

Rossitza Setchi
Ivan Jordanov
Robert J. Howlett
Lakhmi C. Jain (Eds.)

LNAI 6278

Knowledge-Based and Intelligent Information and Engineering Systems

14th International Conference, KES 2010
Cardiff, UK, September 2010
Proceedings, Part III

3
Part III



 Springer

Lecture Notes in Artificial Intelligence 6278

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Rossitza Setchi Ivan Jordanov
Robert J. Howlett Lakhmi C. Jain (Eds.)

Knowledge-Based and Intelligent Information and Engineering Systems

14th International Conference, KES 2010
Cardiff, UK, September 8-10, 2010
Proceedings, Part III

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Rossitza Setchi
Cardiff University, School of Engineering
The Parade, Cardiff CF24 3AA, UK
E-mail: Setchi@cf.ac.uk

Ivan Jordanov
University of Portsmouth, Dept. of Computer Science and Software Engineering
Buckingham Building, Lion Terrace, Portsmouth, PO1 3HE, UK
E-mail: Ivan.Jordanov@port.ac.uk

Robert J. Howlett
KES International
145-157 St. John Street, London EC1V 4PY, UK
E-mail: rjhowlett@kesinternational.org

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

Library of Congress Control Number: 2010932879

CR Subject Classification (1998): I.2, H.4, H.3, I.4, H.5, I.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-15392-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-15392-1 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.com

© Springer-Verlag Berlin Heidelberg 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

The 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems was held during September 8–10, 2010 in Cardiff, UK. The conference was organized by the School of Engineering at Cardiff University, UK and KES International.

KES2010 provided an international scientific forum for the presentation of the results of high-quality research on a broad range of intelligent systems topics. The conference attracted over 360 submissions from 42 countries and 6 continents: Argentina, Australia, Belgium, Brazil, Bulgaria, Canada, Chile, China, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hong Kong ROC, Hungary, India, Iran, Ireland, Israel, Italy, Japan, Korea, Malaysia, Mexico, The Netherlands, New Zealand, Pakistan, Poland, Romania, Singapore, Slovenia, Spain, Sweden, Syria, Taiwan, Tunisia, Turkey, UK, USA and Vietnam.

The conference consisted of 6 keynote talks, 11 general tracks and 29 invited sessions and workshops, on the applications and theory of intelligent systems and related areas. The distinguished keynote speakers were Christopher Bishop, UK, Nikola Kasabov, New Zealand, Saeid Nahavandi, Australia, Tetsuo Sawaragi, Japan, Yuzuru Tanaka, Japan and Roger Whitaker, UK.

Over 240 oral and poster presentations provided excellent opportunities for the presentation of interesting new research results and discussion about them, leading to knowledge transfer and generation of new ideas.

Extended versions of selected papers were considered for publication in the *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *Engineering Applications of Artificial Intelligence*, *Journal of Intelligent Manufacturing*, and *Neural Computing and Applications*.

We would like to acknowledge the contribution of the Track Chairs, Invited Sessions Chairs, all members of the Program Committee and external reviewers for coordinating and monitoring the review process. We are grateful to the editorial team of Springer led by Alfred Hofmann. Our sincere gratitude goes to all participants and the authors of the submitted papers.

September 2010

Rossitza Setchi
Ivan Jordanov
Robert J. Howlett
Lakhmi C. Jain

Organization

KES 2010 was hosted and organized by the School of Engineering at Cardiff University, UK and KES International. The conference was held at the Mercure Holland House Hotel, September 8–10, 2010.

Conference Committee

General Chair	Rossi Setchi, Cardiff University, UK
Conference Co-chair	Lakhmi C. Jain, University of South Australia, Australia
Executive Chair	Robert J. Howlett, University of Brighton, UK
Chair of the Organizing Committee	Y. Hicks, Cardiff University, UK
Program Chair	I. Jordanov, University of Portsmouth, UK

Organizing Committee

KES Operations Manager	Peter Cushion, KES International
Publicity Chairs	D. Todorov, Cardiff University, UK Yu-kun Lai, Cardiff University, UK
KES Systems Support	Shaun Lee, KES International
Members	Engku Fadzli, Cardiff University, UK Lei Shi, Cardiff University, UK Nedyalko Petrov, Portsmouth University, UK Panagiotis Loukakos, Cardiff University, UK

Track Chairs

Bruno Apolloni	University of Milan, Italy
Bojana Dalbelo Basic	University of Zagreb, Croatia
Floriana Esposito	University of Bari, Italy
Anne Hakansson	Stockholm University, Sweden
Ron Hartung	Franklyn University, USA
Honghai Liu	University of Portsmouth, UK
Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
Andreas Nuernberger	University of Magdeburg, Germany
Bernd Reusch	University of Dortmund, Germany
Tuan Pham	University of New South Wales, Australia
Toyohide Watanabe	Nagoya University, Japan

Invited Sessions Chairs

3D Visualization of Natural Language	Minhua Eunice Ma, University of Derby, UK Nikolaos Antonopoulos, University of Derby, UK Bob Coyne, Columbia University, USA
Intelligent Data Processing in Process Systems and Plants	Kazuhiro Takeda, Shizuoka University, Japan Takashi Hamaguchi, Nagoya Institute of Technology, Japan
A Meta-Heuristic Approach to Management Engineering	Junzo Watada, Waseda University, Japan Taki Kanda, Bunri University of Hospitality, Japan Huey-Ming Lee, Chinese Culture University, Taiwan Lily Lin, China University of Technology, Taiwan Cesar Sanin, University of Newcastle, Australia
Knowledge Engineering and Smart Systems	
Skill Acquisition and Ubiquitous Human-Computer Interaction	Hirokazu Taki, Wakayama University, Japan Masato Soga, Wakayama University, Japan
Application of Knowledge Models in Healthcare	István Vassányi, University of Pannonia, Hungary György Surján, National Institute for Strategic Health Research, Hungary
Knowledge Environment for Supporting Creative Learning	Toyohide Watanabe, Nagoya University, Japan Tomoko Kojiri, Nagoya University, Japan
ICT in Innovation and Creativity	Toyohide Watanabe, Nagoya University, Japan Takatoshi Ushiyama, Kyushu University, Japan
Intelligent Support for Designing Social Information Infrastructure	Toyohide Watanabe, Nagoya University, Japan Naoto Mukai, Tokyo University of Science, Japan
Intelligent Systems in Ambient-Assisted Living Environments	Antonio F. Gómez-Skarmeta, Universidad de Murcia, Spain Juan A. Botía, Universidad de Murcia, Spain
Knowledge-Based Systems for e-Business	Kazuhiko Tsuda, University of Tsukuba, Japan Nobuo Suzuki, KDDI Corporation, Japan
Quality Assurance and Intelligent Web-Based Information Technology	Anastasia N. Kastania, Athens University of Economics and Business, Greece Stelios Zimeras, University of the Aegean, Greece
Knowledge-Based Interface Systems	Yuji Iwahori, Chubu University, Japan Naohiro Ishii, Aichi Institute of Technology, Japan Yoshinori Adachi, Chubu University, Japan Nobuhiro Inuzuka, Nagoya Institute of Technology, Japan
Reasoning-Based Intelligent Systems	Kazumi Nakamatsu, University of Hyogo, Japan Jair Minoru Abe, University of Sao Paulo, Brazil
Data Mining and Service Science for Innovation	Katsutoshi Yada, Kansai University, Japan Takahira Yamaguchi, Keio University, Japan Maria Alessandra Torsello, University of Bari, Italy

Web 2.0: Opportunities and Challenges for Social Recommender Systems	Jose J. Pazos-Arias, University of Vigo, Spain Ana Fernandez-Vilas, University of Vigo, Spain
Innovations in Chance Discovery	Akinori Abe, University of Tokyo, Japan
Personalization of Web Contents and Services	In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea Juan D. Velásquez, University of Chile, Chile
Advanced Knowledge-Based Systems	Alfredo Cuzzocrea, ICAR-CNR and University of Calabria, Italy
Knowledge-Based Creativity Support Systems	Susumu Kunifuji, Jaist, Japan Kazuo Misue, University of Tsukuba, Japan Hidehiko Hayashi, Naruto University of Education, Japan Motoki Miura, Kyushu Institute of Technology, Japan Toyohisa Nakada, Niigata University of International and Information Studies, Japan Tessai Hayama, JAIST, Japan
Intelligent Network and Service	Jun Munemori, Wakayama University, Japan
Real-World Data Mining and Digital Intelligence	Takaya Yuizono, JAIST, Japan Rashid Mehmood, Swansea School of Engineering, UK Omer F. Rana, Cardiff University, UK
Advanced Design Techniques for Adaptive Systems	Ziad Salem, Aleppo University, Syria Sorin Hintea, Technical University of Cluj-Napoca, Romania Hernando Fernández-Canque, Glasgow Caledonian University, UK Gabriel Oltean, Technical University of Cluj-Napoca, Romania
Soft Computing Techniques and Their Intelligent Utilizations Toward Gaming, Robotics, Stock Markets etc.	Norio Baba, Osaka Kyoiku University, Japan
Methods and Techniques of Artificial and Computational Intelligence in Engineering Design	Argyris Dentsoras, University of Patras, Greece Nikos Aspragathos, University of Patras, Greece Vassilis Moulianitis, University of the Aegean, Greece
Philosophical and Methodological Aspects of Reasoning and Decision Making	Vesa A. Niskanen, University of Helsinki, Finland

Semantic Technologies for Knowledge Workers	Andreas Dengel, German Research Center for Artificial Intelligence (DFKI), Germany Ansgar Bernardi, German Research Center for Artificial Intelligence (DFKI), Germany
Tools and Techniques for Effective Creation and Exploitation of Biodiversity Knowledge	Andrew C. Jones, Cardiff University, UK Richard J. White, Cardiff University, UK Gabriele Gianini, Università degli Studi di Milan, Italy Antonia Azzini, Università degli Studi di Milan, Italy Stefania Marrara, Università degli Studi di Milan, Italy
Immunity-Based Systems	Yoshiteru Ishida, Toyohashi University of Technology, Japan Takeshi Okamoto, Kanagawa Institute of Technology, Japan Yuji Watanabe, Nagoya City University, Japan Koji Harada, Toyohashi University of Technology, Japan

Program Committee

Abe, Akinori	IREIIMS University, Japan
Alexandre de Matos Araujo	Rui, University of Coimbra, Portugal
Angelov, Plamen	Lancaster University, UK
Anwer, Nabil	LURPA - ENS CACHAN, France
Aoki, Shingo	Osaka Prefecture University, Japan
Apolloni, Bruno	University of Milan, Italy
Aspragathos, Nikos A.	University of Patras, Greece
Bannore, Vivek	University of South Australia, Australia
Barb, Adrian S.	Penn State University, USA
Becker-Asano	Christian, Intelligent Robotics and Communication Labs, Japan
Bianchini, Monica	University of Siena, Italy
Bichindaritz, Isabelle	University of Washington, USA
Boeva, Veselka	Technical University of Sofia, Bulgaria
Boutalis, Yiannis	Democritus University of Thrace, Greece
Brna, Paul	University of Glasgow, UK
Buckingham, Christopher	Aston University, UK
Camastra, Francesco	University of Naples Parthenope, Italy
Cao, Cungen	Chinese Academy of Sciences, China
Ceccarelli, Michele	University of Sannio, Italy
Chalup, Stephan	The University of Newcastle, Australia
Chang, Bao Rong	National University of Kaohsiung, Taiwan
Chen, Lihui	Nanyang Technological University, Singapore
Chen, Toly	Feng Chia University, Taiwan
Cheng, Kai	Brunel University, UK
Cheung, Benny	Honk Kong Polytechnic University, Hong Kong
Cobos Pérez, Ruth	Universidad Autónoma de Madrid, Spain
Crippa, Paolo	Università Politecnica delle Marche, Italy

