a multiple-type object selects one from among its own types depending on an object accessing it. Furthermore, we have proposed and implemented *virtual type* [18] as an amalgamation of *object extension* and *object viewing*. It derives virtual types whose structure and behavior are different from those of base types¹ added to a persistent object.

Separately from the above mentioned works, we have extended *multiple-type object* by enabling persistent objects to obtain a collection of instances of the same type, which is called *repeating type* [19]. However, *virtual type* does not

¹ In this paper, ordinary types are called base types to differentiate them from virtual types.

allow the type which a virtual type is derived from to be a repeating one. To overcome this constraint, this paper proposes *extended virtual type* which allows a virtual type to be derived from a repeating one and presents a method of implementing it in INADA. It differs from *virtual type* in that it enables both of the types which a virtual type is derived from and the derived type itself to be repeating types.

The research [8] proposed another *object extension* facility which takes *repeating type* into consideration like INADA. However, it has not provided *object viewing* facility. While `Object deputy model` [7] and `Galileo97` [15] provide *object extension* and *object viewing* facilities like INADA, their *object extension* facilities have not taken *repeating type* into consideration.

This paper is organized as follows. Chapter 2 introduces *extended virtual type*. Chapter 3 presents a method of implementing *extended virtual type* in INADA. Finally, Chapter 4 summarizes the paper.

2 Extended Virtual Type

This chapter introduces *extended virtual type* by presenting an exemplary definition of a virtual type and by showing how the defined type is manipulated.

2.1 Exemplary Virtual Type Definition

This section presents an example of virtual type definitions.

[`Employee` type and `Manager` type] An object for representing an employee may possess `Employee` type with attributes `empno` for employee number, `name`, `age`, `sect` for section which (s)he belongs to in a company, `salary` and/or `Manager` type with attributes `empno`, `sect` for section which (s)he manages, `base_budget`, `no_of_subordinates` for the number of subordinates whom (s)he supervises. Note that these attributes are assumed to be `public`.

Fig.1 is an example of virtual type definitions using `Employee` type and `Manager` type.

[`ManagerOfLargeScaleSection` virtual type] For a manager responsible for a section which 20 or more subordinates belong to, a virtual type is defined with attributes `empno` for employee number, `name`, `AGE`, `sect` for section which (s)he manages, `total_budget` for `base_budget` plus 10 per subordinate, and with `decrease_subordinates` method to decrease the number of subordinates whom (s)he supervises.

`ManagerOfLargeScaleSection` virtual type is derivable for an object which possesses both `Employee` type and `Manager` type with `no_of_subordinates` attribute being 20 or more (`product` operator and `restrict` operator). The OID of `ManagerOfLargeScaleSection` instance is the same as that of the base object. Some attributes of `Employee` type and `Manager` type are specified to be those of `ManagerOfLargeScaleSection` virtual type (`project` operator). As for `age` attribute of `Employee` type, its name is changed into `AGE` (`rename` operator) and its value type is changed from `integer` to `string` with `itoa` function. `total_budget` attribute is defined by an arithmetic expression with attributes `base_budget` and

```

virtual type ManagerOfLargeScaleSection
on e:Employee, m:Manager{
char* empno() : e->empno;
char* name() : e->name;
char* AGE() : itoa(e->age);
char* sect() : m->sect;
int total_budget() :
    m->base_budget+10*m->no_of_subordinates;
void decrease_subordinates(int no){
    m->no_of_subordinates-=no;}
where m->no_of_subordinates>=20;
}

```

Fig. 1. Definition of `ManagerOfLargeScaleSection` virtual type

`no_of_subordinates` of `Manager` type (attribute extend operator). Method `decrease_subordinates` is added to the virtual type (method extend operator), which updates `no_of_subordinates` attribute of `Manager` base type.

2.2 Exemplary Virtual Type Manipulation

This section presents an example of virtual type manipulation using virtual type `ManagerOfLargeScaleSection`. First, an object of `Employee` type is created in the following way, where `database` is a variable referring to an object of `d.Database` class [1] and `pe` is a variable referring to the object. Each argument of the constructor provides the object with each value of attributes `empno`, `name`, `age`, `sect`, and `salary` in order.

```

d.Ref<Employee>pe=new(database,"Employee")
    Employee("101","Sato",38,"software1",50);

```

Next, `Manager` type is added to the former object. Each argument of the constructor provides the type with each value of attributes `empno`, `sect`, `base_budget`, and `no_of_subordinates` in order. Method `transforms` makes the OID of the object referenced by `pm1` equal to the OID of the object referenced by `pe`.

```

d.Ref<Manager>pm1=new(database,"Manager")
    Manager("101","software1",500,30);
pm1.transforms(pe);

```

Furthermore, two other objects of `Manager` type respectively referenced by `pm2` and `pm3` are added to the object referenced by `pe`, to represent concurrently holding manager for distinct sections.

```

d.Ref<Manager>pm2=new(database,"Manager")
    Manager("101","software2",300,15);
pm2.transforms(pe);
d.Ref<Manager>pm3=new(database,"Manager")
    Manager("101","system3",400,22);
pm3.transforms(pe);

```

Whether or not an object has a specific type is examined with `hastype` method in the following way. Since the object referenced by `pe` possesses both `Employee` type and `Manager` type with attribute `no_of_subordinates` being 20 or more, two instances of `ManagerOfLargeScaleSection` virtual type are derivable. In the case, minus 2 is assigned to `t1`².

² A minus means a virtual type and an absolute value indicates the number of its instances.

```
t1=pe.hastype("ManagerOfLargeScaleSection");
```

Method `create_iterator` is provided to instantiate an iterator [1] for sequentially returning each element from a repeating type to be specified by an argument. Method `next` provides a facility for checking the end of iteration, advancing the iterator and returning the current element, if there is one. In the following statements, each of `ManagerOfLargeScaleSection` virtual type is sequentially accessed with an iterator, its attributes `sect` and `total_budget` are outputted ("`software1:800`" and "`system3:620`"), and `no_of_subordinates` attribute is decreased by 3. After execution, `no_of_subordinates` attribute of `software1` section and that of `system3` section are changed into 27 and 19 respectively. If an iterator is instantiated to access `ManagerOfLargeScaleSection` virtual type again, only a single instance responsible for `software1` section is returned due to the `restrict` condition used to define the type (See Fig.1).

```
d.Iterator <d.Ref<ManagerOfLargeScaleSection>>iter=
    pe.create_iterator("ManagerOfLargeScaleSection");
ManagerOfLargeScaleSection* pml;
while(iter.next(pml)){
    cout<<pml->sect()<<": "<<pml->total_budget()<<"\n";
    pml->decrease_subordinates(3);
}
```

Method `as` is used for returning a reference of only one instance whose type name is specified by its argument. It is uncertain which type instance is to be returned with the method, if the specified type is a repeating one. The following statements output string "`software1:770`".

```
ManagerOfLargeScaleSection* pml1=(ManagerOfLargeScaleSection*)
    pm.as("ManagerOfLargeScaleSection");
cout<<pml1->sect()<<": "<<pml1->total_budget()<<"\n";
```

The following statement deletes the virtual type referenced by `pml1` like volatile objects. This leads to the removal of the area occupied in a temporary heap.

```
delete pml1 ;
```

3 Implementation of Extended Virtual Type

This chapter firstly presents the overview of *extended virtual type* implementation. Then, management of meta-object collection used for virtual type processing is described.

3.1 Overview of Extended Virtual Type Implementation

Fig.2 shows the overview of generation of a schema definition/implementation and method execution for a virtual type. The generator of a schema definition/implementation inputs the definition of virtual type `xxx` and outputs meta-object defining class `Meta_xxx`, iterator defining class `xxx_Iterator`, and virtual type defining class `xxx`. `Meta_xxx` class is a subclass of `Meta_Virtual_Type` class which is a generic class responsible for virtual type meta-processing. Class

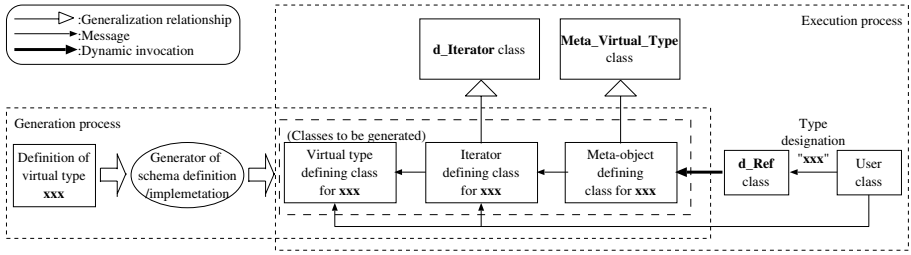


Fig. 2. Overview of generation of schema definition/implementation and method execution for a virtual type

xxx_Iterator is a subclass of *d_Iterator* class [1] and provides a function for sequentially returning each virtual type from a repeating type.

As mentioned in Section 2.2, methods `transforms`, `hastype`, `create_iterator`, and `as` are called from methods which are defined in user classes for manipulating multiple-type objects. Each of them is implemented in the corresponding method of *d_Ref* class [1]. When any of methods `hastype`, `create_iterator`, and `as` is called with virtual type *xxx* as its argument, the corresponding method of *Meta_xxx* class is subsequently called. In the case of `create_iterator` method, a *xxx_Iterator* instance is created which `next` method is called to for sequentially returning each instance of virtual type *xxx*.

Fig.3(a) shows the schema definition of *Meta_Virtual_Type* system-defined class and Fig.3(b)-(d) show the schema definitions which the generator in Fig.2 outputs for *ManagerOfLargeScaleSection* virtual type. *Meta_ManagerOfLargeScaleSection* class in Fig.3(b) is a subclass of *Meta_Virtual_Type* class and its methods `create_iterator`, `as`, and `hastype` are polymorphic. *ManagerOfLargeScaleSection_Iterator* class in Fig.3(c) defines an iterator, whose method `set_iterator` initializes the iterator, and whose another method `next` checks the end of iteration, advances the iterator, and returns the current element, if there is one. *ManagerOfLargeScaleSection* class in Fig.3(d) corresponds to the virtual type with the same name. Due to the space limitation, schema implementations which the generator in Fig.2 outputs for a virtual type cannot be described in this paper.

3.2 Management of Meta-object Collection

As mentioned in Section 3.1, each method of *Meta_xxx* class is called from that of *d_Ref* class. Since a virtual type to be processed is designated by string "*xxx*" to be given to methods of *d_Ref* class, a mechanism is needed in the strongly typed language INADA, which dynamically selects the corresponding object of *Meta_xxx* class for the name of the virtual type (See Fig.2). A collection of virtual type meta-objects is managed to realize the mechanism, as shown in Fig.4.

The collection is implemented by *d_Dictionary*<K,V> class [1] and its element is an instance of *Meta_Virtual_Type* class. *d_Dictionary*<K,V> class

```

template <class T> class Meta_Virtual_Type : public d_Object {
public :
    virtual d_Iterator<d_Ref<T>> create_iterator(d_Ref_Any) ;
    virtual void* as(d_Ref_Any) ;
    virtual int hastype(d_Ref_Any) ;
};

```

(a) Meta-object defining class for a virtual type **Meta_Virtual_Type**

```

class Meta_ManagerOfLargeScaleSection : public Meta_Virtual_Type {
public :
    d_Iterator<d_Ref<ManagerOfLargeScaleSection>> create_iterator(d_Ref_Any) ;
    void* as(d_Ref_Any) ;
    int hastype(d_Ref_Any) ;
};

```

(b) Meta-object defining class for virtual type **ManagerOfLargeScaleSection**

```

class ManagerOfLargeScaleSection_Iterator : public {
private :
    d_Ref_Any ORTE ;
    int vflag_e ;
    int flag_e ;
    int dflag_e ;
    d_Iterator<d_Ref<Employee>> iter_e ;
    Employee* e ;
    int vflag_m ;
    int flag_m ;
    int dflag_m ;
    d_Iterator<d_Ref<Manager>> iter_m ;
    Manager* m ;
public :
    void set_iterator(d_Ref_Any) ;
    d_Boolean next(void*) ;
};

```

(c) Iterator defining class for virtual type **ManagerOfLargeScaleSection**

```

class ManagerOfLargeScaleSection {
private :
    d_Ref_Any ORTE ;
    int dflag_e ;
    Employee* e ;
    int dflag_m ;
    Manager* m ;
public :
    char* empno() ;
    char* name() ;
    char* AGE() ;
    char* sect() ;
    int total_budget() ;
    void decrease_subordinates(int) ;
    void ManagerOfLargeScaleSection(d_Ref_Any, int, Employee*, int, Manager*) ;
    void ~ManagerOfLargeScaleSection() ;
};

```

(d) Virtual type defining class for **ManagerOfLargeScaleSection****Fig. 3.** Schema definitions to be outputted by generator

is an unordered collection of key K and value V pairs³, where at most one value V exists for a specific key K . `d_Dictionary<K,V>` class is derived from `d_Collection<T>` class [1], where a pair of key K and value V is represented by `d_Association<K,V>` class [1]. On defining virtual type `xxx`, the `Meta_xxx` instance is instantiated and `bind` method is invoked for inserting a pair of key "xxx" and the instance into the collection. On deleting virtual type `yyy`, `unbind("yyy")` method is invoked for removing the corresponding instance of `d_Association<K,V>` from the collection. On referencing virtual type `zzz`,

³ Since `d_Dictionary<K,V>` class is implemented with a hashing method in INADA processing system, value V is efficiently retrieved for key K .

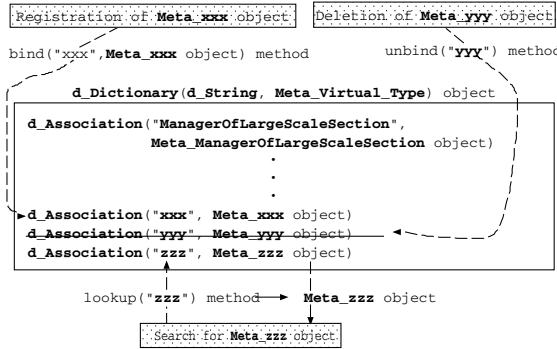


Fig. 4. Management of meta-object collection

lookup("zzz") method is invoked for returning the Meta_zzz instance corresponding to the key.

4 Conclusion

This paper has proposed *extended virtual type* which allows a virtual type to be derived from repeating types that a persistent object possesses. This facility has been implemented in the enhanced C++ persistent programming language INADA, compliant with the ODMG standard. It differs from *virtual type* in that it enables both of the types which a virtual type is derived from and the derived type itself to be repeating types. Since a repeating type is a collection of instances of the same type added to a persistent object, an iterator becomes one of the main points from the standpoint of implementation. With *extended virtual type*, it becomes possible to model a virtual role such as a manager who is responsible for a section which a certain number of subordinates or more belong to, for an employee (one role) who holds some managers (another role) concurrently for distinct sections. It is uniquely formed by the amalgamation of *object extension* facility and *object viewing* facility for realizing intelligent databases.

References

1. Cattel, R.G.G., Barry, D.K.: The Object Data Standard: ODMG3.0. Morgan Kaufmann (2000)
2. Fishman, D.H., Beech, D., Cate, H.P., Chow, E.C., Connors, T., Davis, J.W., Derrett, N., Hoch, C.G., Kent, W., Lyngback, P., Mahbod, B., Neimat, M.A., Ryan, T.A., Shan, M.G.: Iris: An Object-Oriented Database Management System. ACM Trans. Office Information Systems **5** (1987) 48-69
3. Sciore, E.: Object Specialization. ACM Trans. Office Information Systems **7** (1989) 103-122
4. Steing, L.A., Zdonik, S.B.: Clovers: The Dynamic Behavior of Type and Instances. Brown University, Technical Report **CS-89-42** (1989)

5. Richardson, J., P. Schwardz, P.: Aspects: Extending Objects to Support Multiple, Independent Roles. Proc. ACM SIGMOD Int. Conf. on Management of Data (1991) 298-307
6. Albano, A., Bergamini, R., Ghelli, G., Orsini, R.: An Object Data Model with Roles. Proc. Int. Conf. on Very Large Data Bases (1993) 39-51
7. Kambayashi, Y., Peng, Z.: Object Deputy Model and Its Applications. Proc. Int. Conf. on Database Systems for Advanced Applications (1995) 1-15
8. Gottlob, G., Schrefl, M., Rock, B: Extending Object-Oriented Systems with Roles". ACM Trans. Office Inf. Syst. **14** (1996) 268-296
9. Abiteboul, S., Bonner, A.: Objects and Views. Proc. ACM SIGMOD Int. Conf. on Management of Data (1991) 238-247
10. Leung, T., Mitchell, G., Subramanian, B., Vance, B., Vandenberg, S., Zdonik, S.: The AQUA Data Model and Algebra. Proc. Int. Workshop Database Programming Languages (1993) 157-175
11. Ogori, A., Tajima, K.: A Polymorphic Calculus for Views and Object Sharing. Proc. ACM SIGACT-SIGMOD Symp. Principles of Database Systems (1994) 255-266
12. Scholl, M.H., Laasch, C., Rich, C., Scheck, H.J., Tresch, M.: The COCOON Object Model. Department Informatik, ETH, Zurich, Technical Report**211** (1994)
13. Kim, W., Kelly, W.: On View Support in Object-Oriented Database Systems. in Kim., W.(eds.):Modern Database Systems, Addison-Wesley (1995) 108-129
14. Guerrini, G., Bertino, E., Catania, B., Garcia-Molina, J.: A Formal Model of Views for Object-Oriented Database Systems. Theory and Practice of Object Systems (TAPOS), **3** (1997) 103-125
15. Albano, A., Antognoni, G., Ghelli, G.: View Operations on Objects with Roles for a Statically Typed Database Language. IEEE Trans. Knowledge and Data Engineering **12** (2000) 548-567
16. Aritsugi, M., Makinouchi, A.: Multiple-type Objects in an Enhanced C++ Persistent Programming Language. Software-Practice and Experience **30** (2000) 151-174
17. Sato, H., Aritsugi, M.: Accessee Controlled Type Selction for a Multiple-Type Object. Proc. ACM Symp. Applied Computing (2003) 515-521
18. Sato, H., Aritsugi, M.: A Virtual Type for a Multiple-type Object and Its Implementation. IEICE Trans. Information and Systems **J89-D** (2006) (to appear)
19. Sato, H., Aritsugi, M.: Implementation of a Multiple-type Object with Repeating Types. IEICE Trans. Information and Systems **J85-D-I** (2002) 1093-1098

Spatial Relation for Geometrical / Topological Map Retrieval

Toru Shimizu¹, Masakazu Ikezaki¹, Toyohide Watanabe¹,
and Taketoshi Ushiana²

¹ Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{shimizu, mikezaki, watanabe}@watanabe.ss.is.nagoya-u.ac.jp

² Faculty of Design, Kyushu University
9-1 Shiobaru 4-chome, Minami-ku, Fukuoka, 815-0032, Japan
ushiana@design.kyushu-u.ac.jp

Abstract. When we look for locations, we use the names of geographical objects (e.g., landmarks, rivers, lakes). However, we often memorize the locations by using relations among landmarks like remarkable buildings. For this reason, we focus on spatial relations among geographical objects, and we define spatial relations based on the geographical proximity. In this paper, we manipulate two types of geographical proximities: a vertical proximity and a horizontal proximity. The vertical proximity shows two geographical objects are overlapping, and the horizontal proximity shows two geographical objects are adjacent. In order to represent these proximities, we introduce two spatial relations: “overlapping” and “neighboring”. “Overlapping” represents the vertical proximity, and “neighboring” represents the horizontal proximity. Furthermore, we propose a spatial relational graph which is generated by connecting geographical objects with these two spatial relations. In addition, we search maps by inclusive relations among spatial relational graphs, and implement the prototype system to find out locations by using a map.

1 Introduction

A keyword search is used in many information systems, and is also useful in Geographical Information System (GIS). When we look for locations, we usually use names of geographical objects (e.g., landmarks, rivers, lakes) in GIS. In contrast, when we look for a location in maps, we use remarkable landmarks. We often memorize a remarkable geographical object under the relationship for other landmarks like buildings. For this reason, we propose an intuitive approach to represent a map and look for a location.

One of the intuitive approaches focuses on the set operation for regions associated with two geographical objects [1][2]. This approach focuses on intersections between two geographical objects. In particular, this approach defines topological / spatial relations which are based on intersections between two geographical objects. However, using topological / spatial relations is not so enough to distinguish maps, because two geographical objects often do not have intersections.

Therefore, this approach should define other relations which are based on the direction of geographical objects. For example, Spatial-Query-by-Sketch derives the semantic network based on cardinal directions among geographical objects [3]. In this semantic network, all geographical objects have relations with all other objects. Therefore, this approach compares with networks by using weight of links. However, if relations defined as relations have a logical property, we can compare with networks by using structures.

Another intuitive approach focuses on roads, and regards a map as connections of roads. This approach generates a network structure by connecting roads, and that the network structure is called the road network. The road networks are useful in conforming geographical objects in two maps which are made by different cartographers [4]. On the other hand, the road networks allow users to search locations by inputting a simplified route map [5]. However, the road networks do not have enough information to identify maps. Therefore, this approach should use geometric information like the direction of road and the length of road, which depends on the scale and cartographers. For this reason, it is not effective to use connections of roads for the map retrieval which is independent of scales and cartographers.

In this paper, we focus on the geographical proximity which indicates two geographical objects are adjacent. We define two spatial relations to represent the geographical proximity. Moreover, we generate graph structures, and describe these graph structures as a spatial relational graph. The remainder in this paper is organized as follows: Section 2 summarizes the spatial relational graph, and describes a map retrieval. In Section 3, we introduce our prototype system, and evaluate a map retrieval. At last, the conclusions in Section 4 discuss our future research.

2 Framework

2.1 Spatial Relational Graph

Usually, geographical objects have their identifiers, their coordinates, and their attributes. Additionally, geographical objects have relations with each another. In this paper, we focus on spatial relations which are based on the geographical proximities in which two geographical objects are adjacent. We define two geographical proximities: a vertical proximity and a horizontal proximity. In order to represent these proximities, we define two spatial relations among geographical objects: “overlapping” and “neighboring”. We define these relations as logical relations, and generate the spatial relational graph by connecting geographical objects with these relations. Therefore, we compare with spatial relational graphs by using their structures (Fig.1).

In Fig.1, there are three maps: map *A*, map *B*, and map *C*. These maps look like different maps: however, spatial relational graphs which are generated by

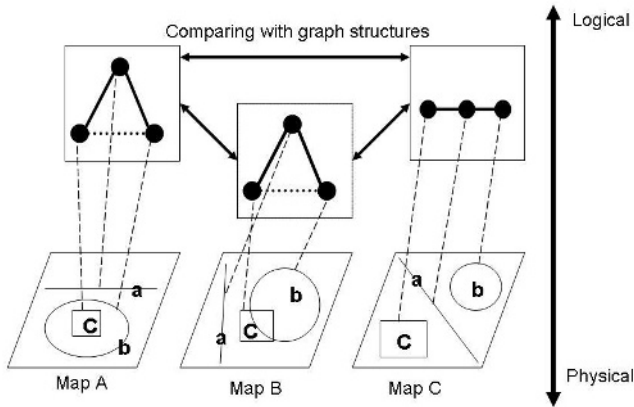


Fig. 1. Using spatial relational graphs to compare with maps

maps *A* and *B* have similar structures. Therefore, we regard maps *A* and *B* as similar maps. In contrast, spatial relational graphs which are generated by maps *A* and *C* do not have similar structures. Therefore, we do not regard maps *A* and *C* as similar maps.

The spatial relational graph structures a map by spatial relations which are based on positions of geographical objects. We assume that our spatial relations among geographical objects do not depend on a kind of maps: that is, spatial relations are independent of scales and cartographers. In contrast, we assume that similar maps generate similar graph structures. If two spatial relational graphs have the similarity, two maps which generate these graphs indicate the same location. For this reason, we propose the map retrieval method based on the comparison among the spatial relational graphs.

2.2 Classification of Geographical Objects

We deal with planar maps which are represented as vector data. Additionally, we classify geographical objects by concepts of geographical objects, because we assume concepts of geographical objects are independent of types of maps. We classify geographical objects into six groups: “road”, “railroad”, “water”, “building”, “ground”, and “district” (Table 1).

There are various types of geographical objects and these types depend on maps. For this reason, in order to deal with maps which are made by different cartographers, we also deal with types of geographical objects similarly.

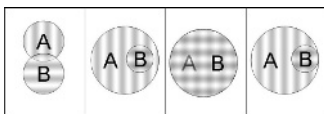
2.3 Derivation of Spatial Relation

The spatial relational graph is based on the proximity of geographical objects. We classify the proximities into vertical and horizontal ones. When there are intersections between geographical objects in planar maps, these intersections

Table 1. Classifying geographical objects into six groups

Name	Representations
Road	Interval between intersections
Railroad	Interval between stations
Water	Rivers, lakes, and coastlines
Building	All of buildings
Ground	Premise of parks, temples, shrines, schools, and so on
District	Nations, states, prefectures, and so on

represent affiliation relations among geographical objects. When two geographical objects have intersections, we assume that these objects have vertical proximities. In order to represent vertical proximities, we define the relation “overlapping”. In addition, we calculate “overlapping” by using intersections among geographical objects. When two geographical objects have an intersection, their relation is called “overlapping” (Fig.2). Fig.2 shows that objects *A* and *B* have “overlapping” with each other.

**Fig. 2.** Overlapping between two geographical objects

On the other hand, if two geographical objects are adjacent, we assume that they have horizontal proximities. We look on the horizontal proximity as a situation in which two geographical objects do not have obstacle objects. In other words, if there is no other objects between two geographical objects, two objects have a horizontal proximity. For this reason, the horizontal proximity does not depend on a physical interval between two geographical objects. The horizontal proximity only expresses that two geographical objects are adjacent. We define the relation “neighboring” to represent a horizontal proximity.

In order to calculate “neighboring”, we focus on barricades. If two geographical objects are contiguous, these two objects do not have obstacle objects, and thus they are “neighboring”. Additionally, if two geographical objects have intersections, these two objects are “overlapping”, and they are not “neighboring”. If two geographical objects are disjoint objects, connecting with two geographical objects, and look for obstacle objects to calculate “neighboring” (Fig.3). Fig.3 shows three situations of “neighboring”. The left side in Fig.3 shows objects *A* and *B* are contiguous. The center in Fig.3 shows that objects *A* and *B* are “neighboring”, and objects *B* and *C* are also “neighboring”. However, because *B* is interfering, *A* and *C* are not “neighboring”. The right side in Fig.3 shows three relations: *A* and *C* are “overlapping”, *A* and *B* are “neighboring”, and *B* and *C* are also “neighboring”. Because *A* and *C* are “overlapping”, *C* cannot interfere in *A* and *B*.

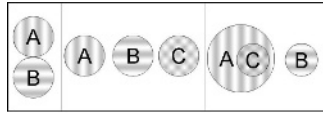


Fig. 3. Neighboring between two geographical objects

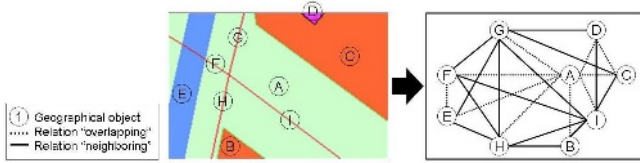


Fig. 4. The map and the spatial relational graph

We show a spatial relational graph which is generated by the sample map (Fig.4). In Fig.4, there are nine geographical objects. Objects *A*, *B*, and *C* are district objects. Then, Object *D* is a building object, and object *E* is a water object. Objects *F*, *G*, *H*, and *I* are road objects. Additionally, there are ten “overlapping” relations and fourteen “neighboring” relations. All of geographical objects are on *A*, therefore: all of geographical objects have “overlapping” between *A*. In addition, there is “overlapping” relation between *E* and *F*, because *F* is a bridge. On the other hands, *G* and *H* are on the river, therefore: there are “neighboring” relations. Moreover, *C* and *G*, *C* and *E* also have “neighboring” relations.

2.4 Map Retrieval by Spatial Relational Graph

In order to compare with maps by using the spatial relational graphs, we focus on inclusive relations of graph structures, because we assume that the same graph structure shows a similar location. We use a map as a query, and look for a location which the map shows in another map.

In order to search inclusive relations of graph structures, we use a simple method which looks like the breadth-first search (Fig.5). First, we find out a starting node which has the largest number of links in a query map. If there are some starting nodes, we use one node which was primarily found. Second, we find out potential starting node sets in a data map. The potential starting node is the same type of geographical objects as the starting node, and it has more links than the root node. Third, we select one of the potential starting nodes, and then we compare with two link node sets: link node set of the selected potential starting node and link node set of the starting node. Through comparing link node sets, we get some potential link node sets which have the same graph structure with link node sets of the starting node. However, selecting potential link node sets generates a combination explosion. In order to avoid a combination explosion, we use relations among potential link nodes. Querying gotten potential link nodes sets, we repeat this node matching process by using the next link nodes. If we

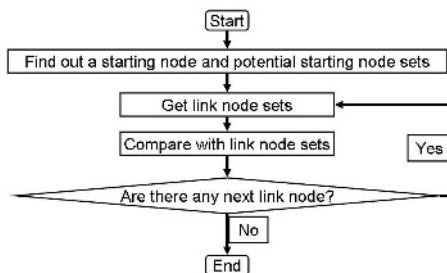


Fig. 5. Method of the map retrieval

check all nodes in the query map, we select the next potential starting node, and repeat the same process. As a result, we get node sets which have the same graph structure of a node set in the query map. However, if we check all nodes in the data map, we fail to find out the location of the query map.

3 Evaluation

3.1 Prototype System

We implemented the prototype system to find out locations by using a map. We use “Numeric Map 2500” which is made by the Geographical Survey Institute in Japan. In “Numeric Map 2500”, there are many types of geographical objects and we classify them into our six groups. Fig.6 shows the interface in our prototype system.

When finding out locations by a map, a user selects two map files: a query map and a data map. Then, our system finds out the same graph structure for the query map in the data map. After our system found out the same graph structures, system shows the result map which is composed of geographical objects in matching node sets. For example, Fig.7 shows three maps: the query map, the data map, and the result map. The result map is a part of the data map, and two spatial relational graphs generated by the query map and the result map has the same structures.

3.2 Experimental Evaluation

In our experiments, we use four maps around Nagoya University, and separate the map to twenty five small maps. Besides, we use four large maps as data maps, and use small maps as query maps. We show properties about four large maps: map *A*, map *B*, map *C*, and map *D* (Table 2). Moreover, Table 2 shows information about twenty five small maps which constitute four large maps.

In our experiments, seventy two small maps were successful in finding out locations (Fig.8). In Fig.8, we show number of result node sets and number of nodes in the query maps. If both the data map and the query map do not

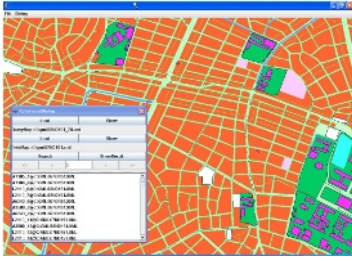


Fig. 6. Interface of prototype system **Fig. 7.** Query map, data map, and result map

Table 2. Map properties

Name of large map	Map A	Map B	Map C	Map D
Number of nodes in large map	2399	965	1772	868
Number of links in large map	12969	5194	9410	4601
Average number of nodes in small maps	124.04	54.12	93.16	48.72
Average number of links in small maps	585.36	237.36	424.8	209.88

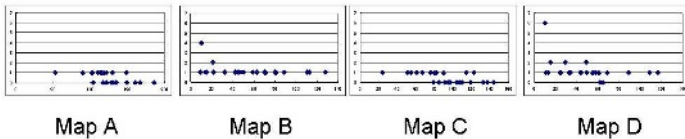


Fig. 8. Number of result node sets and number of nodes in the query maps

have many nodes, we found out locations in the data maps by the query map. However, when we use some query maps which have a lot of nodes, combination explosions occurred. For this reason, we did not get result node sets by some query maps in maps A and C.

4 Conclusions

In this paper, we proposed the spatial relational graph to structurize maps by using the spatial relation among geographical objects. In order to generate spatial relational graph, we define two spatial relations: “overlapping” and “neighboring”. These two relations represent two geographical proximities: a vertical proximity and a horizontal proximity. Additionally, we proposed the map retrieval method by comparing with spatial relational graphs.

In the future work, we must check up the map retrieval is useful in maps which are built by different cartographers. In this paper, we used maps which are made by the same cartographer. However, there are some differences among maps which are built by different cartographers. For example, the representations

of geographical objects depend on a cartographer. Therefore, we use maps which are built by different cartographers, and compare them with spatial relational graphs. Additionally, we must consider map retrieval method. For example, if geographical objects are omitted in the query map, there are difference between the data map and the query map. Therefore, we should get over difference of the spatial relational graphs caused by omitted.

Acknowledgements

The authors would like to thank the 21st Century COE(Center of Excellence) Program for 2002, a project titled Intelligent Media (Speech and Images) Integration for Social Information Infrastructure, proposed by Nagoya University.

References

1. M. Egenhofer : "Spatial Relations: Models, Inferences, and their Future Applications", Advanced Database Symposium (1996).
2. M. Egenhofer, D. Mark, and J. Herring : "The 9-Intersection: Formalism And Its Use For Natural-Language Spatial Predicates", NCGIA Tech. Report (1994).
3. M. Egenhofer : "Query Processing in Spatial-Query-by-Sketch", Journal of Visual Languages and Computing, 8(4), pp.403-424 (1997).
4. V. Walter, and D. Fritsch : "Matching Spatial Data Sets: a Statistical Approach", International Journal of Geographical Information Science, 13(5), pp.445-473 (1999).
5. Y. Kurata, and A. Okabe : "An Algorithm for Identifying Features in A Simplified Route Map", Theory and Applications of GIS, 10, pp.9-17 (2002).

Geographical Information Structure for Managing a Set of Objects as an Event

Masakazu Ikezaki¹, Toyohide Watanabe¹, and Taketoshi Ushiamo²

¹ Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{mikezaki, watanabe}@watanabe.ss.is.nagoya-u.ac.jp

² Faculty of Design, Kyushu University,
9-1 Shiobaru 4-chome, Minami-ku, Fukuoka, 815-0032, Japan
ushiamo@design.kyushu-u.ac.jp

Abstract. This paper addresses a geographic information management to model various kinds of geographic objects with respect to the space and time features. In this model, an event represents the geographic situation in which geographic objects change, and represented as a set of geographic objects which are concerned to the event. It is possible to easily divide an event into sub-events along space and time criteria. Using this model, our geographic information system can provide users with a capability to manipulate an event from various viewpoints.

1 Introduction

Recently, geographic information systems (GISs) are used in various fields such as economics and city administration, etc., and take an important role as the social information infrastructure systems. In GISs, geographic objects such as roads, buildings, etc., are maintained. Geographic objects are changable over time. Appropriate management of information about changes of geographic objects is one of functions expected of GISs. A lot of methods which can manage such functions have been proposed. However, they can represent only local changes such as destructions of a buildings, repairs of roads, etc., because they manage only the version history of geographic objects. GISs are expected to manage local changes from a macro point of view, because local changes are caused by real phenomena such as typhoons, earthquakes and so on. Such phenomena can associate local changes mutually. Therefore, it is important to represent, manage, and utilize such phenomena called an event in a GIS.

There are researches focusing on representation of events. Peuquet et al. proposed an event-based spatio-temporal data model (ESTDM)[1]. In ESTDM, a set of geographic changes at grid-based location for one geographic theme (event) is organized according to temporal axis. With this time-based organization, sequential changes of property values in a location associated with the event are explicitly stored in a GIS database. Chen et al. defined an event as decision-making actions of human beings, and proposed another event-based spatio-temporal database model in which three relations are represented: deterministic relation between events, relation between an event and a state of space, and causal relation between states of space[2]. Their model is capable to represent events hierarchically. In these researches, the event is associated to the point on

time axis. However, there are some mismatches between human recognition and representations in these researches. We usually recognize an event in space and time with a time interval.

In [3], Yuan et al. analyzed physical phenomena such as storms. In their research, events are treated as aggregations of some related processes, and have the starting time and ending time. A process is represented by sequential changes of data measured from the real world, based on grid cells. However, to construct processes from measured data, knowledge of specialists is needed. In addition, the sequential processes can represent only one aspect of an event.

In this paper, we propose a model for representing events from various aspects. In conventional researches, an event is represented as aggregations of sequential changes for geographic information based on the particular viewpoint for the events. However, in such an approach, an obtained event can represent only one aspect of the real phenomenon. We focus on not the partial order among changes of geographic information, but the structure of an event. The real phenomena cause a lot of changes of geographic object. In our approach, an event is defined as a situation of the space in one time interval, and represented as a set of geographic objects. Based on our representation of an event, it is possible to represent features of an event dynamically corresponding to various viewpoints for the event by referring to the set of geographic object concerned to the event.

This paper is organized as follows. In Section 2, we show a framework and provide a formal definition of our model. Then we discuss a handling mechanism of events. A prototype system based on our model is presented in Section 3. Finally, in Section 4 we state the conclusion of this paper and the future work.

2 Representation of Event

We discuss a representation model of events in order to extract features of an event dynamically according to viewpoints for the event. The feature of an event is regarded as one aspect of the event, and the event has various aspects. For example, the precipitation of a typhoon every one hour and the precipitation of the same typhoon every day show different aspects.