Cuzzocrea, Alfredo	University of Calabria, Italy
Damiana, Maria Luisa,	University of Milan, Italy
Dasiopoulou, Stamatia	Informatics and Telematics Institute, Greece
De Cock, Martine	University of Washington, USA
De Wilde, Philippe	Heriot-Watt University, UK
Dengel, Andreas	German Research Center for Artificial Intelligence (DFKI), Germany
Duro, Richard J.	Universidade da Coruña, Spain
Dustdar, Schahram	Vienna University of Technology, Austria
Elomaa, Tapio	Tampere University of Technology, Finland
Fernandez-Canque, Hernando	Glasgow Caledonian University, UK
Georgieva, Petia	University of Aveiro, Portugal
Godoy, Daniela	UNICEN University, Argentina
Grabot, Bernard	LGP-ENIT, France
Graña Romay, Manuel	Universidad del Pais Vasco, Spain
Grecos, Christos	University of West Scotland, UK
Hara, Takahiro	Osaka University, Japan
Hintea, Sorin	Cluj-Napoca University, Romania
Honda, Katsuhiko	Osaka Prefecture University, Japan
Hong, Tzung-Pei	National University of Kaohsiung, Taiwan
Hu, Chenyi	University of Central Arkansas, USA
Hurtado Larrain, Carlos	University of Chile, Chile
Ichalkaranje, Nikhil	University of South Australia, Australia
Ishibuchi, Hisao	Osaka Prefecture University, Japan
Ishida, Yoshiteru	Toyohashi University of Technology, Japan
Ito, Takayuki	Massachusetts Institute of Technology, USA
Ivancevic, Tijana	University of South Australia, Australia
Janicki, Ryszard	McMaster University, Canada
Jastroch, Norbert	MET Communications GmbH, Germany
Jensen, Richard	Aberystwyth University, UK
Jones, Andrew	Cardiff University, UK
Jordanov, Ivan	University of Portsmouth, UK
Jung, Jason J.	Yeungnam University, Korea
Juric, Matjaz B.	University of Maribor, Slovenia
Katagiri, Hideki	Hiroshima University, Japan
Ko, In-Young	KAIST, Korea
Kodogiannis, Vassilis S.	University of Westminster, UK
Koenig, Andreas	Technische Universitaet Kaiserslautern, Germany
Kojadinovic, Ivan	University of Auckland, New Zealand
Kompatiaris, Yiannis	Informatics and Telematics Institute, Greece
Konar, Amit	Jadavpur University, India
Koshizen, Takamasa	Honda R&D Co., Ltd., Japan
Koychev, Ivan	University of Sofia, Bulgaria
Kwong, C.K.	The Hong Kong Polytechnic University, Hong Kong
Lee, Dah-Jye	Brigham Young University, USA
Lee, W.B.	Hong Kong Polytechnic University, Hong Kong
Likas, Aristidis	University of Ioannina, Greece

Lim, C.P.	Universiti Sains Malaysia, Malaysia
Liu, Lei	Beijing University of Technology, China
Maglogiannis, Ilias	University of Central Greece, Greece
Maier, Patrick	The University of Edinburgh, UK
Marinov, Milko T.	University of Ruse, Bulgaria
McCauley Bush, Pamela	University of Central Florida, USA
Montani, Stefania	Università del Piemonte Orientale, Italy
Moreno Jimenez, Ramón	Universidad del Pais Vasco, Spain
Nguyen, Ngoc Thanh	Wroclaw University of Technology, Poland
Nishida, Toyooki	Kyoto University, Japan
Niskanen, Vesa A.	University of Helsinki, Finland
Ohkura, Kazuhiro	Hiroshima University, Japan
Palade, Vasile	Oxford University, UK
Pallares, Alvaro	Plastiasite S.A., Spain
Paranjape, Raman	University of Regina, Canada
Pasek, Zbigniew J.	University of Windsor, Canada
Pasi, Gabriella	University of Milan, Italy
Passerini, Andrea	Università degli Studi di Trento, Italy
Pazos-Arias, Jose	University of Vigo, Spain
Petrosino, Alfredo	Università di Napoli Parthenope, Italy
Prada, Rui	IST-UTL and INESC-ID, Portugal
Pratihari, Dilip Kumar	Osaka Prefecture University, Japan
Putnik, Goran D.	University of Minho, Portugal
Reidsema, Carl	University of New South Wales, Australia
Resconi, Germano	Catholic University in Brescia, Italy
Rovetta, Stefano	University of Genoa, Italy
Sansone, Carlo	Università di Napoli Federico II, Italy
Sarangapani, Jagannathan	Missouri University of Science and Technology, USA
Sato-Ilic, Mika	University of Tsukuba, Japan
Schockaert, Steven	Ghent University, Belgium
Seiffert, Udo	Fraunhofer-Institute IFF Magdeburg, Germany
Simperl, Elena	University of Innsbruck, Austria
Smrz, Pavel	Brno University of Technology, Czech Republic
Soroka, Anthony	Cardiff University, UK
Szczerbicki, Edward	The University of Newcastle, Australia
Tanaka, Takushi	Fukuoka Institute of Technology, Japan
Teng, Wei-Chung	National Taiwan University of Science and Technology, Taiwan
Tichy, Pavel	Rockwell Automation Research Centre, Czech Republic
Tino, Peter	The University of Birmingham, UK
Tolk, Andreas	Old Dominion University, USA
Toro, Carlos	VICOMTech, Spain
Torra, Vicenc	IIIA-CSIC, Spain
Tsihrintzis, George	University of Piraeus, Greece
Tsiporkova, Elena	Sirris, Belgium
Turchetti, Claudio	Università Politecnica delle Marche, Italy

Uchino, Eiji	Yamaguchi University, Japan
Urlings, Pierre	DSTO, Department of Defence, Australia
Vadera, Sunil	University of Salford, UK
Valdéz Vela, Mercedes	Universidad de Murcia, Spain
Vellido, Alfredo	Universitat Politècnica de Catalunya, Spain
Virvou, Maria	University of Piraeus, Greece
Wang, Zidong	Brunel University, UK
Watts, Mike	TBA, New Zealand
White, Richard J.	Cardiff University, UK
Williams, M. Howard	Heriot-Watt University, UK
Yang, Zijiang	York University, Canada
Yoshida, Hiroyuki	Harvard Medical School, USA
Zanni-Merk, Cecilia	LGeCo - INSA de Strasbourg, France
Zheng, Li-Rong	Royal Institute of Technology (KTH), Sweden

Reviewers

Adam Nowak	Bao Rong Chang	David Vallejo
Adam Slowik	Benjamin Adrian	Davor Skrlec
Adrian S. Barb	Bernard Grabot	Dickson Lukose
Akinori Abe	Bernd Reusch	Dilip Pratihar
Akira Hattori	Bettina Waldvogel	Doctor Jair Abe
Alan Paton	Björn Forcher	Don Jeng
Alessandra Micheletti	Bob Coyne	Donggang Yu
Alfredo Cuzzocrea	Bojan Basrak	Doris Csipkes
Ammar Aljer	Bojana Dalbelo Basic	Eduardo Cerqueira
Amparo Vila	Bozidar Ivankovic	Eduardo Merlo
Ana Fernandez-Vilas	Branko Zitko	Edward Szczerbicki
Anastasia Kastania	Bruno Apolloni	Eiji Uchino
Anastasius Moutzoglou	Calin Ciufudean	Elena Pagani
Andrea Visconti	Carlo Sansone	Elena Simperl
Andreas Abecker	Carlos Ocampo	Esmail Bonakdarian
Andreas Dengel	Carlos Pedrinaci	Esmiralda Moradian
Andreas Oikonomou	Carlos Toro	Francesco Camastra
Andrew Jones	Cecilia Zanni-Merk	Frane Saric
Annalisa Appice	Cesar Sanin	Fujiki Morii
Anne Håkansson	Chang-Tien Lu	Fumihiko Anma
Ansgar Bernardi	Christian Becker-Asano	Fumitaka Uchio
Anthony Soroka	Christine Mumford	Gabbar Hossam
Antonio Gomez-Skarmeta	Chunbo Chu	Gabor Csipkes
Antonio Zippo	Costantino Lucisano	Gabriel Oltean
Aristidis Likas	C.P. Lim	Gabriella Pasi
Armando Buzzanca	Cristos Orovos	George Mitchell
Artur Silic	Daniela Godoy	George Tsihrintzis
Athina Lazakidou	Danijel Radosevic	Gergely Héja
Azizul Azhar Ramli	Danilo Dell'Agnello	Gianluca Sforza
Balázs Gaál	David Martens	Giovanna Castellano

Giovanni Gomez Zuluaga	Kazuhiro Ohkura	Narayanan
Gunnar Grimnes	Kazuhiro Takeda	Kulathuramaiyer
Gyorgy Surjan	Kazuhisa Seta	Nikica Hlupi
Haoxi Dorje Zhang	Kazumi Nakamatsu	Nikola Ljubescic
Haruhiko H. Nishimura	Kazunori Nishino	Nikos Tsourveloudis
Haruhiko Haruhiko	Kazuo Misue	Nobuhiro Inuzuka
Nishimura	Keiichiro Mitani	Nobuo Suzuki
Haruki Kawanaka	Kenji Matsuura	Norbert Jastroch
Hector Alvarez	Koji Harada	Norio Baba
Hernando	Kouji Yoshida	Noriyuki Matsuda
Fernandez-Canque	Lars Hildebrand	Omar Rana
Hideaki Ito	Laura Caponetti	Orleo Marinaro
Hidehiko Hayashi	Lei Liu	Paolo Crippa
Hideo Funaoi	Lelia Festila	Pasquale Di Meo
Hideyuki Matsumoto	Leonardo Mancilla	Pavel Tichy
Hirokazu Miura	Amaya	Philippe Wilde
Hisayoshi Kunimune	Lily Lin	Rafael Batres
Hrvoje Markovic	Ljiljana Stojanovic	Raffaele Cannone
Huey-Ming Lee	Lorenzo Magnani	Ramón Jimenez
Ilias Maglogiannis	Lorenzo Valerio	Rashid Mehmood
Ing. Angelo Ciccazzo	Ludger van Elst	Richard Pyle
Ivan Koychev	Manuel Grana	Richard White
Ivan Stajduhar	Marek Malski	Robert Howlett
J. Mattila	Maria Torsello	Roberto Cordone
Jair Abe	Mario Koeppen	Ronald Hartung
Jari Kortelainen	Marko Banek	Roumen Kountchev
Jayanthi Ranjan	Martin Lopez-Nores	Rozália Lakner
Jerome Darmont	Martine Cock	Ruediger Oehlmann
Jessie Kennedy	Masakazu Takahashi	Ruth Cobos
Jesualdo Tomás	Masaru Noda	Ryohei Sakano
Fernández-Breis	Masato Soga	Ryuuki Sakamoto
Jiangtao Cao	Masayoshi Aritsugi	Sachio Hirokawa
Jim Sheng	Mayumi Ueda	Satoru Fujii
Johnson Fader	Melita Hajdinjak	Sebastian Rios
Jose Manuel Molina	Michelangelo Ceci	Sebastian Weber
Juan Botia	Michele Missikoff	Seiji Isotani
Juan Manuel Corchado	Miguel Delgado	Seiki Akama
Juan Pavon	Milko Marinov	Setsuya Kurahashi
Julia Hirschberg	Minhua Ma	Shamshul Bahar Yaakob
Jun Munemori	Minoru Minoru Fukumi	Shinji Fukui
Jun Sawamoto	Monica Bianchini	Shusaku Tsumoto
Junzo Watada	Motoi Iwashita	Shyue-Liang Wang
Jure Mijic	Motoki Miura	Simone Bassis
Katalina Grigorova	Nahla Barakat	Sophia Kossida
Katsuhiro Honda	Naohiro Ishii	Stamatia Dasiopoulou
Katsumi Yamashita	Naoto Mukai	Stefan Zinsmeister
Kazuhiko Tsuda	Naoyuki Naoyuki Kubota	Stefania Marrara

Stefania Montani	Toru Fukumoto	Yiannis Boutalis
Stephan Chalup	Toshihiro Hayashi	Yoshifumi Tsuge
Steven Schockaert	Toshio Mochizuki	Yoshihiro Okada
Sunil Vadera	Toumoto	Yoshihiro Takuya
susumu hashizume	Toyohide Watanabe	Yoshinori Adachi
Susumu Kunifuji	Toyohis Nakada	Yoshiyuki Yamashita
Takanobu Umetsu	Tsuyoshi Nakamura	Youji Ochi
Takashi Hamaguchi	Tuan Pham	Young Ko
Takashi Mitsuishi	Valerio Arnaboldi	Yuichiro Tateiwa
Takashi Yukawa	Vassilis Kodogiannis	Yuji Iwahori
Takaya Yuizono	Vassilis Moulitanitis	Yuji Wada
Takeshi Okamoto	Vesa Niskanen	Yuji Watanabe
Taketoshi Kurooka	Veselka Boeva	Yuki Hayashi
Taketoshi Ushiana	Vivek Bannore	Yukio Ohsawa
Takushi Tanaka	Wataru Sunayama	Yumiko Nara
Tapio Elomaa	Wei-Chung Teng	Yurie Iribe
Tatiana Tambouratzis	William Hochstettler	Zdenek Zdrahal
Tessai Hayama	Winston Jain	Ziad Salem
Thomas Roth-Berghofer	Wolfgang Stock	Zijiang Yang
Tomislav Hrkac	Xiaofei Ji	Zlatko Drmac
Tomoko Kojiri	Yi Xiao	Zuwairie Ibrahim

Table of Contents – Part III

Knowledge-Based Systems for e-Business

A Study on Traveling Purpose Classification Method to Extract Traveling Requests	1
<i>Nobuo Suzuki, Mariko Yamamura, and Kazuhiko Tsuda</i>	
Variable Selection by C_p Statistic in Multiple Responses Regression with Fewer Sample Size Than the Dimension	7
<i>Mariko Yamamura, Hirokazu Yanagihara, and Muni S. Srivastava</i>	
Customer Path Controlling in the Retail Store with the Vertex Dominating Cycle Algorithms	15
<i>Takeshi Sugiyama</i>	

Quality Assurance and Intelligent Web-Based Information Technology

A Framework for the Quality Assurance of Blended E-Learning Communities	23
<i>Iraklis Varlamis and Ioannis Apostolakis</i>	
Quality of Content in Web 2.0 Applications	33
<i>Iraklis Varlamis</i>	
Telepediatrics Education on the Semantic Web	43
<i>Sofia Sidirokastriti and Anastasia N. Kastania</i>	
Web Applications and Public Diplomacy	53
<i>Antigoni Koffa and Anastasia N. Kastania</i>	

Knowledge-Based Interface Systems

A Hybrid Face Recognition System for Managing Time of Going to Work and Getting away from Office	63
<i>Yoshinori Adachi, Zeng Yunfei, Masahiro Ozaki, and Yuji Iwahori</i>	
Multi-Relationa Pattern Mining System for General Database Systems	72
<i>Nobuhiro Inuzuka and Toshiyuki Makino</i>	
Recovering 3-D Shape Based on Light Fall-Off Stereo under Point Light Source Illumination and Perspective Projection	81
<i>Yuji Iwahori, Claire Roweyrol, Robert J. Woodham, Yoshinori Adachi, and Kunio Kasugai</i>	

Shadow Detection Method Based on Dirichlet Process Mixture Model	89
<i>Wataru Kurahashi, Shinji Fukui, Yuji Iwahori, and Robert J. Woodham</i>	
Vowel Sound Recognition Using a Spectrum Envelope Feature Detection Method and Neural Network	97
<i>Masashi Kawaguchi, Naohiro Yonekura, Takashi Jimbo, and Naohiro Ishii</i>	
Information Extraction Using XPath	104
<i>Masashi Okada, Naohiro Ishii, and Ippei Torii</i>	
Information Visualization System for Activation of Shopping Streets	113
<i>Ippei Torii, Yousuke Okada, Takahito Niwa, Manabu Onogi, and Naohiro Ishii</i>	
Reasoning Based Intelligent Systems	
Introduction to Intelligent Network Routing Based on EVALPSN	123
<i>Kazumi Nakamatsu, Jair Minoro Abe, and Takashi Watanabe</i>	
Introduction to Intelligent Elevator Control Based on EVALPSN	133
<i>Kazumi Nakamatsu, Jair Minoro Abe, Seiki Akama, and Roumen Kountchev</i>	
Monadic Curry System N_1^*	143
<i>Jair Minoro Abe, Kazumi Nakamatsu, and Seiki Akama</i>	
A Sensing System for an Autonomous Mobile Robot Based on the Paraconsistent Artificial Neural Network	154
<i>Claudio Rodrigo Torres, Jair Minoro Abe, Germano Lambert-Torres, João Inácio Da Silva Filho, and Helga Gonzaga Martins</i>	
Paraconsistent Artificial Neural Networks and EEG Analysis	164
<i>Jair Minoro Abe, Helder F.S. Lopes, Kazumi Nakamatsu, and Seiki Akama</i>	
A Reasoning-Based Strategy for Exploring the Synergy among Alternative Crops	174
<i>Hércules Antonio do Prado, Edilson Ferneda, and Ricardo Coelho de Faria</i>	
Reasoning Elements for a Vehicle Routing System	182
<i>Edilson Ferneda, Bernardo A. Mello, Janaína A.S. Diniz, and Adelaide Figueiredo</i>	
A Mechanism for Converting Circuit Grammars to Definite Clauses	190
<i>Takushi Tanaka</i>	

Constructive Discursive Reasoning	200
<i>Seiki Akama, Kazumi Nakamatsu, and Jair Minoro Abe</i>	
Formal Concept Analysis of Medical Incident Reports	207
<i>Takahiro Baba, Lucing Liu, and Sachio Hirokawa</i>	
Compression of Multispectral Images with Inverse Pyramid Decomposition	215
<i>Roumen Kountchev and Kazumi Nakamatsu</i>	

Data Mining and Service Science for Innovation

Econometric Approach for Broadband Market in Japan	225
<i>Takeshi Kurosawa, Hiromichi Kawano, Motoi Iwashita, Shinsuke Shimogawa, Shouji Kouno, and Akiya Inoue</i>	
Opinion Exchange Support System by Visualizing Input History	235
<i>Yukihiko Tamura, Yuuki Tomiyama, and Wataru Sunayama</i>	
Extracting Promising Sequential Patterns from RFID Data Using the LCM Sequence	244
<i>Takanobu Nakahara, Takeaki Uno, and Katsutoshi Yada</i>	
Relation between Stay-Time and Purchase Probability Based on RFID Data in a Japanese Supermarket	254
<i>Keiji Takai and Katsutoshi Yada</i>	
Implementing an Image Search System with Integrating Social Tags and DBpedia	264
<i>Chie Iijima, Makito Kimura, and Takahira Yamaguchi</i>	
The Influence of Shopping Path Length on Purchase Behavior in Grocery Store	273
<i>Marina Kholod, Takanobu Nakahara, Haruka Azuma, and Katsutoshi Yada</i>	
Existence of Single Input Rule Modules for Optimal Fuzzy Logic Control	281
<i>Takashi Mitsuishi, Hidefumi Kawakatsu, and Yasunari Shidama</i>	

Innovations in Chance Discovery

Temporality and Reference Place: Discovering Chances for Conflict Avoidance in Teamwork	290
<i>Ruediger Oehlmann</i>	
Discovering Research Key Terms as Temporal Patterns of Importance Indices for Text Mining	297
<i>Hidenao Abe and Shusaku Tsumoto</i>	

Categorized and Integrated Data Mining of Medical Data from the Viewpoint of Chance Discovery	307
<i>Akinori Abe, Norihiro Hagita, Michiko Furutani, Yoshiyuki Furutani, and Rumiko Matsuoka</i>	
Support System for Thinking New Criteria of Unclassified Diseases	315
<i>Yoko Nishihara, Yoshimune Hiratsuka, Akira Murakami, Yukio Ohsawa, and Toshiro Kumakawa</i>	
Interpretation of Chance Discovery in Temporal Logic, Admissible Inference Rules	323
<i>Vladimir Rybakov</i>	
Faking Chance Cognitive Niche Impoverishment	331
<i>Lorenzo Magnani and Emanuele Bardone</i>	
Advanced Knowledge-Based Systems	
Summarization for Geographically Distributed Data Streams	339
<i>Anna Ciampi, Annalisa Appice, and Donato Malerba</i>	
Gradual Data Aggregation in Multi-granular Fact Tables on Resource-Constrained Systems	349
<i>Nadeem Iftikhar and Torben Bach Pedersen</i>	
A Refinement Operator Based Method for Semantic Grouping of Conjunctive Query Results	359
<i>Agnieszka Lawrynowicz, Claudia d'Amato, and Nicola Fanizzi</i>	
Semantic Network of Ground Station-Satellite Communication System	369
<i>Katarzyna Dąbrowska-Kubik</i>	
W-kmeans: Clustering News Articles Using WordNet	379
<i>Christos Bouras and Vassilis Tsogkas</i>	
An Efficient Mechanism for Stemming and Tagging: The Case of Greek language	389
<i>Giorgos Adam, Konstantinos Asimakis, Christos Bouras, and Vassilis Pouloupoulos</i>	
Co-clustering Analysis of Weblogs Using Bipartite Spectral Projection Approach	398
<i>Guandong Xu, Yu Zong, Peter Dolog, and Yanchun Zhang</i>	
Talking Biology in Logic, and Back	408
<i>Hasan Jamil</i>	

Analysis of Medical Pathways by Means of Frequent Closed Sequences	418
<i>Elena Baralis, Giulia Bruno, Silvia Chiusano, Virna C. Domenici, Naeem A. Mahoto, and Caterina Petrigni</i>	
Inheriting Access Control Rules from Large Relational Databases to Materialized Views Automatically	426
<i>Alfredo Cuzzocrea, Mohand-Said Hacid, and Nicola Grillo</i>	
MySQL Data Mining: Extending MySQL to Support Data Mining Primitives (Demo)	438
<i>Alfredo Ferro, Rosalba Giugno, Piera Laura Puglisi, and Alfredo Pulvirenti</i>	
A Genetic Algorithm to Design Industrial Materials	445
<i>E. Tenorio, J. Gómez-Ruiz, J.I. Peláez, and J.M. Doña</i>	
Intelligent Network and Service	
A Proposal of P2P Content Retrieval System Using Access-Based Grouping Technique	455
<i>Takuya Sasaki, Jun Sawamoto, Takashi Katoh, Yuji Wada, Norihisa Segawa, and Eiji Sugino</i>	
The Effects of Individual Differences in Two Persons on the Distributed and Cooperative KJ Method in an Anonymous Environment	464
<i>Takaya Yuizono and Zhe Jin</i>	
Pictograph Chat Communicator III: A Chat System That Embodies Cross-Cultural Communication	473
<i>Jun Munemori, Taro Fukuda, Moonyati Binti Mohd Yatid, Tadashi Nishide, and Junko Itou</i>	
Distance Learning Support System for Game Programming with Java	483
<i>Kouji Yoshida, Takumu Yaoi, Isao Miyaji, Kunihiro Yamada, and Satoru Fujii</i>	
Evidence Analysis Method Using Bloom Filter for MANET Forensics	493
<i>Takashi Mishina, Yoh Shiraishi, and Osamu Takahashi</i>	
Diminished Reality for Landscape Video Sequences with Homographies	501
<i>Kosuke Takeda and Ryuuki Sakamoto</i>	
Prediction of Combinatorial Protein-Protein Interaction Networks from Expression Data Using Statistics on Conditional Probability	509
<i>Takatoshi Fujiki, Etsuko Inoue, Takuya Yoshihiro, and Masaru Nakagawa</i>	