In this paper, the space is defined as a set of spatio-temporal objects. A spatio-temporal object is a geographic object such as a building, a river, etc. It has the lifespan and changes during the lifespan. An event represents a situation in space and time. Therefore, we treat an event as a set of spatio-temporal objects that are concerned to the event, based on the definition of the space. In our representation, a spatio-temporal object concerned to an event is managed as an “aspect” of the spatio-temporal object. In other words, an aspect of a spatio-temporal object indicates a condition of the spatio-temporal object when it is involved in the event. The features of an event are aggregated from a set of aspects concerned to the event dynamically. By using our representation of an event, it is possible to provide users with features reflecting the viewpoint for an event dynamically (Figure 1).

In addition, it is also important to look upon each event as an object.

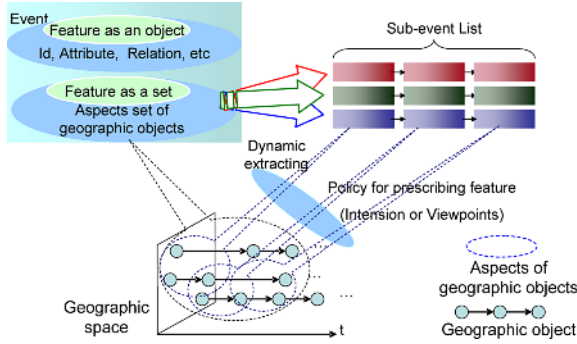


Fig. 1. Conceptual structure of our model

2.1 Formal Definition of Our Model

Spatio-temporal Object. A spatio-temporal object is an object with the shape and the lifespan such as a building, a road, and so on. Properties of a spatio-temporal object can be changed during the lifespan. The values of the properties are managed with the valid time. A spatio-temporal object o is described as follows.

$$\begin{aligned}
 o &= (id_o, t_s, t_e, ATTR) \\
 ATTR &= \{AT_1, \dots, AT_n\} \\
 AT_i &= \{(val_0, t_0, t_1), \dots, (val_{m-1}, t_{m-1}, t_m)\}, t_0 = t_s \wedge t_m = t_e
 \end{aligned} \tag{1}$$

Here, id_o is the identifier of o . t_s and t_e represent the onset time and the termination time of o , respectively. $ATTR$ is the attributes of o . Each AT_i is an attribute corresponding to the type of o , and represented as a set of tuples in which a value of the attribute with the valid time is represented.

Aspect of Spatio-temporal Object. We introduce an aspect that is a segment of a spatio-temporal object. It indicates a condition of a spatio-temporal object in one time interval. In other words, it provides a view for the spatio-temporal object. Assumed that a time interval is (t_{st}, t_{et}) , the aspect as of $o = (id_o, t_s, t_e, ATTR)$ is represented as follows.

$$as = (id_{as}, id_o, t_{st}, t_{et}, ATTR_{as}), t_s < t_{st} \wedge t_{et} < t_e \tag{2}$$

id_{as} is the identifier of as . t_{st} and t_{et} represent a focused time interval in the lifespan of o . as can exist in the lifespan of o . $ATTR_{as}$ is the inheritance of attributes of o , and it can be referred during the focused time interval (t_{st}, t_{et}) .

Event. An event consists of a set of aspects of spatio-temporal objects and represents a situation in space and time. In addition, it is also important to look upon each event as an object. There is very close similarity between the event and the object[4]. An event could

- have property,
- stand in relation to one another,
- possess determinate and objective identity, and
- take their place in predicates.

Namely, an event has some attributes, its lifespan, relationships to another event, and class hierarchy. An event ev is represented as follows.

$$ev = (id_{ev}, AS, attr_{ev}, t_s, t_e) \tag{3}$$

Here, id_{ev} is the identifier of ev . $attr_{ev}$ is the invariable attribute values of ev during the lifespan. AS is a set of aspects of spatio-temporal objects involved in ev . t_s and t_e is the onset time and the termination time of ev .

2.2 Event Handling

There are various aspects for an event, depending on standpoints. In this section, manipulation methods to extract various features reflecting standpoints for an event are described. The feature of an event can be aggregated from a set of aspects of spatio-temporal objects involved in the event. For example, the feature such as “destruction of 50 or more buildings” is extracted from the set of aspects of spatio-temporal objects involved in the event.

In our representation of events, it is possible to easily divide an event into a set of sub-events along time and space units. These functions define the domain for statistical calculation with space and time unites. To manage time and space units, we introduce two structures: the directory tree and the time tree. With these trees, our manipulation functions provide a capability to represent features of an event with various granularities. In addition, we address functions for manipulating events as an object.

Directory Tree and Time Tree. The directory tree represents a hierarchical structure of administrative units. A directory tree is obtained by recursively decomposing the administrative districts into a sequence of increasingly finer districts. Namely, a set of regions corresponding to administrative units is managed by a tree structure with inclusive relation among administrative districts. The directory tree is as follows.

$$DT = (\Sigma_s, \leq_s) \tag{4}$$

Here, $\Sigma_s = \{\sigma_{si} | i = 0, \dots, n\}$ is a set of administrative units, and \leq_s is an inclusive relation on Σ_s . In addition, the depth of σ_{si} from the root is described as $sl(\sigma_{si})$, and defined as a level of space. The level of district corresponds to the level of an administrative unit such as “country”, “prefecture”, “city”, and so on. We give an example in Figure 2. Assumed that there is a set of regions, $\Sigma_s = \{Japan, Aichi, Nagano, Nagoya, Ichinomiya, \dots\}$, these administrative districts are arranged into a directory tree. The levels of spaces are represented such as $sl(Japan) = country$, $sl(Aichi) = sl(Nagano) = prefecture$, $sl(Nagoya) = sl(Ichinomiya) = city$, and so on.

The time tree can also be constructed on time domain (Figure 3). The time intervals corresponding to time units such as “month”, “day”, etc. are structured hierarchically. It is obtained by composing finer time intervals recursively.

$$TT = (\Sigma_t, \leq_t). \tag{5}$$

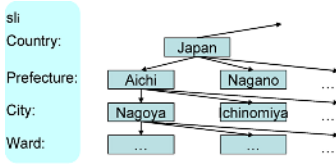


Fig. 2. Directory tree

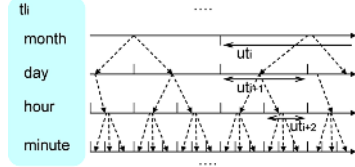


Fig. 3. Time tree

Here, $\Sigma_t = \{\sigma_{ti} | i = 0, \dots, m\}$ is a set of time intervals corresponding to time units. For example, “2006/3/3” is one of time intervals on the unit time “day”, and “2006/3/3 4 : 50 a.m.” is one of time intervals on the unit time “minute”. \leq_t is “contains” relation between time intervals noted in Allen’s Interval Algebra[5]. In addition, the depth of σ_{ti} from the root is described as $tl(\sigma_{ti})$, and defined as a level of time. The level of time corresponds to the unit time such as “month”, “week”, “day”, and so on. The unit time is described as ut_i like $ut_0 = minute$, $ut_1 = hour$, and so on.

Division of Aspect. An aspect $as = (id_{as}, id_o, t_{st}, t_{et}, ATTR_{as})$ of a spatio-temporal object can be divided with the time axis. Given the time level tl_i , and if the minimum time unit of as is finer than the unit time ut_i , then as can be divided into a set of aspects of spatio-temporal objects as follows.

$$\begin{aligned}
 div_{as}(as, tl_i) &= \{as_{i0}, \dots, as_{in}\} \\
 as_{ij} &= (id'_{as}, id_o, t_{ij}, t_{i(j+1)}, ATTR'_{as}) \\
 t_{i0} &= \lfloor t_{st} \rfloor, t_{i1} = t_{i0} + ut_i, \dots, t_{i(n+1)} = \lceil t_{et} \rceil
 \end{aligned}
 \tag{6}$$

Here, the division function for an aspect is described as div_{as} . as_{ij} is one of aspects produced by this function. $ATTR'_{as}$ is the inheritance of attributes of as . $\lfloor t \rfloor$ and $\lceil t \rceil$ are rounding up and rounding off in the time t .

We give an example of this function in Figure 4.

Division of Event on Time Domain. An event $ev = (id_{ev}, AS, attr_{ev}, t_s, t_e)$ can be divided with a time level tl_i , and this function is described as div_t . div_t is defined as follows.

$$\begin{aligned}
 div_t(ev, tl_i) &= \{ev_{ij} | j = 0, \dots, n\} \\
 ev_{ij} &= (id'_{ev}, AS_{ij}, attr_{ev}, t_{ij}, t_{i(j+1)}) \\
 t_{i0} &= \lfloor t_s \rfloor, t_{i1} = t_{i0} + ut_i, \dots, t_{i(n+1)} = \lceil t_e \rceil \\
 AS_{ij} &= \{as_x | as_x \in div_{as}(as \in AS, tl_i) \wedge as_x.t_{st} = t_{ij} \wedge as_x.t_{et} = t_{i(j+1)}\}
 \end{aligned}
 \tag{7}$$

A set of events $\{ev_{ij} | j = 0, \dots, n\}$ is sub-events created by dividing ev with the time level tl_i . An aspect set AS_{ij} which includes elements of each ev_{ij} is a subset of products by dividing each element($as \in AS$) of ev with the the same time interval (Figure 5).

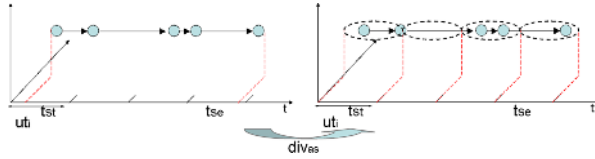


Fig. 4. Division function for aspect with time level

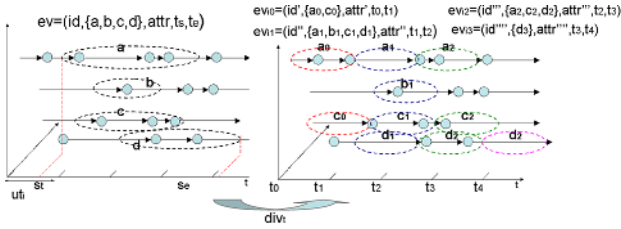


Fig. 5. Division function on time domain

Division of Event on Spatial Domain An event $ev = (id_{ev}, AS, attr_{ev}, t_s, t_e)$ can be divided with a spatial level sl_i . The function to divide an event spatially is described as div_s . This function is described as follows.

$$\begin{aligned}
 div_s(ev, sl_i) &= \{ev_{ij} | j = 0, \dots, n\} \\
 ev_{ij} &= (id'_{ev}, AS_{ij}, attr_{ev}, t_{sij}, t_{eij}) \\
 AS_{ij} &= \{as_x | inclusive(as_x, \sigma_{sj}) \wedge sl(\sigma_{sj}) = sl_i\}
 \end{aligned}
 \tag{8}$$

Here, $\sigma_{sj} \in \Sigma_s$ is an administrative district with spatial level sl_i . A set of events $\{ev_{ij} | j = 0, \dots, n\}$ is sub-events created by dividing ev with a spatial level sl_i . They have the lifespans described as (t_{sij}, t_{eij}) aggregated from AS_{ij} respectively. The elements of a sub-event ev_{ij} are aspects of spatio-temporal objects included by the region σ_{sj} .(Figure 6).

The execution results of functions div_t and div_s become the same, not depending on the application order of these functions.

Relation between Events as Objects. As stated above, the event also has features as an object. Consequently, we could define some relations between events[6].

- Is-a relation
- Part-of relation
- Causal relation.

Is-a relation represents different aspects of a single event. For example, an event “*typhoon*” can be treated as not only “*storm wind*”, but also “*disaster*”. Part-of relation represents a composite structure. For example, an event “*typhoon*” could consist of events “*downpour*”. A causal relation represents a causal affection among events. For instance, an event “*downpour*” may cause “*flooding*”. These relationships make it possible to retrieve and refine targeted events routinely.

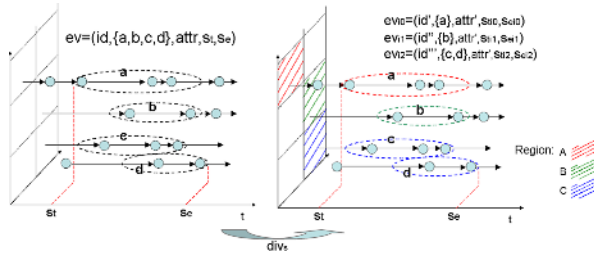


Fig. 6. Division function on spatial domain

Manipulation of Event Data. As mentioned previously, an event can be divided into sub-events with space and time units corresponding to the user’s viewpoints. In addition, selecting events by attribute values of events and relation between events can be performed. Moreover, we can observe a set of aspects of spatio-temporal objects involved in an event. Therefore, we can select the events by referring to the attributes of spatio-temporal objects driving from aspects managed in the event. Consequently, we can perform queries reflecting user’s intention such as “select disastrous fire events in which 50 or more buildings were destroyed at Nagoya city in one hour and continued for 12 or more hours”.

3 Prototype System

We developed a prototype system based on our data representation model. The spatial division function and the temporal division function were mounted on our system. Examples of computational results are presented in Figure 7. A typhoon event which was simulated on the map of Japan (Map 2500 of Geographical Survey Institute in Japan) was divided into sub-events with the time level “hour”.

The spatio-temporal objects involved in each sub-event were displayed emphatically. The tracks of the typhoon every one hour can be observed from these figures. The

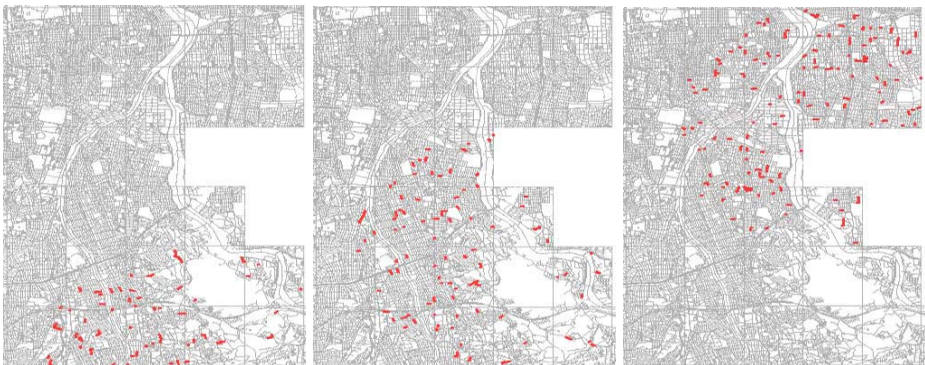


Fig. 7. Calculation results for division function with time level *hour*

feature of the event could be extracted in each time according to the time level “hour”. Consequently, we can obtain the features of the event such as “the maximum number of buildings which was destructed by the event in one hour”.

4 Conclusion

In this paper, we proposed a representation model of events in order to represent features of events dynamically. In our model, an event is represented as both a set and an object. In our representation model, an event can be divided with space and time units. These functions provide a capability to extract features in an event dynamically corresponding to user’s views.

In the future work, we must consider spatio-temporal relations between events topologically for more flexible retrieval. Moreover, we have to develop a data structure for efficient management of events and its elements. The visualization method for multiple events is also needed.

Acknowledgements

This work was supported in part by the 21st Century COE(Center of Excellence) Program for 2002, a project titled Intelligent Media Integration for Social Information Infrastructure, proposed by Nagoya University.

References

1. Peuquet, D., Duan, N.: An event-based spatiotemporal data model (estdm) for temporal analysis of geographical data. *International Journal of Geographical Information Systems* **9** (1995) 7–24.
2. Jiang, J., Chen, J.: Event-based spatio-temporal database design. In: *International Journal of Geographical Information Systems*. Volume 32/4., Germany (1998) 105–109.
3. Yuan, M.: Representing complex geographic phenomena in gis. *Cartography and Geographic Information Science* **28** (2001) 83–96.
4. Worboys, M.: Modelling changes and events in dynamic spatial systems with reference to socio-economic units. In: *Life and Motion of Socio-Economic Units*. Number 8 (2001) 129–138.
5. Allen, J.F.: Towards a general theory of action and time. *Artif. Intell.* **23** (1984) 123–154.
6. Ikezaki, M., Mukai, N., Watanabe, T., Ushiyama, T.: Event-based specification for controlling spatio-temporal changes of geographic situation. In: *International Special Workshop on Databases for Next Generation Researchers*, Japan (2005) 1249.

Using Multi-agent Systems to Manage Community Care

Martin D. Beer and Richard Hill

Web & Multi-Agents Research Group
Faculty of Arts, Computing, Engineering & Sciences
Sheffield Hallam University
Sheffield, United Kingdom
{m.beer, r.hill}@shu.ac.uk

Abstract. This paper discusses the evolution of the INCA demonstrator through a number of re-implementations that have investigated the applicability of various aspects of multi-agent technology to the management of Community Care. The latest experiences are described, making full use of the latest developments in semantic agents to provide a richer, more rigorous and highly scalable implementation than the previous demonstrators. This is presented in the context of a simulated real-world environment, based on knowledge of the actual operational environment within which the fully deployed agents would be expected to work. In particular the grouping of different communities of agents so that scalable solutions can be fully implemented and rigorously tested. So for example the agents that one would normally expect within a single household, including the home unit and the associated sensors, alarms etc. are treated as one group, and a care provider and the associated carers also. Not only does this reduce the communications overhead but also leads to simplifications in implementation as each class of agent only needs to be implemented once and individual instances are characterized by initial configurations and their interactions with their peers.

1 Introduction

The motivation for the INCA demonstrator came from the systematic modeling of Community Care to identify similarities and differences in that provision across Europe [6]. This provided a basic model that was clearly suitable for further analysis using the multi-agent paradigm [1] and led to an initial implementation, which although analyzed and designed on multi-agent principles was implemented as a distributed database application. This provided the basis for analyzing the data requirements using techniques that had been developed as part of the KRAFT project [7]. This led to the development of an INCA ontology that was mapped in this demonstrator into the underlying database schema. Whilst supporting the basic model as it then stood, it was clear that a distributed database approach was not going to be scalable to the levels required to represent real delivery scenarios, in which uncertainty and negotiation always play a major part.

This agent-based demonstrator has undergone a number of further re-implementations intended to test the applicability of different aspects of agent technology to this important and complex domain. These have been designed to test different aspects of the use of agent technology including:

- negotiation protocols as a means of matching capabilities of carers with current needs [2]
- use of large-scale agent networks to build scalable communities of care agents [3] based on the AgentCities [8] network and
- economic models to better represent the complex relationships that exist between carers, care providers and those being cared for [4]

The objective of the current implementation as described in this paper is to investigate the effectiveness of using the Jade Semantic Agent Framework [5] to build a relatively large-scale and diverse agent community and to investigate the properties that that community exhibits.

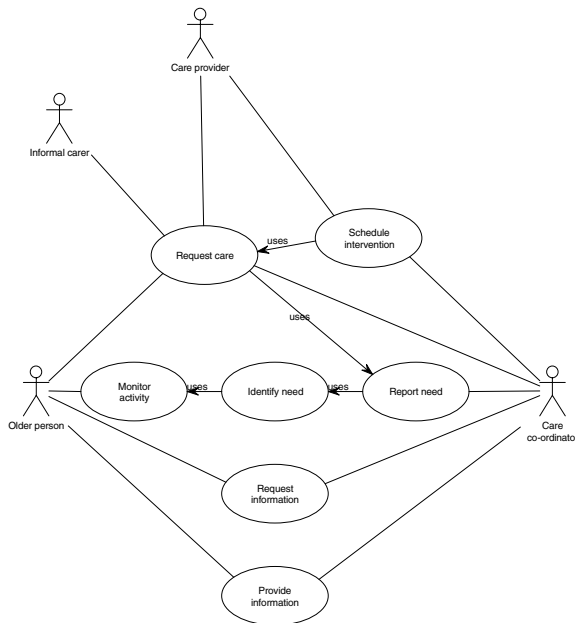


Fig. 1. The Routine Care Use Case

2 The Scenario

The current demonstrator is designed to service the use cases that handle routine (Figure 1) and Emergency (Figure 2) use cases as described in [1] over realistically configured networks. As broadband is spreading rapidly across most urban and semi-rural areas this is a realistic representation of the existing infrastructure in most areas of certainly the UK and increasingly the rest of Europe. These use cases form an important part of the provision of community care and are often treated separately, causing massive inefficiencies and duplication. For our purposes in this scenario we define routine care as the provision of a specified care package on a regular or routine basis based on the provisions of the Individual Care Plan or other recognized agreements and Emergency

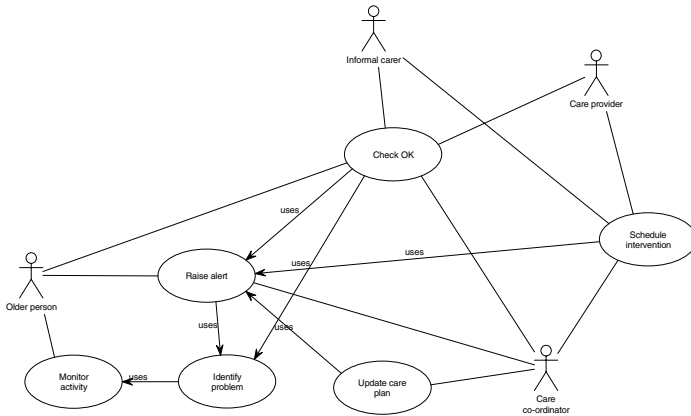


Fig. 2. The Emergency Care Use Case

Care as the unscheduled response alarms and events, however triggered. The first are characterized by their planned nature and predictability whereas the second are by their very nature totally unpredictable. Our hypothesis is that by linking the two considerable efficiencies can be achieved by:

- informing the next carer to visit of any emergency issues if they have the capabilities to deal with the issue and their visit is within a reasonable time, and
- canceling routine care when it is not required, for example when the person requiring it has been transferred to hospital, or when some other competent carer has already provided it while dealing with the emergency.

A typical example of the latter is when a home help regularly visits say at 10.00 am to get the person up and prepare their breakfast. Today some incident has happened and a relative has called in to see that they are all right on the way to work and has got them dressed and prepared their breakfast at the time of the visit. Another example is the provision of ambulances for routine hospital visits that are not needed because either the person is away or has already been delivered to the hospital for some other reason.

The scenario for the current demonstrator is that of the moderately sized town of Axebridge, with its own Social Services Department and a number of suppliers of different types of care and medical services. The Emergency services each have single control centres that cover the town. The system therefore needs to service several thousand home units, several care providers who deploy several hundred professional carers each with specific capabilities, a number of general practitioner surgeries, each of which provide medical and nursing services to their own patients and a number of pharmacies who dispense medicines prescribed by any of the doctors practicing in the town. This information is used to design the basic agents and to define the conversation classes that control the interactions between them.

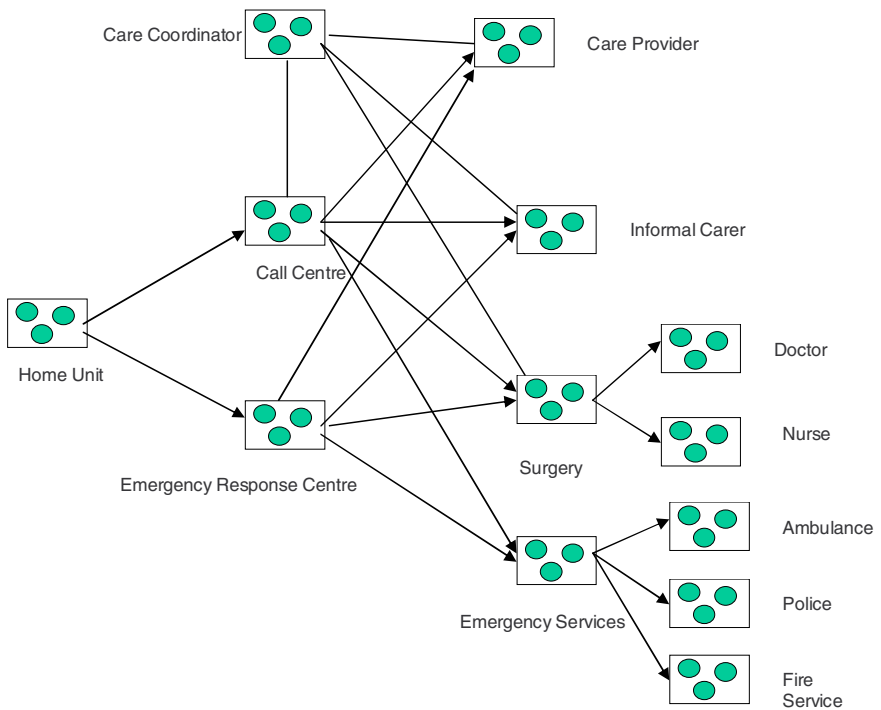


Fig. 3. The INCA Agent Architectural Model

3 The Architecture of the Demonstrator

Figure 3 shows the interactions between the different types of agent within the INCA environment. These include:

- the Home Unit and its associated sensor agents
- the care coordinators that develop and monitor the Individual Care Plans, allocate resources and manage the transfer of financial resources on the basis of the supply contracts etc.
- the care providers who manage these resources and employ the individual carers, and
- the carers who actually provide the care.

3.1 The Home Unit

Each Home unit agent acts as an interface to a collection of sensor and alarm agents associated with it. At present these are of two types:

- Alarm buttons which represent the red buttons provided with alarm systems, and can be both mobile and fixed.

- Temperature sensors that are used to ensure that the environment within the home is appropriate.

Further sensors can be added relatively easily to test further scenarios.

The Home Unit also maintains a list of carers due to visit (and also whether a carer is currently preset) based on information passed to it by the carers as they accept appointments. Should circumstances change, and a particular piece of care is no longer needed, the Home Unit cancels the appointment with the appropriate carer agent, which then notifies its related Care Provider agent that it is now free to take on other commitments.

3.2 Care Coordinators

The Care Coordinator agents monitor the delivery of care against the Individual Care Plans and the contractual arrangements with the various Care Providers. The transactional framework is used to select appropriate care providers when necessary, and to maintain economic balance by ensuring that those services that should be recharged are charged appropriately. As discussed elsewhere [4], this can be a complicated process as often the charging mechanism is not direct in that whether or not the client is charged will depend on factors such as:

- their ability to pay
- the service that they receive (in the UK for example, medical services are generally free at the point of delivery but social services are chargeable at standardized rates)
- the contractual arrangements with the various care providers and care coordinators
- whether informal or professional carers provide the necessary care

When an alarm is raised the care coordinator agents need to:

- identify the capabilities required and the speed of response necessary to respond to it based on the specifications given in the appropriate Individual Care Plan
- negotiate with the various care providers as to who has the capacity to deliver the necessary care within the required time frame
- select a carer and inform both them and the client of the arrangements
- monitor the delivery of both emergency and routine care to ensure that it is actually delivered, and in case of non-delivery restart the process with updated requirements
- manage the transactional arrangements so that the financial provisions of the care delivery are fully complied with

All these are complex tasks that are undertaken by a collection of agents located in a central location that has the necessary communication links with the more localized agents. This is in accordance with the existing organizational models that rely on centralized administrative services to undertake these tasks.

3.3 Carers and Care Providers

It is assumed that carers are available for set one hour slots throughout the day. No account is currently taken of traveling time, which is assumed to be included in the slots. While this is highly ineffective in a practical sense as a five minute visit would

be booked for a whole hour, the demonstrator is intended to investigate the operation of the transaction model at various load levels and to investigate the effects of capacity constraints. This means that each carer is available for a maximum of twenty four slots, for each of which the carer agent maintains the following properties:

- whether the carer is available or not available
- whether the carer is already booked or not
- the home (by means of the home agent identifier) that the carer is to visit and the capability required
- the capabilities of its carer

These give the carer agent sufficient information for it to maintain an accurate diary for the carer. Pre-arranged visits are loaded from configuration files. These are displayed and can be updated by means of a graphical user interface, as shown in Figure 4. Carers

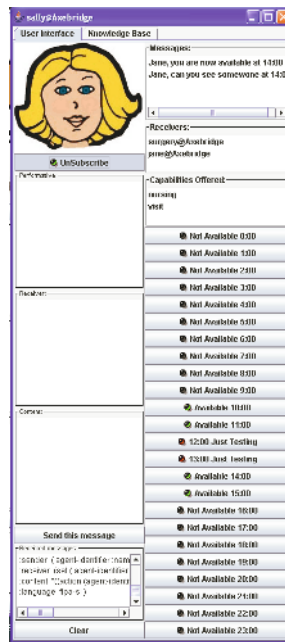


Fig. 4. An Example of the interface to a typical carer agent (in this case the District Nurse, Sally)

subscribe to their appropriate Care Provider so that they can update it on their diary as it changes. The Care Provider agent can therefore maintain a full list of availability and capabilities that it can provide through its associated carers. When a request comes for care, from a Care Coordinator agent, the Care Provider bids so long as it has that capability available within reasonable time on the following basis:

1. If a carer is already due to visit on the basis of $24/n$ where n =the number of slots before the carer is due ($n = 1$ means the next slot). This ensures that the first visiting carer with the appropriate capability accepts the commitment

2. If no carer is scheduled then on the basis of $p * 24/n$ where p is the nominal cost of that carer for the next available carer with the required capabilities. This allows the care provider to select the cheaper carer with a longer delay, if that is economically efficient
3. If no carer with the appropriate capabilities is available, then on the basis of $m * p/n$ where m is the number of appointments that need to be rearranged +1 to provide a means of meeting otherwise impossible requests

If the Care Provider's bid is accepted then it negotiates with its carer agents to make the necessary bookings and confirms the arrangement with the relevant Home Unit. Should it not be possible to fulfill the commitment at any stage, it is the care provider agent's responsibility either to arrange for another carer agent to take it on or to notify the appropriate care coordinator agent so that it can be reallocated. This provides a level of robustness and provides for recovery when plans go wrong.

The difficulty comes with informal carers who do not fit into this organizational structure. They could communicate directly with the Care Coordinator but this reduces flexibility, where for example a family of relatives who share the responsibility for caring. An alternative that is currently being investigated is that they are treated as a small care provider allowing negotiation to be undertaken between family members' agents as to who will respond to the client's needs.

4 Results and Further Work

Initial results with small groups of agents are encouraging. The mechanisms described work effectively to allocate tasks to appropriate carers in an efficient manner, taking into account the requirements for both routine and emergency care. Disturbances to schedules are handled effectively so that the care required is delivered as and when expected. The agents are currently implemented with Version 1.0 of the Jade Semantic Framework which lacks several important features, notably in the handling of the semantic knowledge base. Since this is used extensively to store and share appointment, care requirements and economic information, various work-arounds have been employed. At the time of writing, Version 1.2 has just been released as part of the 3.4 release of the JADE platform. It is intended to upgrade to this as a matter of urgency to take advantage of the additional and improved facilities offered.

The administrative arrangements for Community Care are complex because of the need to manage extremely complex delivery and recharging mechanisms, as care is provided on a means-tested basis. This gives a range of possibilities including:

- providing all care without payment of any sort
- providing different sorts of care with different payment schemes
- recharging the client for all care provided

These all have to be handled effectively by the Care Coordinator in accordance with each individual's circumstances.

5 Conclusions

The initial object of this study was to test the claim that the JADE Semantic Framework could greatly ease the burden of designing, building and deploying agents performing complex sets of tasks. Our development of both practice and economic models of community care provided a useful platform on which to test these issues. The availability of the semantic knowledge base removed the need for most agents to maintain access to separate data stores, considerably reducing the complexity of implementation. It is of particular benefit that each agent, even one that contains no local code, automatically maintains its own knowledge base which can be initialized at startup by querying the appropriate agents, as necessary. Design and implementation then centre on the filters needed to provide the actions required of the agent, rather than having to deal with the complexities of message handling as such, which are now generally hidden from the implementer.

References

1. Beer, M. D., Bench-Capon, T., & Sixsmith, A. (1999b), 'The Delivery of Effective Integrated Community Care with the aid of Agents', Proceedings of *ICSC99*, Hong Kong, December 1999. (Lecture Notes in Computer Science 1749, Springer-Verlag pp303-398)
2. Beer, M. D, Huang W. & Sixsmith, A. "Using Agents to Build a Practical Implementation of the INCA (Intelligent Community Alarm) System", in L. C. Jain & Z. Chen, & N. Ichalkaranje, "Intelligent Agents & their Applications", Springer (2002), pp320-345.
3. Beer, M. D., Hill, R., & Sixsmith, A., "Building an Agent Based Community Care Demonstrator on a Worldwide Agent Platform", in *Agents and Healthcare*, Barcelona, February 2003, pp19-34, published in the *Witterstein series on Multi-agent Systems*.
4. Hill, R., Polovina, S. & Beer, M., (2005), "Managing Community Health-care Information in a Multi-Agent System Environment", Multi-Agent Systems for Medicine, Computational Biology and Bioinformatics (MAS*BIOMED), AAMAS'05, Utrecht, Netherlands, July 2005, 35-49.
5. Louis, V. & Martinez, T (2005), "'An operational model for the FIPA-ACL semantics", AAMAS-05 Agent Communication Workshop, Utrecht NL, July 2004.
6. Lunn, K., Sixsmith, A., Lindsay, A., Vaarama, M., "'Traceability in requirements through process modelling, applied to social care applications". *Information & Software Technology*, **45:15**, (2003), 1045-1052.
7. Preece, A. D., Hui, K-Y., Gray, W. A., Marti, P., Bench-Capon, T. J. M., Jones, D. M., & Cu, Z., (1999) 'The KRAFT Architecture for Knowledge Fusion and Transformation', *19th SGES International Conference on Knowledge-based Systems and Applied Artificial Intelligence (ES'99)*, Springer, Berlin.
8. Willmott, S. et al, (2002), "Agentcities Network Architecture", in *Proceedings of the first International Workshop on Challenges in Open Agent Systems*, July 2002

Predictive Adaptive Control of the Bispectral Index of the EEG (BIS) – Using the Intravenous Anaesthetic Drug Propofol

Catarina S. Nunes¹, Teresa F. Mendonça^{1,2}, Hugo Magalhães^{1,2},
João M. Lemos³, and Pedro Amorim⁴

¹ Faculdade de Ciências da Universidade do Porto, Departamento de Matemática Aplicada, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
ccnunes@fc.up.pt

² UI&D Matemática e Aplicações, Universidade de Aveiro, Portugal

³ INESC-ID/IST, Portugal

⁴ Serviço de Anestesiologia, Hospital Geral de Santo António, Portugal

Abstract. The problem of controlling the level of unconsciousness measured by the Bispectral Index of the EEG (BIS) of patients under anaesthesia, is considered. It is assumed that the manipulated variable is the infusion rate of the hypnotic drug propofol, while the drug remifentanyl is also administered for analgesia. Since these two drugs interact, the administration rate of remifentanyl is considered as an accessible disturbance. In order to tackle the high uncertain present on the system, the predictive adaptive controller MUSMAR is used. The performance of the controller is illustrated by means of simulation with 45 patient individual adjusted models, which incorporate the effect of the drugs interaction on BIS. This controller structure proved to be robust to the remifentanyl disturbance, different reference values and noise. A reduction of propofol consumption was also observed when comparing to the real clinical dose used for a similar BIS trend.

1 Introduction

Anaesthesia can be defined as the lack of response and recall to noxious stimuli, involving the use of three drugs, a muscle relaxant, an anaesthetic (hypnotic) and an analgesic. The bispectral index of the EEG (BIS) is a numerical processed, clinically-validated EEG parameter, used as an indicator of the level of hypnosis, measuring the degree of depression in the central nervous system. The BIS is a number between 0 and 100, where values near 100 represent an "awake" clinical state while 0 denotes the maximal EEG effect possible (i.e., an isoelectric EEG). In a surgery, the level of hypnosis should be driven to a value between 40-60 in a few (3 ~ 5) minutes, and kept there.

Overall, general anaesthesia consists of both loss of consciousness through the action of anaesthetic drugs, and the inhibition of noxious stimuli reaching the brain through the acting of the analgesics. The analgesic drug is of great importance since it affects the pharmacodynamics of the anaesthetic drug and

there is no clear indicator of the degree of pain. The analgesic and anaesthetic drugs are interconnected, since they interact with each other so as to achieve an adequate level of hypnosis (unconsciousness) and analgesia. It is known that remifentanyl (analgesic) and propofol (hypnotic, i.e. anaesthetic) potentiate their effects when applied together. In what concerns control, this means that, if the level of hypnosis is controlled by selecting the dose of propofol (manipulated variable), the dose of remifentanyl being administered may be considered as an accessible disturbance and its knowledge may be used to increase the controller's performance. Automatic control is playing an increasing role in biomedical applications in a diversity of fields [1,2]. This paper presents a feasibility study of the control of the BIS (level of hypnosis) exploring the above ideas. A simulation study of the control of BIS taking the dose of propofol as manipulated variable and the dose of remifentanyl as an accessible disturbance is presented. In order to tackle the high uncertain present on the system, the predictive adaptive controller MUSMAR [3] is used. Simulations performed on a nonlinear model relating BIS with the doses of propofol and remifentanyl yield results complying with the specifications. The BIS model is presented in section 2, including the pharmacokinetic and pharmacodynamic structure. Section 3 describes the structure of the adaptive controller. Section 4 presents the results of the simulations under different conditions. The conclusions are presented in section 5.