Development and Evaluation of a Historical Tour Support System Using 3D Graphics and Mobile Terminal	519
<i>Satoru Fujii, Takahiro Shima, Megumi Takahashi, and Koji Yoshida</i>	
Repetition of Dialogue Atmosphere Using Characters Based on Face-to-Face Dialogue	527
<i>Junko Ito and Jun Munemori</i>	
Soft Computing Techniques and Their Intelligent Utilizations Toward Gaming, Robotics, Stock Markets etc.	
CMOS-Based Radiation Movie and Still Image Pickup System with a Phototimer Using Smart Pattern Recognition	535
<i>Osamu Yuuki, Hiroshi Mineno, Kunihiko Yamada, and Tadanori Mizuno</i>	
Optimal H ₂ Integral Controller Design with Derivative State Constraints for Torsional Vibration Model	545
<i>Noriyuki Komine and Kunihiko Yamada</i>	
Utilization of Evolutionary Algorithms for Making COMMONS GAME Much More Exciting	555
<i>Norio Baba, Hisashi Handa, Mariko Kusaka, Masaki Takeda, Yuriko Yoshihara, and Keisuke Kogawa</i>	
Education of Embedded System by Using Electric Fan	562
<i>Osamu Yuuki, Junji Namiki, and Kunihiko Yamada</i>	
Development and Evaluation of a Routing Simulator for a Mutually Complementary Network Incorporating Wired and Wireless Components	572
<i>Hiroki Morita, Naoki Yusa, Noriyuki Komine, Kouji Yoshida, Masanori Kojima, Tadanori Mizuno, and Kunihiko Yamada</i>	
On the Impact of the Metrics Choice in SOM Learning: Some Empirical Results from Financial Data	583
<i>Marina Resta</i>	
Reinforcement Learning Scheme for Grouping and Characterization of Multi-agent Network	592
<i>Koichiro Morihiro, Nobuyuki Matsui, Teijiro Isokawa, and Haruhiko Nishimura</i>	
Extracting Principal Components from Pseudo-random Data by Using Random Matrix Theory	602
<i>Mieko Tanaka-Yamawaki</i>	

Music Impression Detection Method for User Independent Music Retrieval System	612
<i>Masato Miyoshi, Satoru Tsuge, Hillary Kipsang Choge, Tadahiro Oyama, Momoyo Ito, and Minoru Fukumi</i>	
Applying Fuzzy Sets to Composite Algorithm for Remote Sensing Data	622
<i>Kenneth J. Mackin, Takashi Yamaguchi, Jong Geol Park, Eiji Nunohiro, Kotaro Matsushita, Yukio Yanagisawa, and Masao Igarashi</i>	
Immunity-Based Systems	
Evaluations of Immunity-Based Diagnosis for a Motherboard	628
<i>Haruki Shida, Takeshi Okamoto, and Yoshiteru Ishida</i>	
A Note on Dynamical Behaviors of a Spatial Game Operated on Intercrossed Rules	637
<i>Kouji Harada and Yoshiteru Ishida</i>	
Asymmetry in Repairing and Infection: The Case of a Self-repair Network	645
<i>Yoshiteru Ishida and Kei-ichi Tanabe</i>	
A Note on Symmetry in Logic of Self-repair: The Case of a Self-repair Network	652
<i>Yoshiteru Ishida</i>	
An Immunity-Based Scheme for Statistical En-route Filtering in Wireless Sensor Networks	660
<i>Yuji Watanabe</i>	
Author Index	667

A Study on Traveling Purpose Classification Method to Extract Traveling Requests

Nobuo Suzuki¹, Mariko Yamamura², and Kazuhiko Tsuda²

¹KDDI Corporation

Iidabashi 3-10-10, Chiyoda, Tokyo 102-8460, Japan
nu-suzuki@kddi.com

²Graduate School of Business Sciences, University of Tsukuba
Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
{yamamura, tsuda}@gssm.otuka.tsukuba.ac.jp

Abstract. It is possible to get the flow information of people and transportation efficiently by collecting travel information of people and vehicles in the Internet. In the meantime, Q&A Web sites on the Internet express human requests more directly and they are the useful resources to extract many kinds of knowledge and intentions. This paper proposes the classification method for Japanese speeches in the sites based on the road traffic census by MLIT in Japan to extract traveling requests. Specifically, this method presumes the traveling purposes by SVM. Learning with the frequency of parts of speech that express traveling requests and TFIDF as the features of SVM is carried out. We also evaluated the performance by the experiment and got the result the accuracy 45.5%.

Keywords: Traveling Purpose, SVM, Q&A site.

1 Introduction

In recent years, the information of the road congestion are generated from sensors on some roads and location information collected by each vehicles. However, it is difficult to manage the global congestion situation includes flows of human and trains so on with only these information. In other hand, a large amount of information opened on the Internet include geometric traveling requests. It has a high possibility to manage the traffic flow effectively with extracting such information and collecting the traveling information of human and vehicles. On the other hand, Q&A Web sites that are used widely on the Internet express the human requests information directly well and are useful resources on the Internet for extracting many kind of knowledges and intentions. Therefore, the goal of this study is to manage the traffic flow by extracting the traveling request information that appears in the sites and collecting the human traveling conditions.

This paper proposes an extracting method of Japanese traveling requests from the sites to realize the goal. Specifically, we considered the traveling purpose classification method by a speech in the sites based on the traveling purpose classification of “Road Traffic Census” defined at MLIT in Japan[1]. This method classified with SVM (Support Vector Machine) and used the frequent order of the morphemes and TFIDF of the kind of speeches appeared distinctively in the traveling requests as its features. This paper also describes the classification method details and the result of the evaluation experiment and its analysis.

2 The Classification of the Traveling Purposes

Table 1 shows the examples of the traveling request speeches in a Q&A Web site. It is regarded to extract the information such as a traveling purpose, an age, a sex, and an occupation of a traveling person, number of persons, a route of travel and a strength of a traveling request from such speeches as known each speeches. We focused the traveling purpose that was the fundamental information of the traveling request in those information.

The definition of the referenced classification is required to classify the traveling purpose. Many classifications are proposed and this study uses the classification defined in “Road Traffic Census” of MLIT in Japan because of treating the traffic information mainly.

Table 1. The examples of the traveling request speeches in Q&A sites

I'm going to travel to Fuji-Hakone-Izu with my parents and grand parents at September. Please tell me how long it will take to Izu-Kitakawa from Hakone(around Ashino Lake) by a car.
I'm going to go to Tsumakago from Chubu International Airport. Please tell me which route is faster, Nagoya Highway and Chuo Highway at Komaki JCT or Tokai Circle Highway and Chuo Highway. I'm going to leave from Chubu International Airport at 10 o'clock for your reference.
I am a countryman who goes to Tokyo at 2 nd week in August. I'm going to tour Asakusa, Odaiba, Shibuya and Harajuku. It's a royal road of the sightseeing at Tokyo. Do you know the shops open even in Japanese Obon Season? I'm wondering they are not opened at the season.

Table 2 shows this traveling purpose classification of Road Traffic Census. We realized a lot of returning to the family home speeches at analyzing the traveling request speeches. Therefore, “Returning to the family home” was added into this traveling purpose classification of Road Traffic Census. And speeches didn't include the traveling purposes were classified to “Unknown”.

Table 2. The traveling purpose classifications of Road Traffic Sensous

No.	Items	Descriptions
1	Going to work	Going to the office.
2	Going to school	Going to the school and extracurricular activities. It doesn't include supplementary private schools.
3	Housework and Shopping	It does not include a shopping for business.
4	Meal, Social Contact and Entertainment	It includes private acquaintance in one's scope of daily life. Films and Restraints so on.
5	Sightseeing	Sightseeing to places of natural beauties and historic interests.
6	Recreation	Hot spring and exchange with families and friends.
7	Sports	Hiking, golf, sports day and other sports.
8	Experience type leisure	Amusement park, going for a drive, eating local specialties and so on.
9	Other private business	Seeing a doctor regularly, going to lessons and so on.
10	Picking up	It does not include a business.
11	Business without transportation	It does not include slight one's belongings such as bags.
12	Business with transportation	It includes using vehicles due to difficulties to transport for the business.
13	Backing to the office	Moving to back to the office after finishing the business.
14	Backing to home	Backing to home from the office, the school, shopping and others.
15	Returning to the family home	Moving for returning to the family home.
16	Others	Others except for described above.
17	Unknown	Unknown traveling purposes.

3 Selecting the Features

Each speech often does not express the traveling purpose clearly. For instance, if the speech includes "We are going to travel to Izu district", the possibility of its traveling purpose is "Recreation" in table 2 is high even it does not mention a trip to a hot spring clearly. Therefore, our method uses SVM that is based on machine learning method focused in recent years without classifying by construction words dictionary directly[2]. It also tried two types of data mentioned under below as the SVM features.

(Type 1) The order of the frequency at each classification

We analyzed the frequency ratio of parts of speech in 225 speeches included the traveling requests collected for learning. The result of this analysis is table 3. Although these parts of speech always appear high frequently, its order of the frequency is different with each speech. For example, if one speech is classified to "Recreation", a

place name and a hotel name appear in high order. If one speech is classified to “Going to work”, the time and the transportation methods appear in high order. Therefore, our study uses the order of the frequency of morphemes for certain parts of speech in each traveling purpose classification as SVM features. Specifically, ten parts of speech in table 3 were used.

Table 3. The frequency of the parts of speech related to traveling purpose in traveling request speeches

Parts of speech	Ratio	Examples of words
Verb – Independence	10.1%	Reach, Go, Back
Noun – General	10.1%	Taxi, Shinkansen, Highway, Hotel, Concert
Particle – Case particle – General	9.5%	Kara, Ni (Japanese particles)
Noun – Proper noun – Area – General	4.6%	Shinagawa, Omotesando (Japanese place name)
Noun – Sa-conjunction	4.4%	Departure, Arrival, Detour, Traffic jam
Unknown word	3.0%	ETC, 160km, Aqualine
Noun – Adverb possible	2.3%	Tuesday, Morning
Noun – Proper noun – General	0.7%	Tomei-highway, Komaki, Disneyland
Noun – Proper noun – Parson’s name – Last name	0.4%	Mishima, Toyoshina, Kumamoto
Noun – Proper noun – Organization	0.3%	Ohiso Prince hotel, JR, Kitasato university

(Type 2) The TFIDF in the parts of speech appear distinctively

TFIDF is often used as an index to extract representative keywords for the document[3]. It is expected that these representative keywords classify more accurately for traveling request speeches. However, only calculating TFIDF of words will have a lot of noises of particles and make to fall the accuracy. Therefore, we decided to use the TFIDF of morphemes by parts of speech with “Noun-General” and “Noun-Sa-conjunction” that express the features of speeches include the traveling purposes well. The TDIDF of morphemes by a part of speech is expressed with the equation (1). “tf” is the frequency of morphemes for the part of speech included in a speech. “df” is the number of speeches included the morphemes of the parts of speech in the traveling request speeches. “N” is total number of the speeches for learning. “i” is the number of the speeches and “j” is the number of the part of speech.

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

4 Evaluation Experiment

We evaluated the classification method mentioned above by the experiment with actual speech data in Q&A sites. 450 speeches were collected as the evaluation

experimental data. 225 speeches were used for the learning and 225 speeches for the evaluation. The positive examples were 68% and the negative examples were 84% for the learning, and the positive examples were 84% and the negative examples were 16% for the evaluation. TinySVM was used as a SVM tool[4]. The kernel function of SVM was set with linear as the default setting and the cost margin parameter C also was set with 1.0 as the default one.

At first, the experiment calculated two kinds of the features by a speech from the learning data showed chapter 3 and built the model. Specifically, the morphemes of the learning data collected were calculated with ChaSen, Japanese morpheme analysis tool, and the frequency of the morphemes by a speech. Next, the TFIDF values by a morpheme of the certain part of speech were calculated. It learned these two values with SVM by a speech and made the models. Moreover, the classification was presumed by a speech of the evaluation data with these models. The evaluation made the correct answer if the classification each speech was distinguished by the person was same as the presumption result with this method. As the result of the experiment, the precision was 45.5%. The examples of the correct and incorrect answers are shown at table 4.

Table 4. The examples of the correct and incorrect answers

	Contents of the Speeches	Presumed Classification	Correct Classification
Correct	We would like to go to eat grapes on this weekend or next week. Please tell me a grape garden where we can eat as much as we like at Fukuoka or Ohita prefecture.	Experience type Leisure	Experience type Leisure
Incorrect	I'm going to go on a business and go around Sakai city and Higashi-Ohsaka city with a rental car. I will stay Sakai city on the previous day and get a rental car at Sakai station on the next morning. Should I return it at Osaka station or the suburban store?	Going to work	Business without transportation

It was realized that the incorrect data were classified to three kinds of data as table 5 through the analysis. The classification of each incorrect data is described below.

Table 5. The classifications of the incorrect data

	Classification	Rate
1	Incorrect with the characteristic morphemes	52.8%
2	Incorrect with the characteristic parts of speech	33.7%
3	Incorrect traveling purposes without expressions of traveling purposes in the speech	13.5%

First of all, “Incorrect with the characteristic morphemes” may become incorrect because the correct learning data don’t include the characteristic morphemes represent the traveling purposes. Those speeches may be presumed the traveling purposes with the parts of speech in table 3 defined by the present features, and become incorrect because they don’t include the characteristic morphemes in the positive learning data. The data, for instance, can be presumed to “Meal, Social Contact and Entertainment” classification from the morpheme of “Live”. We think these kinds of data will be presumed correctly by increasing the positive learning data.

Next, “Incorrect with the characteristic parts of speech” may become incorrect because the positive learning data don’t include the characteristic parts of speech represent the traveling purposes for TFIDF. The data, for instance, are “Noun – Proper noun” of “Ise-Jingu” and “Verb – Independence” of “Tabe-rareru Tokoro (The place we can eat)”. We think these kinds of data will be presumed correctly by adding following new parts of speech. “Noun – Proper noun”, “Noun – Suffix”, “Noun – Independence”, “Verb – Dependence”, “Verb – Suffix” and “Adjective – Dependence”.

Finally, “Incorrect traveling purposes without expressions of traveling purposes in the speech” includes the data classified to incorrect traveling purposes without expressions of traveling purposes themselves. They may become incorrect because of a lack of the negative learning data don’t represent the traveling purposes. The characteristic morphemes are appeared now and then the characteristic morphemes apply to any classification in these speeches. However, they represent the traveling purposes as a whole. We think these kinds of data will be presumed to “Unknown” correctly by increasing the negative learning data.

5 Conclusion

In this paper, we propose the classification method to extract the traveling requests for speeches in Q&A sites include the traveling requests. The classification processing uses SVM and adopts the order of the frequency and TFIDF for certain parts of speech as its features. Moreover, we confirmed its performance with the evaluated experiment. As the result, the accuracy was 45.5% and we couldn’t get enough performance. The reasons of this result are found that a lack of learning data and the validity for the kinds of parts of speech used as the features after analyzing the incorrect data. We are planning to evaluate these concerns with building the system concretely by the experiment.

References

1. MLIT: Road Traffic Census (2009), <http://www.mlit.go.jp/road/census/h21/index.html>
2. Vapnik, N.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1998)
3. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York (1986)
4. TynySVM, <http://chasen.org/~taku/software/TinySVM/>

Variable Selection by C_p Statistic in Multiple Responses Regression with Fewer Sample Size Than the Dimension

Mariko Yamamura¹, Hirokazu Yanagihara^{2,*}, and Muni S. Srivastava³

¹ Graduate School of Business Sciences, University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-0012, Japan
yamamura@gssm.otsuka.tsukuba.ac.jp

² Department of Mathematics, Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan
yanagi@math.sci.hiroshima-u.ac.jp

³ Department of Statistics, University of Toronto
100 St. George Street, Toronto, Ontario ON M5S 3G3, Canada
srivasta@utstat.toronto.edu

Abstract. In this paper, we introduce a better statistical method about model selection, and contribute to updating data mining technique. We consider the problem of selecting q explanatory variables out of k ($q \leq k$), when the dimension p of the response variables is larger than the sample size n in the multiple responses regression. We consider C_p statistic which is an estimator of the sum of standardized mean square errors. The standardization uses the inverse of the variance-covariance matrix of p response variables and thus the estimator of the inverse of the sample variance-covariance matrix. However, since $n < p$, such an inverse matrix cannot be used. Thus, we use the Moore-Penrose inverse and define the C_p statistic. Such a statistic will be denoted by C_p^+ . An example is given to illustrate the use of C_p^+ statistic. The performance is demonstrated by simulation result and real data study.

Keywords: High dimensional data, Mallows' C_p statistic, Model selection, Moore-Penrose inverse, Multivariate linear regression model.

1 Introduction

A statistical analysis is a powerful tool to understand and explain about our interest in many fields, i.e., it is well used for marketing in business. A data mining technique, especially statistical analysis, depends on a statistical software, and the statistical software has been updated after some new or better statistical analysis methods are introduced. In these days, we often see a high dimensional data that the dimension of vector of mutually correlated response variables is

* This research was supported by the Japan Society for the Promotion of Science, Excellent Young Researchers Overseas Visit Problem, #21-2086.

larger than sample size. Methods of analyzing the high dimensional data is recently started studying, and updating statistical software has not been done yet. Therefore, in this paper, we introduce a statistical method about model selection when data is the high dimension, and contribute to updating statistical software.

Suppose that k -variate explanatory variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})'$ and p -variate mutually correlated response variables $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})'$ ($i = 1, \dots, n$) are observed, where n is the sample size. A linear regression model is useful to predict \mathbf{y}_i by \mathbf{x}_i . Such a model is practically called a multivariate liner regression (MLR) model. The MLR model is one of the basic models in multivariate analysis. It is introduced in many textbooks on applied multivariate statistical analysis (see e.g., [9], Chapter 9), [14], Chapter 4], and even now it is widely applied in chemometrics, engineering, econometrics, psychometrics and other many fields for the prediction of correlated multiple responses using a set of explanatory variables (e.g., [1], [6], [7], [8] and [15]).

The n vectors of response variables $\mathbf{y}_1, \dots, \mathbf{y}_n$ and the n vectors of k explanatory variables $\mathbf{x}_1, \dots, \mathbf{x}_n$ are written in a matrix notation as an $n \times p$ matrix $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)'$ and an $n \times k$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$, respectively. Here, we assume that \mathbf{X} is of full rank, i.e., $\text{rank}(\mathbf{X}) = k$. A matrix form of the MLR model is given by

$$\mathbf{Y} \sim N_{n \times p}(\mathbf{X}\boldsymbol{\Xi}, \boldsymbol{\Sigma} \otimes \mathbf{I}_n). \quad (1)$$

Here $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$ is called the Kronecker product of $\boldsymbol{\Sigma}$ and \mathbf{I}_n and its (i, j) th block element is given by $\sigma_{ij}\mathbf{I}_n$, where σ_{ij} is the (i, j) th element of $\boldsymbol{\Sigma}$.

It is desirable to have as few explanatory variables as possible for ease of interpretation, and after all not all the k explanatory variables are needed for a good prediction. Although several methods are available, most applied researchers use, C_p statistic proposed by [4]. It is based on an estimate of the standardized version of mean square errors (MSE). Suppose we choose a subset of q explanatory variables out of k explanatory variables ($q \leq k$), i.e., we use an $n \times q$ matrix \mathbf{X}_q consisting of the q columns of \mathbf{X} for the prediction. Then the predicted value of \mathbf{Y} will be given by

$$\hat{\mathbf{Y}}_q = \mathbf{X}_q \hat{\boldsymbol{\Xi}}_q, \quad \hat{\boldsymbol{\Xi}}_q = (\mathbf{X}'_q \mathbf{X}_q)^{-1} \mathbf{X}'_q \mathbf{Y}, \quad (2)$$

and the MSE is given by

$$\begin{aligned} \text{MSE} &= E \left[\text{tr} \left\{ \boldsymbol{\Sigma}^{-1} (\mathbf{X}\boldsymbol{\Xi} - \hat{\mathbf{Y}}_q)' (\mathbf{X}\boldsymbol{\Xi} - \hat{\mathbf{Y}}_q) \right\} \right] \\ &= pq + \text{tr} \left\{ \boldsymbol{\Sigma}^{-1} \boldsymbol{\Xi}' \mathbf{X}' (\mathbf{I}_n - \mathbf{H}_q) \mathbf{X} \boldsymbol{\Xi} \right\}, \end{aligned} \quad (3)$$

where $\mathbf{H}_q = \mathbf{X}_q (\mathbf{X}'_q \mathbf{X}_q)^{-1} \mathbf{X}'_q$. When $n > p$, an unbiased estimator of this MSE is given by

$$\{1 - (p+1)/(n-k)\} \text{tr}(\mathbf{S}^{-1} \mathbf{V}_q) - np + 2kq + p(p+1),$$

(see [13]), and is called C_p statistic. Here

$$\mathbf{S} = \frac{1}{n-k} \mathbf{V}, \quad \mathbf{V} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}, \quad \mathbf{V}_q = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H}_q)\mathbf{Y}, \quad (4)$$

and $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. However, when p is close to n , the estimator \mathbf{S} is not a stable estimator of $\mathbf{\Sigma}$. And when $n < p$, the inverse matrix of \mathbf{S} does not even exist. In this case, we use \mathbf{S}^+ , the Moore-Penrose inverse of \mathbf{S} as has recently been done by [11]. The Moore-Penrose inverse of any matrix is unique and satisfies the following four conditions:

$$(i) \mathbf{S}\mathbf{S}^+\mathbf{S} = \mathbf{S}, (ii) \mathbf{S}^+\mathbf{S}\mathbf{S}^+ = \mathbf{S}^+, (iii) (\mathbf{S}\mathbf{S}^+)' = \mathbf{S}\mathbf{S}^+, (iv) (\mathbf{S}^+\mathbf{S})' = \mathbf{S}^+\mathbf{S},$$

(see [5, p. 26]). The objective of this paper is to obtain an asymptotically unbiased estimator of MSE when $n < p$, and $(n, p) \rightarrow \infty$. Such an estimator will be denoted by C_p^+ .

This paper is organized in the following ways: In Section 2, we propose the new C_p^+ when $p > n$. In Section 3 and 4, we verify performances of proposed criteria by conducting studies with numerical simulation and real data, respectively. In Section 5, we give discussions and conclusions. Technical details are provided in the Appendix.