2 Bispectral Index (BIS) Model

The clinical data of 45 neurosurgeries were used in a previous studies [4,5] to test the model structure. The model parameters were adjusted to the individual patients during the first 15 minutes of induction of anaesthesia, and used to predict the BIS signal during surgery. The model results were validated for the 45 cases, using the real propofol and remifentanyl doses (ml/h). Figure 1 shows the real propofol and remifentanyl doses (ml/h), and BIS signal (filtered with a Butterworth filter of order 2) for Patient 8 as an example. In this case the BIS signal was maintained by the clinician around the level of 40, and the infusion rates of the drugs were changed accordingly. The maximum rate allowed by the syringe pumps is $1200 ml/h$. Figure 2 shows the block diagram of the BIS model. The objective is to describe the relationship between the drugs effect concentrations and its effect. The pharmacokinetic/pharmacodynamic (PKD) models of the two drugs use a 3-compartment model structure. For propofol, the PKD parameters from Marsh [6] were used, whereas for remifentanyl the parameters from Minto [7] were used. The PKD model for remifentanyl has its parameters adjusted to age, gender and lean body mass of the patients, whereas the PKD model for propofol only takes into consideration the patient's weight.

Bruhn et al. [8] used an interaction model to relate the electroencephalographic parameter values (including BIS) to the effect concentrations of propofol ($C_{e_p}(t)$ $\mu g/ml$) and remifentanyl ($C_{e_r}(t)$ ng/ml). This model was developed by Minto et al. in a previous study [9]. First, the effect concentrations were normalised to their respective potencies (EC_{50p} and EC_{50r} for propofol and

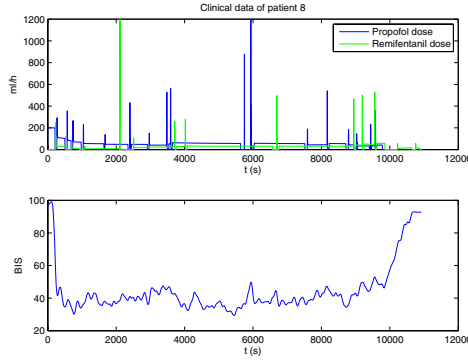


Fig. 1. Real propofol and remifentanyl doses (*ml/h*), and BIS signal (filtered) for the clinical case of Patient 8

remifentanyl, respectively), i.e. the effect concentration at half the maximal effect:

$$U_{remi}(t) = \frac{Ce_r(t)}{EC_{50r}} , \quad U_{prop}(t) = \frac{Ce_p(t)}{EC_{50p}} \tag{1}$$

where Ce_r and Ce_p are the respective effect concentrations of remifentanyl and propofol. The potency of the drug mixture depending on the ratio of the interacting drugs is modelled as (2).

$$\theta(t) = \frac{U_{prop}(t)}{U_{prop}(t) + U_{remi}(t)} \tag{2}$$

By definition, θ ranges from 0 (remifentanyl only) to 1 (propofol only). Thus, the concentration-response relationship for any ratio of the two drugs regardless of the type of interaction can be described as (3).

$$BIS(t) = BIS_0 \left(1 - \frac{((U_{prop}(t) + U_{remi}(t)) / U_{50(\theta)}(t))^\gamma}{1 + ((U_{prop}(t) + U_{remi}(t)) / U_{50(\theta)}(t))^\gamma} \right) \tag{3}$$

where BIS_0 is the effect at zero concentrations (e.g. $BIS_0 = 97.7$ for the case of BIS - monitor restriction), γ is the steepness of the concentration-response relation, and $U_{50(\theta)}$ is the number of units (U) associated with 50% of maximum effect at ratio θ . According to [9], (2) can be simplified to a quadratic polynomial (4).

$$U_{50(\theta)}(t) = 1 - \beta_{2,U50}\theta(t) + \beta_{2,U50}\theta^2(t) \tag{4}$$

3 The Adaptive Control Algorithm

The algorithm used is the predictive adaptive controller MUSMAR [3] that aims at minimizing a quadratic cost and reads as follows.

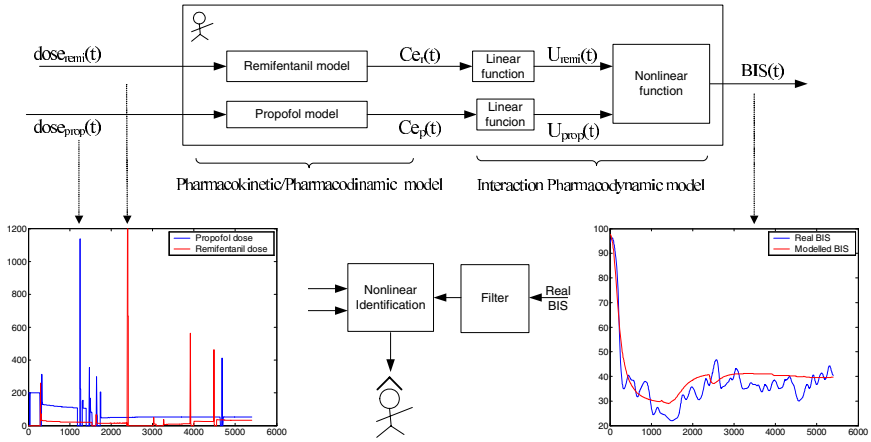


Fig. 2. Block diagram of the Bispectral Index (BIS) model, including the individual pharmacokinetic/pharmacodynamic models for propofol and remifentanyl, and the nonlinear interaction model describing the drugs synergistic effect on BIS. The graphs show the remifentanyl and propofol doses (on the left) and the real BIS signal versus the modelled BIS (on the right), for patient 23.

At the beginning of each sampling interval t (discrete time), recursively perform the following steps:

1. Sample plant output, $y(t)$ and compute the tracking error \tilde{y} , with respect to the desired set-point $ref(t)$, by:

$$\tilde{y}(t) = ref(t) - y(t) \tag{5}$$

2. Using Recursive Least Squares (RLS), update the estimates of the parameters $\theta_j, \psi_j, \mu_{j-1}$ and ϕ_{j-1} in the following sets of predictive models ($j = 1, \dots, T$):

$$\tilde{y}(t+j) \approx \theta_j u(t) + \psi'_j s(t) \quad , \quad u(t+j-1) \approx \mu_{j-1} u(t) + \phi'_{j-1} s(t) \tag{6}$$

where \approx denotes equality in least squares sense and $s(t)$ is a sufficient statistic for computing the control, hereafter referred to as the pseudo-state, given by

$$s(t) = [\tilde{y}(t) \dots \tilde{y}(t-n_a+1) \quad u(t-1) \dots u(t-n_b) \quad ref(t) \dots ref(t-n_g+1) \quad v(t) \dots v(t-n_v+1)] \tag{7}$$

with $v(t)$ as the accessible disturbance, and $u(t)$ as the controller output [3]. Since, at time t , $\tilde{y}(t+j)$ and $u(t+j)$ are not available for $j \geq 1$, for the purpose of estimating the parameters, the variables in (6) are delayed in block of T samples.

3. Apply to the plant the control given by

$$u(t) = f' s(t) + \eta(t) \tag{8}$$

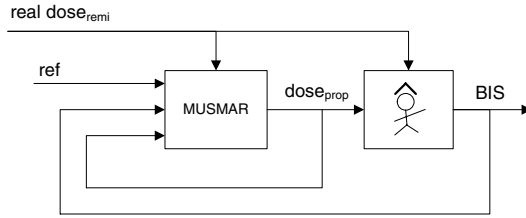


Fig. 3. Block diagram of the control system structure: the dose of remifentanil $dose_{remi}(t)$ is the real dose used for the identified patient (accessible disturbance - $v(t)$), the dose of propofol $dose_{prop}(t)$ is obtained from the MUSMAR controller (controller output - $u(t)$), the Bispectral Index $BIS(t)$ signal is the patient model output (controller input - $y(t)$), and $ref(t)$ is the reference target for BIS.

where η is a white dither noise of small amplitude and f is the vector of controller gains, computed from the estimates of the predictive models by

$$f = -\frac{1}{\alpha} \left(\sum_{j=1}^T \theta_j \psi_j + \rho \sum_{j=1}^{T-1} \mu_j \phi_j \right) , \quad \alpha = \sum_{j=1}^T \theta_j^2 + \rho \left(1 + \sum_{j=1}^{T-1} \mu_j^2 \right) \quad (9)$$

where ρ is a positive weight on the control action and α is the normalization factor. The choice of the variables and the number of their past samples entering $s(t)$ defines the structure of the controller. The choice of n_a and n_b should be such that it allows to capture the dominant dynamics of the system. It should be kept in mind that too big values of n_a and n_b imply more parameters to estimate and this may lead to identifiability problems, in turn causing loss of control performance. The pseudo state $s(t)$ includes samples accessible disturbances to embody feedforward action. This will be further discussed below. Figure 3 shows the block diagram of the control system structure used.

4 Results

A number of simulations [10], with a specific patient model (representing patient 23) have been conducted in order to find the best configuration defined by the MUSMAR parameters T , n_a , n_b , n_g and n_v with a sampling interval of 5s (the one used for real data collection). This lead to the choice of $T = 5$, $n_a = 9$, $n_b = 10$, $n_g = 1$, $n_v = 1$, $\rho = 0.0001$, $\sigma_\eta = 0.02$. The reasons for these values are presented and analysed in [10]. The controller gains obtained using the model of Patient 23 without the disturbance of remifentanil, were used as the initial gains for all the other 44 patient models. But this time, the real remifentanil dose (per patient) was used as the accessible disturbance. The cost function J_k was calculated for each simulation ($k = 1, \dots, 45$) after the initial 5min (10).

$$J_k = \frac{1}{n} \sum_{t=0}^n ((ref(t) - BIS(t))^2 + \rho dose_{prop}^2(t)) \quad k = 1, \dots, 45 \quad (10)$$

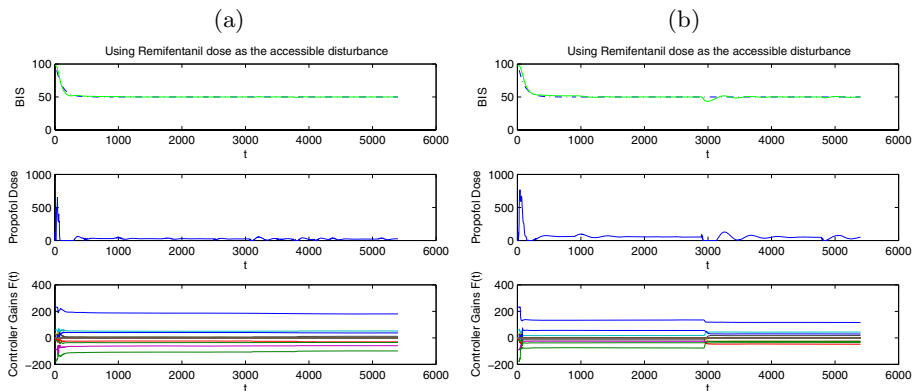


Fig. 4. BIS (model output), propofol dose (ml/h) (controller output), and the MUSMAR controller gains, for a reference BIS target value of 50: (a) using patient model 34 (b) using patient model 11

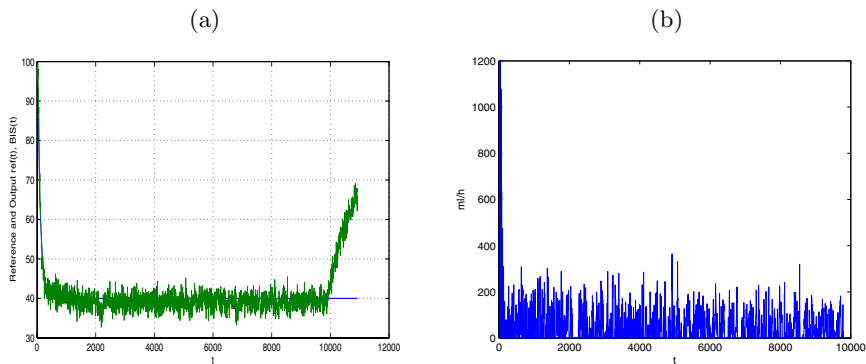


Fig. 5. Simulation using patient model 8 for a reference BIS target value of 40. Gaussian noise (zero mean and variance 3) was added to the model output (BIS) before the feedback to the MUSMAR controller: (a) Bispectral Index (BIS - model output) (b) Propofol dose (ml/h) (controller output).

The minimum and maximum values of J were 0.16 and 2.05 (mean value of 0.66), for patient model 34 and patient model 11, respectively. Figure 4 shows the BIS *versus* target reference, the propofol dose ($dose_{prop}(t)$ -controller output) and the MUSMAR controller gains (9) for the simulations with patient model 34 and 11. The controller is able to follow the BIS reference value of 50 in all simulations, and adequately adjusts its gains to cope with the patient’s intervariability and individual remifentanyl infusion scheme. In Patient 11 (figure 4 (b)) there was a big remifentanyl bolus at around 3000s which made the BIS signal decrease. Nevertheless, the controller is able to respond adequately to such a big disturbance and brings the BIS value back to the reference, adjusting the gains accordingly.

To further evaluate the performance of the MUSMAR controller, the target BIS reference value was changed so as to compared with similar BIS level in the individual clinical cases. In addition, gaussian noise (zero mean and variance 3) was added to the model output (BIS) before the feedback to the MUSMAR controller. Figure 5 shows the BIS signal for the simulation with patient model 8, considering a reference value of 40. Comparing figure 5 with figure 1, the BIS level is very similar. Figure 5 also shows the propofol dose (controller output) for the same simulation with patient model 8. The controller gains were initialised in the same way as the previous simulations, i.e. with the values adapted for patient model 23 without any disturbance. The real remifentanil dose of Patient 8 (figure 1) was used as the accessible disturbance.

The total amount of propofol used in the surgery of Patient 8 was 153.02ml , while the total amount of propofol used in the simulation was 146.41ml to reach a similar BIS value.

5 Conclusions

This control structure (MUSMAR) proved to be efficient to control the BIS value using the dose of the anaesthetic propofol, in extensive simulations using individual patient models adjusted to real clinical data. The fact that the controller gains can be initialised successfully using a specific patient model, with no over/undershoot about the reference value, and adapt to the individual patient intervariability and remifentanil dose (disturbance), shows the robustness of the overall structure. The controller responds to the remifentanil interference and is able to control the BIS value effectively, allowing for the effect of the synergistic interaction between the two drugs. The control structure was also robust to the addition of noise in the BIS signal. This is an important aspect, since in the real clinical data setup noise is present in all signals and a robust controller is necessary. Although the gains were initialised with values adjusted to a specific patient with a reference value of 50, the controller adjusted well to a different reference value and different patients. In the simulation with patient model 8, there was a reduction of 6.61ml (4.3%) in the total amount of propofol used, when comparing with the real amount used for that patient during surgery.

These results show that such a control structure could be adequate to control the BIS signal, during total intravenous anaesthesia with a combination of two drugs. It is important, not just that the patient model incorporates the remifentanil effect, but also that the controller is robust to such interaction and changes the propofol dose accordingly. Future work, should show the performance of the controller with dynamic changes in the reference level. Furthermore, for a clinical implementation of the controller an empirical system needs to be incorporated, so as to supervise the online adaptation of the parameters.

A key feature in the control of anaesthesia is to achieve a good rejection of disturbances caused by interfering actions from various sources. The practitioner knows what the surgeon is doing, as well as his own actions, and is therefore able to anticipate the corresponding induced disturbances, acting to counteract

them even before their effects are visible. A major challenge for the automation of anaesthesia consists in replicating similar performances. Clearly, this calls for the use of feedforward from measurable signals correlated with disturbances. The use of predictive control laws is also an immediate suggestion in this respect.

Acknowledgments

The authors wish to acknowledge the Portuguese Foundation for Science and Technology, for their support also under project POSC/EEA-SRI/57607/2004.

References

1. Bailey, J. M., Haddad, W. M.: Drug dosing control in clinical pharmacology. *IEEE Control Syst. Mag.* **25**(2005) 35–51.
2. Araki, M., Furutani, E.: Computer control of physiological states of patients under and after surgical operation. *Annual Reviews in Control* **29** (2005) 229–236.
3. Mosca, E., Zappa, G., Lemos, J.M.: Robustness of multipredictive adaptive regulators: MUSMAR. *Automatica* **25** (1989) 521–529.
4. Nunes, C.S., Mendonça, T., Ferreira, D.A., Antunes, L., Amorim, P.: Propofol and remifentanil pharmacokinetics/pharmacodynamics during induction may predict recovery of anaesthesia. *Anesthesiology* **103** (2005) A801.
5. Nunes, C.S., Mendonça, T., Antunes, L., Ferreira, D.A., Lobo, F., Amorim, P.: Modelling Drugs' Pharmacodynamic Interaction during General Anaesthesia: The Choice of Pharmacokinetic Model. Submitted to 6th IFAC Symposium on Modelling and Control in Biomedical Systems, MCBMS'06, Reims, France, September 20-22 (2006).
6. Marsh, B., White, M., Morton, N., Kenny, G.: Pharmacokinetic model driven infusion of propofol in children. *British Journal of Anaesthesia* **67** (1991) 41–48.
7. Minto, C., Schnider, T., Egan, T., Youngs, E., Lemmens, H., Gambus, P., Billard, V., Hoke, J., Moore, K., Hermann, D., Muir, K., Shafer, S.: Influence of age and gender on the pharmacokinetics and pharmacodynamics of remifentanil. I. Model development. *Anesthesiology* **86** (1997) 10–23.
8. Bruhn, J., Bouillon, T., Radulesco, L., Hoeft, A., Bertaccini, E., Shafer, S.: Correlation of approximate entropy, Bispectral index, and spectral edge frequency 95 (SEF95) with clinical signs of "anesthetic depth" during coadministration of propofol and remifentanil. *Anesthesiology* **98** (2003) 621–627.
9. Minto, C., Schnider, T., Short, T., Gregg, K., Gentilini, A., Shafer, S.: Response surface model for anesthetic drug interactions. *Anesthesiology* **92** (2000) 1603–1616.
10. Mendonça, T., Nunes, C.S., Magalhães, H., Lemos, J.M., Amorim, P.: Predictive Adaptive Control of Unconsciousness – Exploring Remifentanil as an Accessible Disturbance. Accepted in IEEE International Conference on Control Applications, CCA06, Munich, Germany, October 4-6 (2006).

Using Aggregation Operators to Personalize Agent-Based Medical Services

David Isern, Aïda Valls, and Antonio Moreno

Universitat Rovira i Virgili (URV)
Department of Computer Science and Mathematics
Artificial Intelligence Research Group, Banzai
43007 Tarragona, Catalonia (Spain)
{david.isern, aida.valls, antonio.moreno}@urv.cat

Abstract. In previous papers we introduced *HeCaSe2*, a multi-agent system that helps doctors to follow the automatic application of clinical guidelines to patients. In this paper we show how aggregation operators, based on fuzzy logic, may be integrated in this system in order to personalize some of its tasks. These operators take into account the patient preferences when several medical services propose different conditions under which a specific medical test can be performed. The paper describes how different proposals can be rated and ranked, and discusses the influence of two parameters (the set of linguistic preference values and the rating policy) on the results of the aggregation procedure.

1 Introduction

Any computer system designed to work in a medical setting has to take into account different issues; in [1] it was argued that the following ones suggest the appropriateness of the use of *agent technology* in the health care area:

- *Heterogeneous data*: medical centres generate data from very different sources (*e.g.* an X-ray image, a blood test, the result of a medical visit, etc.) and it is necessary to integrate them smoothly (*e.g.* new data should be added easily into the patient’s electronic medical record).
- *Autonomy*: services, departments, medical practitioners and patients are autonomous entities with their own knowledge, beliefs and goals. Any model of the activities within a medical centre should allow these entities to keep their autonomous behavior.
- *Distributed data*: the data related to a patient is usually distributed among different units (services, departments) of a hospital.
- *Complex coordination*: a medical centre has a large number of (human and physical) resources that have to be managed during *careflow* (*i.e.* the workflow processes involved in the provision of care, [2, 3]). All of them play a specific role within the medical centre organisation, and they must coordinate their activities to provide the best possible care to patients.

A *clinical guideline (CG)* indicates the protocol to be followed when a patient is diagnosed a certain illness (*e.g.* which medical tests have to be performed on the patient to get further data, or what steps have to be taken according to the results of the tests). Therefore, they provide very detailed information about the resources needed in the treatment of the patient [4]. Its adoption could improve the quality of patient assistance. Unfortunately, guidelines are not used extensively by practitioners due to two main reasons: *a)* the difficulty to adapt standard CGs to the particularizations of each sanitary centre, and *b)* the little tuning between the CG and the workflow of the professionals [5]. It can be argued that an agent-based automation of CGs could bring interesting benefits to the health care area, such as the following:

- Doctors are automatically reminded about the steps that should be followed in the treatment of a certain disease, and that reduces the possibility of them making errors or forgetting tasks to be done.
- Agents representing patients, doctors, departments and hospital services can automatically coordinate their activities to provide a fast care (*e.g.* by scheduling different tests to be performed on the patient on his behalf).
- These agents can apply AI techniques to solve their tasks, adding an intelligent component that improves their performance during negotiation and coordination processes.

The aim of the paper is to show how we can integrate aggregation techniques, based on fuzzy logic, into an agent-based system that provides health care services. A Multicriteria Decision Making process has been implemented in order to rank a set of alternatives (*e.g.* different possible dates in which a medical test can be performed) depending on the user's preferences [6].

The rest of the paper is organised as follows. The next section describes an agent-based system (*HeCaSe2*)¹, in which a set of agents has been designed to help doctors to follow the application of guidelines to particular patients. Section 3 explains how agents can use aggregation procedures to analyse a set of appointment proposals from the patient's point of view in order to personalize the system's performance. Section 4 shows the application of the aggregation method to some example data, and it analyzes how the change of some parameters influences in the final results. Finally, the last section details some lines of future work.

2 Guideline-Based Distributed Healthcare System

In previous works ([7]) we presented the main ideas underlying *HeCaSe2*, an agent-based distributed system that has the aim of easing the application of computer interpretable guidelines on particular patients. *HeCaSe2* is a multi-agent system that maps different entities in a healthcare organization (medical centres, departments, services, doctors, patients) as agents with different roles

¹ The work has been partially supported by K4CARE Project (IST-2004-026968).

and goals. This system provides interesting services both to patients (*e.g.* booking a visit with a doctor, or looking up the medical record) and to doctors (*e.g.* support in the application of a CG to a patient).

Guidelines are used to provide a high level supervision of the activities to be carried out to address a specific pathology. We use *PROforma* as the language to represent and share guidelines [8]. It defines four types of tasks: *i) actions*, that are procedures that have to be executed outside the computer, *ii) decisions*, that are used to choose a candidate from a given set of options using arguments pro and con, *iii) inquiries*, that are requests for information needed to execute a certain procedure, and *iv) plans*, that are a sequence of sub-tasks taking into account logical or temporal constraints. Thus, a guideline can be defined as a set of plans that are composed by actions, decisions and inquiries.

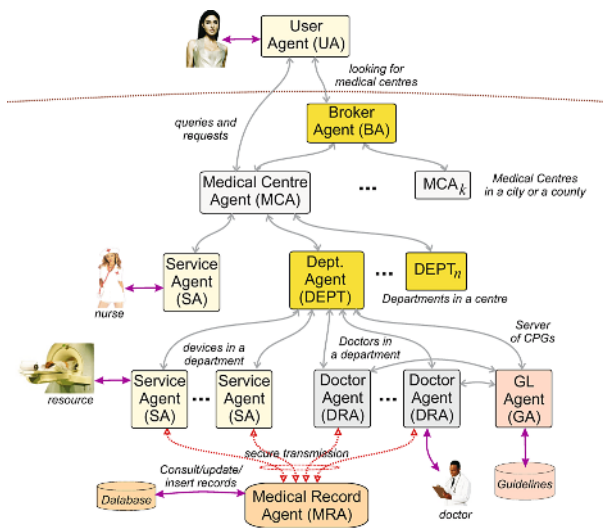


Fig. 1. HeCaSe2 agent-based architecture

The agents in the system (see Fig. 1) coordinate their activities in order to apply a guideline to a patient (always under the supervision of a doctor). The basic steps are the following (more details are given in [7]):

- When the doctor diagnoses that the patient has a certain disease, its associated *Doctor Agent* (DRA) requests from the *Guideline Agent* (GA) the guideline associated to that condition, and it starts to execute it.
- If the guideline needs some medical data, the DRA can request it from the *Medical Record Agent* (MRA), that has access to the electronic medical record of the patient.
- Sometimes an action has to be performed on the patient, or the guideline needs data that is not included in the medical record (*e.g.* the level of glucose in blood, which can be known with a blood analysis). In these cases, the

DRA has to contact *Service Agents* (SAs), from the same medical centre or other medical centres, that can provide a certain action or clinical test. As there will be different options for each action, we propose in the next section a personalization mechanism, based on aggregation procedures, that receives the service proposals from different SAs and analyses them, taking into account the user's preferences on several criteria. The user receives a ranked list of alternatives, from which he can choose the one that he prefers.

- Once a test has been performed, the result can be sent directly from the SA to the MRA, to be included in the patient's medical record, and the DRA that requested that data is informed so it can follow the application of the guideline on the patient with the new available data.

3 A Patient-Centered Ranking of Appointments

In order to build patient-oriented medical systems, we propose the use of intelligent decision-making techniques to help the patient as [9]. In particular, when the doctor, following a guideline, decides that some test must be performed, usually it is the patient who has to find an appointment with an external medical unit that can perform this test, and this is a problem that requires a lot of time and effort from the patient. We propose to automatize this process using the facilities of multi-agent technology and multicriteria decision analysis. In *HeCaSe2*, when some test t must be performed in one patient, the agent DRA_d begins a call for proposals with the SAs that can do task t . That message is sent to different agents in the medical centre and to SAs in other centres by means of the *broker agent* (Fig. 1). All SAs that can perform task t seek in its own agenda and send k proposals of possible appointments to the initiator agent DRA_d . A proposal contains the day, location and hour. After receiving all the answers from the SAs, DRA_d completes the proposal with some additional information, building a tuple $p_i = \langle day_of_week, centre, period_day, distance, delay_days \rangle$. In this tuple, we have three linguistic variables: *day_of_week*, *centre* (destination medical centre) and *period_day* (morning, afternoon, night), and two numerical ones: *distance* (kilometres from the origin centre to the destination) and *delay_days* (days to wait before performing t). Once we have a list of proposals, the system ranks them and only the best n options are shown to the patient. Then, the user can select the most appropriate appointment. The ranking is based on the user's preferences, which are stored in his profile. In the following sections we will describe the user's profile and the ranking technique, which is based on the Linguistic Ordered Weighted Averaging (LOWA) operator [6].

3.1 User's Profile

The user's profile is stored in each *User Agent* (UA). That profile contains the user's preference information for each attribute. This preference information is given by a utility function, such that $profile_i = \{U_{atr_h}, h = 1..m\}$. In the profile we can have linguistic and numerical preferences. Linguistic values are given to categorical attributes, and numerical scores to numerical attributes.

P=Perfect (0.925,0.95,1.0,1.0)
 VH=Very_High (0.8,0.825,0.925,0.95)
 H=High (0.675,0.7,0.8,0.825)
 AH=Almost_High (0.55,0.575,0.675,0.7)
 M=Medium (0.425,0.45,0.55,0.575)
 AM=Almost_Medium (0.3,0.325,0.425,0.45)
 L=Low(0.175,0.2,0.3,0.325)
 VL=Very_Low(0.05,0.075,0.175,0.2)
 N=None (0.0,0.0,0.05,0.075)

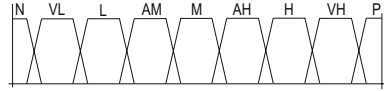


Fig. 2. A uniformly distributed ordered set of nine labels with its semantics

We will denote $S = \{s_i\}$, and $i \in \{0, \dots, T\}$ a finite ordered set of T linguistic labels whose semantics is given by fuzzy sets. Each label s_i is defined by a 4-tuple (x_0, x_1, x_2, x_3) , where x_1 and x_2 indicate the interval in which the membership function value is 1, and x_0 and x_3 are the bounds of the definition of a trapezoidal fuzzy membership function. For example, Fig. 2 shows an example considering nine symmetrically distributed fuzzy linguistic labels. Then, in the user’s profile we have a utility function $U_{atr_i}^L$ that associates each possible value of the categorical attribute atr_i to a label in S , indicating its preference score. For numerical attributes, we have a utility function $U_{atr_i}^N$ that receives the numerical value r of the corresponding attribute, and compares r with the preferred value of the user r_{user} . The utility function of the i^{th} attribute takes $k_i \approx \frac{10}{(max_{atr_i} - min_{atr_i})}$.

$$\begin{aligned}
 U_{atr_i}^N : \mathbb{R} &\rightarrow [0, 1] & U_{atr_j}^L : String &\rightarrow S \\
 r &\rightarrow 1/e^{k|r_{user}-r|} & str &\rightarrow s_i
 \end{aligned}$$

3.2 The Aggregation Operator

To rank the set of alternatives we use a decision-making process with two stages: rating and ranking. The rating of each alternative is done in three steps (Fig. 3):

- Step 1)* All the values describing an alternative, p_i , are transformed into preference values in the domain $([0, 1] \cup S)$ by applying the appropriate utility functions U_{atr_h} as defined in §3.1.
- Step 2)* The numerical preferences in $[0, 1]$ are transformed into the linguistic domain S by means of a particular numerical-linguistic transformation function defined in [10] (linguistic preferences are left without changes), obtaining the transformed vector called $alt_i = \{a_k\}$ ($a_k \in S$).
- Step 3)* The linguistic preferences in alt_i are aggregated using the LOWA operator, obtaining a linguistic rating.

Finally, all alternatives can be ranked using the rating values. Then, a filtering is performed to show to the user only the best alternatives, so that he can confirm one of them (the communication from the UA to the selected SA through the DRA is shown in Fig. 3).

The problem of aggregating information has been widely studied [11]. There exist several methods to aggregate numerical values as well as linguistic terms.

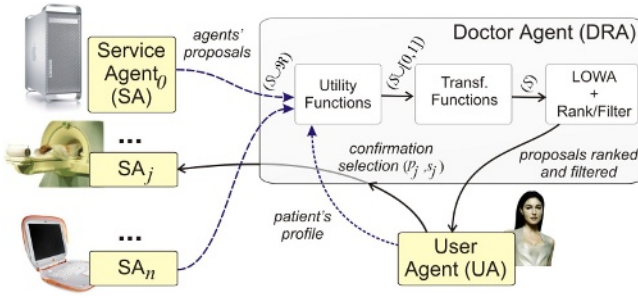


Fig. 3. Aggregation of information process

The family of OWA operators are in the class of mean operators, they are idempotent, monotonic and commutative. They are useful to adjust the degree of conjunction and disjunction implicit in any aggregation. This is done by means the use of linguistic quantifiers (expressed as a set of weights) that permits to define different aggregation policies. In [12] different fuzzy majority-based policies are identified, such as “most”, “at least half” or “as many as possible”.

The LOWA aggregation operator ϕ was defined in [6]. It is an extension of the OWA operator to deal with linguistic variables. The operator ϕ aggregates a set of labels $A = \{a_1, \dots, a_m\}$, where $a_i \in S$, with respect to a set of weights $W = \{w_1, \dots, w_m\}$ such that $w_i \in [0, 1]$ and $\sum_i w_i = 1$. Those weights specify the decision-maker policy.

$$\begin{aligned} \phi(a_1, \dots, a_m) &= W \cdot B^T = \mathcal{C}^m \{w_k, b_k, k = 1, \dots, m\} \\ &= w_1 \odot b_1 \oplus (1 - w_1) \odot \mathcal{C}^{m-1} \{\beta_h, b_h, h = 2, \dots, m\} \end{aligned}$$

where $\beta_h = w_h / \sum_2^m w_h, h = \{2, \dots, m\}$ and $B = \{b_1, \dots, b_m\}$ is a permutation of the elements of A, such that $B = \sigma(A) = \{a_{\sigma(1)}, \dots, a_{\sigma(m)}\}$, where $a_{\sigma(j)} \leq a_{\sigma(i)} \forall i \leq j$. \mathcal{C}^m is the convex combination operator of m labels; if $m = 2$, then $\mathcal{C}^2 \{w_i, b_i, i = 1, 2\} = w_1 \odot s_j \oplus (1 - w_1) \odot s_i = s_k, s_i, s_j \in S, (i \leq j)$ such that, $k = \min\{T, i + \text{round}(w_i \cdot (j - i))\}$. If $w_j = 1$ and $w_i = 0$ with $i \neq j$, then $\mathcal{C}^m \{w_i, b_i, i = 1, m\} = b_j$

4 Analysis of the Parameters and Results

In this section we give some insight in some points of our personalization method for this particular application. First of all, we will consider the initialization of the utility functions used in the patient’s profile $\mathcal{U}_{atr} = (\mathcal{U}_{atr}^N \cup \mathcal{U}_{atr}^L)$. When a profile is created, the system does not know any preference, so preferred time and days is considered to be zero and a *medium* preference score is given to each of the qualitative terms. After modifying the patient’s profile, we could have a situation like this:

$$\begin{aligned} & \text{delay_days} = \langle 0.0 \rangle; \text{distance} = \langle 0.0 \rangle; \text{centre} = \langle (MCBona, N) (MCBorges, M) \\ & (MCConst, VH) (MCMorell, P) (MCGimb, M) (MCHospi, AM) (MCJaume, L) \end{aligned}$$

$(MCLlib, L)$; $day_of_week = \langle (Sun, VL) (Mon, L) (Tue, M) (Wed, VH) (Thu, VH) (Fri, H) (Sat, VL) \rangle$; $period_day = \langle (Morning, H) (Afternoon, L) (Night, VL) \rangle$

The second aspect to analyse is the selection of the set of labels S . We tested the system with different number of labels and with different membership functions (see Table 1). The last point to consider is the LOWA weighting vector W . In Table 1 different configurations of that vector can be observed. Different policies giving priority to the low or to the high values give different results. For instance, alternatives 1 and 3 show the influence of W obtaining both different ratings and positions. Now, we are going to explain the results obtained in the following scenario. Let's consider the previous patient profile and the following 5 possible appointments:

- $p_0 : \langle 2.0, 1.2, MCBorges, Wed, Morning \rangle$ $p_3 : \langle 5.0, 1.2, MCBorges, Sat, Nighth \rangle$
- $p_1 : \langle 1.0, 8.0, MCCConst, Mon, Afternoon \rangle$ $p_4 : \langle 9.0, 17.0, MCHospi, Fri, Afternoon \rangle$
- $p_2 : \langle 4.0, 9.0, MCBona, Thu, Afternoon \rangle$

Table 1. Aggregation-based results obtained in different scenarios

Entry	Alternatives (A)	Conditions	Rating($\phi(alt_i)$)	Ranking
1	$alt_0 = \langle AH H M VH H \rangle$	$W, most$	$\langle H \rangle$	H: alt_0
	$alt_1 = \langle H VL VH L L \rangle$	$W_5 = \langle .0, .2, .4, .4, .0 \rangle$	$\langle AM \rangle$	AM: alt_1, alt_3
	$alt_2 = \langle AM VL N VH L \rangle$	$S_9, symmetric$	$\langle L \rangle$	L: alt_2, alt_4
	$alt_3 = \langle L H M AM VL \rangle$		$\langle AM \rangle$	
	$alt_4 = \langle VL N AM H L \rangle$		$\langle L \rangle$	
2	$alt_0 = \langle AH H M VH H \rangle$	$W, at least half$	$\langle H \rangle$	H: alt_0, alt_1
	$alt_1 = \langle H VL VH L L \rangle$	$W_5 = \langle .4, .4, .2, .0, .0 \rangle$	$\langle H \rangle$	AH: alt_2, alt_3
	$alt_2 = \langle AM VL N VH L \rangle$	$S_9, symmetric$	$\langle AH \rangle$	M: alt_4
	$alt_3 = \langle L H M AM VL \rangle$		$\langle AH \rangle$	
	$alt_4 = \langle VL N AM H L \rangle$		$\langle M \rangle$	
3	$alt_0 = \langle AH H M VH H \rangle$	$W, mean$	$\langle H \rangle$	H: alt_0
	$alt_1 = \langle H VL VH L L \rangle$	$W_5 = \langle .2, .2, .2, .2, .2 \rangle$	$\langle M \rangle$	M: alt_1, alt_3
	$alt_2 = \langle AM VL N VH L \rangle$	$S_9, symmetric$	$\langle AM \rangle$	AM: alt_2, alt_4
	$alt_3 = \langle L H M AM VL \rangle$		$\langle M \rangle$	
	$alt_4 = \langle VL N AM H L \rangle$		$\langle AM \rangle$	
4	$alt_0 = \langle H VH H VH P \rangle$	$W, as many as possible$	$\langle VH \rangle$	VH: alt_0
	$alt_1 = \langle VH VL VH L L \rangle$	$W_5 = \langle .0, .0, .2, .4, .4 \rangle$	$\langle H \rangle$	H: alt_1, alt_3
	$alt_2 = \langle L VL N VH L \rangle$	$S_7, symmetric$	$\langle M \rangle$	M: alt_2, alt_4
	$alt_3 = \langle L VH H M VL \rangle$		$\langle H \rangle$	
	$alt_4 = \langle VL N M H L \rangle$		$\langle M \rangle$	
5	$alt_0 = \langle H H M H VH \rangle$	$W, as many as possible$	$\langle H \rangle$	H: alt_0
	$alt_1 = \langle H VL H QL QL \rangle$	$W_5 = \langle .0, .0, .2, .4, .4 \rangle$	$\langle M \rangle$	M: alt_1, alt_2, alt_3
	$alt_2 = \langle L VL N H QL \rangle$	$S_7, non-symmetric$	$\langle M \rangle$	L: alt_4
	$alt_3 = \langle QL H M L VL \rangle$		$\langle M \rangle$	
	$alt_4 = \langle VL N L M QL \rangle$		$\langle L \rangle$	

The results of applying the LOWA operator are given in the rating column of Table 1. We compare different situations by changing both the weighted vector W and the label set S . In all cases we obtain the same *best* alternative: alt_0 . If we observe the first column, this tuple has a lot of good rated labels, obtaining a good rate at the end. In contrast, the worst alternative is alt_4 , because it has most of the attributes bad labelled. The rest of alternatives change the ranking position according to the conditions set by the parameters, because they have a mixture between good and bad labelled attributes.

5 Conclusions and Future Work

This paper presents a distributed patient-centered system that facilitates to the user the selection of appointments when a clinical test is required. We have explained how we combine (1) the use of a multi-agent system based on medical guidelines with (2) decision-making techniques that use two types of values, linguistic and numerical. The main part of the paper has been devoted to explaining the rating and ranking of the appointments. We proposed the use of the LOWA operator that uses fuzzy linguistic variables.

We are now busy designing a method to monitor the behavior of the patient (which of the proposed appointments is selected) in order to learn how the user's preferences evolve and improve the ranking over time. At the moment, the profile is updated by hand. Another future research line is the study of the adequacy of giving different weights to the attributes.

References

- [1] Nealon, J.L., Moreno, A.: Agent-Based Applications in Health Care. In: Applications of Software Agent Technology in the Health Care Domain. Birkhäuser Verlag (2003) 3–18
- [2] Quaglini, S., Stefanelli, M., Cavallini, A., Micieli, G., Fassino, C., Mossa, C.: Guideline-based careflow systems. *Artif Intell Med* **20** (2000) 5–22
- [3] Seroussi, B., Bouaud, J., Chatellier, G.: Guideline-based modeling of therapeutic strategies in the case of chronic diseases. *Int J Med Inform* **74** (2005) 89–99
- [4] IOM: Clinical Practice Guidelines: Directions for a New Program. Institute of Medicine (IOM), National Academy Press, Washington, D.C (1990)
- [5] Quaglini, S., Ciccarese, P., Micieli, G., Cavallini, A.: Non-Compliance with Guidelines: Motivations and Consequences in a case study. In: Symposium on Computerized Guidelines and Protocols, CGP 2004, IOS Press (2004) 75–87
- [6] Herrera, F., Herrera-Viedma, E.: Linguistic decision analysis: steps for solving decision problems under linguistic information. *Fuzzy Sets and Systems* **115** (2000) 67–82
- [7] Isern, D., Moreno, A.: Distributed guideline-based health care system. In: Intel. Systems Design and Applications, ISDA-2004, IEEE Press (2004) 145–150
- [8] Peleg, M., Tu, S., et. al: Comparing Computer-Interpretable Guideline Models: A Case-Study Approach. *J Am Med Inf Ass* **10** (2003) 52–68
- [9] Ghinea, G., Magoulas, G.: Intelligent Protocol Adaptation in an Enhanced Medical e-Collaboration Environment. *Int J Art Int Tools* **13** (2004) 199–218
- [10] Delgado, M., Herrera, F., Herrera-Viedma, E., Martínez, L.: Combining numerical and linguistic information in group decision making. *Inf Sci* **107** (1998) 177–194
- [11] Grabisch, M., Orlovski, S., Yager, R.: Fuzzy aggregation of numerical preferences. In Slowinski, R., ed.: *Fuzzy sets in decision analysis, operations research and statistics*. Kluwer Academic (1999) 31–68
- [12] Yager, D.R.: On Ordered Weighted Averaging Aggregation. Operators in Multi-criteria Decision making. *IEEE Trans Syst Man Cybern* **18** (1988) 183–190

Shifting Patterns Discovery in Microarrays with Evolutionary Algorithms*

Beatriz Pontes¹, Raúl Giráldez², and Jesús S. Aguilar–Ruiz²

¹ Department of Computer Science, University of Seville
Avenida Reina Mercedes s/n, 41012 Sevilla, Spain
bepontes@lsi.us.es

² Area of Computer Science, University of Pablo de Olavide
Ctra. de Utrera, km. 1, 41013, Sevilla, Spain
{rgirroj, sjsagurui}@upo.es

Abstract. In recent years, the interest in extracting useful knowledge from gene expression data has experimented an enormous increase with the development of microarray technique. Biclustering is a recent technique that aims at extracting a subset of genes that show a similar behaviour for a subset conditions. It is important, therefore, to measure the quality of a bicluster, and a way to do that would be checking if each data submatrix follows a specific trend, represented by a pattern. In this work, we present an evolutionary algorithm for finding significant shifting patterns which depict the general behaviour within each bicluster. The empirical results we have obtained confirm the quality of our proposal, obtaining very accurate solutions for the biclusters used.

Keywords: Gene Expression Data, Biclustering, Evolutionary Algorithm, Shifting Pattern.

1 Introduction

Microarray data are widely used due to the great potential in different biomedical fields as gene expression profiling, facilitating the prognosis and the discovering of subtypes of diseases. A microarray is a set of DNA/RNA sequences, where the gene expression data are organized in a two-dimensional array. Columns represent genes and rows represent experimental conditions, so that, each element in the matrix refers to the expression level of a particular gen under specific conditions.

In order to extract relevant knowledge from microarray expression data, clustering techniques have been applied [4]. The main application of this techniques is to group genes together according to any specific algorithm or mathematical formula related to their functional similarities over all conditions. However, relevant genes are not necessarily related to every condition [15]. Thus, biclustering [12] is a variation of clustering where the process consist of simultaneously

* This research was supported by the Spanish Research Agency CICYT under grants TIN2004-00159 and TIN2004-06689C0303.

mining columns and rows of the matrix. In the context of microarrays study, it is applied to identify groups of genes which exhibit similar behaviour under a specific subset of experimental conditions [8]. Bicluster analysis [14] takes into account the fact that not every gene in a microarray may be relevant for all the conditions, thus addressing in the two dimensions simultaneously the clustering problem. Biclustering methods for biological data analysis have been widely studied in the literature [5,6,13].

In [8], Cheng and Church showed that some biclusters should contain a subset of genes showing similar behaviour and not necessarily similar values, or in other words, such genes could follow a pattern of behaviour. Thus, two types of patterns [1], such as shifting and scaling patterns, should be found in biclusters. These patterns can be very useful for different aspects as to find more genes or conditions that should be included in a bicluster, or simply to describe the common conduct of the genes belonging to a certain bicluster.

In this work, we address the finding pattern problem with Evolutionary Algorithms (AE), which has been proven to have an excellent performance on highly complex optimization problems. Thus, we present a new EA-based tool for finding the shifting patterns which represents more accurately the behaviour of the genes in a given bicluster. The experimental results show that our approach obtains shifting patterns with an excellent performance.

The paper is organized as follows: in Section 2, an overview on patterns from gene expression data is presented. We provide a description of our algorithm in Section 3 and the experimental results are shown in Section 4. Finally, the last section summarises the main conclusions of this work.

2 Patterns from Biclustering in Microarrays

The genes included in a bicluster could follow a pattern of behaviour [8]. This idea was formally described in [1], where two kind of patterns were defined.

Let \mathcal{M} be a microarray with N rows (conditions c_i , with $1 \leq i \leq N$) and M conditions (genes g_j , with $1 \leq j \leq M$). Each element in the matrix will be represented as $v_{ij} \in \mathcal{M}$. Also, let $\mathcal{B} \subseteq \mathcal{M}$ be a bicluster made up of $n \leq N$ conditions and $m \leq M$ genes. Each element in the bicluster will be represented as $w_{ij} \in \mathcal{B}$. With these premises, shifting and scaling patterns are defined as follows [1]:

A bicluster \mathcal{B} shows a *shifting pattern* when the values w_{ij} can be obtained by adding a certain value β_i , constant for the i^{th} condition, to a typical value (π_j) for the j^{th} gene. Analogously, the definition of scaling pattern is similar to the scaling by replacing the additive factor β_i with multiplicative value α_i . Formally, a bicluster follows a shifting pattern (Equation 1) or a scaling pattern (Equation 2) when it follows the expressions:

$$w_{ij} = \pi_j + \beta_i + \xi_{ij} \quad (1)$$

$$w_{ij} = \pi_j \times \alpha_i + \xi_{ij} \quad (2)$$

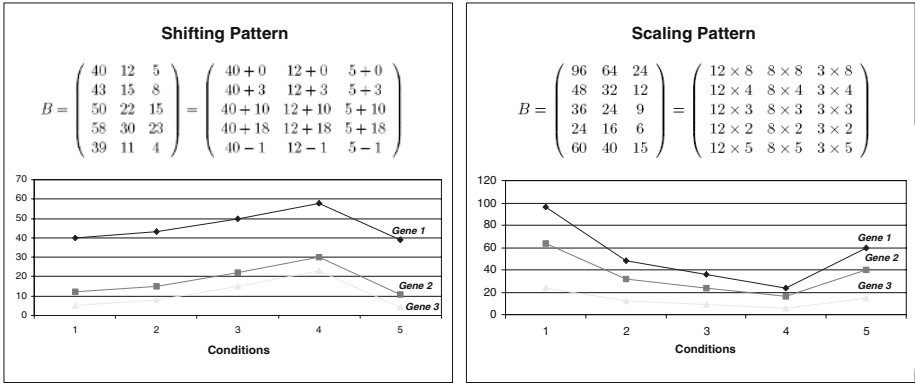


Fig. 1. Examples of biclusters with shifting and scaling patterns

where w_{ij} is the value for gen j and condition i within a bicluster; π_j is the fixed value for j^{th} gene; β_i in Equation 1 is the shifting value for condition i ; α_i is the scaling factor for i^{th} condition in Equation 2; finally, ξ_{ij} is the error that the pattern makes for w_{ij} value. In both cases, when such error is 0 for all bicluster’s values, we say we have a *perfect bicluster*.

In order to illustrate these definitions, Figure 1 shows an example of two biclusters that follow a perfect shifting pattern (on the left) and a perfect scaling pattern (on the right). In the shifting case, if we represent all the values for each gene, all the charts have the same shape and slope, but in a different range (they are parallel). However, in the scaling case, all the charts have similar shape but different slopes.

In this work we only propose an algorithm for finding shifting patterns, although both shifting and scaling patterns can be present in the data matrix simultaneously. In any case, we are working in order to suggest a scaling approach for future works.

3 Algorithm

A family of computational techniques inspired by the concept of evolution is known as Evolutionary Algorithms (EAs). These algorithms find the solutions to a particular problem by applying a random search on a set of possible solutions [7,11]. EAs use a finite subset of the search space, called population, in each iteration. Previously, these possible solutions were encoded according to the selected coding. The coding is the internal representation of the search space that the algorithm uses. Each encoded element of the population is an individual. Thus, beginning by a pseudo-randomly generated initial population, the evolutionary algorithm selects some individuals and recombine them to generate a new generation of individuals. This process is repeated for a number of generations until the algorithm converges. The selection of individuals is carried out according to their fitness, that is a measurement of the quality of each individual with regards

to the remaining ones. The process of calculating the fitness of the individuals is called evaluation. The evaluation consists in assigning a fitness value to every individual by applying a fitness function.

As aforementioned, our goal is to find the best shifting pattern which represents the general trend within a bicluster. In this work, we address this problem with EAs. Thus, we propose an algorithm which takes as inputs a bicluster and various configuration parameters, returning a set of β_i values (henceforth *beta set*) for such bicluster.

Each chromosome or individual (\mathcal{I}) is made up of a set of real numbers that represent the beta values in real coding ($\mathcal{I} = \{\beta_1, \beta_2, \dots, \beta_i, \dots, \beta_n\}$), corresponding to a shifting pattern proposal. All the individuals have the same length and equal to the number of conditions in the bicluster.

The initial population can be built in two different ways, generating the initial solutions randomly or by using an algorithm based on mutations of the values of the bicluster. For the experiments we present in this work, the second option has been used. In each iteration, each individual of the population is evaluated according to the fitness function defined in Equation 3, and based on the *Mean Absolute Error* (MAE) of each individual, that is, the mean of $|\xi_{ij}|$ values (Equation 4). We have also implemented other alternatives for the fitness function such as the mean squared error, but the obtained results were similar.

$$\phi(\mathcal{I}) = \frac{\sum_{i=1}^n \sum_{j=1}^m \xi_{ij}}{n \times m} \quad (3)$$

$$\xi_{ij} = w_{ij} - \pi_j - \beta_i \quad (4)$$

At the end of each generation, the best individual is replicated to the next one (elitism). Later, a set of individuals (the number of individuals is given by the replication percentage) are selected through the roulette wheel method [10] and replicated to the next generation. Afterwards, the use of recombination and mutation operators allow us to combine a percentage of solutions selected by the roulette wheel for producing new individuals [11]. These operators modify the individuals in a random way. Crossover operator takes as input the number of points for the recombination and create offspring by exchanging the substrings of both parents, thus producing individuals in which the beta values are from both of them. The mutation operator is applied to each individual depending on the mutation probability. Whenever a solution is chosen for a mutation, a random beta value is selected for being changed. The new value is calculated by adding a value between zero and the mean of the error values committed for this beta. Note that this mean can be either positive or negative, depending on the range of the values in the bicluster. After a preset number of generations, the algorithm return the best found beta set.

4 Experimental Results

To show the quality of our tool, we conducted experiments on the biclusters obtained in previous work [2]. These biclusters were obtained by means of an

Table 1. Parameters values of the EA

Parameter	Value
Population size	100
Number of generations	100
Crossover probability	0.80
Mutation probability	0.50
Replication probability	0.20
Number of points in crosses	2

EA from two well-known datasets: yeast *Saccharomyces cerevisiae* cell cycle expression dataset [9]; and the human B-cells expression data [3]. In this section, we expose the empirical results obtained. In Table 1 the parameter settings for all the experiments presented here are shown.

The algorithm was applied with several kinds of biclusters. Thus, for instance, there are biclusters containing different number of genes and conditions or showing different grades of shifting behaviour. Of course, as all of them are obtained from real data, no of them exhibit a perfect shifting pattern, manifesting also scaling trends. In the case of finding a shifting trend within a perfect bicluster, the tool will perform an error value equal to zero in the first iteration.

We have performed our approach over all the biclusters obtained in [2], displaying here the most relevant results. While testing the algorithm, several parameters configuration were used, presenting in this work the ones which performed more interesting results. In all cases we have obtained a set of beta values corresponding to the best found shifting pattern.

4.1 Yeast Dataset

The graphics for five yeast biclusters are represented in Figure 2. In this figure, we expose two charts for each bicluster, the original bicluster (on the left) and the pattern shifted to the range of each of the genes (on the right). We can appreciate how the quality of the found pattern depends on the shifting trend followed by the bicluster. In general terms, we could say that the pattern tries to uniform the behaviour, thus ignoring some isolated local shapes. For instance, in the bicluster labeled 99_1 , the global trend has been perfectly simulated by the result pattern.

Note that the less uniform the genes are, the worse the pattern will be. It means that if we run the algorithm on a bicluster with these characteristic, the shape of the pattern could be very different from some of the genes shapes. Nevertheless, a great number of genes does not implies a bad quality of the found pattern (see Figure 2, bicluster 1_1). Another important characteristic of our algorithm is its rapid convergence; for almost every bicluster in the data set we have experimented with, all of them present a similar convergence throughout the generations. The bicluster labeled 64_1 has the best final error value ($MAE = 10.8$), meaning that the pattern our approach has found for this bicluster is closer to the behaviour of the genes than the pattern in other biclusters. The worst value

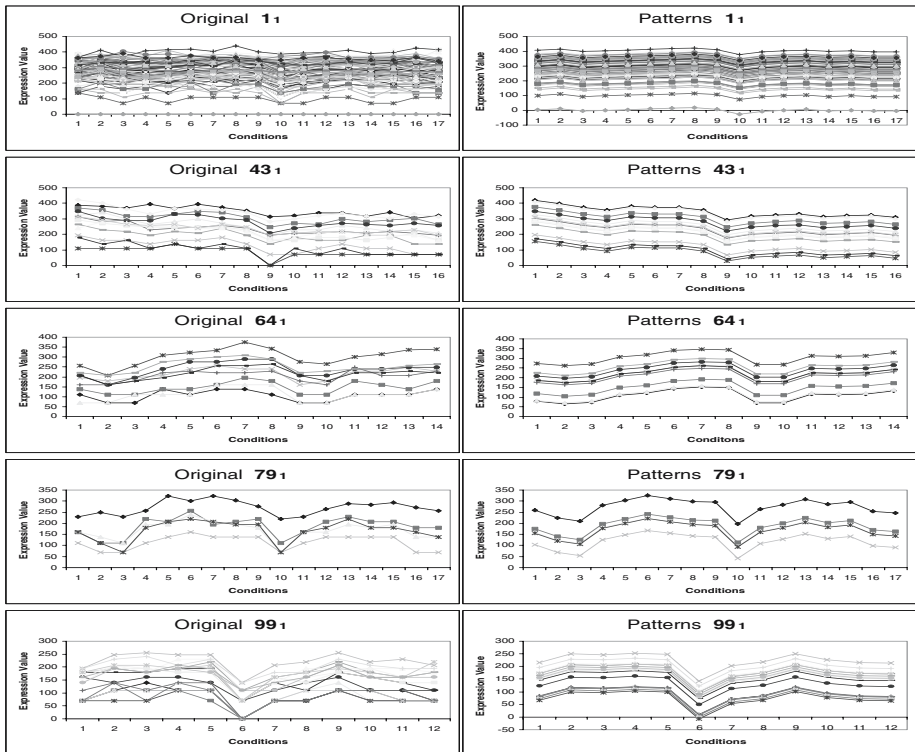


Fig. 2. Yeast biclusters analysed and their pattern result

of the fitness function in the last iteration is for bicluster 79_1 ($MAE = 11.7$), due to this bicluster has few genes but they are quite different one from each other. However, the differences among the error rates are not significant.

4.2 Human Dataset

Five out of hundred bicluster analysed are shown in Figure 3. This figure have a similar structure that the previous one. There exists some differences from the case of the yeast dataset. One point is that now the data include non-positive values, although it makes no difference for our method. But another issue is that the biclusters of the human dataset contain much more conditions. From this point of view, we could expect the fitness function values to be worse than for the previous dataset. Furthermore, the genes represented here are closer than in the previous case, thus the result patterns are closer too, as we can easily appreciate comparing the results in Figures 2 and 3, where we can see how the range of the outcome patterns is bigger in the first case.

Figure 3 shown different kinds of biclusters. For instance, the one labeled 101_1 contains only 3 genes but the mayor number of conditions (72). A medium-length bicluster would be 50_1 , which is made up of 11 genes and 58 conditions.

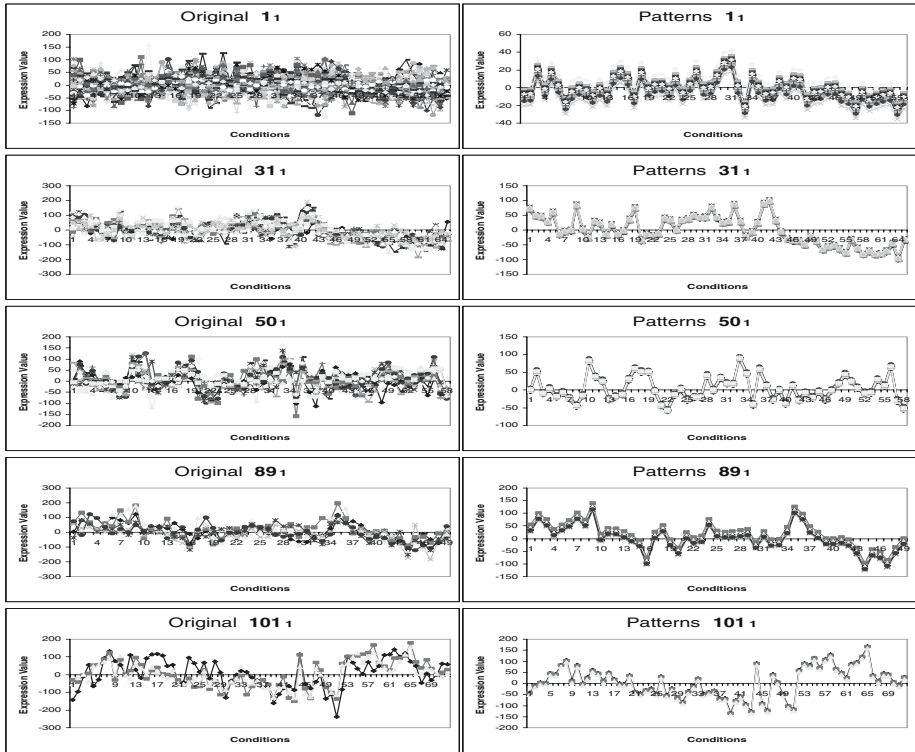


Fig. 3. Human biclusters analysed and their pattern result

Nevertheless, the final error value depends on the quality of each bicluster and not on its size. As we had predicted, the error values are greater than in the case of the yeast dataset, due mainly to the great number of conditions. However, although most of the biclusters tested by the tool show a similar fitness function behaviour (the best MAE was 24.7 for 101₁, and the worst was 27.4 for 31₁), the convergence is not so quickly as for the yeast dataset, although an established value has almost been reached in the last iterations.

5 Conclusions

This work has been developed on the idea that every gene in some types of biclusters follows a similar behaviour, and their graphical representations follow a similar trend with similar slopes. This behaviour is called shifting patterns. In this paper we have presented a novel EA-based tool capable of finding shifting patterns representing the general trend within a bicluster. Beginning from a given bicluster, our approach applies a typical EA to obtain the β_i coefficients that define the pattern. Experimental results over hundred of samples confirm

the quality of our approach for finding this kind of patterns, obtaining very accurate solutions for the biclusters used.

Future works will focus on finding both shifting and scaling patterns simultaneously. A first approach to the scaling problem would consist of considering it as a shifting problem, using the properties of the logarithms.

References

1. J. S. Aguilar-Ruiz. Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21:3840–3845, 2005.
2. J. S. Aguilar-Ruiz and F. Divina. Biclustering of expression data with evolutionary computation. *IEEE Transactions on Knowledge & Data Engineering*, to be published.
3. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.
4. A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297, 1999.
5. S. Bleuler, A. Prelić, and E. Zitzler. An ea framework for biclustering of gene expression data. pages 166–173, Piscataway, NJ, 2000.
6. K. Bryan, P. Cunningham, and N. Bolshakova. Biclustering of expression data using simulated annealing. In *18th IEEE Symposium on Computer-Based Medical Systems*, pages 383–388, Dublin, Ireland, 2005.
7. L. D. Chambers et al. *Practical Handbook of Genetic Algorithms, volume III*. CRC Press, 1999.
8. Y. Cheng and G. M. Church. Biclustering of expression data. In *In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, La Jolla, CA, 2000.
9. R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfberg, A. Gabrielian, D. Landsman, D. Lockhart, , and R. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.
10. K. A. DeJong. *An analysis of the behavior of a class of genetic adaptive systems*. PhD thesis, University of Michigan, 1975.
11. D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison Wesley, 1989.
12. J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.
13. S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics*, 1:24–25, 2004.
14. A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:136–144, 2002.
15. H. Wang, W. Wang, J. Yang., and P. S. Yu. Clustering by pattern similarity in large data sets. In *ACM SIGMOD International Conference on Management of Data*, page 394–405, Madison, WI, 2002.

Gene Ranking from Microarray Data for Cancer Classification—A Machine Learning Approach*

Roberto Ruiz¹, Beatriz Pontes¹, Raúl Giráldez², and Jesús S. Aguilar–Ruiz²

¹ Department of Computer Science, University of Seville
Avenida Reina Mercedes s/n, 41012 Sevilla, Spain
{rruiz, bepontes}@lsi.us.es

² Area of Computer Science, University of Pablo de Olavide
Ctra. de Utrera, km. 1, 41013, Sevilla, Spain
{rgirroj, jsagurui}@upo.es

Abstract. Traditional gene selection methods often select the top-ranked genes according to their individual discriminative power. We propose to apply feature evaluation measure broadly used in the machine learning field and not so popular in the DNA microarray field. Besides, the application of sequential gene subset selection approaches is included. In our study, we propose some well-known criteria (filters and wrappers) to rank attributes, and a greedy search procedure combined with three subset evaluation measures. Two completely different machine learning classifiers are applied to perform the class prediction. The comparison is performed on two well-known DNA microarray data sets. We notice that most of the top-ranked genes appear in the list of relevant-informative genes detected by previous studies over these data sets.

1 Introduction

The gene expression data are typically organized in microarrays. These are matrices where columns represent genes and rows represent experimental conditions (henceforth samples). Each element in the matrix refers to the expression level of a particular gene under a specific condition.

Analysis of microarray data presents unprecedented opportunities and challenges for data mining in areas such as gene clustering [1], sample clustering and class discovery [1,4], sample classification [4] and gene selection [6,9,16,18]. In this work, we address the gene selection issue under a classification framework. The task is to build a classifier that accurately predicts the classes (diseases or phenotypes) of new unlabeled samples. A typical data set may contain thousand of genes but only small number of samples (often less than two hundred). Theoretically, having more features should give us more discriminating power. However, this can cause several problems: increase computational complexity and cost; too many redundant or irrelevant genes; and degradation of the estimation of the classification error. In addition to reducing noise and improving

* This research was supported by the Spanish Research Agency CICYT under grants TIN2004-00159 and TIN2004-06689C0303.

the accuracy of classification, the selected subsets of genes may have important biological interpretation and may be used for drug target discovery or identifying future possible research directions.

In this work, we carry out a study of the performance that several feature selection methods show with two microarrays: Colon Cancer [1] and Leukemia [4]. Although such methods are widely applied in machine learning area, they are not so popular in the DNA microarray field. The application of sequential gene subset selection approaches is included too. In particular, we used six filter and three wrapper methods to rank attributes, and a greedy search procedure combined with three subset evaluation measures. Two well-known machine learning classifiers (naive Bayes and C4.5 [13]), with completely different approaches to learning, are applied to perform the class prediction. This analysis shows that most of the top-ranked genes appear in the list of relevant-informative genes detected by previous studies over these data sets.

The paper is organized as follows. We introduce feature (gene) selection for classification and related work in the next section. Experimental results are shown in Section 3, and the most interesting conclusions are summarized in Section 4.

2 Feature Selection for Classification

The problem of feature selection received a thorough treatment in pattern recognition and machine learning [12]. The gene expression data sets are problematic in that they contain a large number of genes (features) and thus methods that search over subsets of features can be prohibitively expensive. Moreover, these data sets contain only a small number of samples, so the detection of irrelevant genes can suffer from statistical instabilities. Feature selection is reviewed in two ways according to the evaluation measure: depending on their dependency on mining algorithms or based on the way that features are evaluated.

2.1 Filter and Wrapper Model

Feature selection algorithms designed with different evaluation criteria broadly fall into two categories [12]: the filter model and the wrapper model. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than filter model. A hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages [16,18].

As described in [12], some popular criteria are distance, information, dependency, consistency and performance of a classifier measures. A large number of

measures have been proposed for scoring genes in the microarray field: Golub et al. [4] proposed PS (Prediction Strength); Ben-Dor et al. [2] TNoM score (Threshold Number of Misclassification); information gain [16]; t-score [17]; and LDA (Linear Discriminant Analysis), LR (Logistic Regression) and SVM (Support Vector Machine).

2.2 Individual and Subset Evaluation

There exist two major approaches in gene/feature selection from the method's output point of view: feature ranking (FR) and feature subset selection (FSS), depending on the way that features are evaluated. The first one, also called feature weighting [5], assesses individual features and assigns them weights according to their degrees of relevance, while the second one evaluates the goodness of each found feature subset.

In the FR algorithms category, one can expect a ranked list of features which are ordered according to evaluation measures. A subset of features is often selected from the top of a ranking list. A feature is good and thus will be selected if its weight of relevance is greater than a user-specified threshold value, or we can simply select the first k features from the ranked list. This approach is efficient due to its linear time complexity in terms of dimensionality.

In the FSS algorithms category, candidate feature subsets are generated based on a certain search strategy. Each candidate subset is evaluated by a certain evaluation measure and compared with the previous best one with respect to this measure. If a new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Different algorithms address these issues differently. In [12], a great number of selection methods are categorized. We found different search strategies, namely exhaustive, heuristic and random search, combined with several type of measures to form different algorithms. The time complexity is exponential in terms of data dimensionality for exhaustive search and quadratic for heuristic search. The complexity can be linear to the number of iterations in a random search, but experiments show that in order to find the best feature subset, the number of iterations required is mostly at least quadratic to the number of features [3].

Most popular search methods in machine learning can not be applied to microarray data sets due to the large number of genes. Usually, existing algorithms rank genes according to their individual relevance or discriminative power to the targeted classes and select top-ranked genes.

Some existing subset evaluation measures in machine learning that have been shown effective in removing both irrelevant and redundant features include the consistency measure [3], the estimated accuracy of a learning algorithm, and the correlation measure [7]. Above-mentioned are the two first, and correlation measure evaluates the goodness of feature subsets based on the hypothesis that good feature subsets contain features highly correlated to the class, yet uncorrelated to each other.

3 Experiments and Results

In this section, a comparison among a group of different filter and wrapper metrics is carried out. Besides, we empirically evaluate the efficiency and effectiveness of three FSS approaches on gene expression microarray data. Descriptions of the two data sets are studied follow.

Colon cancer data set. This data set is a collection of expression measurements from colon biopsy samples reported by Alon et al. [1]. The data set consists of 62 samples of colon epithelial cells. These samples were collected from colon-cancer patients. The $\$tumor\bar{T}$ biopsies were collected from tumors, and the $\$normal\bar{T}$ biopsies were collected from healthy parts of the colons of the same patients. The final assignments of the status of biopsy samples were made by pathological examination. Of the ≈ 6000 genes represented in these arrays, 2000 genes were selected based on the confidence in the measured expression levels.

Leukemia data set. This data set is a collection of expression measurements reported by Golub et al. [4]. The data set contains 72 samples. These samples are divided to two variants of leukemia: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). The source of the gene expression measurements was taken from 63 bone marrow samples and 9 peripheral blood samples. The expression levels of 7129 genes are reported.

The experiments are conducted using the WEKA's implementation of all these existing algorithms[15]. In order to apply some of the measures (ig, cn and c4), the expression values of each gene are discretized previously.

3.1 Classification with the Genes of Highest Scoring Value

In our study, we apply nine well-known criteria to rank attributes, each of them has a long tradition in feature selection and statistics literature. Six of them are filters: information gain (IG), non-linear correlation (CR) and consistency (CN) are mentioned in Section 2.1 (information, dependency and consistency measures), and ReliefF (RL) [10], Soap (SP) [14] and Chi2(CH) [11]. In the three wrapper approaches applied, naive Bayes (WNB), instance-based (WIB) and c4.5 (WC4) classifiers are used to provide the ranked list. For each metric, we construct the classification models with three, five, ten and twenty genes of highest scoring value. Thus, for each ranked-list, the same subset of genes is used to build the two classification models with Bayesian classifier (NB) and C4.5 (C4).

The main contribution of this study is the use of some criteria for ranking genes that have rarely been used in the biological context. While some filters (CR, IG) are broadly mentioned in the literature [18,16], some others, such as RL or CH, have been applied in the machine learning field but they are not so popular in genomic databases. Furthermore, we present here the results obtained by means of five criteria (filters CN, SP and wrappers WNB, WIB, WC4) that are barely used in this kind of data.

Table 1 reports the leave-one-out cross-validation (LOOCV) accuracy for each metric in Colon and Leukemia data sets. In the table, the first row shows

Table 1. LOOCV accuracy results for each classifier and gene selection technique

	Colon								Leukemia							
	NB (full 58.06)				C4 (full 80.65)				NB (full 100%)				C4 (full 73.61%)			
	(3)	(5)	(10)	(20)	(3)	(5)	(10)	(20)	(3)	(5)	(10)	(20)	(3)	(5)	(10)	(20)
SP	79.0 ⁺	80.6 ⁺	69.3	80.6 ⁺	64.5	83.8	80.6	93.5 ⁺	98.6	94.4	94.4	95.8	88.8 ⁺	87.5 ⁺	84.7	81.9
IG	85.4 ⁺	85.4 ⁺	85.4 ⁺	80.6 ⁺	85.4	74.1	85.4	85.4	94.4	93.0	94.4	95.8	90.2 ⁺	87.5 ⁺	86.1 ⁺	81.9
RL	82.2 ⁺	85.4 ⁺	85.4 ⁺	83.8 ⁺	85.4	85.4	79.0	83.8	90.2	94.4	95.8	95.8	91.6 ⁺	94.4 ⁺	88.8 ⁺	86.1 ⁺
CH	85.4 ⁺	85.4 ⁺	87.1 ⁺	88.7 ⁺	85.4	85.4	85.4	83.8	98.6	97.2	95.8	97.2	88.8 ⁺	84.7 ⁺	83.3	81.9
CR	88.7 ⁺	87.1 ⁺	87.1 ⁺	82.2 ⁺	85.4	85.4	85.4	85.4	98.6	94.4	95.8	95.8	88.8 ⁺	87.5 ⁺	83.3	81.9
CN	85.4 ⁺	85.4 ⁺	87.1 ⁺	87.1 ⁺	85.4	85.4	83.8	85.4	98.6	97.2	95.8	97.2	88.8 ⁺	88.8 ⁺	83.3	81.9
WNB	82.2 ⁺	87.1 ⁺	85.4 ⁺	87.1 ⁺	85.4	80.6	69.3	74.1	95.8	95.8	95.8	95.8	88.8 ⁺	91.6 ⁺	83.3	81.9
WIB	62.9	67.7	77.4 ⁺	79.0 ⁺	82.2	77.4	77.4	88.7	98.6	95.8	94.4	97.2	88.8 ⁺	88.8 ⁺	86.1 ⁺	84.7
WC4	88.7 ⁺	85.4 ⁺	85.4 ⁺	85.4 ⁺	85.4	83.8	83.8	85.4	94.4	95.8	95.8	95.8	88.8 ⁺	91.6 ⁺	83.3	81.9

the dataset and the second one the classifier next to the LOOCV percentage accuracies for non-gene selection for each classifier (in brackets). The rest of rows show the LOOCV values obtained by each method (first column) for each specified gene subset cardinality (3, 5, 10, 20). Furthermore, we conduct Student's paired two-tailed t-test in order to evaluate the statistical significance of the difference between the accuracy of each approach with gene selection and the result of the full set. Thus, the symbol " + " and " - " respectively identify statistically significant, at 0.05 level, wins or losses over the full set.

The top-ranked genes using the nine measures in each data set is listed next. All showed genes appear in the top-20-scoring lists of five ranking at least.

- **Colon:** R87126, M76378(1), M63391, M76378(2), J02854, M76378(3), X12671, M22382, T96873, M26383.
- **Leukemia:** X95735_at, M23197_at, M27891_at, U46499_at, M84526_at, L09209_s_at, D88422_at, M31523_at, M83652_s_at, M92287_at.

With regard to Colon domain, as we can see from table 1, for NB classifier, in all cases, except for SP(10) (i.e. subset with the top-10 genes from Soap list) and WIB(3)(5), these accuracy differences between the non-gene selection and the gene subset selected are statistically significant at 0.05 level. For C4 classifier, no statistical significant differences are shown between the accuracy of all the gene subsets selected by ranking metrics, except SP(20) for C4 classifier, and the accuracy of whole gene set. In some cases (most of then for C4 classifier), the classification accuracy is not improved when the number of genes of the subset is increased. In most of the cases, the accuracy obtained with the three first genes is the same or better than that obtained with the full set. Ranking provided by CR measure obtain the best averaged performance for the two classifier. An analysis of the genes selected by different approaches reveals interesting questions:

- Among the first 20 genes scored by the nine measures, the following two genes appear in the top-20-scoring lists of all scores (GenBank number): R87126, M76378(1).
- The following three genes appear eight times in the top-20: M63391, M76378(2), J02854.

- The following four genes appear seven times in the top-20: M76378(3), X12671, M22382, T96873.
- M63391 and M26383 (human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds.) appear seven and five times respectively in the top-3. The three clones of M76378 appear in the top-20 of seven rankings, and one version in two rankings.
- Most of the genes selected by evaluation measures appear in the lists of relevant genes detected by previous studies over this data set [9,8,2].
- *ib* and *rl* are measures of the same type (distance measures), but their lists are different. The same occurs with *sp* and *cn*, both of them are consistency measures and they have different rankings. However, *ig* and *c4* using information measures and *ch* and *cn* have almost the same top-10.

Regarding Leukemia domain, for *C4* classifier, when the cardinality of the subset is 3 or 5, the accuracy differences between the non-gene selection and the gene subset selected are statistically significant at 0.05 level for all cases. For NB classifier, no statistical significant differences are shown between the accuracy of all the gene subsets selected by ranking metrics and the accuracy of whole gene set. Also, as we can observe, most of the top-3 subsets obtain better results than the rest. There is not difference at the averaged performance of the nine ranked-lists. NB classifier obtains better results than *C4*, although such results are not statistically significant. An analysis of the genes selected by different approaches in Leukemia data set reveals the following interesting questions:

- Among the first 20 genes scored by the nine measures, the following six genes appear in the top-20-scoring lists of all scores (GenBank number): X95735_at, M23197_at, M27891_at, U46499_at, M84526_at, L09209_s_at.
- The following six genes appear eight times in the top-20: D88422_at, M31523_at, M83652_s_at, M92287_at, X62320_at, M11722_at.
- X95735, M23197 and M27891 appear seven, six and seven times respectively in the top-3.
- Most of the genes selected by proposed evaluation measures appear in the lists of relevant genes detected by previous studies over this data set [9,8,2]. Note that these twelve genes are located almost at the same position in [2] with TNoM score.
- The all top-20 are very similar, emphasizing *sp*, *ig*, *ch*, *cr* and *cn* with 10 genes.