2 C_p^+ Statistics

When $p < n$, a rough estimator of MSE in (3) is given by

$$\text{tr}(\mathbf{S}^{-1}\mathbf{V}_q) - np + 2pq,$$

where \mathbf{S} and \mathbf{V}_q are given by (4). Hence, when $p > n$, a rough estimator of MSE is defined by replacing \mathbf{S}^{-1} with \mathbf{S}^+ as

$$C_p^+ = \text{tr}(\mathbf{S}^+\mathbf{V}_q) - np + 2pq. \quad (5)$$

However, C_p^+ has a constant bias in estimating MSE. Such a bias may become large when the dimension p is large. Hence we try to remove such a bias by evaluating the bias from an asymptotic theory based on the dimension and the sample size approaching to ∞ simultaneously. Suppose that the following three conditions hold: (C.1) $0 < \lim_{p \rightarrow \infty} \text{tr}(\mathbf{\Sigma})/p (= \alpha_i) < \infty$ ($i = 1, 2$ and 4), (C.2) $n - k = O(p^\delta)$ for $0 < \delta \leq 1/2$, (C.3) the maximum eigenvalue of $\mathbf{\Sigma}$ is bounded in large p . Under the three conditions and $\mathbf{H}_q\mathbf{X} = \mathbf{X}$, the bias $\Delta = \text{MSE} - E[C_p^+]$ can be expanded as

$$\Delta = \{(n - k)^2\gamma/p - p\}q + np - (n - k)^2(1 + k\gamma/p) + o(1), \quad (6)$$

where $\gamma = \alpha_2/\alpha_1^2$ (the proof is given in the appendix). Let

$$\hat{\gamma} = \frac{p(n - k)^2}{(n - k - 1)(n - k + 2)} \left\{ \frac{\text{tr}(\mathbf{S}^2)}{(\text{tr}\mathbf{S})^2} - \frac{1}{n - k} \right\}. \quad (7)$$

The estimator $\hat{\gamma}$ is a consistent estimator of γ when the conditions (C.1), (C.2) and (C.3) are satisfied (see [10]). By using (6) and (7), we propose a new estimator of MSE as

$$C_{p,\hat{\gamma}}^+ = \text{tr}(\mathbf{S}^+\mathbf{V}_q) + \{(n - k)^2\hat{\gamma}/p + p\}q - (n - k)^2(1 + k\hat{\gamma}/p). \quad (8)$$

3 A Simulation Study

In the previous section, we avoid the nonexistence of the matrix to standardize \mathbf{V}_q by using the Moore-Penrose inverse. However, we can avoid the singularity by another way if we allow model misspecification. Another choice is to omit correlations between \mathbf{y}_i tentatively, namely, we use $\mathbf{S}_{(d)} = \text{diag}(s_1, \dots, s_p)$ to standardize \mathbf{V}_q , where s_i ($i = 1, \dots, p$) is the i th diagonal element of \mathbf{S} . This has been done by [2] in discriminant analysis and by [12] in testing the equality of two mean values. Thus, we can also define the estimator of MSE as

$$C_p^{(d)} = \text{tr}(\mathbf{S}_{(d)}^{-1} \mathbf{V}_q) - np + 2pq. \quad (9)$$

The effect of correlations between \mathbf{y}_i to model selection can be studied by comparing with proposed two C_p^+ and $C_p^{(d)}$.

We evaluate the proposed C_p statistics applied numerically to the regression model in (II) with $n = 20$, $k = 8$ and $p = 30$ and 100. Here, we assume that $\mathbf{\Sigma} = \text{diag}(\psi_1, \dots, \psi_p) \mathbf{\Phi} \text{diag}(\psi_1, \dots, \psi_p)$. In this numerical study, we chose \mathbf{X} : the first column vector was $\mathbf{1}_n$ and the others were generated from $U(-1, 1)$, $\mathbf{\Xi}$: the first, second, third and fourth rows were $-\tau(1 - a_j)$, $\tau(1 + a_j)$, $-\tau(2 - a_j)$ and $\tau(1 + a_j)$ ($j = 1, \dots, p$), respectively, and the others are 0, ψ_j : $\psi_j = 2 + a_j$ ($j = 1, \dots, p$), $\mathbf{\Phi}$: the (i, j) th element is $\rho^{|i-j|^{1/\tau}}$ ($i = 1, \dots, p; j = 1, \dots, p$), where $\mathbf{1}_n$ is an n -dimensional vector of ones and $a_j = (p - j + 1)/p$. Let M_j denote the j th candidate model with the matrix of explanatory variables, which is consisting of the first j columns of \mathbf{X} ($j = 1, \dots, k$). This means that the candidate models are nested. Moreover, we chose $\tau = 0.0$ or 8.0. It means that there are two types of true model, i.e., the true model is M_1 ($\tau = 0.0$) and M_4 ($\tau = 8.0$), respectively.

We compared C_p^+ , $C_{p, \hat{\gamma}}^+$ and $C_p^{(d)}$ with respect to the following two properties: (i) the selection probability of the model chosen by minimizing the criterion, (ii) the true MSE of the predicted values of the best model chosen by minimizing the criterion, which is defined by

$$\text{MSE}_B = \frac{1}{np} E \left[\text{tr} \left\{ \mathbf{\Sigma}^{-1} (\mathbf{X} \mathbf{\Xi} - \hat{\mathbf{Y}}_B)' (\mathbf{X} \mathbf{\Xi} - \hat{\mathbf{Y}}_B) \right\} \right], \quad (10)$$

where $\hat{\mathbf{Y}}_B$ is the predictor of \mathbf{Y} based on the best model chosen by each C_p . Since the prediction error has to be measured by the same measurement as the goodness of fit of the model, the prediction error of the best model has to be defined by (10).

These two properties were evaluated by the Monte Carlo simulation with 1,000 iterations. Since $\mathbf{\Xi}$ and $\mathbf{\Sigma}$ are known in the simulation study, we can evaluate MSE_B by the Monte Carlo simulation. Tables 1 and 2 show obtained properties (i) and (ii), respectively. From tables, we can see that $C_{p, \hat{\gamma}}^+$ had good performances in all cases. Performances of C_p^+ were also good, however, these became bad when $\tau = 8.0$, $\rho = 0.8$ and $p = 100$. Furthermore, we can see that performances of $C_p^{(d)}$ were not too bad when ρ is low. However, when ρ is

Table 1. Selection probabilities of three C_p statistics

	$\tau = 0.0$				$\tau = 8.0$			
	$\rho = 0.2$		$\rho = 0.8$		$\rho = 0.2$		$\rho = 0.8$	
	$p = 30$	$p = 100$	$p = 30$	$p = 100$	$p = 30$	$p = 100$	$p = 30$	$p = 100$
C_p^+	100.0	100.0	100.0	100.0	99.80	100.0	99.90	89.00
$C_{p,\hat{\gamma}}^+$	100.0	100.0	98.90	100.0	99.80	100.0	99.30	99.90
$C_p^{(d)}$	98.70	99.90	70.90	74.20	98.70	99.80	73.80	76.30

Table 2. True MSEs of three C_p statistics

	$\tau = 0.0$				$\tau = 8.0$			
	$\rho = 0.2$		$\rho = 0.8$		$\rho = 0.2$		$\rho = 0.8$	
	$p = 30$	$p = 100$	$p = 30$	$p = 100$	$p = 30$	$p = 100$	$p = 30$	$p = 100$
C_p^+	0.050	0.050	0.050	0.050	0.200	0.200	0.200	0.210
$C_{p,\hat{\gamma}}^+$	0.050	0.050	0.050	0.050	0.200	0.200	0.200	0.201
$C_p^{(d)}$	0.051	0.050	0.095	0.085	0.201	0.200	0.227	0.231

high, the performances became bad. This result means that we should consider correlations to evaluate the goodness of fit of a statistical model correctly if response variables are not independent. We have studied several other settings for simulation, and have obtained similar results.

4 An Example Study

We show an example of model selection by using real data in [3]. This data gives 21 body dimension measurements in cm such as biacromial diameter, biiliac diameter or pelvic breadth, bitrochanteric diameter, chest depth, chest diameter, elbow diameter, wrist diameter, knee diameter, ankle diameter, shoulder girth, chest girth, waist girth, navel girth, hip girth, thigh girth, bicep girth, forearm girth, knee girth, calf girth, ankle girth, wrist girth. The data also gives age in years, weight in kg, height in cm, gender. Observations are 507 individuals in their 20s and 30s. We applied multivariate linear regression to see performances of C_p^+ , $C_{p,\hat{\gamma}}^+$ and $C_p^{(d)}$. Response variables were 21 body dimension measurements and 4 explanatory variables were age, weight, height, gender taking 1 for males and 0 for females. A best model was selected from 16 models having different combinations of 4 explanatory variables by C_p^+ , $C_{p,\hat{\gamma}}^+$ or $C_p^{(d)}$.

We divided data to three samples, 10(= $n_{(1)}$), 10(= $n_{(2)}$) and 487(= $n_{(3)}$), randomly, and repeated such division 1,000 times. Divided samples were denoted by $(\mathbf{Y}_{(1)}, \mathbf{X}_{(1)})$, $(\mathbf{Y}_{(2)}, \mathbf{X}_{(2)})$ and $(\mathbf{Y}_{(3)}, \mathbf{X}_{(3)})$, respectively. To calculate the mean squared error M , we used $(\mathbf{Y}_{(3)}, \mathbf{X}_{(3)})$ to estimate the covariance matrix by $\hat{\Sigma}_{(3)} = \mathbf{Y}_{(3)}' \{ \mathbf{I}_{n_{(3)}} - \mathbf{X}_{(3)}(\mathbf{X}_{(3)}' \mathbf{X}_{(3)})^{-1} \mathbf{X}_{(3)}' \} \mathbf{Y}_{(3)} / (n_{(3)} - 5)$. We used $(\mathbf{Y}_{(1)}, \mathbf{X}_{(1)})$ for the model selection by C_p^+ , $C_{p,\hat{\gamma}}^+$ and $C_p^{(d)}$. We also used $(\mathbf{Y}_{(1)}, \mathbf{X}_{(1)})$ for estimating the regression parameters Ξ . Thus, the predicted value of $\mathbf{Y}_{(2)}$ is

Table 3. Results of real data

Variables	Frequencies			Variables	Frequencies		
	C_p^+	$C_{p,\hat{\gamma}}^+$	$C_p^{(d)}$		C_p^+	$C_{p,\hat{\gamma}}^+$	$C_p^{(d)}$
{}	47	7	0	{2, 3}	19	63	35
{1}	0	2	0	{2, 4}	82	237	233
{2}	781	567	11	{3, 4}	0	0	0
{3}	2	1	0	{1, 2, 3}	0	3	34
{4}	56	43	0	{1, 2, 4}	1	14	204
{1, 2}	10	50	15	{1, 3, 4}	0	0	3
{1, 3}	0	1	0	{2, 3, 4}	1	7	251
{1, 4}	1	3	0	{1, 2, 3, 4}	0	2	214
				MSE	0.753	0.695	0.948

given by $\hat{\mathbf{Y}}_{(2)B} = \mathbf{X}_{(2)B}(\mathbf{X}'_{(1)B}\mathbf{X}_{(1)B})^{-1}\mathbf{X}'_{(1)B}\mathbf{Y}_{(1)}$, where $\mathbf{X}_{(1)B}$ and $\mathbf{X}_{(2)B}$ are matrices of best explanatory variables chosen by C_p^+ , $C_{p,\hat{\gamma}}^+$ and $C_p^{(d)}$. Thus, M is given by

$$M = \frac{1}{n_{(1)}p} \text{tr} \left\{ \hat{\Sigma}_{(3)}^{-1} \left(\mathbf{Y}_{(2)} - \hat{\mathbf{Y}}_{(2)B} \right)' \left(\mathbf{Y}_{(2)} - \hat{\mathbf{Y}}_{(2)B} \right) \right\} - 1,$$

We may regard that sample average of M in 1,000 repetitions is the MSE of the best model.

Results of calculations are in Table 3. ‘‘Variables’’ shows 16 models and numbers in brace are used explanatory variables. ‘‘1’’, ‘‘2’’, ‘‘3’’, and ‘‘4’’ mean age, weight, height, and gender, respectively. All models contain constant terms, therefore model {} has only constant term. ‘‘Frequencies’’ shows the number of times that the model was selected as the best model in 1,000 iterations. C_p^+ selected the model {2} 781 times out of 1,000 repetitions. The model {2} was also selected frequently by $C_{p,\hat{\gamma}}^+$, however $C_{p,\hat{\gamma}}^+$ selected the model {2,4} 237 times which was more than C_p^+ . The result of $C_p^{(d)}$ was different from those of C_p^+ and $C_{p,\hat{\gamma}}^+$. The $C_p^{(d)}$ frequently selected models with many explanatory variables such as models {1,2,4}, {2,3,4}, and {1,2,3,4} in 204, 251, and 214 times, respectively. The $C_p^{(d)}$ did not consider correlations between $\mathbf{Y}_{(1)}$. Thus the result indicates importance of considering the correlations. In ‘‘MSE’’, $C_{p,\hat{\gamma}}^+$ was 0.695 and the smallest among 3 statistics. From this, we understood that a performance of $C_{p,\hat{\gamma}}^+$ was better than those of C_p^+ and $C_p^{(d)}$.

5 Conclusion and Discussion

In this paper, we proposed new three C_p statistics for selecting variables in the multivariate linear model with $p > n$. These are defined by replacing \mathbf{S}^{-1} with \mathbf{S}^+ , and $C_{p,\hat{\gamma}}^+$ is constructed by adding renewal bias correction terms evaluated from an asymptotic theory, which is based on $p \rightarrow \infty$ and $n \rightarrow \infty$ simultaneously.

A simulation shows that performances of C_p^+ and $C_{p,\hat{\gamma}}^+$ were better than those of $C_p^{(d)}$. Especially, in all cases, $C_{p,\hat{\gamma}}^+$ had good performance. An example of model selection using the real data shows that the importance of considering correlations between response variables and the performance of $C_{p,\hat{\gamma}}^+$ is better than C_p^+ and $C_p^{(d)}$. Hence, we recommend the use of $C_{p,\hat{\gamma}}^+$ for selecting variables in multivariate linear regression model with $p > n$. $C_{p,\hat{\gamma}}^+$ could help to update statistical software in high dimensional data analysis.

References

1. van Dien, S.J., Iwatani, S., Usuda, Y., Matsui, K.: Theoretical analysis of amino acid-producing *Escherichia coli* using a stoichiometric model and multivariate linear regression. *J. Biosci. Bioeng.* 102, 34–40 (2006)
2. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* 97, 77–87 (2002)
3. Grete, H., Louis, J.P., Roger, W.J., Carter, J.K.: Exploring relationships in body dimensions. *J. Statist. Educ.* 11 (2003)
4. Mallows, C.L.: Some comments on C_p . *Technometrics* 15, 661–675 (1973)
5. Rao, C.R.: *Linear Statistical Inference and Its Applications* (Paper back ed). John Wiley & Sons, New York (2002)
6. Sârbu, C., Onișor, C., Posa, M., Kevresan, S., Kuhajda, K.: Modeling and prediction (correction) of partition coefficients of bile acids and their derivatives by multivariate regression methods. *Talanta* 75, 651–657 (2008)
7. Saxén, R., Sundell, J.: ^{137}Cs in freshwater fish in Finland since 1986 – a statistical analysis with multivariate linear regression models. *J. Environ. Radioactiv.* 87, 62–76 (2006)
8. Skagerberg, B., Macgregor, J.F., Kiparissides, C.: Multivariate data analysis applied to low-density polyethylene reactors. *Chemometr. Intell. Lab.* 14, 341–356 (1992)
9. Srivastava, M.S.: *Methods of Multivariate Statistics*. John Wiley & Sons, New York (2002)
10. Srivastava, M.S.: Some tests concerning the covariance matrix in high dimensional data. *J. Japan Statist. Soc.* 35, 251–272 (2005)
11. Srivastava, M.S.: Multivariate theory for analyzing high dimensional data. *J. Japan Statist. Soc.* 37, 53–86 (2007)
12. Srivastava, M.S., Du, M.: A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* 99, 386–402 (2008)
13. Srivastava, M.S., Kubokawa, T.: Selection of variables in multivariate regression models for large dimensions. In: *CIRJE Discussion Papers CIRJE-F-709*, University of Tokyo, Japan (2010)
14. Timm, N.H.: *Applied Multivariate Analysis*. Springer, New York (2002)
15. Yoshimoto, A., Yanagihara, H., Ninomiya, Y.: Finding factors affecting a forest stand growth through multivariate linear modeling. *J. Jpn. For. Soc.* 87, 504–512 (2005) (in Japanese)

Appendix

Under the assumption that $\mathbf{H}_q \mathbf{X} = \mathbf{X}$, $\text{MSE} = pq$ holds. Hence, the bias of C_p^+ for MSE is rewritten as

$$\Delta = (n - q)p - E[\text{tr}(\mathbf{S}^+ \mathbf{V}_q)]. \quad (11)$$

Note that $\mathbf{V}_q = \mathbf{V} + \mathbf{Y}'(\mathbf{H} - \mathbf{H}_q)\mathbf{Y}$, $\mathbf{S}^+ = (n - k)\mathbf{V}^+$ and $\text{tr}(\mathbf{V}^+ \mathbf{V}) = n - k$. Hence, we derive

$$\text{tr}(\mathbf{S}^+ \mathbf{V}_q) = (n - k)^2 + (n - k)\text{tr}\{\mathbf{V}^+ \mathbf{Y}'(\mathbf{H} - \mathbf{H}_q)\mathbf{Y}\}.$$

Since \mathbf{V}^+ and $\mathbf{Y}'(\mathbf{H} - \mathbf{H}_q)\mathbf{Y}$ are mutually independent, and $E[\mathbf{Y}'(\mathbf{H} - \mathbf{H}_q)\mathbf{Y}] = (k - q)\boldsymbol{\Sigma}$ holds. Then, the expectation of $\text{tr}(\mathbf{S}^+ \mathbf{V}_q)$ is expressed as

$$E[\text{tr}(\mathbf{S}^+ \mathbf{V}_q)] = (n - k)^2 + (n - k)(k - q)E_{\mathbf{Y}}^*[\text{tr}(\mathbf{V}^+ \boldsymbol{\Sigma})]. \quad (12)$$

Let \mathbf{L} be a $(n - k) \times (n - k)$ diagonal matrix $\mathbf{L} = \text{diag}(\ell_1, \dots, \ell_{n-k})$, where $\ell_1, \dots, \ell_{n-k}$ are positive eigenvalues of \mathbf{V} , and $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ be a $p \times p$ orthogonal matrix such that $\mathbf{Q}'\mathbf{V}\mathbf{Q} = \text{diag}(\mathbf{L}, \mathbf{O}_{p-n+k})$, where \mathbf{Q}_1 and \mathbf{Q}_2 are $p \times (n - k)$ and $p \times (p - n + k)$ matrices, respectively, and \mathbf{O}_{p-n+k} is a $(p - n + k) \times (p - n + k)$ matrix of zeros. Note that

$$\mathbf{V}^+ = \mathbf{Q}\text{diag}(\mathbf{L}^{-1}, \mathbf{O}_{p-n+k})\mathbf{Q}' = \mathbf{Q}_1\mathbf{L}^{-1}\mathbf{Q}_1'. \quad (13)$$

By using (13) and applying the simple transformation, we derive

$$\begin{aligned} E[\text{tr}(\mathbf{V}^+ \boldsymbol{\Sigma})] &= E[\text{tr}(\mathbf{L}^{-1}\mathbf{Q}_1'\boldsymbol{\Sigma}\mathbf{Q}_1)] \\ &= \frac{\alpha_2}{p\alpha_1^2}E\left[\text{tr}\left\{\left(\mathbf{L}/(p\alpha)\right)^{-1}(\alpha_1/\alpha_2)\mathbf{Q}_1'\boldsymbol{\Sigma}\mathbf{Q}_1\right\}\right]. \end{aligned} \quad (14)$$

Since \mathbf{V} is distributed according to the Wishart distribution, from (11), we obtain $\lim_{p \rightarrow \infty} \mathbf{L}/(p\alpha_1) = \mathbf{I}_{n-k}$ and $\lim_{p \rightarrow \infty} (\alpha_1/\alpha_2)\mathbf{Q}_1'\boldsymbol{\Sigma}\mathbf{Q}_1 = \mathbf{I}_{n-k}$ in probability. Therefore, we have $E[\text{tr}(\mathbf{L}^{-1}\mathbf{Q}_1'\boldsymbol{\Sigma}\mathbf{Q}_1)] = (n - k)\gamma/p + o(p^{-1+\delta})$. By combining this result and (14), (12) is expanded as

$$E[\text{tr}(\mathbf{S}^+ \mathbf{V}_q)] = (n - k)^2 + \frac{1}{p}(n - k)^2(k - q)\gamma + o(1). \quad (15)$$

Finally, substituting (15) into (11) yields (6).

Customer Path Controlling in the Retail Store with the Vertex Dominating Cycle Algorithms

Takeshi Sugiyama

Graduate School of Business Sciences, University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
sugiyama@gssm.otsuka.tsukuba.ac.jp

Abstract. In this paper, we propose the customers' path control method with the dominating cycle algorithms. It is one of the important issues to control the customers' path among the retail businesses. Even for the customer, it is natural to minimize route in the store to find and purchase the objectives. On the other hand, the store operator would like to maximize the route on the contrary. According to the conventional store management theory, the items with the volume purchased set in the isle with the lots of traffic. Whereas concerning to the research for the traffic controlling with the geometrical theory in the store is less. Therefore, in this paper, we regard customers' paths as the structure of graph and make use of the characteristics for the vertex dominating cycle. As the result, we find out the route of the saddle point for both customer and the store.

Keywords: Graph, Vertex Dominating cycle, Minimum Degree Condition, Degree Sum Condition, Customers' Path.

1 Introduction

It is difficult to control the walking paths in the store with the intentions of the store operator, some of the effective methods of controlling the paths is put the items that are price valuables or in fashion. And as for the structure of the store layout, it is not sufficient to analyze geometrically, even though the item location data stored in the transaction system such as POS (Point-of-Sales) system are widely adopted among the retail industries.

The graph is an effective method to analyze the structure of the object. For example, we can translate the walking path of the customer in the store for the graph structure. On the other hand, the dominating problem has long been fundamental in graph theory. Among many sufficient conditions for a graph to contain dominating cycle, the following sufficient condition is well-known such as the following theorems.

Theorem A. Let G be a 2-connected graph of order n . If for any x (an element of $V(G)$), $d_G(x) \geq (|G|+2)/3$, then every longest cycle in G is edge dominating [1].

A cycle C in a graph G is called edge dominating if $E(G-H)$ is an emptyset, and vertex dominating if each vertex u in $V(G-C)$ has a neighbor in H .

On the other hand, compared with an edge dominating cycle, there are a few results on a vertex dominating cycle. Among them, we introduce two results. The first one was proved by Bondy and Fan. A vertex set S is called r -stable if $\text{dist}(x,y) \geq r$ for every x,y in S , where $\text{dist}(x,y) := \min\{|E(P)|: P \text{ is an } xy\text{-path}\}$.

Theorem B. Let G be a k -connected graph on n vertices. Suppose that for any 3-stable set S of order $k+1$, we have $d_G(x) \geq (n-2k)/(k+1)$. Then there exists a vertex dominating cycle [2].

In [2], the length of a vertex dominating cycle is not taken into account. In [3], this paper consider the existence of a longest cycle which is vertex dominating. In Section 3, we show the following theorems' algorithm. The following theorem shows the existence of a vertex dominating cycle.

Theorem C. Let G be a 2-connected graph on n vertices with n . If for every x , $d_G(x) \geq (n-4)/3$, then every longest cycle in G is vertex dominating (except for some specified graphs) [3].