3.2 Classification with FSS Approaches

In this section, we empirically evaluate the efficiency and effectiveness of three subset evaluation measures (see Section 2.2) combined with a sequential forward search engine. LOOCV accuracy results for each gene selection algorithm are: 1) For Colon data set and NB classifier, 85.48⁺, 85.48⁺ and 91.94⁺, with correlation, consistency and wrapper subset evaluation measure respectively; and with *C4*, 88.71, 91.94 and 96, 77⁺, respectively. 2) For Leukemia and NB, 98.61, 94.44 and 98.61 for the three measure respectively; and with *C4*, 81.94, 94.44⁺ and

94.44⁺. With the aid of the wrapper gene selection technique, the two classifiers improve their results in the two data sets with respect to the ranking approach. In most of the cases, except in colon data set for NB classifier, accuracy differences between the wrapper procedure and the full set are statistically significant at 0.05 level. Besides, the consistency approach of the sequential search procedure wins over the full set in two cases, while correlation approach once. Results obtained with the wrapper approach are better than those obtained with the two filter techniques in all the cases except two on leukemia data set. This is due to the fact that subsets obtained by wrapper approaches will be better suited to the subsequent classification. The novelty of the application of wrapper approaches within biological data sets constitute a technique that has been proved to have a very good performance.

In both data sets, we notice the low number of genes selected by the consistency and wrapper approaches. In Colon domain, wrapper algorithm choose seven (H20709, M84326, H50623, M63391, H78386, R80427 and H23975) and six (R39465, H08156, J02854, D00860, R08021 and M26383) genes for NB and C4 classifiers respectively. We obtain two different subsets with wrapper approach because the process depend on the employed classifier, but only one subset with filter approaches, consistency five genes (M63391, D14812, T52015, K03460 and R87126), while correlation subset evaluation provide twenty-six genes. In Leukemia domain, wrapper choose three genes (D49950, D88422 and V68162) and two (M27891 and M195507) for NB and C4, three for consistency (M23197, AF009426 and AC002115) and fifty-one for correlation approach.

Sequential forward search procedure starts with an empty set and evaluates each gene individually to find the best single gene. It then tries each of the remaining genes in conjunction with the best to find the most suited pair of genes. In the next iteration each of the remaining genes are tried in conjunction with the best pair to find the most suited group of three genes. This process continues until no single gene addition improves the evaluation of the subset. Therefore, always choose the gene with the best individual evaluation, but generally the rest of the genes are not located at first positions of any ranked list of genes. Gene interactions can be captain for the subset selection approaches. All gene subset selection techniques are able to considerably reduce the huge number of genes to small informative and accurate subsets of components.

However, these accuracy improvements of wrapper procedures are couple with demanding computer-load necessities. In most of the cases, the computer-load necessities of ranking procedures can be considered as negligible with respect to wrapper ones. Consistency approach took 3 and 14 seconds to produce results on colon and leukemia domain respectively, correlation 26 and 1440 seconds, and wrapper took 165 and 1156 for NB classifier, and 520 and 309 seconds for C4.

4 Conclusions

Traditional feature selection methods often select the top-ranked features according to their individual discriminative power. When the number of features

is high, about thousands, as it happens in the microarray gene expression data sets, there are many irrelevant and/or redundant genes. For this reason, gene rankings might not be useful to select the best k genes from that ranked-list.

In this paper, we show that the classification accuracy may vary depending on the number of genes selected from the ranked-list, and not always is better when more genes are involved. In fact, it depends on the feature ranking method and also on the classifier. To show this situation, we have used nine feature ranking methods together with two different classifiers.

In addition, due to the effect of irrelevant and redundant genes in microarray gene expression data sets, those rankings might provide some noise to the classifier when we select the k top-ranked genes. This reason motivated us to study an algorithm to extract a subset of genes, trying to avoid the influence of unnecessary genes on the later classification. The wrapper approach of this algorithm shows an excellent performance, obtaining subsets better suited to the subsequent classification.

References

1. U. Alon et. al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–50, 1999.
2. A. Ben-Dor et. al. Tissue classification with gene expression profiles. *Proc. Natl. Acad. Sci. USA*, 98(26):15149–54, 2001.
3. M. Dash, H. Liu, and H. Motoda. Consistency based feature selection. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 98–109, 2000.
4. T. Golub et. al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–37, 1999.
5. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
6. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machine. *Machine Learning*, 46(1-3):389–422, 2002.
7. M. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, Dept Computer Science, Hamilton, New Zealand, 1999.
8. T. Hellem and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4):0017.1–0017.11, 2002.
9. I. Inza et. al. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence in Medicine*, 31:91–103, 2004.
10. I. Kononenko. Estimating attributes: Analysis and estensions of relief. In *European Conf. on Machine Learning*, pages 171–182, Vienna, 1994. Springer.
11. H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *7th IEEE Int. Conf. on Tools with Artificial Intelligence*, 1995.
12. H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. on Knowledge and Data Eng.*, 17(3):1–12, 2005.
13. J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, California, 1993.
14. R. Ruiz, J. Riquelme, and J. Aguilar-Ruiz. Projection-based measure for efficient feature selection. *Journal of Intelligent and Fuzzy System*, 12(3–4):175–183, 2002.

15. I. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2005.
16. E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. In *Proc. 18th Int. Conf. on Machine Learning*, pages 601–608. Morgan Kaufmann, San Francisco, CA, 2001.
17. M. Xiong, L. Jin, W. Li, and E. Boerwinkle. Computatinal methods for gene expression-based tumor classification. *BioTechniques*, 29:1264–70, 2000.
18. L. Yu and H. Liu. Redundancy based feature selection for microarry data. In *10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2004.

H^3 : A Hybrid Handheld Healthcare Framework

Seung-won Hwang

Pohang University of Science and Technology

Abstract. Handheld devices, which have been widely adopted to clinical environments for medical data archiving, have revolutionized error-prone manual processes of the past. Meanwhile, the use of devices has been reported to be limited to data entries and archiving, without fully leveraging their computing and retrieval capabilities. This paper studies a hybrid system which complements current state-of-art, by combining intelligent retrieval techniques developed over middleware environments for *retrieval effectiveness* and flash-aware data management techniques for *retrieval efficiency*. By enabling intelligent ranked retrieval, the limited resources of handheld devices, *e.g.*, limited display and computation capabilities, can be utilized effectively, by selectively retrieving the few most relevant results. However, to achieve this goal, we need a hybrid approach, bridging middleware-based ranked retrieval techniques to optimize for flash memory storage, as typically adopted by handheld devices. We address this newly emerging challenge and propose a flash-aware framework H^3 , which we empirically validate its effectiveness over baseline alternatives.

1 Introduction

As handheld computing devices are becoming a common feature on the medical landscape, medical data are being stored and accessible from such devices. However, their uses reported to be restricted mostly to the entry of digitized healthcare information [1], replacing the previous error-prone hand-written or transcribed prescriptions from physicians.

However, current use mostly stores the digitized information in the centralized server, from which, the analysis, computation, and retrieval are executed. Meanwhile the computing and storage capabilities of handheld devices explode, such that an affordable \$300 handheld devices are now equipped with 312MHz processor, which can be extended with a 8GB flash storage. Considering these trends, restricting the use of handheld healthcare devices to anything less than a computer is simply a waste of computing and storage resources. Our goal is thus to support medical activities, including computing, analysis, and retrieval, at the device, fully leveraging the underlying computing and storage potential, as the following example illustrates.

Example 1 (Handheld Healthcare Data Management). *Physicians in St. Vincent Hospital, a major, 397-bed acute care hospital and a leading referral center for open-heart surgery, recently adopted a data management software, which*

supports physician's entire day in varying care settings. For example, the device allows physicians to access their patients electronic records, write prescriptions, enter charges for services, document patient encounters, place orders, and securely send messages to other caregivers all in a single device, which enables physicians to receive and respond information on-site, immediately as soon as they are available.

The goal of this paper is to develop data management techniques to successfully support handheld healthcare applications, retrieving varying types of medical data on-site within the handheld storage, as the above example illustrates. Toward this goal, we identify two challenges:

- **Intelligent data management:** Complex nature of medical data, *e.g.*, medical images, and limited display of handheld devices naturally call for ranked retrieval of retrieving “finding top 5 closest images” or “finding top 5 most relevant prescriptions”, to support complex query semantics or to focus display resources to the few most relevant data respectively. In a clear contrast, current support from an embedded operating system and databases are limited to traditional Boolean query accesses. Evaluating a rank query over Boolean databases would require to perform a sequential scan over all pages stored on the external flash memory. For *retrieval effectiveness*, it is thus crucial to build an cost-effective index and intelligent ranking algorithms, on top of current data management layers.
- **Handheld storage support:** Portability requirement of handheld devices call for a small, power-efficient, and non-volatile storage media. Such characteristics have established flash media as an ideal choice, and thus, flash media have been prevalent in current mobile and wireless devices. While ranked retrieval have been actively studied lately [2, 3, 4], existing works are designed with middleware subsystems as access scenarios in mind. Consequently, adopting them for handheld devices incur a prohibitive cost, since flash memory has significantly different storage characteristics. For *retrieval efficiency*, it is thus crucial to build a intelligent ranking framework, to optimize the cost, according to the flash memory storage characteristics.

Our goal is to address these two challenges together and pursue both retrieval effectiveness and efficiency, complementing current state-of-art addressing exclusively one of the two. More specifically, we develop Framework H^3 , which supports efficient ranking, optimizing the cost over flash memory storage characteristics. Our novelty is thus our *hybrid* efforts to bridge the intelligent data management techniques developed for middleware cost settings, to the flash-specific storage environments.

This paper is organized as follows. We briefly review related works in Section 2. Section 3 overviews the preliminaries in supporting flash-aware rank processing. Section 4 then develops our cost-effective framework H^3 . Section 5 reports our preliminary evaluation results. Finally, Section 6 concludes our work.

2 Related Work

Ranked retrieval has been actively studied recently as a means of effectively retrieving only the most relevant results [2, 3, 4]. However, existing works build upon access models of middleware subsystems. In contrast, handheld devices typically build on flash memory with significantly different storage characteristics from disks or middleware access scenarios. Existing ranked retrieval techniques thus cannot be cost-effective for handheld devices. Recently, data management techniques building on flash storage characteristics have been studied for commercial file systems, *e.g.*, YAFFS or JFFS [5, 6], and a simple Boolean data retrieval, *e.g.*, B-tree [7].

This paper develops a flash-aware rank retrieval framework, complementing existing works in both directions. Our proposed framework extends the applicability of ranked retrieval to new access scenarios of flash memory, yet complements flash-aware data management techniques to support advanced query semantics such as ranking, for the first time to our best knowledge. Our work enables effective data retrieval by ranking, efficiently optimized specifically for flash-specific storage characteristics.

3 Preliminaries

To establish context, this section discusses preliminaries and challenges in supporting ranked retrieval over flash memory storage.

Flash memory has been widely adopted to handheld devices lately and its capacity has been rapidly increasing up to the point to replace disks in many devices. However its data management capacity remains largely unexplored compared to disks, due to its distinct characteristics: Unlike disks, in flash memory, erasing is extremely costly and even deteriorates the lifetime of the storage. Data management techniques over flash memory thus need to minimize the erase operations and distribute such operations evenly over all memory segments, to achieve a longer overall lifetime. Toward this goal, recently developed flash-aware data management techniques [5, 6, 7] build on “out-place” update strategy. That is, to minimize erase operations, when data item x is updated, instead of erasing x in-place requires a costly erase operation, we simply mark x as invalid and use *pointer redirection* to refer the updated x' located in another segment. While such out-place update strategy significantly reduces the needs of erase operations, it introduces a hidden overhead of a pointer redirection to reach the fresh copy x' and also ruins the locality of having objects with similar values together.

Understanding such distinct characteristics hints why applying an existing ranked algorithm “as is” incur a higher access cost. To retrieve top- k objects with the highest *query score* that combines several attributes t_1, \dots, t_n (or, more generally, *predicates* mapping attributes into some meaningful scores) by a monotonic scoring function $\mathcal{F}(t_1, \dots, t_n)$, where t_1, \dots, t_n all contribute positively toward the \mathcal{F} score, there have been many existing ranking algorithms [2, 3, 4] leveraging underlying “sorted access” to selectively retrieve “promising” objects

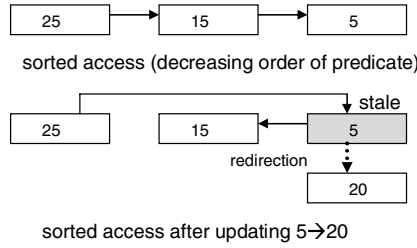


Fig. 1. Sorted access on flash storage

in the decreasing order of predicate scores. While we can build such accesses upon out-place update strategy, as Figure 1 illustrates, the cost characteristics will not be the same— Unlike middleware scenarios where sorted accesses incur a lower amortized cost by providing access to promising objects located together in batch, sorted access over flash storage scenarios can be expensive with hidden pointer redirections (as noted with a dotted arrow in Figure 1) to reach fresh values of promising objects scattered across blocks. In summary, an effective ranked retrieval algorithm for flash storage should leverage locality as much as possible, as we will develop in Section 4.

4 Framework

This section develops a flash-aware ranked retrieval framework H^3 , (1) being aware of and (2) optimizing for flash-specific storage characteristics.

First, for *awareness*, we distinguish locality-preserving in-block pointer from out-place redirections pointing fresh copies, as denoted by solid and dotted lines respectively in Figure 1. When objects are added, removed, or deleted, we only change in-block pointers to keep the sorted order, as Figure 1 illustrated. As a result, following in-block pointers will visit objects in the order of fresh scores, with occasional stale vales requiring redirections— To minimize such redirections, instead of redirecting for each stale object visited, we adopt *lazy redirection* to delay them, as we will develop below, until it is absolutely necessary to return the correct results.

Second, for *optimization*, we need to decide when is the point that a redirection is absolutely necessary. Toward this goal, we expand the concept of *necessary probe principle* [3] developed for minimizing expensive random accesses. This principle determines if a random access (or a “probe”) to evaluate p on u is absolutely necessary, such that no algorithm can ensure correctness without performing such a probe. In other words, an algorithm performing only a necessary probe is guaranteed to perform the minimal probes. We can similarly decide whether to perform a expensive redirection or not, to redirect as lazily as possible, waiting until the redirection is absolutely necessary— With that assurance, our framework is guaranteed to perform minimal redirections necessary in ensuring the correctness, which we state formally below.

Theorem 1 (Lazy Redirection). *Consider a ranked query with scoring function \mathcal{F} and retrieval size k . A redirection on object u be an object is necessary, if there do not exist k objects v_1, \dots, v_k with higher upper bound scores.*

Proof 1. *We will show by contradiction that a redirection on u is necessary such that no algorithms can determine correct top- k answers without it. To contradict, suppose that there exists some algorithm \mathcal{A} to the contrary. Let \mathcal{K} be the top- k output of \mathcal{A} . We show that \mathcal{K} can either be incomplete or incorrect— and thus such \mathcal{A} does not exist.*

- *Suppose that $u \notin \mathcal{K}$: Let $\mathcal{H}(p)$ be predicates scheduled in prior to predicate p in schedule \mathcal{H} . Since p is the next predicate for u , it must have $T_u = \mathcal{H}(p)$ as the evaluated predicates. There exist less than k objects v_1, \dots, v_g ($g < k$) such that the upper bound of u is no more than that of v_i , which we denote as $\overline{\mathcal{F}}_{T_u}[u] < \overline{\mathcal{F}}_{T_{v_i}}[v_i]$.*

Suppose that \mathcal{A} returns answers $\mathcal{K} = (a_1, \dots, a_k)$. Since $g < k$ (i.e., there are more a_j than v_i), there exists some a_j such that a_j is not among v_1, \dots, v_g . That is, for some T_{a_j} (as some subset of the schedule), u has a higher upper bound.

$$\overline{\mathcal{F}}_{T_u}[u] > \overline{\mathcal{F}}_{T_{a_j}}[a_j] \tag{1}$$

We show that, if the unevaluated predicates of u are all perfectly scored, or, 1.0, then \mathcal{K} will be incorrect, and thus \mathcal{A} does not always return correct answers. In this case, u has predicate scores (in addition to those already probed in T_u) $\forall t \notin T_u : t[u] = 1.0$ and thus $\mathcal{F}[u] = \overline{\mathcal{F}}_{T_u}[u]$. It thus follows that $\mathcal{F}[u] > \overline{\mathcal{F}}_{T_{a_j}}[a_j]$. Since \mathcal{F} is monotonic and $\overline{\mathcal{F}}_{T_{a_j}}[a_j]$ upper bounds $\mathcal{F}[a_j]$, it follows that $\mathcal{F}[u] > \mathcal{F}[a_j]$. That is, u outperforms some top- k answer, and thus \mathcal{K} is incorrect.

- *Suppose that $u \in \mathcal{K}$: Since without the redirection, \mathcal{A} cannot determine the predicate score $p[u]$ and thus neither the query score $\mathcal{F}[u]$. Consequently, \mathcal{K} does not return all the query scores as required; it is thus incomplete.*

To make such an essential decision, our theorem hints that a redirection on object u is necessary if the object is currently ranked at the top- k in terms of upper bound. The intuition behind using upper bound score as a guideline is simple to understand— With its upper bound ranked top- k , object u has a potential to make the top- k answers if it ends up scoring the bound score. It is thus absolutely necessary to evaluate further on the object to refine the bound, without which we cannot determine whether u is a part of top- k results and thus cannot ensure the correctness.

Putting together, we develop a framework in Figure 4. For now, we assume there is one sorted access x for simplicity, which can be straightforwardly extended for multiple sorted streams, by merging them into a single combined stream x using existing top- k algorithms. To illustrate, a representative ranking algorithm TA [2] scans multiple sorted streams in decreasing order of predicates and terminates when top- k objects in the merged stream are identified. Such a

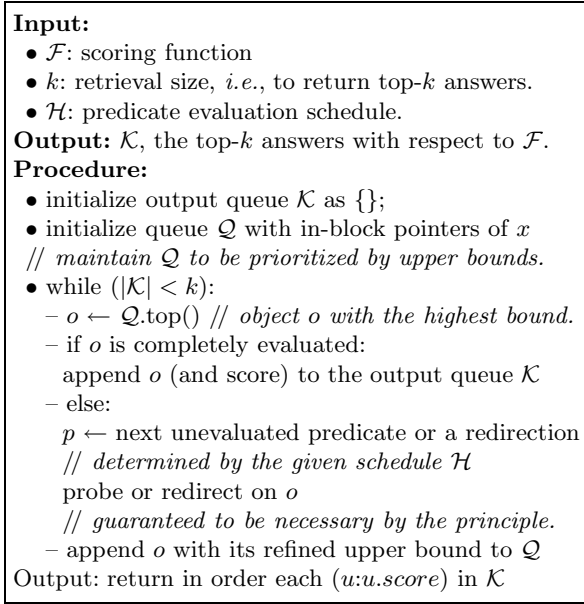


Fig. 2. H^3 with minimal redirections

scheme can be straightforwardly extended to retrieve the top object incrementally, by asking top-1 at a time— By doing so, TA generates a merged stream, which will be an input of our framework H^3 .

Framework H^3 scans a sorted stream x , following only the in-block pointers and delaying the redirection to the stale data— Redirections are enqueued only with the object ID (obtained from in-block access to a stale copy) with a trivial upper bound of the highest possible x score. Framework H^3 keeps on requesting the top-priority object u with the highest ceiling score from sorted access— According to the necessary probe principle, if u has an unknown score, it is necessary to evaluate further, using either a random access or a redirection to fresh value according to the given schedule \mathcal{H} .

To further optimize by identifying a desirable schedule \mathcal{H} , we can adopt *a priori* knowledge on data, ranking function, or cost characteristics. To illustrate, suppose ranking function used is a weighted average— Different weights indicate different impacts of each attribute toward the overall score. In particular, scheduler can take advantage of such hints to schedule the predicate with high impact first. Further, comparing a probe and a redirection, a redirection, by being a part of sorted access, is expected to score as high as (or slightly lower than) the last seen object from the sorted access. In other words, when accessing top objects in the sorted access, the expected score of a redirection is higher than a probe we have no idea how it scores. It is thus effective to favor a redirection over a probe in general. When such *a priori* information is not available, existing schemes [3] dynamically sample objects to obtain statistical information can be

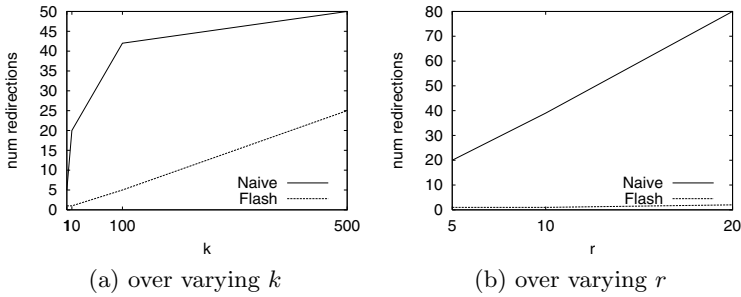


Fig. 3. Performance analysis

adopted, while such a random sampling may not be trivial for flash memory systems.

Framework H^3 continues until all top- k objects are fully evaluated. The correctness of H^3 is ensured as all k objects in \mathcal{K} has final scores higher than the ceiling scores of object o still in \mathcal{Q} . Further, Algorithm H^3 performs the minimal random access and redirections, by performing only those absolutely necessary to ensure correctness.

5 Evaluation Results

To validate the effectiveness of our flash-aware ranked retrieval, this section reports our experimental setup and preliminary results. All evaluations were carried out on a Pentium 4 PC with a 3 GHz processor, over synthetic data and queries to control performance parameters, such as an average update ratio r . In particular, we generated normally distributed synthetic data (with mean 0.5 and variance 0.16) with $N = 1000$ objects and one attribute with a sorted access and two without, all three within range $[0, 1]$ (with no loss of generality) with randomly generated updates of $r = 5\%$ of data as a default, which we later vary for the sensitivity analysis. We implemented our Framework H^3 and validated the effectiveness of flash-aware algorithms by comparing with a naive adoption of MPro [3] (which we call Framework Naive).

As a performance metric, we measured the number of redirections (the y -axis) with respect to varying retrieval size k or update ratios r (the x -axis). Figure 3(a) first reports our evaluation results with respect to varying k , when a ranking function is an average of three attributes. Observe that Framework H^3 significantly outperforms Naive— For instance, when $k = 10$, the number of redirections of H^3 is 1, which saves 95% from Naive with 20 directions. Note, such cost gap only increases as the update ratio r increases, as Figure 3(b) demonstrates for $k = 10$ — The cost of Naive proportionally increases over the increase of r , while H^3 stays more or less unchanged. As a result, for $r = 20\%$, cost saving of H^3 reaches up to 98% from the cost of Naive.

6 Conclusion

This paper studies how to support ranked retrieval over handheld devices. In particular, we build over flash memory storage, as typically adopted by such devices. We develop a framework that enables effective ranked retrieval, fully leveraging flash-aware specific storage characteristics. Our preliminary evaluation suggests that, by supporting flash-awareness, our proposed framework H^3 enables an effective and efficient handheld data retrieval, by developing hybrid techniques bridging effective middleware-based retrieval techniques with flash-aware data management techniques. Note that, enabling an effective ranked retrieval framework is an essential foundation for enabling rank-based textual and multimedia retrieval, as heavily demanded in medical domain with heavy multimedia resources and textual clinical notes— Existing ranked-based textual and multimedia retrieval systems mostly base on the same middleware-based ranking query models.

References

- [1] Peter J. Embi. Information at hand:using handheld computers in medicine. *Cleveland clinical journal of medicine*, 68, 2001.
- [2] Ronald Fagin, Amnon Lote, and Moni Naor. Optimal aggregation algorithms for middleware. In *PODS 2001*, 2001.
- [3] Kevin C. Chang and Seung-won Hwang. Minimal probing: Supporting expensive predicates for top-k queries. In *SIGMOD 2002*, pages 346–357, 2002.
- [4] Seung-won Hwang and Kevin C. Chang. Optimizing access cost for top-k queries over web sources. In *ICDE 2005*, 2005.
- [5] Aleph one. Yaffs: Yet another flash filing system. In <http://www.aleph1.co.uk/yaffs/index.html>, 2002.
- [6] Axis Communication. Jffs home page. In <http://developer.axis.com/software/jffs/>, 2004.
- [7] C.-H. Chang, L.-P. Chang, and T.-W. Kuo. An efficient b-tree layer for flash-memory storage systems. In *ACM RTCSA*, 2003.

Hybrid Intelligent Medical Tutor for Atheromatosis*

Katerina Kabassi^{1,2}, Maria Virvou¹, and George Tsihrintzis¹

¹ Department of Informatics, University of Piraeus,
80 Karaoli & Dimitriou St., 18534 Piraeus, Greece
{kkabassi, mvirvou, geoatsi}@unipi.gr

² Department of Ecology and the Environment,
Technological Educational Institute of the Ionian Islands,
2 Kalvou Sq., 29100 Zakynthos, Greece

Abstract. This paper describes a hybrid intelligent medical tutor for atheromatosis. The tutor is called INTATU (INTElligent Atheromatosis TUTOR). INTATU provides adaptive tutoring on Atheromatosis to various classes of users depending on their interests, background medical knowledge and computer skills. The adaptivity results from user modelling that is based on stereotypical knowledge about the potential users (patients, patients' relatives, doctors, medical students, etc.). The inference mechanism uses a hybrid combination of rule-based reasoning of double stereotypes and decision making techniques.

1 Introduction

Education on medical domains is of interest to many categories of people such as medical students, pharmaceutical students, nurses, other health professionals, patients or people with an interest in health and medicine. A new trend in medical education is the implementation of Information and Communication Technologies (ICT) to support student learning [2]. Additionally, research in Artificial Intelligence in Education demonstrates that the use of artificial intelligence can enhance learning in medical domains [9]. As a result, a lot of research energy has been put in developing learning environments that incorporate some kind of artificial intelligence technique for supporting learners in medical domains (e.g. COMET [13]; Ines [5]; CIRCSIM-Tutor [4]). However, systems such as COMET and CIRCSIM-Tutor refer only to medical students. Ines, on the other hand, is addressed to nurses. If a doctor tries to interact with such a system to learn about a medical problem, it is possible that s/he would find it boring as his/her background medical knowledge is quite different. A quite interesting approach is that of Koutsojiannis et al. [8] who use AI techniques to specify user models and, therefore, adapt the interaction in an Intelligent Tutoring System (ITS) that is addressed to medical students as well as to other health professionals.

However, none of the above mentioned systems addresses the needs of patients or other users without any medical background that want to learn about a particular disease. Such users have different needs and do not have the medical background required to understand the medical terminology provided in medical texts. In view of

* This work has been funded by the Greek Ministry of Education, as part of the PYTHAGORAS II basic research program.

the above, Tweddle et al. [14] designed a web site about cancer, presenting information in small chunks, using clear “everyday” language, so that it would be an educational resource for people with little or no formal knowledge about cancer. Of course, the problem still exists, as information in such form is not useful to medical students or doctors. A remedy to this problem is the development of systems with an ability to adapt their behaviour to the goals, tasks, interests and other features of individual users and groups of users [1; 15].

In view of the above, we have developed an ITS for the medical domain of Atheromatosis, which is a topic that is of interest to many categories of people. Atheromatosis of the aortic arch has been recognized as an important source of embolism. System embolism is a frequent cause of stroke. The severity of Atheromatosis is granted by the fact that aortic atheromas are found in about one quarter of patients presenting with embolic events [12]. Information about Atheromatosis is considered crucial because the diagnosis of this particular disease is mostly established after an embolic event has already occurred.

The ITS developed is called INTATU (INTelligent Atheromatosis TUtor) and maintains and processes information about its users so that the system can adapt its interaction to each user dynamically. INTATU is based on hybrid intelligence that uses a novel combination of user stereotypes with a decision making theory in order to provide personalised interaction of learners with the system. The user stereotypes constitute rule-based reasoning that is widely used in user modelling systems for drawing inferences about users based on a small set of observations [10; 11]. The information of the stereotypes is used in combination with a multi-criteria decision making theory called Simple Additive Weighting (SAW) [3, 6] in order to evaluate each theory topic on Atheromatosis and present the information that would be of interest to the user interacting with the system and in a way that it would be appropriate for him/her.

More specifically, INTATU makes use of stereotypes for providing default assumptions about the interests, background knowledge and needs of the users belonging to a certain group until the user model acquires sufficient information about each individual user. In INTATU, users are classified into six categories, namely, patients, patients’ relatives, users with simple concern to medical problems, doctors, medical students and medical researchers.

2 Simple Additive Weighting

The Simple Additive Weighting (SAW) [3, 6] method is among the best known and most widely used decision making method. SAW consists of two basic steps:

1. **Scale the values of the n criteria to make them comparable.** There are cases where some criteria assume values in the $[0,1]$ interval, whereas in other cases they assume values in the $[0,1000]$ interval. Such values are not easily comparable. A solution to this problem is given by transforming the values of criteria so that they are limited in the same interval. If the values of the criteria are already scaled up, this step is omitted.
2. **Sum up the values of the n criteria for each alternative.** As soon as the weights and the values of the n criteria have been defined, the value of a

multi-criteria function is calculated for each alternative as a linear combination of the values of the n criteria.

The SAW approach consists of translating a decision problem into the optimisation of some multi-criteria utility function U defined on A . The decision maker estimates the value of function $U(X_j)$ for every alternative X_j and selects the one with the highest value. The multi-criteria utility function U can be calculated in the SAW method as a linear combination of the values of the n criteria:

$$U(X_j) = \sum_{i=1}^n w_i x_{ij} \quad (1)$$

where X_j is one alternative and x_{ij} is the value of the i criterion for the X_j alternative.

3 Intelligent Atheromatosis Tutor

INTATU (Intelligent Atheromatosis Tutor) is an Intelligent Tutoring System about Atheromatosis. The system addresses a variety of users, such as patients, patients' relatives, doctors, medical students, etc. The main goal of INTATU is to adapt dynamically its interaction to each user. For this purpose, the system faces the decision problem about which theory topic might be of interest to the user interacting with it. Therefore, INTATU incorporates a user modelling component. This component maintains information about the interests, needs and background knowledge of all categories of potential users.

In order to locate which theory topic is to be presented to a user, each theory topic is evaluated on a set of criteria that reflect the user's interests, previous knowledge and computer skills. The user model that the system maintains provides continuously the evaluation data of the theory topics against the criteria.

In order to determine the criteria, we interviewed 10 human experts. These human experts were medical tutors in well established educational institutions. These tutors were asked about the criteria that they take into account when making decisions about what the student should learn. After analysing their answers, we concluded that the attributes that most human tutors take into account are the following:

- **The degree of acquisition of prerequisite knowledge (p).** If the users belonging to a particular stereotype have acquired the prerequisite knowledge needed for comprehending the theory topic that is evaluated, then the likelihood that this theory topic is suitable for the user interacting with the system increases.
- **The degree of user's interest (i).** The higher a user's interest for a theory topic the greater is the possibility that the particular user wants to learn about that theory topic.
- **The degree of suitability of the theory topic for the particular user (s).** Some parts of the theory are suitable for doctors and not comprehensible by patients because medical terminology is used. The parts of the theory that are addressed to users without a medical background are too simplified for doctors and researchers. This degree shows how suitable the style of text is for the user interacting with the ITS.

- **The degree of difficulty (d).** It has been observed that some parts of the theory are not easily comprehensible by the user. Therefore, the higher the degree of difficulty of a theory the higher the need for tutoring on that particular subject.
- **The degree of appropriateness to his/her computer skills (c).** The way a subject is presented to a user may vary considerably in order to make the topic comprehensible. For example, novice users contrary to expert users cannot handle large amounts of data presented in a screen and prefer information given in small chunks.

The above mentioned criteria acquire their values based on the information maintained in the user modelling component of the system. More specifically, INTATU uses stereotypes to categorise users according to their status and the level of expertise in computers and the Internet. However, the system is also constantly collecting information about a particular user's behaviour and errors and updates the individual user model of the user.

4 Stereotypes for User Modelling

INTATU makes use of a double-stereotype system for constructing user model in the beginning of a user's interaction with the system. More specifically, stereotypes are used for giving information about the users' interests and background knowledge in medical matters and their level of expertise in computers and the Internet. Stereotypes constitute a powerful mechanism for building user models [7]. This is due to the fact that stereotypes represent information that enables the system to make a large number of plausible inferences on the basis of a substantially smaller number of observations [10; 11].

Users are classified into six major classes with respect to their motive to learn about the disease. The six stereotypes that have been identified are: User with simple concern, Patient, Patients' Relative, Doctor, Medical student and Medical researcher. Such a categorisation of users was considered important because users belonging in different categories have different interests, background knowledge and needs. For example, simple users usually seek a general description of the disease and ways to prevent it. Patients and their relatives, on the other hand, would seek a complete presentation of the disease in plain text without special medical terminology. From these two categories of potential learners, patients may look for new treatments and a more detailed description of Atheromatosis than their relatives.

The users belonging to the last three stereotypes are completely different from the first three. Their main difference is that they have a medical background and, as a result, they use special terminology and usually seek more specific and scientific information about Atheromatosis. Despite the above mentioned similarities, the categorisation of the users into three stereotypes was considered crucial because users of different categories have different interests and background medical knowledge. For example, a doctor and a researcher have more advanced background knowledge than a medical student. However, medical researchers probably seek innovative information about the disease and its treatment whereas a doctor would seek tested and totally accepted treatments of Atheromatosis.

Additionally, users can be classified into three major classes according to their level of expertise, namely, novice, intermediate and expert. Each one of these classes represents an increasing mastery in computer skills. Such a classification was considered important because it would enable the system to draw initial inferences about the usual errors and misconceptions of a user, belonging to a group.

Every time a new user is connected to INTATU, s/he has to answer some questions about his/her personal data (name, surname) and select what kind of interest s/he has in the disease, as well as his/her level of expertise in computers and the Internet. This information is used for activating the appropriate stereotypes. More specifically, two stereotypes are activated, one categorising the user according to his/her relation to the disease and one indicating what his/her level of expertise in computers and the Internet is.

5 Acquisition of Criteria's Values

The default assumptions of the stereotypes activated for the particular user provide the values for the criteria used for the application of SAW. More specifically, the stereotypes that show the learner's kind of interest in the disease provide the values for the criteria p , i , s and d and the stereotype that is connected with the user's mastery in computer skills provides the value of the criterion c . However, these inferences must be treated as defaults, which can be overridden by specific observations as the user's knowledge, interests and needs may change over time. Therefore, the system is also constantly collecting information about a particular user's behaviour and interests and informs the individual user model of the user.

When the individual user model has enough information about the user, the criteria acquire their values dynamically. More specifically, the criterion p is calculated by dividing the number of theory topics the user knows with the number of topics needed as prerequisite knowledge for a user to comprehend another theory topic. The criterion i shows the user's interest for each topic and is calculated by taking into account the frequency of visits of the particular user to theory topics that relate to the topic that is evaluated. The value of the criterion d that represents the difficulty of each theory topic diminishes as a user acquires more and more knowledge about a subject. The criterion s represents the degree of suitability of the theory topic for the particular user and is always required by the stereotype that is activated for a user. Similarly, the criterion c that shows how appropriate the way of presentation of a theory topic is for a particular user always acquires its values by the stereotype that is connected with the user's mastery in computer skills. However, the activated stereotypes for a user may change over time as s/he acquires more knowledge about Atheromatosis and improves his/her computer skills.

6 Application of the Simple Additive Weighting

SAW is used to evaluate the alternative theory topics based on some criteria and select the one that seems more suitable for the particular user. For this purpose, INTATU calculates a multi-criteria utility function for each theory topic. The theory

topic that maximises the function U is selected by the system in order to be presented to the user. The function U is calculated as a linear combination of the five criteria presented above: $U_{SAW}(T_j) = w_p p_j + w_i i_j + w_c c_j + w_d d_j + w_s s_j$ (1)

where T_j is the evaluated theory topic, w_p, w_i, w_c, w_d, w_s are the weights of the criteria and p_j, i_j, c_j, d_j, s_j are the values of the criteria for the j theory topic. The values of the criteria are acquired by the stereotype. The weights of the criteria, on the other hand, are not known.

The decision about which theory topic is to be proposed to the user and in what priority relates to the reasoning of a human expert who would watch a user work over his/her shoulder. Therefore, in order to calculate the weights of the five criteria, the 10 experts that selected these criteria, were also asked to define their relative importance in their reasoning process.

More specifically, human experts were asked to rank the five criteria with respect to how important they are in their reasoning process. Each human expert was asked to share 15 points into the 5 different criteria. For example, three different human experts thought the criteria p, i and s were equally important and, therefore, assigned 4 points to each one. Finally, they assigned the rest of their 3 points in the other two criteria but not equally. They stated that the criterion d was most important for them and assigned 2 points to it and only 1 point to the criterion c .

As soon as the scores of all human experts were collected, they were used to calculate the weights of the certainty parameters. The scores assigned to each criterion by all human experts were summed up and then divided to the sum of scores of all criteria (15 points assigned to all criteria by each human expert * 10 human experts = 150 points assigned to all criteria by all human experts). In this way the sum of all weights could be equal to 1.

As a result, the calculated weights for the criteria were the following:

- The weight for the criterion p : $w_p = \frac{41}{150} = 0.27$
- The weight for the criterion i : $w_i = \frac{39}{150} = 0.26$
- The weight for the criterion s : $w_s = \frac{39}{150} = 0.26$
- The weight for the criterion d : $w_d = \frac{20}{150} = 0.13$
- The weight for the criterion c : $w_c = \frac{11}{150} = 0.07$

In view of above, the formula for the calculation of the multi-criteria utility function U is: $U_{SAW}(T_j) = 0.27 p_j + 0.26 i_j + 0.26 s_j + 0.13 d_j + 0.07 c_j$ (2)

A simple example of the evaluation of the theory topics in INTATU is presented below. The user of our example is a 60-year-old male that has been assigned to the stereotypes 'Patient' and 'Expert'. When the user interacts with the system, INTATU evaluates all theory topics by estimating the value of the multi-criteria utility function

($U_{SAW}(T_j)$) for each one. For example, the value of the multi-criteria utility function for T_1 , which corresponds to the theory topic ‘Symptom – SL’ (SL stands for Simple Language), has been estimated to 0.87 and has been selected to be proposed to the user as it maximised the values of $U_{SAW}(T_j)$. The value of $U_{SAW}(T_1)$ has been estimated by applying the values of the criteria that are acquired by the activated stereotypes. As the user has only interacted with the system once before, INTATU uses stereotypes for acquiring the values of the criteria. More specifically, the stereotype ‘Patient’ provides the values of the criteria $p_1 = 1$, $i_1 = 0.8$, $s_1 = 1$ and $d_1 = 0.5$ whereas the stereotype ‘Expert’ provides the value of the criterion $c_1 = 1$. Applying these values to the formula (1) the value of the multi-criteria utility function is calculated:

$$U_{SAW}(T_1) = 0.27 * 1 + 0.26 * 0.8 + 0.26 * 1 + 0.13 * 0.5 + 0.07 * 1 = 0.87 .$$

On the other hand, one of the topics that have been evaluated but is not proposed to the user, because the value of the multi-criteria utility function ($U_{SAW}(T_9) = 0.25$) is very low, is the theory topic T_9 , which corresponds to the theory topic ‘Research and Progress in surgical procedures – MT’ (MT stands for Medical Terminology).

7 Conclusions

In this paper, we described INTATU, an ITS about Atheromatosis. The main characteristic of the system is that it can adapt its interaction to individual users dynamically. In order to achieve that, INTATU’s inference mechanism uses double stereotypes and a Multi-Criteria Decision Making technique called SAW. More specifically, the system uses stereotypes to maintain information about the interests, background knowledge and needs of all targeted user groups.

The main reason for the application of stereotypes is that they provide a set of default assumptions, which can be very useful during hypotheses generation about the users’ interests. Generation of default assumptions can prove very effective for modelling a large proportion of users. The default assumptions of the stereotypes are used in combination with SAW in order to evaluate all alternative theory topics and select the one that seems more appropriate for the user interacting with the system. The system selects to present to the user the theory topic that maximises the multi-criteria utility function.

References

1. Brusilovsky, P., Maybury, M. T.: From Adaptive Hypermedia to the Adaptive Web. *Communications of the ACM*, 45(5) (2002) 31-33.
2. Govindasamy, T.: Successful implementation of e-learning; pedagogical considerations. *The Internet and Higher Education*, 4(3-4) (2001) 287-299.

3. Fishburn, P.C.: Additive Utilities with Incomplete Product Set: Applications to Priorities and Assignments, *Operations Research* (1967).
4. Freedman, R., Cho, B.I., Glass, M., Zhou, Y., Kim, J.H., Mills, B., Yang, F.J., Evens, M.W.: Adaptive Processing in a Medical Intelligent Tutoring System, *Workshop on Adaptation in Dialogue Systems, NAACL 2001, Pittsburgh* (1967).
5. Hospers, M., Kroezen, E., Nijholt, A., op den Akker, H.J.A., Heylen, D.: An agent-based intelligent tutoring system for nurse education. In Care, J. Nealon and A. Moreno (eds): *Applications of Intelligent Agents in Health*, (2003) 143-159.
6. Hwang, C.L., Yoon, K.: *Multiple Attribute Decision Making: Methods and Applications. Lecture Notes in Economics and Mathematical Systems, Vol 186*, Springer, Berlin/Heidelberg/New York (1981).
7. Kay, J.: Stereotypes, Student Models and Scrutability. In G., Gautier, C., Frasson and K., VanLehn (Eds.) *Lecture Notes in Computer Science, Intelligent Tutoring Systems, Proceedings of the 5th International Conference on Intelligent Tutoring Systems* (2000) 19-30.
8. Koutsojannis, C., Hatzilygeroudis I., Prentzas, J.: A Web-Based Intelligent Tutoring System Teaching Health Care Technology, *Proceedings of the IASTED International Conference on Web-Based Education (WBE-2004), Feb. 16-18, 2004, Innsbruck, Austria* (2004) 607-612.
9. Lillehaug, S.I., Lajoie, S.P.: AI in medical education – another grand challenge for medical informatics. *Artificial Intelligence in Medicine* 12 (1998) 197-225.
10. Rich, E.: Stereotypes and User Modeling. In A. Kobsa & W. Wahlster (eds.), *User Models in Dialog Systems* (1989) 199-214.
11. Rich, E.: Users are individuals: individualizing user models. *International Journal of Human-Computer Studies* 51 (1999) 323-338.
12. Sheikhzadeh, A., Ehlermann, P.: Atheromatosis disease of the thoracic aorta and systemic embolism – Clinical picture and therapeutic challenge. *Zeitschrift fur kardiologie* 93 (2004) 10-17.
13. Suebnukarn, S., Haddawy, P.: *Modeling Individual and Collaborative Problem Solving in Medical Problem-Based Learning* (2005)
14. Tweddle, S., James, C., Daniels, H., Davies, D., Harvey, P., James, N., Mossman, J., Woolf, E.: Use of a Web site for learning about cancer, *Computers & Education* 35 (2000) 309-325.
15. Virvou M.: Intelligence and Adaptivity in Human-Computer Interaction concerning Biotechnological Software Users. *Proceedings of the 5th International Workshop on Mathematical Methods in Scattering Theory and Biomedical Technology* (2001).

Evolutionary Tuning of Combined Multiple Models

Gregor Stiglic and Peter Kokol

Faculty of Electrical Engineering and Computer Science, University of Maribor,
2000 Maribor, Slovenia

{Gregor.Stiglic, Kokol}@uni-mb.si

Abstract. In data mining, hybrid intelligent systems present a synergistic combination of multiple approaches to develop the next generation of intelligent systems. Our paper presents an integration of a Combined Multiple Models (CMM) technique with an evolutionary approach that is used for tuning of parameters. Proposed hybrid classifier was tested in microarray analysis domain. This domain was chosen intentionally, because of the nature of Combined Multiple Models classifiers that are specialized in solving problems with high dimensionality and contain low number of samples. Evolutionary tuning of parameters in combination with validation dataset enables fine tuning of parameters that are usually set to pre-defined values. Using this technique we made another step in leveling the accuracy of comprehensible classifiers to those represented by ensembles of classifiers.

1 Introduction

Recently developed microarray technology allows measurement of expression levels for thousands of genes simultaneously. Using classification or clustering techniques we are able to search for significant genes that can help us identifying different clinical states of the patient. To improve the accuracy of classification in microarray data many new methods have been developed. Unfortunately the majority of these approaches were black box methods, so the next logical step was the development of techniques that would improve comprehensibility of almost incomprehensible classification algorithms like neural networks or ensembles of classifiers.

The main scheme for such methods is rule extraction, that is, symbolic rules are extracted from the ‘black-box’ model. An example of such system was presented by Domingos in [1] where he generalized the concept of oracle queries. His idea was that any ‘black-box’ model can be captured in a comprehensible classification model by using additional ‘artificial’ examples that are labeled according to the original model.

So why would this idea work? To answer this question we should return to the initial problem in the classification of microarray data – i.e. small number of samples on one side and enormous number of attributes on the other side. The CMM method proposed by Domingos represents a solution to the small number of samples problem

by multiplying them in form of artificial data points. The only problem here is that CMM was not optimized for continuous values, but was meant to be used on all kinds of datasets in the machine learning domain. In case of microarray analysis datasets we always work only with continuous values that represent gene expressions. To solve this problem we must optimize the CMM to be effectively applied on continuous values and because of the complexity of this optimization if performed manually, we decided to use evolutionary algorithms for the tuning of CMM parameters.

Evolutionary algorithms are proven powerful tools for solving optimization problems using mimicking mechanisms of natural evolution. The most popular type of evolutionary algorithms are genetic algorithms where approximate solutions to optimization and search problems are found using techniques inspired by evolutionary biology such as inheritance, mutation, natural selection, and recombination (or crossover).

In this paper we propose a novel hybrid classification technique that uses artificial data points generation optimized for continuous values combined with evolutionary parameter tuning.

The rest of this paper is organized as follows. In Section 2, our methods for Combined Multiple Models and Evolutionary Tuning of Parameters are presented. This section also includes subsections on our proposed modifications to CMM and the implementation of Evolutionary Parameter Tuning. Next Section presents the experimental settings where we describe experimental environment and is followed by Section 4 where the results and comparison to other classification methods are presented. In the last section, the main contribution of this paper is summarized and several issues for future works are indicated.

2 Evolutionary Combined Multiple Models

Combined Multiple Models (CMM) is one of the names used for specific methods that are able to build single comprehensible models from an ensemble of models and was presented in [2]. CMM was later studied and improved by Estruch et al. in [3]. Estruch used a new term for CMM - “mimetic classifiers”, because of their ability to imitate more complex and more accurate classifiers.

2.1 Combined Multiple Models

Basic idea of CMM is to build a single classifier that would retain most of the accuracy gains of the ensemble models. This is done by adding additional artificial data points to the learning dataset. Those additional data points are then classified (i.e. labeled) by applying the ensemble of classifiers that was trained on the learning dataset. The next step is joining the original training dataset examples with the new “artificial” dataset examples. This final dataset is used to build a single comprehensible classifier. The whole process is shown in Fig. 1.

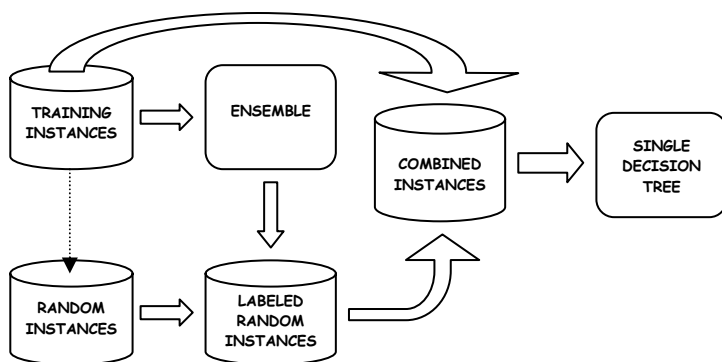


Fig. 1. Combined Multiple Models

The idea of generating the “artificial data points” when building classifiers, was already used in several papers. One of the first such methods is active learning method proposed by Cohn et al. [4]. Another application of artificial examples was presented by Craven and Shavlik in [5] where they describe the learning of decision trees from neural networks. This approach was later used in several papers on neural networks knowledge extraction.

2.2 Proposed Modifications

This paper presents the CMM based method that is specialized for microarray dataset classification problems. Optimization of the original method was done on artificial data points creation due to specific structure of the microarray datasets. Opposite to the original research [2], where most of the best results were achieved on the datasets containing nominal values, microarray analysis presents continuous-valued datasets. Therefore a new method called Combined Multiple Models for Continuous values (CMMC) is proposed. In this paper two new artificial data points creation techniques are examined which will be referred to as CMMC-1 and CMMC-2.

Both methods are based on multiplication of the original training set instances. Data points are generated from original training set by creating copies of original training set instances by minimally changing the values of attributes.

First method is based on the variance of the gene expression values and each attribute can be changed by adding the random value from the interval $\pm[\sigma, (1/3)\sigma]$ to the original gene expression value. Because of the large number of attributes we change only 50% randomly selected attributes. Result of such data point multiplication is a wide dispersion of the points around their base data point, but original training set distribution of the samples is still preserved.

Second method tries to maintain the original distribution on even tighter area than the first one, especially when data points lie tightly together. This is done by generating the random points in the interval $x \pm d$, where x is the value of the attribute and d is the distance to the nearest neighbor value of this attribute. Again only 50% of

attributes are randomly selected for modification. From here on our proposed methods follow the steps proposed in [3] – i.e. labeling of the artificial data points and building of the final classifier. The only difference is in final step where we build a decision tree instead of a set of rules.

2.3 Evolutionary Tuning of Parameters

Evolutionary computation (EC) mimics the processes of biological evolution with its ideas of natural selection and survival of the fittest to provide solutions for global optimization problems. EC uses algorithms based on natural evolution principles explained by Darwinian Natural Selection to solve a wide range of problems, which usually cannot be solved by conventional optimization techniques. In its primary forms EC included Evolutionary Strategies proposed by Schwefel et al. [6], followed by Genetic Algorithms proposed by Holland in [7], and Genetic Programming by Koza [8].

This research uses genetic algorithms for tuning of parameters that cannot be deterministically set when optimal performance from CMM method is requested. Our proposed modification to original CMM method requires even more “fine-tuning” of additional parameters when using CMM for continuous values.

Therefore we propose genetic algorithm to find the optimal balance between the boundary values of CMMC artificial data points creation process. Additional to boundary values we try to optimize the pruning level of final CMMC decision trees. Fig. 2 represents a graphical representation of the genotype that was used during evolutionary parameter tuning. First bit represents one of the two possible CMMC subtypes described in section 2.1. The other three values are coded as two-bit genes and represent lower and upper CMMC boundary values and pruning level of the resulting C4.5 decision tree. In the lowest row of Table 1 possible gene values are presented – for example lower boundary can represent 0, 10, 20 or 30 percent of CMMC’s maximal boundary value.

1		2		3		4		5		6		7	
CMMC Type		CMMC Lower Boundary				CMMC Upper Boundary				C4.5 Pruning Confidence Threshold			
0	1	0.0	0.1	0.2	0.3	0.4	0.6	0.8	1.0	0.0	0.1	0.2	0.3

Fig. 2. Genotype of CMMC parameter tuning

To set the parameters on-line during the classification process we use a model presented in Fig. 3 which uses a part of training dataset as a parameter tuning validation dataset. Termination conditions for genetic algorithm can be time or iteration based. Our proposed model uses 10 iterations for improvement of the initial population. Because of computational limitations our model uses only 20 chromosomes; selection parameter was set to 5 chromosomes per generation to breed a new generation.

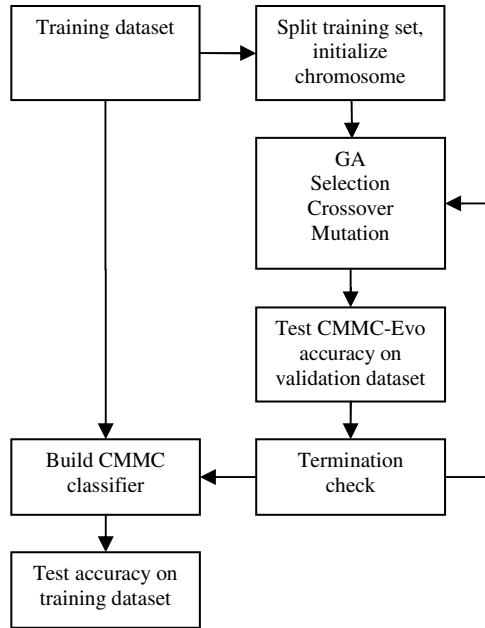


Fig. 3. Genetic parameter tuning model

3 Experimental Settings

To estimate the accuracy of proposed CMMC-Evo hybrid classifier and compare it to other classifiers, five microarray analysis datasets from Kent Ridge Bio-medical Data Set Repository [9] were used. The accuracy of our method is compared to J48 decision trees [10], default CMMC algorithms and Ada-boosting [11] using 100 iterations of J48 decision trees. All experiments were done using WEKA machine learning toolkit [10], which was used for comparison of the proposed method with the other machine learning algorithms. Five widely used publicly available gene expression datasets that were used in evaluation of the proposed method are presented in this chapter.

Leukemia dataset (amlall) comes from the research on acute leukemia by Golub et al. [12]. Dataset consists of 38 bone marrow samples from which 27 belong to acute lymphoblastic leukemia (ALL) and 11 to acute myeloid leukemia (AML). Each sample consists of probes for 6817 human genes. Golub used this dataset for training. Another 34 samples of testing data were used consisting of 20 ALL and 14 AML samples. Because we used leave-one-out cross-validation, we were able to make tests on all samples together (72).

Breast cancer dataset (breast) was published in [13] and consists of extremely large number of scanned gene expressions. It includes data on 24481 genes for 78 patients, 34 of which are from patients who had developed distance metastases within

5 years, the rest 44 samples are from patients who remained healthy from the disease after their initial diagnosis for interval of at least 5 years.

Lung cancer dataset (lung) includes the largest number of samples in our experiment. It includes 12533 gene expression measurements for each of 181 tissue samples. The initial research was done by Gordon et al. [14] where they try to classify malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung.

Leukemia2 dataset (mll) tries to discern between 3 types of leukemia (ALL, MLL, AML). Dataset contains 72 patient samples, each of them containing 12582 gene expression measurements. Data was collected by Armstrong et al. and results published in [15].

Prostate Tumor dataset (prostate) contains 52 prostate tumor samples and 50 non-tumor (labeled as "Normal") prostate samples with around 12600 genes. The original study was conducted by Singh et al. in [16].

4 Results

In order to evaluate our proposed method all experiments consisted of 10-fold cross-validation tests that were additionally repeated 20 times, because of the randomness present in some of the used classification methods. The results of average accuracy for four different classifiers are presented in Table 1. It can be seen that our CMMC-Evo classifier clearly outperformed the classical decision tree, but still lacks some accuracy to perform at the level of ensemble based classifiers. This fact only shows that combining multiple models together still gains the most in terms of accuracy.

Table 1. Comparison of accuracy

Dataset	J48	CMMC (Average)	CMMC-Evo	Ada-Boost (J48 trees)
<i>amlall</i>	83.58	89.73	90.17	91.20
<i>breast</i>	64.81	67.45	74.11	86.97
<i>lung</i>	97.06	97.27	98.33	97.51
<i>mll</i>	85.11	88.79	92.86	89.55
<i>prostate</i>	86.83	87.13	87.36	94.25
Average	83.47	86.07	88.61	91.89

Fig. 4 presents the results of evolutionary parameter tuning for two parameters that have direct influence on performance of CMMC classifier. In our previous experiments using CMMC [17], we usually used the default values for those two parameters that were based on a few runs of the algorithm. The lower boundary was usually set to 0.33 and the upper to 1.00. Our evolutionary parameter tuning confirmed the upper boundary value and slightly changed our opinion on the most appropriate lower boundary value that should lie around 0.2.

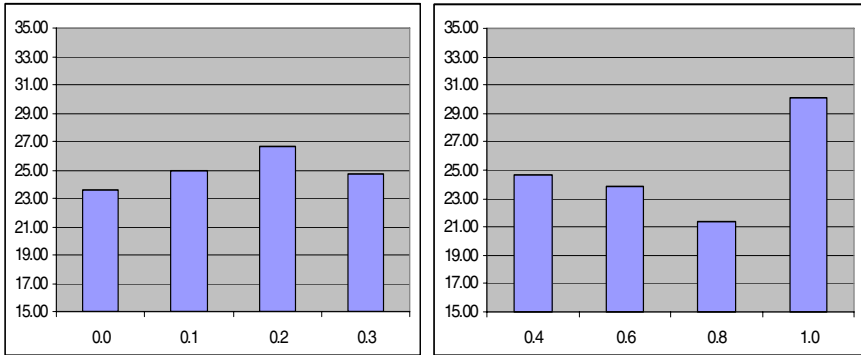


Fig. 4. Distribution of lower (left) and upper (right) boundary CMMC parameter after optimizations

5 Discussion

This paper presents a hybrid classification method based on modified CMM classification algorithm that was combined with evolutionary parameter tuning using genetic algorithm. The main advantage of the proposed CMMC-Evo over similar decision tree ensemble classification methods is not accuracy, but combination of improved accuracy and good comprehensibility of the classifier. The accuracy of the proposed classifier building technique still cannot be compared to ensembles of similar classifiers, but we get improved comprehensibility in addition to significant advantage over classical single decision trees in terms of accuracy.

An interesting question for the future research is whether the size of genome that was used for parameter optimization can be extended with more parameters that have direct impact on accuracy of the CMM classification method. That kind of parameter could be the number of artificial data points generated that is usually set to some pre-defined value.

References

1. P. Domingos, Knowledge acquisition from examples via multiple models. In Proc. of the 14th International Conference on Machine Learning, pp. 98 – 106, Morgan Kauffman, 1997.
2. P. Domingos, “Knowledge Discovery Via Multiple Models”, *Intelligent Data Analysis*, vol. 2 no.1-4, pp. 187-202, 1998.
3. V. Estruch, C. Ferri, J. Hernández-Orallo, M. J. Ramírez-Quintana, “Simple Mimetic Classifiers,” in Proc. IAPR International Conference on Machine Learning and Data Mining (MLDM2003), pp. 156-171, 2003.
4. D. Cohn, L. Atlas and R. Ladner, “Improving generalization with active learning,” *Machine Learning*, vol. 15, pp. 201-221, 1994.

5. M.W. Craven and J. W. Shavlik, "Extracting comprehensible concept representations from trained neural networks," in Working Notes on the IJCAI'95 Workshop on Comprehensibility in Machine Learning, Montreal, Canada, pp.61-75, 1995.
6. H.P. Schwefel, *Kybernetische Evolution als Strategie der experimentellen Forschung in der Strömungstechnik*, Master's thesis, Technical University of Berlin, 1965.
7. J.H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press, 1975.
8. J.R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press, 1992.
9. J. Li and H. Liu, "Ensembles of cascading trees," in Proc. IEEE International Conference on Data Mining, IEEE Computer Society, Melbourne, Florida, pp. 585, 2003.
10. I. H. Witten, E. Frank, *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 1999.
11. Y. Freund, R.E. Schapire, "Experiments with a New Boosting Algorithm," in Proc. of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann, pp. 148-156, 1996.
12. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, 1999.
13. L. J. van 't Veer, H. Dai, M. J. van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer", *Nature*, no. 415 pp. 530-536, 2002.
14. G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswami, W. G. Richards, D. J. Sugarbaker, R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, no. 62, pp. 4963-4967, 2002.
15. S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia," *Nat Genet.*, vol. 30 no. 1, pp. 41-47, 2002.
16. D. Singh et al. , "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*. Vol. 1, pp. 203-209, 2002.
17. G. Stiglic, P. Kokol, "Knowledge extraction from microarray datasets using combined multiple models," in Proc. IEEE International Conference on Data Mining Workshop on Foundation of semantic oriented data and web mining, Houston, Texas, pp. 75-80, 2005.

A Similarity Search Algorithm to Predict Protein Structures

Jiyuan An¹ and Yi-Ping Phoebe Chen^{1,2}

¹ School of Information Technology, Deakin University, VIC 3125, Australia

² Australian Research Council Centre in Bioinformatics

{jiyuan, phoebe}@deakin.edu.au

Abstract. Accurate prediction of protein structures is very important for many applications such as drug discovery and biotechnology. Building side chains is an essential to get any reliable prediction of the protein structure for any given a protein main chain conformation. Most of the methods that predict side chain conformations use statistically generated data from known protein structures. It is a computationally intractable problem to search suitable side chains from all possible rotamers simultaneously using information of known protein structures. Reducing the number of possibility is a main issue to predict side chain conformation. This paper proposes an enumeration based similarity search algorithm to predict side chain conformations. By introducing “beam search” technique, a significant number of unrelated side chain rotamers can easily be eliminated. As a result, we can search for suitable residue side chains from all possible side chain conformations.

1 Introduction

Although over one million protein primary sequences have so far been identified, 3D coordinate information is determined for only around 28 thousand proteins using experimental techniques [10]. Voluminous work has been done in bioinformatics to predict protein structure to complement ineffective experimental techniques for protein structure. More than 80% of the protein structures were found using X-ray crystallography method and 16% using nuclear magnetic resonance (NMR); Theoretical modeling method account for 2% for the predicted structure. Only a limited number of structures were determined by other methods [10]. Experimental techniques have so far been mainly used to predict protein structure.

It is a vital part in predicting accurate protein structures to build side chains for a protein main chain conformation. Majority of the methods that predict side chain conformations rely on statistical data generated from known protein structures. However, it is a computationally intractable problem for building side chains from all possible rotamers simultaneously using information of known protein structures. Reducing the number of possible conformations is a main issue to predict side chain conformation. Among several methods that are explored to reduce the possible number of side chain rotamers, the following two methods are prominent: 1) building of side chains from the conformational searching of rotamer libraries [2], and 2) creation of discriminatory function [9][8].

In this paper, we analyze the all-atom distances from high-resolution protein structures to find the conformational preferences of amino acid residues. A score function proposed by Samudrala and Moult [9] is used to search the suitable side chain rotamers. To avoid exponential increase of side chain rotamer candidates, this paper adopts “beam search” technique [6]. The algorithm of “beam search” contains a list of the k -best candidates at each searching step, rather than all candidates or a single best candidate. We used a native protein structure to demonstrate the effectiveness of our approach in the prediction of side chain conformation.

2 Background

Proteins are polypeptide chains made up of amino acid residues. The amino acids are linked together in a definite sequence. The sequence of side chains determines the unique characters about a particular protein, which includes its biological function and its specific three-dimensional structure. Even though the sequences of many proteins are known, their structures are to be determined as accurate as possible.

2.1 The Problem Descriptions

To build side chains on a fixed main chain, most of the existing methods use algorithm to search for a suitable side chain conformation from all possible rotamer conformations. Figure 1 (A) shows three amino acids in a protein. All amino acid with exception of glycine, has side chain called R_α , which can rotate in different angles such as χ_1 and χ_2 . Different angles cause different structures of side chain. For example, In Figure 1 (B), the side chain structure is determined by rotamer $\chi_1, \chi_2, \chi_3, \dots$. Our algorithm finds out the suitable rotamers from all possible angles: $\chi_1, \chi_2, \chi_3, \dots$.

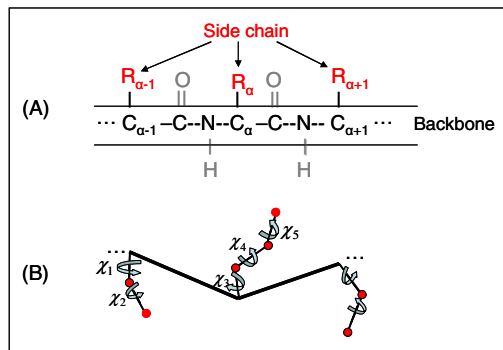


Fig. 1. Side chain

The traditional similarity search is used to find near objects from a database. The answer of the k^{th} nearest neighbor (k-NN) is the objects that have the shortest distance from the query object. However, we do not have any knowledge of the query protein structure during the whole process of predicting protein structure. Otherwise, if we

know the exact side chain conformations, we do not need to do prediction. So we have to create a reasonable criterion to evaluate the side chain conformations “good” or “bad”.

A discriminatory function proposed by Samudrala et al.[9], is a quantitative formula that reflects the preference of all-atoms in protein structures. We have adopted this as a criterion to evaluate the side chain conformations. Effectiveness of the discriminatory function has been further investigated [1] [8].

The discriminatory function principally calculates the probabilities of all residue-specific atoms in different distance intervals. All non-hydrogen atoms are considered. The atom types are considered to be residue specific; for example, the C α of a glycine is different from that of an alanine. The total number of atom types is 167. The detail for the atom types can be found in [9]. The distance that ranges from 0Å to 20 Å is considered and is divided into 18 distance bins. The atom pairs that have more than 20 Å, are excluded. Every atom pair has a score value. For an atom pair, a and b, the score is represented as below:

$$S(d_{ab}) = -\ln \frac{N(d_{ab}) / \sum_d N(d_{ab})}{\sum_{ab} N(d_{ab}) / \sum_d \sum_{ab} N(d_{ab})} \quad (1)$$

where $N(d_{ab})$ is the number of observations of atom types a and b in a particular distance bin d, $\sum_d N(d_{ab})$ is the total number of observations of atom types, a and b for all distance bins d, $\sum_{ab} N(d_{ab})$ is the total number of observations of atom types, a and b in a particular distance bin d, summed up over each atom types, and $\sum_d \sum_{ab} N(d_{ab})$ is the total number of observations of all pairs of atom types a and b, summed over all distance bins d.

3 The Preferences of Side Chain Rotamers

To find similar objects, the reasonable definition of distance is an important step. Using “good” definition, we can get appropriate results from the spatial retrieval. If a “bad” distance is defined, good results we can not be expected as the incorrect criterion can not produce correct results. Therefore, selection of a suitable criteria function is essential to evaluate the correctness of protein structures and thereby predicts protein structure more effectively.

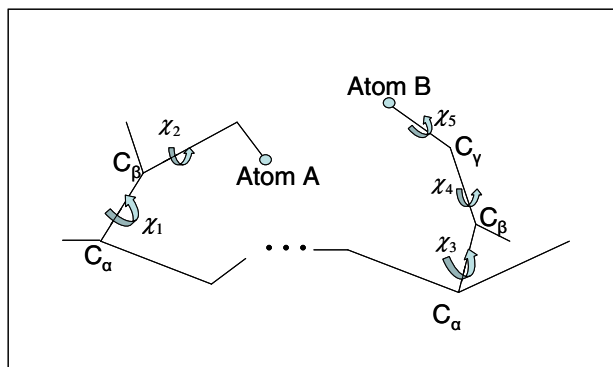
In order to test the preferences of amino acid residues, we use the same data set of [8]. We select the proteins with amino acid sequences less than 25% identical to each other. All the structures were determined using NMR methods and have either a resolution greater than 1.5Å or R-factor greater than 0.2.

Table 1 lists the proteins used to test the preferences of all pairs of amino acid residues.

The different pairs of atom types have different preferred distances. This reflects the importance of the side chain conformations. Figure 2 shows two amino acid residues. Different angles such as χ_1 , and χ_2 produce different residue rotamers. If atom type A (in the left amino acid) prefers to be close to the atom type B (in the right other amino acid), their angles $\chi_1, \chi_2, \dots, \chi_5$ are changed to make two atoms A and B come close as shown in the figure.

Table 1. The proteins used to calculate in discriminatory function

PDB code	# of residues	Resolution(\AA)	R-Value	Name
1cbn	48	0.83	0.16	Crambin
1ccr	111	1.5	0.19	Cytochrome C
1cus	197	1.25	0.16	Cutinas
1pmy	123	1.5	0.2	Pseudoazurin(Cupredoxin)
1ptx	64	1.3	0.15	Scorpion toxin II
1xnb	185	1.49	0.17	Xylanase
2end	137	1.45	0.16	Endonuclease V
2hbg	147	1.5	0.13	Hemoglobin(bloodworm)
2ihl	129	1.4	0.17	Lysozyme(Japanese Quail)
2sga	181	1.5	0.13	Proteinase A
9rnt	104	1.5	0.14	Ribonuclease T1

**Fig. 2.** Side chain rotamers

All residues have more than one atom, therefore the residue rotamers are impacted by all the atoms in all different amino acid residues. The two strongest pair usually decides the distance of the two residues. For example, cysteine sulphur linkage is very strong and so two residue rotamers of cysteine change their angles such as χ_1 and χ_2 to let the distance of two sulphur atoms very close to form sulphur bridge. Note that it does not mean that every pair of cysteine forms Sulphur Bridge.

To confirm the preferences of side chain residues, we test the proteins listed in Table 1. We test the preferences of 20 amino acid residues in 4 distance bins as shown in **Table 2.**

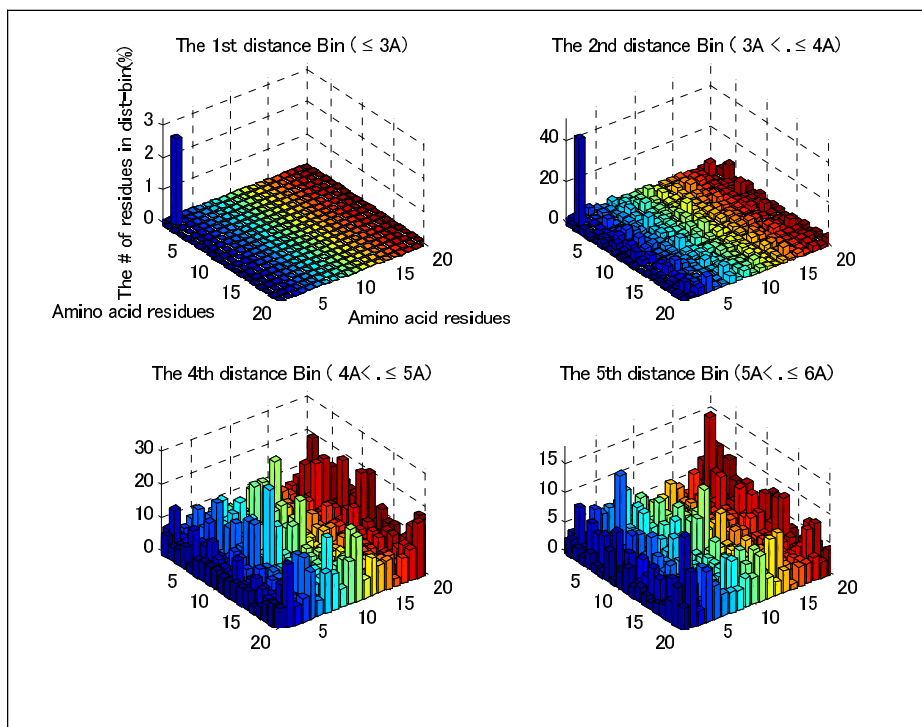


Fig. 3. The preferences of 20-amino acid residues

Table 2. The intervals of distance bins

Distance Bin No.	The interval
1 st	0-3 Å
2 nd	3-4 Å
3 rd	4-5 Å
4 th	5-6 Å

Figure 3 presents the histograms of the pairs of residues appeared in a distance bin. The two horizontal axes present 20 X 20 amino acid residues. There are 400 (20X20) small squares. Each square presents a pair of amino acid residue. The vertical axis presents the percentage of a pair of amino acid residues appeared in the distance bin. For example, in the first distance bin, most of the pairs of amino acid residues do not appear in the distance interval of (0, 3) but only one pair of amino acid residues. Cysteine-cysteine pairs are nearly 3% in the distance range of (0, 3). It is because they form sulphur bridge. It is very interesting to note that more than 40% cysteine-cysteine linkage produce sulphur bridge. We can use this information to enhance the accuracy of protein prediction.

From the Figure 3, we can infer that the different amino acid residues have different preferences. This allows us to predict side chain conformations using the discriminatory

function. Samudrala and Moult [8] use the discriminatory function to predict side chain conformation, but they did not give the distributions of preferences of the pairs of amino acid residues. By using the distribution information of each pair of amino acid residues, many unrelated side chain rotamers can be pruned. The possible number of side chain rotamers can then be reduced. As a result, the protein prediction algorithms can become more effective.

4 Our Approach

As discussed in the introduction, we use the discriminatory function [9] as a criterion to search suitable side chain rotamers. We employ the “beam search” technique to reduce the huge number of side chain rotamer’s candidates.

As described in [10], to reflect the preference of all items reasonably, the total number of residue-specific atom types is 167; hydrogen atom is not accounted. We use scores to evaluate the preferences of all pairs of atom types using eq.(1). The residue side chain conformation is predicted using one amino acid rotamer at a time. Table 3 shows the algorithm for predicting side chain conformations. In this paper, we only explain the major steps involved in the algorithm in Table 3:

Step 4:

All first side chain rotamers are stored into Π which remembers rotamer combinations.

Step 6-10:

All the next side chain rotamers are combined into each combination in Π . The score of each new combination is calculated using score table. After the loop, the number of queue becomes very large. The “beam search” technique is used to select most possible candidates.

Step 11 and 12:

The N-best candidates are moved to Π .

Table 3. Algorithm: Prediction of side chain conformation

<ol style="list-style-type: none"> 1. Algorithm prediction side chain conformation 2. Parameter 3. N : the number of candidates 4. All residue rotamers of 1st side chain $\rightarrow \Pi$ {initializing Π} 5. For $i = 2$: number of side chain residues 6. Foreach side chain rotamer R_i of residue i. 7. Foreach residue rotamer combination in Π 8. $\Pi \cup R_i \cup \text{score}(\Pi, R_i) \rightarrow \text{queue}$ 9. End 10. End 11. Sort queue 12. Select N-best candidates $\rightarrow \Pi$ 13. End 14. The best candidate is the answer of side chain conformation.
--

5 Experiment

We tested with 1crn protein. The protein 1crn.pdb has 46 side chain residues. Its resolution is 1.5 Å. For every atom type, we calculate its preference using eq.(1) and form a score table (167 X 167 X number of distance bin). The scores are calculated for the proteins listed in Table 1. In our experiment, we set torsion angles 3 and 5 for all side chain rotamers. The parameter, N, in the algorithm listed in Table 3, was set at 300. The experimental results show that most side chains match their original side chain conformations; only side chains 2 and 1 go to wrong side chain rotamers and the correct rates are 44/46 and 45/46 in 3, and 5-torsion angles, respectively.

Even though several attempts [1][4] were made to predict side chain conformation, the problem of dealing with enormous number of side chain rotamer candidates is not fully solved. Their methods can not be applied to any large sized protein. In our method, the number of candidates is controlled using “beam search” technique. In principle, our method can deal with any size of proteins as only the best candidates are kept at every step. Figure 4 shows the size of side chain candidates which are put in queue as shown in the algorithm of Table 3. From Figure 4, we can find the number of side chain rotamers is under control. This size of memory can be implemented in any personal computers.

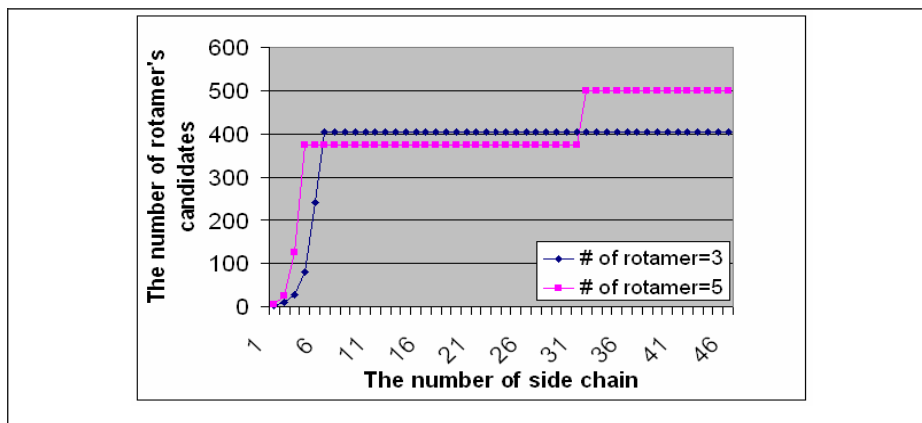


Fig. 4. The number of side chain residue combinations in queue

6 Conclusion

This paper has analyzed the preferences of all amino acid residues using nature protein structures. Based on the distance distribution of all atom types, we proposed a new method to predict side chain conformations. In order to counter the exponential expansion of the number of side chain rotamer candidates, we have employed “beam search” technique to limit the number of candidates without loss of much accuracy. The result of our experiment and analysis indicate that most correct answers are in the candidate set even when the candidate size is small. As a future work, we propose to use our method to test large protein structures.

References

- [1] Bahadur, D., Tomita, E., Suzuki, J. and Akutsu, T. Protein side-chain packing problem: A maximum common edge-weight clique algorithmic approach, *Journal of Bioinformatics and Computational Biology*, 2005.
- [2] Bower, M.J., Cohen, F.E., and Dunbrack Jr., R.L. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. (1997) *J. Mol. Biol.* 267: 1268–1282.
- [3] Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L.: A graph theory algorithm for protein side-chain prediction. (2003). *Protein Science* 12, 2001-2014.
- [4] Loughton, A., C., Prediction of protein side-chain conformations from local three-dimensional Homology relationships. *Journal of Mol. Biol.* (1994) 235 1088-1097.
- [5] Lovell, S. C., Word J. M., Richardson, J. S., and Richardson D. C.: The penultimate rotamer library. *Protein: structure, function and genetics.* (2000) 40 389-408.
- [6] Mitchell, T. T., *Machine Learning*. McGraw Hill, 1997.
- [7] Mosimann, S. Meleshko, R. and James, M A critical assessment of comparative molecular modeling of tertiary structures of proteins, *Protein.*(1995) 23 301 317.
- [8] Swiss-Prot Protein knowledgebase. The universal protein resource. <http://au.expasy.org/sprot/>
- [9] Samudrala, R. and Moulton, J.: Determinants of side chain conformational preferences in protein structures. *Protein Engineering.* (1998) 11(11). 991-997.
- [10] Samudrala, R. and Moulton, J.: An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Mol. Biol.* (1998) 275 893-914.
- [11] RCSB Protein Data Bank. <http://www.rcsb.org/pdb/>

Fuzzy-Evolutionary Synergism in an Intelligent Medical Diagnosis System

Constantinos Koutsojannis and Ioannis Hatzilygeroudis

Dep of Computer Engineering & Informatics, School of Engineering,
265 00 Patras, Hellas (Greece)
{ckoutsog, ihatz}@ceid.upatras.gr

Abstract. In this paper, we present the design, implementation and evaluation of HIGAS, a hybrid intelligent system that deals with diagnosis and treatment consultation of acid-base disturbances based on blood gas analysis data. The system mainly consists of a fuzzy expert system that incorporates an evolutionary algorithm in an off-line mode. The diagnosis process, the input variables and their values were modeled based on expert's knowledge and existing literature. The fuzzy rules are organized in groups to be able to simulate the diagnosis process. Differential evolution algorithm is used to fine-tune the membership functions of the fuzzy variables. Medium scale experimental results show that HIGAS does better than its non-hybrid version, non-experts and other previous computer-based approaches.

Keywords: acid-base disturbances, hybrid expert systems, fuzzy-evolutionary synergism, computer diagnosis.

1 Introduction

Accurate diagnosis and treatment of electrolyte disturbances is an ability that only experienced doctors that have faced many patients for many years can perform. Unfortunately, even undergraduate studies on human physiology do not pay attention to this topic, so that many difficulties are faced during every-day clinical practice by general doctors, intensive care personnel, etc. Understanding a disturbance occurred after a cardiovascular shock or after an operation is crucial for a clinician who has to treat this serious situation for patients' life. There are two main acid-base balance disturbances, acidosis and alkalosis, further distinguished into metabolic acidosis, metabolic alkalosis, respiratory acidosis, respiratory alkalosis and their combinations (mixed disorders).

A number of attempts to tackle disturbances have already been proved successful, but not for all the types of them. Simple calculations [1, 2], diagrams [3, 4, 5] and other computer-based methods [6] are used to help doctors to evaluate and treat the mixed or plain disturbances. In [5], in order to examine the diagnostic validity of the proposed diagrammatic methods, arterial blood gas samples drawn from 114 Intensive Care Unit (ICU) patients were used. The samples were interpreted using the Grogono diagram [4] and the following approaches, as comparators: (a) The Siggaard-Andersen (S-A) chart [3], (b) the Oxygen Status Algorithm (OSA) [2] and (c) two

physicians with more than 10 years of experience in ICUs, considered as experts in acid-base balance disturbances. There, has been proved that the Grogono diagram gives better results than OSA, and both better than the S-A chart. However, the authors conclude that Grogono diagram cannot be safely used for the diagnosis of acid-base balance disturbances in everyday clinical practice, because it has been shown to provide inaccurate diagnoses in at least 25% of the cases. So, they suggest that the creation of a better computer-based system to assist at least non-expert doctors in making an initial diagnosis is still very desirable.

There have been some efforts to use intelligent systems to deal with aspects of the above problem [7, 8, 9, 10]. From them, [8] and [10] refer to infants related aspects. The rest deal with what we call ‘disturbances’ and not with all aspects of what we call ‘disorders’ (the causes of disturbances). They also don’t deal with treatment proposals.

Hybrid intelligent systems are systems that mix different intelligent methods and make them “work together” to achieve a better solution to a problem, compared to using a single method for the same problem. During last decade hybrid intelligent systems have been used to tackle medical problems [11]. From the above efforts, [7] and [10] use hybrid intelligent systems. In [6] a combination of frames and rules is used, whereas in [10] a combination of a back-propagation neural net and decision algorithms. Neither uses fuzzy sets to represent the inherent vagueness in some of the parameters.

In this paper, we present HIGAS, a Hybrid Intelligent system for the diagnosis and treatment of acid-base disturbances based on blood GAS analysis data. Diagnosis is achieved in two stages. In the first stage, diagnosis of the disturbance is made, whereas in the second, diagnosis of the possible disorder and corresponding treatment is made.

The paper is organized as follows: in Section 2 we introduce the medical knowledge involved and the diagnosis process model we designed. In Section 3, development issues of our intelligent system are presented. Section 4 presents evaluation results for the system and finally Section 5 concludes the paper.

2 Medical Knowledge Modelling

Acid-base state in a body fluid is physically determined by several independent variables. In blood plasma *in vivo*, the independent variables are: (1) PCO_2 ; (2) the ‘strong ion difference’ (SID), i.e. the difference between the sums of all the strong (fully dissociated, chemically non-reacting) cations (Na^+ , K^+ , Ca^{2+} , Mg^{2+}) and all the strong anions (Cl^- and other strong anions) and (3) concentrations of nonvolatile weak acids (i.e., for each of them, the sum of its dissociated and undissociated forms, Stewart’s symbol A_{tot}). Normal acid-base status is obtained when the independent variables have normal (empirically established) values. Abnormality of one or more of the independent variables underlies all acid-base disturbances. Adjustment of the independent variables is the essence of all therapeutic interventions, because none of the “dependent variables” (e.g., pH, $[HCO_3^-]$) can be changed primarily or individually: all dependent variables change simultaneously, if and only if one or more of the independent variables changes.

A classification of acid-base disturbances based on this view is shown in Table 1a. Metabolic acid-base disturbances can be caused by two types of abnormalities: mixed and unmixed abnormal concentrations of nonvolatile weak acids.

Table 1a. Examples of acid-base disturbances

Disturbance	pH	Primary Disturbance	Expected response
Unmixed Disturbances			
A. Metabolic acidosis	< 7.38	HCO ₃ < 22 meq/l	DPCO ₂ (↓) = (1.0-1.3) DHCO ₃
B. Metabolic alkalosis	> 7.42	HCO ₃ > 25 meq/l	DPCO ₂ (↑) = (0.4-0.9) DHCO ₃
C. Respiratory acidosis C1. Acute C C2. Chronic C C3. C1 and A or B C4. C1 → C2	< 7.38	PCO ₂ > 43 mmHg	DHCO ₃ (↑) = (0.08-0.12)DPCO ₂ DHCO ₃ (↑) = (0.25-0.55)DPCO ₂ DHCO ₃ (↑) = (0.12-0.19)DPCO ₂ DHCO ₃ (↑) = (0.16-0.25)DPCO ₂
...			
Mixed disturbances (primary and secondary)			
1) A and C	< 7.38	HCO ₃ < 22 meq/l	DPCO ₂ (↓) < 1.0 DHCO ₃ or PCO ₂ > 40 mmHg
2) A and D	< 7.38	HCO ₃ < 22 meq/l	DPCO ₂ (↓) > 1.3 DHCO ₃
...			
9) A and B	7.38-7.42 and AG > 14		

Table 1b. Examples of disorders that cause acid-base disturbances

Metabolic Alkalosis with AG = normal (10-14)				
	Na + K < Cl in urine samples	K > 5.5	HCO ₃ in dose of 0.5-2 meq/Kg	Urine PCO ₂ if alkalic
RTA II	No	No	Urine pH > 7.4 and HCO ₃ < 24 meq/l (FE) HCO ₃ > 15%	
RTA III	No	No	Urine pH > 7.4 and HCO ₃ = 24 meq/l (FE) HCO ₃ = 1-3 %	PCO ₂ > 70 mmHg
RTA I ...	No	No	Urine pH > 7.4 and HCO ₃ = 24 meq/l (FE) HCO ₃ = 1-3 %	Urine PCO ₂ = plasma PCO ₂

The initial knowledge on the field of acid-base disturbances has been acquired from experts as well as from the existing literature. Based on that, we constructed the model of Fig. 1 for the diagnosis and treatment process. According to that, initially, an expert clinician requires the following information from the blood gas analyser: (a) the pH value, (b) the HCO₃ concentration, (c) the partial pressure value of CO₂ (PCO₂) and (d) the Anion Gap value, to make an initial diagnosis, concerning the type of disturbance.

To confirm the disturbance diagnosis, it goes a step further by diagnosing the underlying disorder, which causes the disturbance. In this second stage, further information related to specific diagnostic laboratory tests is required. Variable dependences and diagnostic rules can be seen in Table 1b.

In blood-gas interpretation there is a complete interdependence of laboratory and clinical data and, although the former is often quite precise, the latter may not be, yet both must be accounted for and reconciled for a successful diagnostic solution. Dependencies also exist in that, whenever a major system is affected, the effect on the body may be global and other systems may follow into the destabilisation spiral, if the situation is not quickly rectified. Thus, measurement trends are closely observed.

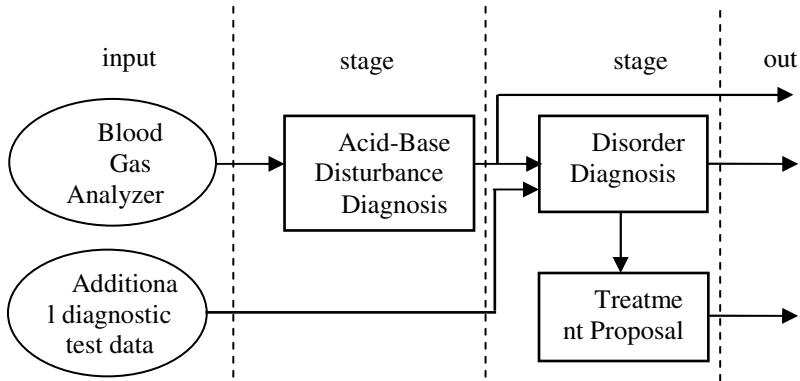


Fig. 1. General Model for Blood Gas Disturbance Diagnosis and Treatment Process

3 HIGAS Architecture and Development

3.1 Fuzzy Variables and Values

As it is known, real world medical knowledge is often characterized by inaccuracy. Medical terms do not usually have a clear-cut interpretation. Fuzzy logic makes possible to define inexact medical entities via fuzzy sets. During last decade, a number of hybrid techniques based on fuzzy sets and rules have appeared which have been applied to medical systems [12, 13]. One of the reasons is that fuzzy logic provides capabilities for approximate reasoning, which is reasoning with inaccurate (or fuzzy) values, expressed as linguistic terms.

Based on our expert, we specified a set of parameters that play a role in diagnosis for each of the entities in the process model (Fig. 1). Finally, we resulted in a number of parameters, which are distinguished in:

Input parameters: pH, HCO_3 concentration, partial pressure of CO_2 (PCO_2) and anion gap (which represent gas analyzer data). They are used in the form of some ratios (see Table 1a), which are represented as fuzzy variables (see Fig. 2).

Intermediate output parameters: disturbance_diagnosis (which represents the possible disturbance, i.e. one of: metabolic acidosis, metabolic alkalosis, respiratory acidosis, respiratory alkalosis and their combinations).

Intermediate input parameters: urine pH, plasma pH, Standard Base Excess, etc (which represent laboratory test data). They are also represented as fuzzy variables.

Output parameters: disorder_diagnosis (which represents the diagnosed disorder, which can be one of RTA II, gastric fluid loss, etc) and proposed_treatment (with as possible values: intravenous dilute hydrochloric acid, ammonium chloride, etc.).

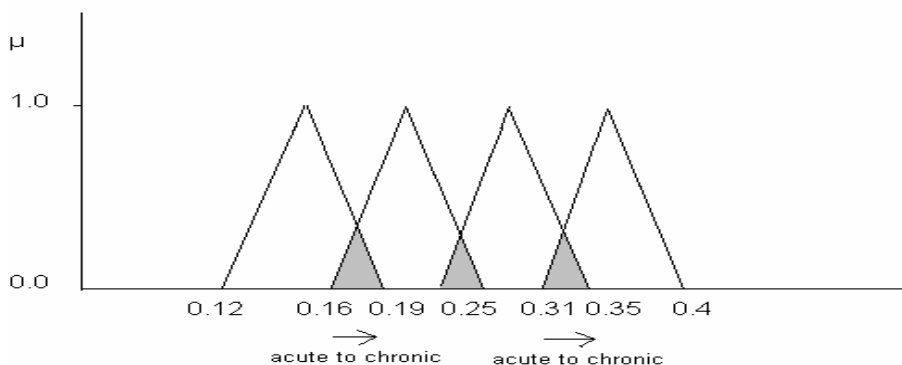


Fig. 2. Fuzzy values and membership function for 'DHCO3/DPCO2'

Fuzzy values and corresponding membership functions have been determined by the aid of the expert and the literature. Examples of values and corresponding membership functions are shown in Fig. 2. Due to the nature of the values and for better performance of the evolution algorithm, used to fine-tune them, we use only triangles to represent membership functions.

3.2 The Fuzzy Expert System

The developed fuzzy expert system has the typical structure of such systems [12, 13]. The *rule base* of the expert system includes (actually) *crisp* and *fuzzy rules*. A fuzzy rule includes one or more fuzzy variables. Definition of each fuzzy variable consists of definitions of its values. Each fuzzy value is represented by a *fuzzy set*, a range of crisp (i.e. non-linguistic) values with different degrees of membership to the set. The degrees are specified via a *membership function*.

Reasoning in such a system includes three stages: fuzzification, inference, defuzzification. In *fuzzification*, the crisp input values (from the fact database) are converted to membership degrees (fuzzy values). In the *inference* stage, the MIN method is used for the combination of a rule's conditions, to produce the membership value of the conclusion, and the MAX method is used to combine the conclusions of the rules. In *defuzzification*, the centroid method is used to convert a fuzzy output to a crisp value, where applicable.

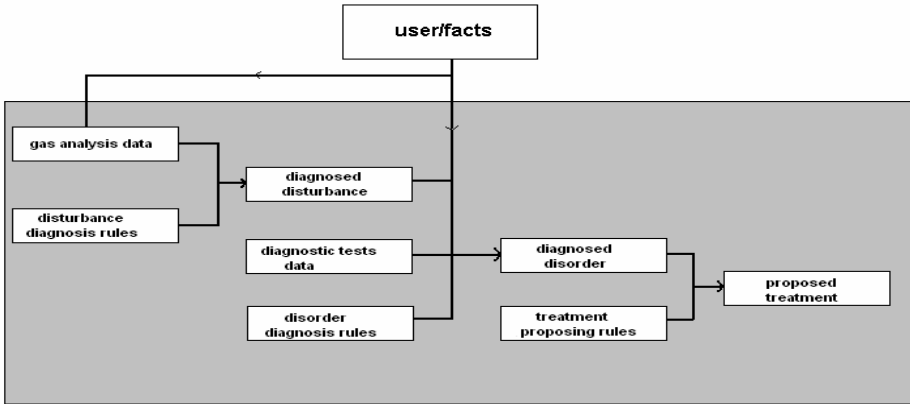


Fig. 3. Inference flow in HIGAS

The system gives its outputs in a semi-fuzzy form. E.g. the values of the diagnosed disorder with their corresponding membership values are presented to the user alongside system’s decision. This gives the user the opportunity to decide by himself/herself something different from the system in some special cases and also acts as some kind of explanation for the final decision of the system, given that membership values are presented as degrees of certainty.

To represent the diagnosis process model of Fig. 1, we organized rules in the rule base into three groups: *disturbance diagnosis rules*, *disorder diagnosis rules* and *treatment proposing rules*. The current patient data are stored as *facts*. Each time the reasoning process requires a value, it gets it from the facts list. In an interactive mode, it could be given by the user. Figure 3 presents how the rule groups and the facts/user are used/participates during the reasoning process to simulate the diagnosis process, whereas Figure 4 presents the architecture of HIGAS, where apart from the expert systems modules an evolutionary algorithm module is used off-line to fine-tune membership functions, as explained in the next section.

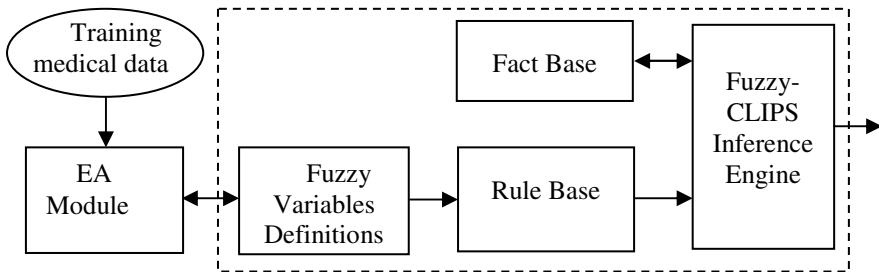


Fig. 4. The architecture of HIGAS

3.3 The EA Module

An evolutionary algorithm (EA) is used for membership functions optimization, which are initially intuitively chosen. Given that the optimization of fuzzy membership functions may involve many changes to many different functions, and that a change to one function may effect others, the large possible solution space for this problem is a natural candidate for an EA based approach despite many neuro-fuzzy approaches that use a gradient descent-learning algorithm to fine-tune the parameters of the fuzzy systems. The evolutionary algorithm optimises the antecedent and consequent membership functions of a number of fuzzy rules.

We use the ‘differential evolution’ (DE) algorithm [14, 15] to achieve that. As other evolutionary algorithms, DE maintains a population (a set) of solutions to the optimization problem at hand. The main idea in DE is to use vector differences in the creation of new candidate solutions, whereas traditional EAs rely on random perturbation (mutation) of a solution and mixing of two or more solutions (recombination). Another major difference is that the three phases of a standard EA (selection, recombination, and mutation) are combined to one operation, which is carried out for each individual. In the standard EA, each phase is performed on the entire population. In contrast, the DE algorithm iterates through the population and creates, for each population index i , a potential candidate $C[i]$ by vector addition (mutation) and a variant of uniform crossover (recombination). Selection is straightforward and very simple; the candidate solution $C[i]$ replaces $P[i]$ if it is better.

DE algorithm runs separately for each fuzzy variable. Let v be a fuzzy variable that has three fuzzy sets/values (A, B and C) and $\mu_A(v)$, $\mu_B(v)$, $\mu_C(v)$ be the three membership distributions over fuzzy sets A, B and C respectively. All membership curves are isosceles triangles (see e.g. Fig. 2). For optimization purposes, we code each triangle via six pairs of values (x_i, y_i) , i.e. two pairs for each edge of the triangle or one pair for each node of an edge. Each pair consists of the two co-ordinates of the corresponding node of an edge. Thus, the genome in the corresponding population has 19 fields, 18 ($=3 \times 6$) for the three fuzzy sets and one for the expected output. A number of training medical data is used to specify the “better” genome, the one that satisfies more training examples. The result of the DE algorithm application is changes to the co-ordinate pairs of the triangles of the membership functions of the values of a fuzzy variable.

3.4 Implementation Issues

The system has been implemented in FuzzyCLIPS 6.1b expert system shell [16]. Finally, about 72 rules and 10 templates have been constructed. Patient data is organized by using CLIPS templates. To implement reasoning flow, different priorities have been used for different rule groups. The EA module has been implemented using the software provided in [17].

4 System Evaluation

To evaluate HIGAS, we used 200 patient cases, successfully diagnosed and treated by the experts in critical care, from the database of the University Hospital of Patras,

Greece. We used two versions of HIGAS, one without the use of the EA module results (untuned) and the other after having tuned the membership functions of the fuzzy values of the fuzzy variables of the system via the EA module. We used 40% of the cases as the training data set for the DE algorithm needs and the rest 60% as the test data set. We also used a third participant in the experiment, a group of three non-expert clinical doctors in critical care.

4.1 HIGAS vs Clinical Doctors

The results are presented in Tables 2a and 2b. Table 2a refers to disorder diagnosis, whereas Table 2b to treatment proposal. We used ‘accuracy’ as the main metric accompanied by ‘specificity’ and ‘sensitivity’ for better interpretation of the results. The results show that the tuned version of HIGAS did better than any other participant (85% and 87%) as far as accuracy is concerned with a good balance between specificity and sensitivity.

Table 2a. Comparison of HIGAS with clinicians (disorder diagnosis)

DISORDER DIAGNOSIS	CLINICIANS			HIGAS	
	1st	2nd	3rd	TUN ED	UNTUN ED
Specificity	0.6	0.74	0.67	0.83	0.79
Sensitivity	0.6	0.70	0.65	0.88	0.83
Accuracy	0.6	0.72	0.66	0.85	0.81

Table 2b. Comparison of HYGAS with clinicians (proposed treatment)

TREATMENT	CLINICIANS			HIGAS	
	1st	2nd	3rd	TUN ED	UNTUN ED
Specificity	0.6	0.75	0.66	0.89	0.80
Sensitivity	0.7	0.80	0.69	0.83	0.84
Accuracy	0.7	0.77	0.67	0.87	0.82

4.2 HIGAS vs Computer-Based Systems

We also compared the tuned version of HIGAS with other two classical computer-based methods, the Grogono diagram and the Oxygen Status Algorithm (OSA), whose implementations are available in the web [4, 18]. The results, presented in Table 3, show the superiority of HIGAS. Notice, that the results concern only disturbance diagnosis, because those systems do not support disorder diagnosis and treatment proposal.

Table 3. Comparison of the HIGAS and other systems (disturbance diagnosis)

POSSIBLE DIAGNOSIS	GROGONO DIAGRAM	HIGAS	OXYGEN STATUS ALGORITHM
Specificity	0.75	0.85	0.71
Sensitivity	0.77	0.89	0.69
Accuracy	0.76	0.87	0.70

5 Conclusions

In this paper, we present HIGAS, a hybrid intelligent system that deals with diagnosis and treatment of blood gas (acid-base) disturbances and disorders. The diagnosis process was modeled based on expert's knowledge and the existing literature. Fuzzy variables were specified based again on expert's knowledge. A characteristic of the system is the synergism between an EA module and a fuzzy rule base. The DE algorithm is used to tune the membership functions of the fuzzy values of the fuzzy variables. This improves the accuracy of the system. Medium scale experimental results showed that HIGAS did quite better than non-experts and other systems, but worse than the expert.

There are two directions that the system can be further improved. First, concerning its performance, a more drastic way of tuning could be applied. Instead of tuning the limits of the membership functions, we could also change the number of fuzzy values for some or all of the fuzzy variables, using e.g. a machine learning technique. Second, it could be enhanced with an explanation facility or/and other modules to be used as an educational system for non-experts.

Acknowledgements

We thank the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II), and particularly the Program PYTHAGORAS I, for funding the above work. We would also like to thank Mr. Apostolos Kandyliis, an MSc student, who implemented a core of the fuzzy expert system.

References

- [1] Williams A. ABC of oxygen: assessing and interpreting arterial blood gases and acid-base balance, *BMJ* 1998, 317:1213-1216.
- [2] Larsen V, Siggaard-Andersen O. The Oxygen Status Algorithm on-line with the pH-blood gas analyzer, *Scand J Clin Lab Invest Suppl* 1996, 224:9-19.
- [3] Siggaard-Andersen O., An acid-base chart for arterial blood with normal and pathophysiological reference areas, *Scand J Clin Lab Invest* 1971, 27:239-245.
- [4] Grogono A., <http://www.acid-base.com/diagram.php> (accessed in March 2006).
- [5] Theakos N., Loukos A., Vassileiou M., Bechrakis P., Latest developments in graphic diagnostic approach of arterial blood gases disturbances, *PNEUMON* Number 2, Vol. 17, May-August 2004, 159-166.

- [6] Zarkadakis G, Carson ER, Cramp DG, Finkelstein L., ANABEL: intelligent blood-gas analysis in the intensive care unit, *Int J Clin Monit Comput.* 1989 Jul, 6(3):167-71.
- [7] Pince H, Verberckmoes R, Willems JL., Computer aided interpretation of acid-base disorders, *Int J Biomed Comput.* 1990 Apr, 25(2-3):177-92.
- [8] Garibaldi JM, Westgate JA, Ifeachor EC., The evaluation of an expert system for the analysis of umbilical cord blood. *Artif Intell Med*, 1999 Oct, 17(2):109-30.
- [9] Malolepszy A, Kacki E, Dogdanik T., Application of genetic programming for the differential diagnosis of acid-base and anion gap disorders. *Stud. Health Technol. Inform.* 2000, 77:388-92.
- [10] Mongelli M, Chang A, Sahota D., The development of a hybrid expert system for the interpretation of fetal acid-base status, *Int J Med Inform.* 1997 Apr, 44(2):135-44.
- [11] Dounias G. and Derek A. Linkens (Eds.), (2001), *Adaptive Systems and Hybrid Computational Intelligence in Medicine*, Special Session Proceedings of the EUNITE 2001 Symposium, Tenerife, Spain, December 13-14, 2001, A Publication of the University of the Aegean, ISBN 960-7475-19-4
- [12] Abbod M. F., von Keyserlingk D. G., Linkens D. A., and Mahfouf M., Survey of Utilization of Fuzzy Technology in Medicine and Healthcare, *Fuzzy Sets and Systems*, 120, 331–349, 2001.
- [13] Nguyen H. P., and Kreinovich V., “Fuzzy Logic and Its Applications in Medicine”, *International Journal of Medical Informatics*, 62, 165–173, 2001.
- [14] Storn, R., "System Design by Constraint Adaptation and Differential Evolution", *IEEE Trans. on Evolutionary Computation*, 1999, Vol. 3, No. 1, 22-34.
- [15] Storn, R. and Price, K. (1997). Differential evolution a simple and efficient heuristic for global optimisation over continuous spaces, *Journal of Global Optimization*, 11:341–359.
- [16] http://www.iit.nrc.ca/IR_public/fuzzy/fuzzyClips/fuzzyCLIPSIndex2.html
- [17] <http://www.icsi.berkeley.edu/~storn/code.html#c++c>
- [18] <http://www.osa.suite.dk/AboutOSA.htm>

Author Index

- Abe, Akinori III-22
Abe, Jair Minoro II-844, II-851,
II-858, II-871
Abe, Norihiro II-620
Abe, Noriyuki II-628
Abraham, Ajith II-500, III-398,
III-677, III-1128
Adachi, Naoto III-166
Adachi, Yoshinori II-401, II-1170,
II-1176
Aftarczuk, Kamila III-805
Agell, Nria II-425
Aguilar-Ruiz, Jess S. II-1264,
II-1272
Aguirre, Acacia III-1264
Ahn, Byung-Ha I-122, I-130
Ahn, Chan-Min III-84
Ahn, Tae-Chon I-52, III-219
Akamatsu, Norio III-692
Akashi, Takuya III-692
Alepis, Efthymios I-435
Alexopoulos, Angelos II-1152
Alfpio, Pedro I-1083
Amaral, Jos Franco M. III-307
Amorim, Pedro II-1248
An, Jiyuan II-1305
Anastasopoulos, Dionysios III-633
Anderson, Mary III-1184
Anguita, Davide II-442
Angulo, Cecilio II-425
Anshin, Peter III-988
Aoe, Jun-ichi II-275, II-303, II-317,
II-325
Aoki, Terumasa II-371
Aoyama, Atsushi II-553
Apeh, Edward Tersoo I-1216
Aquino, Andreia C. de I-268
Araki, Takayoshi I-506
Araki, Yutaka II-212
Arita, Daisaku II-212
Ariton, Viorel II-569
Aritsugi, Masayoshi II-1216
Arroyo-Figueroa, Gustavo I-943
Atkinson, John I-1190
Atlam, El-Sayed II-275, II-303, II-317,
II-325
Augusto, Juan Carlos II-171
Awcock, Graeme J. I-1226
Azzini, Antonia III-1111
Baba, Norio III-663
Babiak, Emilia III-797
Back, Barbro I-720
Bae, Dae Jin I-401
Bae, Hyeon-Deok I-808, I-817
Bae, Ihn-Han I-483
Baek, Jonghun I-1011
Baek, Joong Hwan I-866
Baek, Yong Sun III-248
Bagis, Aytekin I-94
Baik, Ran III-284
Baik, Sung Wook III-284
Bajaj, Preeti I-46
Balachandran, M. Bala III-1192
Balfanz, Dirk I-753
Balvig, Jens J. II-603
Bannore, Vivek II-36
Bao, Yongguang II-393
Barbancho, Antonio I-475
Barbancho, Julio I-475
Bardis, Georgios I-425
Barriga, Angel II-348
Barros, Allan Kardec III-1232
Baruque, Bruno III-432
Bashar, Md. Rezaul I-532, III-116,
III-124
Bařtrk, Alper II-331
Batten, Adam III-349
Baturone, Iluminada II-363
Becker, Matthias I-706
Beer, Martin D. II-1240
Beligiannis, Grigorios I-968
Bellas, F. III-292
Bellik, Yacine II-154
Beřdok, Erkan I-606
Bi, Jun II-678
Bi, Leo I-220
Bian, Zhengzhong III-507

- Bianco, Adriana C. I-268
 Bien, Zeungnam III-248
 Bilgen, Bilge I-37
 Bilgin, Mehmet Zeki I-1075
 Bingul, Zafer I-138
 Blázquez-del-Toro, José M. III-580
 Bolat, Emine Dođru I-841
 Borzemski, Leszek II-195, III-789
 Bouhafis, Lyamine I-409
 Boukis, Christos III-1216, III-1224
 Bourda, Yolaine II-154
 Bradley, Jeremy III-366, III-374
 Brdiczka, Oliver II-162
 Brown, David J. I-639, I-825, I-1198
 Brox, Piedad II-363
 Brunner, Levin III-614
 Bu, Jiajun I-417
 Bugarín, Alberto III-623
 Burguillo-Rial, J.C. II-659, III-263
 Burša, Miroslav II-409
 Byun, Yong Ki III-907
- Cai, Keke I-417
 Calbi, Alessandro II-179
 Cao, Jianting III-1240
 Caponetti, Laura II-340
 Cardoso, Luciana S. I-268
 Carvalho, Paulo I-1083
 Castells, Pablo III-598
 Castiello, Ciro II-340
 Ceravolo, Paolo III-1111
 Cerekovic, Aleksandra II-220
 Cha, Chang-II I-443
 Cha, Jae-Sang III-883
 Cha, Jaehyuk I-1043
 Chambers, Jonathon A. III-1208
 Chamorro-Martínez, Jesús II-355
 Chan, Chee Seng I-639
 Chan, Eddie Chun Lun II-652
 Chang, Byoungchol I-1043
 Chang, Elizabeth I-728, III-1119
 Chang, Hoon III-541
 Chang, Jae-Woo I-1067, III-76
 Chang, KyungHi III-457
 Chang, Sekchin III-449
 Chang, Tae-Gyu II-500, III-677,
 III-1128
 Chang, Ying-Chen II-1
 Chao, Pei-Ju II-1
- Chen, Chen-Wen II-888
 Chen, Chun I-417
 Chen, Guobin I-631
 Chen, Jiangyan III-1201
 Chen, Kuang-Ku I-300
 Chen, Liming II-171
 Chen, Mo III-1216
 Chen, Yen-Wei II-55, II-63
 Chen, Yi-Ping Phoebe II-1305
 Chen, Yuehui III-398
 Chen, Yumei I-145
 Chetty, Girija III-1168
 Chi, Sheng-Chai I-1
 Chis, Monica III-677
 Cho, Dongyoung I-110
 Cho, Jae-Soo I-1011
 Cho, Jungwon I-392
 Cho, Ming-Yuan I-179
 Cho, Nam-deok II-819
 Choi, Byung-Uk I-559
 Choi, Chang-Seok III-108
 Choi, Dae-Young I-490, I-984
 Choi, Dong You I-1242
 Choi, Eui-in III-1058
 Choi, Hun I-808, I-817
 Choi, Sang-Yule III-883
 Choi, Soo-Mi I-753
 Choi, Woo-kyung III-101
 Choi, Yoo-Joo I-753
 Chong, Zhang I-459
 Chou, Ming-Tao II-902
 Chou, Pao-Hua I-300
 Chun, Seok-Ju III-84
 Chung, Jaehak III-449, III-473, III-480
 Chung, Yongwha I-906
 Cichocki, Andrzej III-1232, III-1248
 Cigale, Boris III-515
 Cisternino, Virginia III-1102
 Çiviciođlu, Pmar I-606
 Consoli, Angela III-497
 Corallo, Angelo III-1083, III-1092,
 III-1102
 Corchado, Emilio II-433, III-432
 Corcho, Oscar III-588
 Corella, Miguel Ángel III-598
 Costa-Montenegro, E. II-659
 Couce, Beatriz III-300
 Cox, Robert J. III-1143
 Cox, Trevor J. III-1208
 Crowley, James L. II-162

- Crowther, Patricia S. III-1143
 Crua, Cyril I-1179
 Cutler, Philip II-479

 Daggard, Grant I-976
 Datta, Avijit I-825
 Dawson, Todd III-1264
 Debenham, John I-228
 Deep, Kusum I-951
 Degemmis, Marco III-606
 Deshmukh, Amol I-46
 Dias, Douglas Mota III-307
 Dillon, Tharam S. I-728, III-1119
 Djaiz, Chaker I-687
 Dosil-Outes, I. III-263
 Drapała, Jarosław III-1012
 Druszcz, Adam III-789
 Duanmu, C.J. II-11
 Duro, Richard J. III-292

 Edwards, Graeme C. III-349
 Egri-Nagy, Attila III-333
 Endo, Mamoru II-1045
 Eto, Kaoru II-977
 Eto, Tsutomu III-166

 Falcón, Juanjo I-679
 Fan, Shaofeng II-1144
 Fanelli, Anna Maria II-340
 Fang, Fu-Min II-1
 Fang, Hsin-Hsiung II-922
 Fang, Zhijun I-631
 Feng, Honghai I-145, I-498, I-714,
 I-1029
 Feng, Jun I-1037, I-1059, II-1095,
 II-1191, II-1199
 Fernández-García, Norberto III-580
 Fernández-Hernández, Felipe II-363
 Ferreira, Mauricio G.V. I-268
 Figueroa, Alejandro I-1190
 Finn, Anthony II-537
 Finnie, Gavin I-1115
 Fisteus, Jesús Arias III-580
 Fitch, Phillip II-523
 Flórez-Revuelta, Francisco III-424
 Freeman, Michael III-415
 Fuchino, Tetsuo II-553
 Fuente, Raúl de la III-300
 Fujii, Kunihiro I-1021
 Fujii, Kunikazu III-205

 Fujita, Yoshikatsu II-281
 Fukazawa, Yusuke I-1021
 Fuketa, Masao II-275, II-303, II-325
 Fukue, Yoshinori II-289
 Fukui, Ken-ichi II-929
 Fukumi, Minoru III-692
 Furumura, Takashi III-150, III-189
 Fuwa, Yasushi II-994
 Fyfe, Colin II-472, II-508

 Gabrys, Bogdan I-1216, III-432
 Galán-Perales, Elena II-355
 Gallego, Josune III-277
 Garcia-Almanza, Alma Lilia III-30
 García Chamizo, Juan Manuel III-424
 García Rodríguez, José III-424
 Garcia-Sebastian, M. Teresa III-277
 Gau, Shih-Wei I-179
 Gautama, Temujin III-1216
 Geem, Zong Woo I-86
 Georgoulas, George I-515
 Gerasimov, Vadim III-315
 Ghada, Elmarhomy II-303, II-317,
 II-325
 Gil, Joon-Min I-1147
 Gil-Castañeira, F. III-263
 Giráldez, Raúl II-1264, II-1272
 Go, Kentaro II-1027
 Goel, Piyush I-825
 Goh, Su Lee III-1216
 Golz, Martin III-1256
 Gómez-Pérez, Asunción III-588
 Gong, Peng III-441
 González-Castaño, Francisco J.
 II-659, III-263
 Górecki, Przemyslaw II-340
 Goto, Masato II-1079
 Graña, Manuel III-277
 Griffith, Josephine III-766
 Grosan, Crina III-677, III-1128
 Groumpos, Peter I-515
 Gu, Jinguang I-738
 Gu, Qin I-1106
 Guo, Lei II-19
 Gutiérrez-Ríos, Julio II-363
 Guzmán, Giovanni I-550, I-614

 Ha, Sang-Hyung III-101
 Håkansson, Anne I-342, I-352

- Hadzilacos, Thanasis II-1136
 Hajjam, Amir I-409
 Hall, Richard I-102, I-220
 Hamaguchi, Takashi II-553, II-579,
 II-587
 Han, Chang-Wook I-850
 Han, Dongsoo I-260
 Han, Hee-Seop I-746
 Han, Xuming I-21
 Harada, Jun II-275
 Harada, Kouji II-115
 Harashina, Naoki II-1119
 Hartung, Ronald I-342, I-352
 Hasegawa, Tsutomu II-212
 Hashimoto, Setsuo III-684
 Hashimoto, Yoshihiro II-587
 Hassan, Nashaat M. Hussein II-348
 Hattori, Akira II-1079
 Hatzilygeroudis, Ioannis I-968,
 II-1313
 He, LiYun I-145, I-498, I-714,
 I-1029
 Healy, Gerry III-315
 Heo, Joo III-457
 Hernández, Carmen III-277
 Hernandez, Yasmin I-943
 Herrero, Álvaro II-433
 Higuchi, Yuki II-1027
 Hiissa, Marketta I-720
 Hill, Richard II-1240
 Hirabayashi, Akira III-1272
 Hiraishi, Wataru II-275
 Hochin, Teruhisa II-1182
 Hong, Chao-Fu III-1, III-46
 Hong, Gyo Young I-29
 Hong, Jun Sik III-84
 Hong, Kwang-Seok I-788, I-798
 Hong, Minsuk II-545
 Hong, Sungeon I-203
 Hong, Yeh Sun I-866
 Höppne, Frank II-70
 Hori, Satoshi II-611, II-620, II-628
 Horie, Kenichi III-38
 Hoschke, Nigel III-349
 Hoshio, Kenji III-212
 Hou, Yimin II-19
 Howlett, Robert J. I-1179, I-1206,
 I-1226
 Hsu, Chia-Ling III-1
 Hsu, Mu-Hsiu III-938
 Hu, Hong I-976
 Hu, Meng II-587
 Huang, Eng-Yen II-1
 Huang, Hung-Hsuan II-220
 Huang, Peng I-417
 Huang, Xinyin II-55
 Huang, Xu III-1150, III-1157, III-1163
 Hussain, Farookh Khadeer III-1119
 Hussain, Omar Khadeer III-1119
 Hwang, Chong-Sun I-1124, I-1139,
 I-1155
 Hwang, Kyu-Jeong I-443
 Hwang, Seung-won II-1281
 Hwang, Sun-myoung II-745
 Hwang, Suntae III-248
 Iannone, Luigi III-606
 Ichalkaranje, Nikhil II-450, II-486
 Ichikawa, Sachiyoshi II-387
 Ichimura, Takumi III-742, III-749
 Igarashi, Emi II-1035
 Ikehara, Satoru III-715
 Ikezaki, Masakazu II-1224, II-1232
 Im, SeokJin I-1124, I-1139, I-1155
 In, Jang-uk III-1058
 Inokuchi, Ryo II-78
 Inuzuka, Nobuhiro II-1162
 Iribe, Yurie II-1010
 Irie, Masayuki III-205
 Iritani, Takeshi II-953
 Isern, David II-1256, III-758
 Ishibuchi, Hisao II-86
 Ishida, Yoshiteru II-123, II-131, II-139,
 II-146
 Ishii, Naohiro II-379, II-387, II-393,
 II-1170
 Ishikawa, Norihiro III-159, III-166
 Isokawa, Teijiro III-699
 Isomoto, Yukuo II-985
 Ito, Hideaki II-1208
 Ito, Kei III-813
 Ito, Maiko II-1035
 Ito, Masahiro II-953
 Itoh, Hidenori II-401, II-961
 Itoh, Kohji II-1071
 Itoh, Toshiaki II-587
 Itou, Junko III-212
 Ivancevic, Vladimir II-537
 Iwahori, Yuji II-401, II-1176
 Iwamura, Norikazu III-835

- Iwata, Masayuki III-647
 Iwazaki, Kumiko II-1045

 Jacquet, Christophe II-154
 Jain, Lakhmi C. II-110, II-450,
 II-458, II-472, II-531, II-537,
 III-497, III-504
 Jamali, Mohammed A. I-252
 Jang, Eun Sill I-992
 Jang, Ik-Jin I-1011
 Jang, Jae-Hyuk II-777
 Jang, Kyung-Won III-219
 Jang, Min-Soo I-590
 Jang, SangHyun I-1163
 Jarvis, Bevan II-458
 Jayasooriya, Thimal III-415
 Jee, KyengWhan II-492
 Jelfs, Beth III-1216
 Jelonek, Jacek III-341
 Jeng, Don Jyh-Fu III-922, III-964
 Jeng, LiDer II-879
 Jeng, MuDer II-879
 Jeon, Hong-Tae III-101
 Jeon, Jae Wook I-401
 Jeon, Jaewook II-545
 Jeon, Jin-Hong II-812
 Jeon, M.G. I-130
 Jeong, Hong I-1090
 Jeong, Moon Seok III-284
 Jeong, Seungdo I-559
 Jeong, Seung-Hee II-829
 Jevtic, Dragan I-284
 Jezic, Gordan I-236
 Ji, Younggun III-473
 Jie, Min Seok I-29, I-858, I-866
 Jifeng, He I-459
 Jimbo, Takashi II-387
 Jin, Chunming III-197
 Jin, SongGuo III-124
 Jo, Sun-Moon I-451
 Jones, Jason I-212
 Joo, Hyun Jea III-707
 Jung, Eun Sung I-68, I-78, III-124
 Jung, Kyung-Yong I-163, I-310
 Juszczyszyn, Krzysztof II-243,
 III-1020

 Kabasawa, Yasuo II-977
 Kabassi, Katerina II-1289
 Kakegawa, Jun'ichi II-1071

 Kakehi, Masahide II-1062
 Kalles, Dimitris II-1136
 Kamoshita, Yasuhiro II-1002
 Kanagawa, Koji III-725
 Kaneyama, Takashi III-150
 Kang, Dazhou I-647, I-655
 Kang, Eui-young I-392
 Kang, Joonhyuk III-465
 Kang, Mingyun III-1075
 Kang, Sanggil I-260
 Kang, Sang-Won I-1124, I-1139, I-1155
 Kang, Seo Il II-793
 Kang, Sukhoon II-769, III-248
 Kang, Yeon Gu I-582, III-707
 Kang, YunHee I-203, I-1163
 Karel, Filip I-195
 Karras, Dimitrios A. I-9
 Karsten, Helena I-720
 Karungaru, Stephen III-692
 Kashihara, Akihiro II-1002
 Katarzyniak, Radosław Piotr III-1027
 Kato, Kei III-827
 Kato, Shohei II-961
 Kato, Toshikazu III-16
 Kato, Yoshikiyo III-8
 Katsurada, Kouichi II-1010
 Kawaguchi, Masashi II-387
 Kawanaka, Haruki II-401
 Kawaoka, Tsukasa I-506, I-882, I-1002
 Kawasaki, Takashi II-289
 Kazienko, Przemysław II-417
 Keegan, Stephen II-686
 Kendrick, Paul III-1208
 Kerre, Etienne I-623
 Keskar, A.G. I-46
 Khwan-on, Sudarat I-833
 Kil, Min Wook II-701, III-1075
 Kim, ChangHwan I-874
 Kim, Dong Seong I-935
 Kim, Dongwon III-255, III-670
 Kim, Dong Wook III-859
 Kim, Duk Kyung III-441
 Kim, Haeng-Kon II-751, II-760
 Kim, Hak-Man II-812, III-900
 Kim, Hanil I-392
 Kim, Hong Sok III-541
 Kim, Howon I-924, I-1250
 Kim, Hwangrae III-1068
 Kim, Hyeoncheol I-746
 Kim, Hyun Deok I-276, I-1011

- Kim, Hyung-Jun I-443
 Kim, Ikno III-922, III-964
 Kim, In-Cheol I-244
 Kim, Jaehwan III-465
 Kim, Jeom Goo III-564
 Kim, Jeong Hyun III-556
 Kim, Jin-Geol III-233
 Kim, Jong Tae I-401, III-907, III-914
 Kim, Jong-Yul III-900
 Kim, Jongwan I-1124, I-1139, I-1155
 Kim, Joohee III-480
 Kim, Jung-Hyun I-788, I-798
 Kim, Jungyeop I-203
 Kim, Sang Tae I-276
 Kim, Sang-Wook I-443, I-1043
 Kim, Sangkyun I-1234, I-1259
 Kim, Seokhyun III-473
 Kim, Seoksoo II-694, II-701, II-718,
 III-1075
 Kim, Seong-Joo I-924, III-101
 Kim, SungBu I-890
 Kim, Sung-Hwan I-935
 Kim, Taehee III-556
 Kim, Tai Hoon II-701, II-745, II-836
 Kim, Wankyung II-709, III-1068
 Kim, Woong-Sik I-898
 Kim, Yong I-1090
 Kim, Yong-Guk I-590
 Kim, Yong-Ki I-1067, III-76
 Kim, Yong Soo III-248
 Kim, Young-Chang I-1067
 Kim, Young-Joong III-225
 Kim, Youngsup III-1042
 Kimura, Masahiro II-929, II-937
 Kitajima, Teiji II-553
 Kitakoshi, Daisuke II-969
 Kitazawa, Kenichi III-189
 Klawonn, Frank II-70
 Ko, Il Seok III-1035, III-1050
 Ko, Min Jung I-292
 Ko, SangJun III-457
 Kobayashi, Kanami II-55
 Kogure, Kiyoshi III-22
 Koh, Byoung-Soo II-777
 Koh, Eun Jin I-368, I-569
 Koh, Jae Young III-572
 Kojima, Fumio III-684
 Kojima, Masanori III-189
 Kojiri, Tomoko I-771, II-1053,
 II-1062
 Kokol, Peter II-1297
 Kolaczek, Grzegorz II-243
 Komedani, Akira I-771
 Komiya, Kaori III-16
 Komosinski, Maciej III-341
 Kompatsiaris, Yiannis III-633
 Kondo, Michiro II-851, II-858, II-871
 Kong, Jung-Shik III-233
 Konishi, Osamu III-813
 Koo, Jung Doo III-548
 Koo, Jung Sook III-548
 Koshimizu, Hiroyasu II-1208
 Koukam, Abder I-409
 Koutsojannis, Constantinos I-968,
 II-1313
 Kowalczuk, Zdzisław I-671
 Kozakura, Shigeki II-620
 Kozierekiewicz, Adrianna III-805
 Król, Dariusz II-259, III-774
 Kubota, Naoyuki III-684
 Kucuk, Serdar I-138
 Kuh, Anthony III-1280
 Kuh, Tony III-1216
 Kukkurainen, Paavo III-383
 Kukla, Grzegorz Stanisław
 II-259, III-774
 Kulikowski, Juliusz L. II-235
 Kunchev, Voemir II-537
 Kunifuji, Susumu II-1019, III-851,
 III-859, III-867
 Kunimune, Hisayoshi II-994
 Kunstic, Marijan I-284
 Kurakake, Shoji I-1021
 Kurazume, Ryo II-212
 Kurimoto, Ikusaburo III-742
 Kuroda, Chiaki II-561
 Kusek, Mario I-236
 Kusumoto, Yoshiki III-205
 Kuwahara, Jungo III-159
 Kuwahara, Noriaki III-22
 Kuwajima, Isao II-86
 Kuwata, Tomoyuki II-102
 Kwak, Hoon Sung I-542
 Kwak, Jin I-924
 Kwak, Jong Min III-1050
 Kwon, Taekyoung I-916
 Kwon, Yangsoo III-480
 Kye, Bokyung I-1163

- Laflaquière, Julien I-1171
 Lam, Toby H.W. II-637, II-644
 Lama, Manuel III-623
 Lampropoulos, Aristomenis S. I-376,
 I-384
 Lampropoulou, Paraskevi S. I-384
 Lasota, Tadeusz III-774
 Lee, Boo-Hyung III-135
 Lee, Byoung-Kuk III-875
 Lee, Chang-Hwan I-187
 Lee, Chil-Woo I-598
 Lee, Chungwon III-556
 Lee, Deok-Gyu II-793
 Lee, Do Hyeon III-564
 Lee, Dong Chun III-548
 Lee, Eun-ser II-819
 Lee, Geuk II-701, III-1042, III-1075
 Lee, Hanho III-108
 Lee, Hong Joo I-1234, I-1259
 Lee, Hongwon III-473
 Lee, Hsuan-Shih II-896, II-902, II-910,
 II-917, II-922
 Lee, Huey-Ming III-938
 Lee, Hwa-Ju I-483
 Lee, Hyun-Gu I-590
 Lee, Hyun-Jae II-829
 Lee, Im-Yeong II-793
 Lee, JangMyung I-890
 Lee, Jee Hyung I-401
 Lee, Jeong-On III-225
 Lee, Jong Hyuk Park Sangjin II-777
 Lee, JooYoung I-1250
 Lee, Joon-Sung II-829
 Lee, Ju-Hong III-84
 Lee, Junkyu III-465
 Lee, Kang Woong I-29, I-858, I-866
 Lee, Keon Myung I-401
 Lee, Kyoung Jun I-1267
 Lee, Kyung-Sook I-483
 Lee, Moo-hun III-1058
 Lee, Raymond S.T. II-637, II-644,
 II-652
 Lee, Sangjin II-777
 Lee, Sang Wan III-248
 Lee, Sang-Joong III-893
 Lee, Sang-Wook I-122, I-130
 Lee, Sang-Yun I-443
 Lee, Seok-Joo I-590
 Lee, SeongHoon I-110, I-1124, I-1139,
 I-1147, I-1155
 Lee, Seung Wook I-401
 Lee, Seungjae III-556
 Lee, Shaun H. I-1179, I-1206
 Lee, Soon Woong I-78
 Lee, Sungdoke I-260
 Lee, Tsair-Fwu I-179, II-1
 Lee, Tsang-Yean III-938
 Lee, Yang-Weon I-598
 Lee, Yong Kyu I-292, I-992
 Lee, Young-Kyun III-541
 Leem, Choon Seong I-1259
 Lefkaditis, Dionisios I-1226
 Lehtikunnas, Tuija I-720
 Lemos, João M. II-1248
 Leng, Jinsong II-472
 Lenič, Mitja III-515
 León, Carlos I-475
 Levachkine, Sergei I-698
 Levachkine, Serguei I-550
 Lewis, Chris J. III-349
 Lewkowicz, Myriam I-1131
 Lhotská, Lenka II-409
 Li, Francis F. III-1208
 Li, Jiuyong I-212, I-976
 Li, Ming I-21
 Li, Peng III-507
 Li, Xun III-90
 Li, Yanhui I-647, I-655
 Li, Yueli I-498, I-714, I-1029
 Li, Zhonghua I-153
 Liang, Liang I-1106
 Liang, Yanchun I-21
 Licchelli, Oriana III-606
 Ligon, Gopinathan L. III-1192
 Lim, Jounghoon I-559
 Lim, Myo-Taeg III-225
 Lin, Cheng II-561
 Lin, Geng-Sian III-46
 Lin, Kuang II-922
 Lin, Lily III-930, III-956
 Lin, Mu-Hua III-46
 Lin, Zuoquan I-459
 Liu, Baoyan I-145, I-498, I-714, I-1029
 Liu, Hongbo II-500
 Liu, Honghai I-639, I-825, I-1198
 Liu, Ju II-28
 Liu, Jun II-171
 Liu, Qingshan II-47
 Liu, Xianxing II-204
 López-Peña, Fernando III-292

- Lops, Pasquale III-606
 Lorenzo, Gianluca III-1092
 Lorkiewicz, Wojciech III-1004
 Lortal, Gaëlle I-1131
 Lovrek, Ignac I-318
 Lu, Hanqing II-47
 Lu, Jianjiang I-647, I-655
 Lu, Peng I-780
 Lun, Xiangmin II-19
 Luukka, Pasi III-383
- Ma, Songde II-47
 Ma, Wanli III-1176, III-1184
 Macaš, Martin II-409
 Mackin, Kenneth J. III-820
 Magalhães, Hugo II-1248
 Mahalik, Nitaigour Premchand
 I-122, I-130
 Maisonnasse, Jérôme II-162
 Mak, Raymond Yiu Wai II-652
 Malski, Michał II-251
 Mandić, Danilo P. III-1216, III-1232,
 III-1248, III-1280
 Mao, Yong I-171
 Marcenaro, Lucio II-179
 Martinez, Miguel I-698
 Masuda, Tsuyoshi II-220
 Mathur, Abhishek III-1176
 Matijasevic, Stjepan I-284
 Matsuda, Noriyuki II-620, II-628
 Matsui, Nobuyuki III-699
 Matsui, Tatsunori II-977
 Matsumoto, Hideyuki II-561
 Matsuno, Takuma III-181
 Matsuura, Kyoichi II-1010
 Matsuyama, Hisayoshi II-579
 Matta, Nada I-687
 Mattila, Jorma K. III-358
 Mendonça, Teresa F. II-1248
 Mera, Kazuya III-749
 Miaoulis, Georgios I-425
 Mikac, Branko I-318
 Mille, Alain I-1171
 Mineno, Hiroshi III-150, III-159,
 III-166, III-189
 Misue, Kazuo III-835, III-843
 Mitani, Tsubasa II-1103
 Mitsuishi, Takashi II-1027
 Miura, Hirokazu II-620, II-628
 Miura, Motoki II-1019
- Miyachi, Taizo II-603
 Miyadera, Youzou II-1035
 Miyaji, Masako III-725
 Miyamoto, Sadaaki II-78
 Mizuno, Tadanori III-150, III-159,
 III-166, III-189
 Mo, Eun Jong I-29
 Molan, Gregor I-360
 Molan, Marija I-360
 Molina, Javier I-475
 Montero-Orille, Carlos III-300
 Moon, Daesung I-906
 Moon, Il-Young I-467
 Moreno, Antonio II-1256, III-758
 Moreno, Francisco III-406
 Moreno, Marco I-550, I-614, I-698
 Moret-Bonillo, Vicente III-1136
 Morihiro, Koichiro III-699
 Morita, Kazuhiro II-303, II-317, II-325
 Moriyama, Jun II-603
 Mosqueira-Rey, Eduardo III-1136
 Motoyama, Jun-ichi II-1162
 Mouri, Katsushiro II-1045
 Mouzakidis, Alexandros II-1152
 Mukai, Naoto I-1059, II-1095
 Munemori, Jun III-174, III-212
 Murai, Takeshi II-393
 Murakami, Jin'ichi III-715
 Murase, Yosuke II-1053
 Mure, Yuji II-1103
 Musiał, Katarzyna II-417
- Na, Yun Ji III-1035, III-1050
 Naganuma, Takefumi I-1021
 Nagar, Atulya K. I-1051
 Nagasawa, Isao II-1103
 Nagasawa, Shin'ya III-980
 Naito, Takeshi III-1272
 Nakamatsu, Kazumi II-844, II-851,
 II-858, II-871
 Nakamura, Hiroshi II-1071
 Nakamura, Shoichi II-1035
 Nakano, Ryohei II-945, II-969
 Nakano, Tomofumi II-1162
 Nakano, Yukiko II-220
 Nakao, Zensho II-55
 Nakazono, Nagayoshi III-843
 Nam, Mi Young I-368, I-532,
 III-116, III-124
 Naoe, Yukihisa III-189

- Nara, Yumiko III-64
 Nehaniv, Chrystopher L. III-333
 Neves, José I-1083
 Nguyen, Ngoc Thanh II-267, III-805
 Niimi, Ayahiko III-813
 Niimura, Masaaki II-994
 Nikiforidis, George I-515
 Nishida, Toyoaki II-220
 Nishikawa, Ikuko II-953
 Nishimoto, Kazushi III-859
 Nishimura, Haruhiko III-699
 Nishimura, Shinichi III-174
 Nitta, Tsuneo II-1010
 Noda, Manabu II-1045
 Nojima, Yusuke II-86
 Nouno, Ikue II-953
 Ntoutsis, Christos II-1152
 Numao, Masayuki II-929
 Nunes, Catarina S. II-1248
 Nunohiro, Eiji III-820
 Nürnberger, Andreas I-763
- O'Grady, Michael J. II-686, III-1201
 O'Hare, Gregory M.P. II-686, III-1201
 O'Riordan, Colm III-766
 Odagiri, Kazuya II-379
 Oeda, Shinichi III-742
 Oehlmann, Ruediger III-57
 Ogawa, Hisashi II-620
 Oh, Chang-Heon II-829
 Oh, Jae-Yong I-598
 Oh, Tae-Kyoo II-812, III-900
 Ohsawa, Yukio III-38
 Ohshiro, Masanori III-820
 Okamoto, Takeshi II-123
 Okumura, Noriyuki I-506
 Okuno, Masaaki II-387
 Orłowski, Cezary I-671
 Oyarzun, Joaquín I-679
 Ozaki, Masahiro II-1170, II-1176
 Ozaku, Hiromi Itoh III-22
 Ozkarahan, Irem I-37
- Pacheco, Marco Aurélio C. III-307
 Pahikkala, Tapio I-720
 Palade, Vasile III-487
 Paliouras, Georgios II-1152
 Palmisano, Ignazio III-606
 Pandzic, Igor S. II-220
- Pant, Millie I-951
 Papageorgiou, Elpiniki I-515
 Park, Byungkwan I-906
 Park, Chang-Hyun III-241
 Park, Choung-Hwan III-533
 Park, Gil-Cheol II-694, III-1075
 Park, Gwi-Tae I-590, III-255, III-670
 Park, Hee-Un II-726
 Park, Heejun I-1234
 Park, Hyun-gun II-819
 Park, Jeong-Hyun III-135
 Park, Jin-Won I-906
 Park, Jinsub III-1068
 Park, Jong Hyuk II-777
 Park, Jong Kang III-907
 Park, Jong Sou I-935
 Park, Jung-Il I-850
 Park, KeeHyun I-276
 Park, Kiheon II-545
 Park, Namje I-924
 Park, Soohong I-203
 Park, Sun III-84
 Patel, Meera III-57
 Pei-Shu, Fan II-879
 Peng, Lizhi III-398
 Petridis, Kosmas III-633
 Pham, Tuan D. I-524
 Philips, Wilfried I-623
 Phillips-Wren, Gloria II-515, II-531,
 III-504
 Pi, Daoying I-171
 Pieczyńska, Agnieszka II-227, III-1012
 Plemenos, Dimitri I-425
 Polymenakos, Lazaros C. III-1224
 Pontes, Beatriz II-1264, II-1272
 Popek, Grzegorz III-997
 Posada, Jorge I-679
 Pousada-Carballo, J.M. III-263
 Prado, João Carlos Almeida II-844
 Prentzas, Jim I-968
 Price, Don C. III-349
 Prié, Yannick I-1171
 Prieto, Abraham III-292
 Prieto-Blanco, Xesus III-300
 Prokopenko, Mikhail III-315, III-324
 Pu, Geguang I-459
- Qian, Zuoqin I-1198
 Qiao, Jianping II-28
 Qin, Jun II-1087

- Qiu, Zongyan I-459
 Qudeiri, Jaber Abu I-252
 Quintero, Rolando I-550, I-614

 Ratanamahatana, Chotirat Ann
 III-733
 Redavid, Domenico III-606
 Regazzoni, Carlo S. II-179
 Reignier, Patrick II-162
 Ren, Xiang II-1191
 Resta, Marina III-641
 Rhee, Phill Kyu I-68, I-78, I-368,
 I-532, I-569, I-582, III-116, III-124,
 III-707
 Ridgewell, Alexander III-1163
 Ríos, Sebastián A. II-371
 Rodríguez-Hernández, P.S. III-263
 Rodríguez, Jesús Barrasa III-588
 Roh, Seok-Beom III-219
 Romero, Sixto III-406
 Ross, Robert I-102
 Rousselot, François I-1098
 Ruiz, Francisco J. II-425
 Ruiz, Roberto II-1272
 Rutkowski, Tomasz M. III-1216,
 III-1232

 Saathoff, Carsten III-633
 Sadi, Mohammed Golam I-874
 Saito, Kazumi II-929, II-937,
 II-945, II-969
 Sáiz, José Manuel II-433
 Sakakibara, Kazutoshi II-953
 Sakamoto, Hirotaka II-953
 Sakamoto, Junichi II-628
 Sakurai, Kouichi II-737
 Salakoski, Tapio I-720
 Salanterä, Sanna I-720
 Sánchez, Daniel II-355
 Sánchez, David III-758
 Sánchez-Fernández, Luis III-580
 Sánchez-Solano, Santiago II-363
 Sánchez, Omar III-406
 Sanin, Cesar I-663
 Sarker, M. Omar Faruque I-874
 Saruwatari, Yasufumi II-281
 Sasaki, Mizuho II-611
 Sato, Eri III-725
 Sato, Hideki II-1216
 Sato, Yoshiharu II-94

 Sato-Ilic, Mika II-102, II-110
 Savvopoulos, A. I-960
 Schetinin, Vitaly III-523
 Schmidt, Rainer I-326, I-334
 Schulz, Klaus U. III-614
 Scott, D. Andrew III-349
 Seising, Rudolf III-366, III-374
 Semeraro, Giovanni III-606
 Seno, Yasuhiro III-189
 Seo, Duck Won I-542
 Seo, Young Hwan I-1267
 Settouti, Lotfi S. I-1171
 Sharma, Dharmendra III-1150,
 III-1163, III-1168, III-1176, III-1184,
 III-1192
 Shi, Ruihong I-780
 Shi, Xiaohu I-21
 Shih, Ching-Nan I-179
 Shiizuka, Hisao III-988
 Shim, Bo-Yeon II-803
 Shim, Donghee I-110
 Shimada, Yukiyasu II-553, II-579,
 II-587
 Shimizu, Toru III-189, II-1224
 Shimodaira, Chie III-867
 Shimodaira, Hiroshi III-867
 Shin, Dong-Myung II-726
 Shin, Ho-Jun II-803
 Shin, Miyoung I-1043
 Shin, Myong-Chul II-812, III-883,
 III-900
 Shin, Woochul III-90
 Shinohara, Shuji II-1010
 Shirai, Hirokazu II-1035
 Shoji, Hiroko III-8, III-16
 Sidhu, Amandeep S. I-728
 Siemiński, Andrzej III-782
 Silva, José Demisio S. da I-268
 Sim, Kwee-Bo III-241
 Simoff, Simeon I-228
 Sinkovic, Vjekoslav I-236
 Sioutis, Christos II-450, II-464
 Sirois, Bill III-1264
 Skourlas, Christos II-1152
 Stanina, Marta III-797
 Soak, Sang-Moon I-122, I-130
 Sobecki, Janusz III-797
 Soh, Wooyoung II-709, III-1068
 Sohn, Hong-Gyoo III-533
 Sohn, Sang-Wook I-817

- Sokhi, Dilbag I-1051
 Solazzo, Gianluca III-1092, III-1102
 Sommer, David III-1256, III-1264
 Song, Hyunsoo I-916
 Song, MoonBae I-1139
 Song, Yeong-Sun III-533
 Sorensen, Humphrey III-766
 Sotiropoulos, D.N. I-960
 Soto-Hidalgo, Jose M. II-355
 Staab, Steffen III-633
 Stathopoulou, Ioanna-Ourania II-1128
 Sterpi, Dario II-442
 Stiglic, Gregor II-1297
 Stober, Sebastian I-763
 Streichert, Felix III-647, III-655
 Stylios, Chrysostomos I-515
 Sucar, Enrique I-943
 Sugano, Naotoshi III-948
 Sugiki, Daigo II-63
 Sugino, Yoshiki II-961
 Suh, Il Hong I-559
 Suh, Jae Won I-808, I-817
 Suh, Jungwon III-480
 Sujitjorn, Sarawut I-833
 Sumitomo, Toru II-275
 Sun, Yong I-171
 Sun, Zhaohao I-1115
 Suominen, Hanna I-720
 Suzuki, Hideharu III-159, III-166
 Suzuki, Nobuo II-296
 Suzuki, Susumu II-393
 Szczerbicka, Helena I-706
 Szczerbicki, Edward I-663
- Tabakow, Iwan II-187
 Tabata, Toshihiro II-737
 Tadauchi, Masaharu II-379
 Takahashi, Hiroshi II-310
 Takahashi, Satoru II-289, II-310
 Takahashi, Shin III-197
 Takahashi, Masakazu II-289, II-310
 Takai, Toshihiro II-401
 Takata, Osamu II-1103
 Takeda, Kazuhiro II-553, II-579, II-587
 Taki, Hirokazu II-611, II-620, II-628
 Tan, Hong-Zhou I-153
 Tanahashi, Yusuke II-969
 Tanaka, Jiro III-197, III-835, III-843
 Tanaka, Kaoru III-851
 Tanaka, Kiyoko III-159, III-166
- Tanaka, Toshihisa III-1248
 Tanaka-Yamawaki, Mieko III-647, III-655
 Tang, Ming Xi II-670
 Taniguchi, Rin-ichiro II-212
 Tarasenko, Kateryna II-220
 Tatara, Kohei II-737
 Tawfik, Hissam I-1051
 Termenón, Maite I-679
 Thatcher, Steve II-508
 Tigan, Stefan III-1128
 Timmermann, Norman III-633
 Todirascu-Courtier, Amalia I-1131
 Tokuhisa, Masato III-715
 Tommasi, Maurizio De III-1083
 Toro, Carlos I-679
 Torres, Cláudio Rodrigo II-851
 Torres, Germano L. II-851
 Torres, Miguel I-550, I-614, I-698
 Torres-Schumann, Eduardo III-614
 Tran, Dat T. I-524, III-1176, III-1184
 Trawiński, Bogdan III-774
 Trutschel, Udo III-1264
 Trzec, Krunoslav I-318
 Tsang, Edward P.K. III-30
 Tseng, Wei-Kuo II-910
 Tsihrintzis, George A. I-376, I-384, I-960, II-1128, II-1289
 Tsuchiya, Seiji I-1002
 Tsuda, Kazuhiko II-281, II-296, II-310
 Tsuge, Yoshifumi II-579
 Tweedale, Jeffrey II-450, II-464, II-479, II-486, III-497
- Uchida, Seiichi II-212
 Umeda, Masanobu II-1103
 Unno, Shunsuke II-1071
 Urlings, Pierre II-450
 Ushiyama, Taketoshi II-1111, II-1224, II-1232
 Ushiro, Mika II-994
- Vales-Alonso, J. II-659
 Valls, Aida II-1256
 Van der Weken, Dietrich I-623
 Vansteenkiste, Ewout I-623
 Vayanos, Phebe III-1216
 Velásquez, Juan D. II-371, III-487
 Vélez, Miguel A. III-406
 Verastegui, Karina I-698

- Vialatte, Francois III-1232
 Vidal, Juan Carlos III-623
 Virvou, Maria I-376, I-435,
 I-960, II-1289
 Vorobieva, Olga I-334
- Wada, Masaaki III-813
 Waldeck, Carsten I-753
 Waligora, Tina I-326
 Walters, Simon D. I-1179, I-1206
 Wang, Hongmoon I-401
 Wang, Hua I-976
 Wang, Jian Xun II-670
 Wang, Jin-Long II-888
 Wang, Leuo-Hong III-1
 Wang, Limin I-21
 Wang, Peter III-315
 Wang, Xiaodong II-11
 Wang, Yupeng III-457
 Wang, Zhengyou I-631
 Wang, Zhijian II-1199
 Wang, Zhong-xian I-52
 Washizawa, Yoshikazu III-1248
 Watabe, Hirokazu I-506, I-882, I-1002
 Watada, Junzo III-922, III-964, III-972
 Watanabe, Toyohide I-771, I-1037,
 I-1059, II-1053, II-1062, II-1095,
 II-1111, II-1119, II-1224,
 II-1232, III-827
 Watanabe, Yuji II-131
 Weigel, Felix III-614
 Wen, YuanLin II-879
 Won, Chung-Yuen III-875
 Won, Dongho I-924
 Won, Hee-Sun I-443
 Wong, Ka Yan III-269
 Wong, S.T.C. I-171
 Wu, Linyan I-1037
 Wu, Menq-Jiun I-300
 Wu, Shiqian I-631
- Xia, Zheng I-171
 Xu, Baowen I-647, I-655
 Xu, Hao I-714
 Xu, Zhiping III-390
 Xue, Peng III-441
 Xue, Quan I-631
- Yaegashi, Rihito II-379
 Yamada, Kunihiko III-150, III-189
 Yamada, Takahiro II-393
 Yamaguchi, Toru III-725
 Yamakami, Toshihiko III-143
 Yamamoto, Hidehiko I-252
 Yamasaki, Kazuko III-820
 Yamashita, Yoshiyuki II-595
 Yamawaki, Shigenobu II-866
 Yan, Wang II-47
 Yang, Bingru I-145, I-498, I-714, I-1029
 Yang, Byoungnak I-60
 Yang, Chih Chieh I-1
 Yang, Guosheng II-204
 Yang, Hsiao-Fang III-46
 Yang, Jueng-je I-52
 Yang, Jung-Jin II-492
 Yang, Yongqing II-1199
 Yang, Yuxiang III-507
 Yasuda, Hiroshi II-371
 Yasuda, Takami II-1045, II-1079
 Yeh, Chen-Huei II-917
 Yildiz, Ali Bekir I-1075
 Yip, Chi Lap III-269
 Yokoi, Shigeki II-1045, II-1079
 Yokoyama, Setsuo II-1035
 Yoo, Dong-Wook III-875
 Yoo, Sang Bong III-90
 Yoo, Seong Joon I-753
 Yoo, Weon-Hee I-451, I-898
 Yoon, Chang-Dae III-900
 Yoon, Seok Min III-914
 Yoshida, Jun II-281
 Yoshida, Koji III-189
 Yoshino, Takashi III-181, III-205
 You, Bum-Jae I-874
 You, Ilsun II-785
 You, Kang Soo I-542
 You, Kwanho II-545
 Youk, Sang Jo III-1042
 Yu, Gang III-507
 Yu, Jae-Sung III-875
 Yuizono, Takaya III-174
 Yukimura, Yoko III-205
 Yüksel, Mehmet Emin II-331
 Yun, Al-Chan I-817
 Yun, Byoung-Ju I-1011
 Yun, JaeMu I-890
- Zanni, Cecilia I-1098
 Zatwarnicki, Krzysztof II-195
 Zazula, Damjan III-515

- Zeman, Astrid III-315, III-324
Zeng, Weiming I-631
Zhang, Changjiang II-11
Zhang, Fan II-204
Zhang, Guangde I-1198
Zhang, Miao II-678
Zhang, Naixiao II-1144
Zhang, Shiyong III-390
Zhang, Weishi II-500
Zhang, Xinhong II-204
Zhang, Yonggang III-1208
Zhao, Lei II-678
Zhao, Shuo I-145, I-498
Zharkov, Sergei III-523
Zharkova, Valentina III-523
Zhong, Yiping III-390
Zhou, Xia I-171
Zhou, Yi I-738
Zhu, Chaopin III-1280
Zhu, Yuelong I-1037, II-1191
Zong, Ping II-1087