In Section 4, by using the above algorithm, we propose the customers' path control method with the vertex dominating cycle algorithms.

2 Related Works; Customer Paths

The walking distances for the customer in the store are called the length for the customer paths. The longer for the customers' paths the more increased the sales. As for the management for the customers' paths are usually influenced by the gross sales that closely related in the display position for the items. Therefore, it is usually displayed the location that the customer mostly passes called the main passage. As the length of the customers' paths, it is closely related with the purchase for non-planning such as the decision making in the store [4] [5]. Therefore, it is important for making the saddle points both of the sales policy and the customers' needs to control the customers' paths as long as possible for maximize the sales.

As for the research for the customers' paths, Larson *et al.* analyzed of an extraordinary new dataset that reveals the path taken by individual shoppers in an actual grocery store and used data collected from a grocery store where shopping carts were equipped with RFID tags and tracked throughout the store's aisles [6]. This paper also offers enlightening conclusions about shopping patterns, some of which fly in the face of conventional wisdom. For example, shoppers were historically thought to weave up and down aisles, starting at one side of the store and ending at the other. The new research states, however, that shoppers only visit those aisles that interest them. Furthermore, they do not go down the entire aisle; instead they enter the aisle, select their item, and reverse direction back out of the aisle to the perimeter of the store.

Fig. 1 indicates the store layout example from a certain retail store in Japan. From the layout, we can figure out the intentions of the store operator that induces the customers walk along with the wall side, from the display of the items with the high performance profit, such as Vegetables, Fish, and Meat. On the other hands, the customers intend to visit the site that needs with the shortest route. Therefore, there need to the solution for the saddle point both of them.

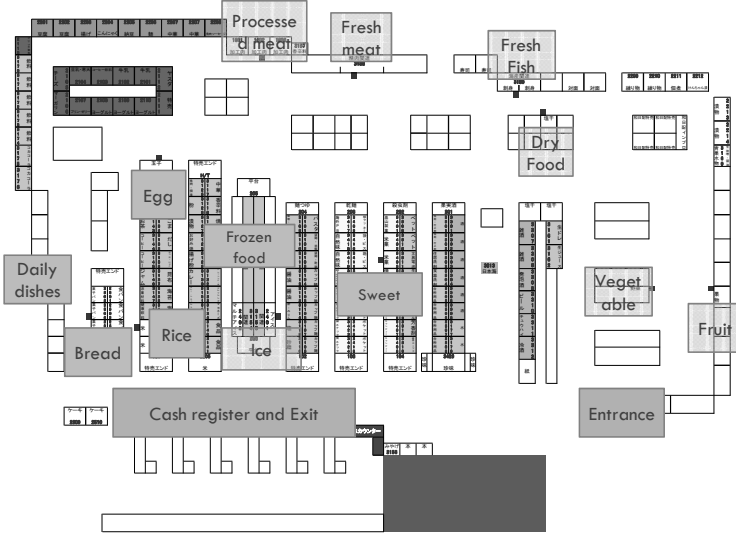


Fig. 1. Store Layout Examples

3 Algorithms

3.1 Graph-Theoretic Terminology

For graph-theoretic terminology not explained in this paper, we refer the reader to [7]. Let u be an element of $V(G)$. A neighborhood of u is denoted by $N_G(u)$. For U which is a subset of $V(G)$, we define $N_G(U)$ by $N_G(U) := \bigcup_{u \in U} N_G(u)$. For a subgraph H of G , we write $N_H(u)$, $N_H(U)$ and $d_H(u)$ instead of $N_G(u) \cap V(H)$, $N_G(U) \cap V(H)$ and $|N_H(u)|$, respectively. We write $N_G(H)$ instead of $N_G(V(H))$. A path joining u and v is called a uv -path. For a subgraph H , a uv -path P is called an H -path if $V(P) \cap V(H) = \{u, v\}$ and $E(P) \cap E(H) = \emptyset$. Let C be a cycle in G .

We give an orientation to C and write the oriented cycle C by C^{\rightarrow} . For $x, y \in V(C)$, we denote an xy -path along C^{\rightarrow} by $x C^{\rightarrow} y$, and write the reverse sequence of $x C^{\rightarrow} y$ by $y C^{\leftarrow} x$. For $x \in V(C)$, we denote the h -th successor and the h -th predecessor of x on C^{\rightarrow} by x^{+h} and x^{-h} , respectively. For X which is a subset of $V(C)$, we define $X^{+h} := \{x^{+h} : x \in X\}$ and $X^{-h} := \{x^{-h} : x \in X\}$. We often write x^+ , x^- , X^+ and X^- for x^{+1} , x^{-1} , X^{+1} and X^{-1} , respectively. For a path P , we sometimes give an orientation to P and at that time define the same terminology as we did for a cycle C .

3.2 Theorems

We introduce the following theorems and this theorem is used for the linear time algorithm.

Theorem 1. Let G be a 2-connected graph on n vertices with n . If for every x , $d_G(G) \geq (n-4)/3$, then every longest cycle in G is vertex dominating (except for some specified graphs) [3].

By using this algorithm, we can find a long vertex dominating cycle in linear times.

For the proof of Theorem 1, we use the following Theorems and Lemma. The following theorems and Lemma use the linear time algorithm, too.

Theorem E. Suppose that G is a 2-connected graph. Then there exists a cycle with length at least $\min\{2d, |V(G)|\}$ [8].

Theorem F. Suppose that G is a 3-connected graph on n vertices. If for any vertex $v \in V(G)$, $d_G(x) \geq (n+3)/4$, then every longest cycle in G is vertex dominating [3].

Theorem G. Let G be a 2-connected graph on n vertices. If $\delta(G) \geq (n+1)/2$, then G is hamilton-connected [9].

Lemma H. Let G be a 2-connected graph on n vertices with $n \geq 27$ and let $u, v \in V(G)$ such that $G - \{u, v\}$ is connected. If $d_G(x) \geq (n-5)/2$ for every $x \in V(G) - \{u, v\}$, then every longest uv -path is dominating or G is a spanning subgraph of G_1 .

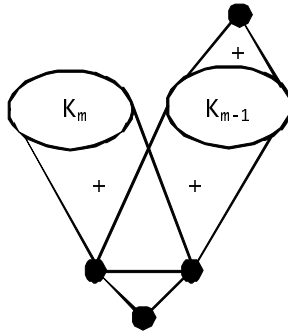


Fig. 2. The graph G_1

3.3 Outline of Proof of Theorem 1

Suppose that G is 3-connected. Since $n \geq 52$, we have $(n-4)/3 \geq (n+3)$. Then by Theorem E, every longest cycle is dominating. Thus, we may assume $\kappa(G) = 2$. Let $S := \{u, v\}$ subset be a 2-cut set of G .

Claim 1. $\omega(G-S) = 2$.

Proof of Claim 1

Suppose that $\omega(G-S) \geq 3$ and let D_1, D_2, \dots, D_j ($j \geq 3$) be components of $G-S$ so that $|V(D_1)| \leq |V(D_2)| \leq \dots \leq |V(D_j)|$. For each $1 \leq i \leq j$ and $x_i \in V(D_i)$, $N_G(x_i)$ is subset of $V(D_i) - \{x_i\} \cup S$ and hence $|V(D_i)| \geq d_G(x_i) - |S| + 1 \geq (n-7)/3$. If $j > 3$, by above condition, $|V(G)| \geq j|V(D_1)| + |S| = j\{(n-7)/3\} + 2 \geq 4n/3 - 5 \geq n$. We obtain $n \leq 15$, a contradiction. We have $j = 3$. Therefore, $|V(D_i)| = n - |S| - (\sum_{j \neq i} |V(D_j)|) \leq (n+8)/3$ and

$$\delta(D_i) \geq \delta(G) - |S| \geq (n-10)/3 > (n+11)/6 \geq (|V(D_i)| + 1)/2.$$

Hence by Theorem G, each D_i is hamilton-connected for $1 \leq i \leq j$. If $|V(D_1)| \geq (n-1)/3$, then $n \geq 2 + \sum_{i=1,2,3} |V(D_i)| \geq n+1$, a contradiction. Moreover, if there exists $x_1 \in V(D_1)$

such that $N_S(x_1)$ is an empty set then $|V(D_1)| \geq d_G(x_1) + 1 \geq (n-1)/3$. Hence $|V(D_1)| < (n-1)/3$ and $N_S(x)$ is not an empty set for every $x \in V(D_1)$.

Suppose that $|V(D_2)| > |V(D_1)|$. Then for every longest cycle C in G , we have $V(C) = S \cup V(D_2 \cup D_3)$ because D_2 and D_3 are hamilton-connected. Since $N_S(x)$ is not an empty set for every $x \in V(D_1)$, then C is a dominating cycle.

Suppose that $|V(D_i)| = |V(D_1)|$ for $i=2$ or $i=3$. Then by the same argument as above, $N_S(x)$ is not an empty set for every $x \in V(D_i)$. Therefore every longest cycle C in G passes u and v , and hence C is a dominating cycle, again.

By Claim 1, we have $\omega(G-S) = 2$ for every 2-cut set S . Let D_1 and D_2 be components of $G-S$ with $|V(D_1)| \leq |V(D_2)|$. By the assumption,

$$(n-7)/3 \leq |V(D_1)| \leq (n-|S|)/2 = (n-2)/2. \quad (1)$$

Then,

$$\delta(D_1) \geq \delta(G) - |S| \geq (n-10)/3 \geq n/4 \geq (|V(D_1)| + 1)/2.$$

Hence by Theorem G, D_1 is hamilton-connected.

Let C be a longest cycle in G . By Theorem E, we may assume that $|V(C)| \geq 2\delta(G) \geq 2(n-4)/3$. Therefore, $V(C) \cap V(D_2)$ is not an empty set. We consider the following three cases.

Case 1. $V(C) \cap S$ is an empty set.

Since G is 2-connected, there are two disjoint paths such that one joins u and w_1 and the other joins v and w_2 , where $w_1, w_2 \in V(C)$ and no vertices are contained in C except for w_1 and w_2 . By (1), $|V(w_1^+C^-w_2^-)|, |V(w_2^+C^-w_1^-)| \geq |V(D_1) \cup S| \geq (n-7)/3 + 2 = (n-1)/3$. Hence $n \geq |V(D_1)| + |S| + |V(w_1^+C^-w_2^-)| + |V(w_2^+C^-w_1^-)| + |\{w_1, w_2\}| \geq (n-7)/3 + 2 + 2(n-1)/3 + 2 = n + 1$, a contradiction.

Case 2. $|V(C) \cap S| = 1$.

Without loss of generality, we may assume that u is an element of $V(C)$ and v is not an element of $V(C)$. By $|V(D_1)| \geq (n-7)/3$, we have $|V(D_2)| \leq n - (n-7)/3 - 2 = 2(n+1)/3$. On the other hand $\delta(D_2 \cup \{u\}) \geq (n-10)/3$ and $|V(C)| \geq 2(n-4)/3$ are hold. Hence C dominates $V(D_2)$. Suppose that $N_C(v)$ is an empty set. Since G is 2-connected, there is a vw -path P such that u is not an element of $V(P)$, where $w \in V(C)$. Since $N_C(v)$ is an empty set, $|V(P)| \geq 3$. Hence $|V(u^+C^-w^-)|, |V(w^+C^-u^-)| \geq |V(D_1) \cup V(P) - \{w\}| \geq (n-1)/3$. Then $n \geq |V(D_1)| + |V(P)| + |V(u^+C^-w^-)| + |V(w^+C^-u^-)| + |\{u\}| \geq (n-7)/3 + 3 + 2(n-1)/3 + 1 = n + 1$, a contradiction. Therefore, $N_C(v)$ is not an empty set and let w' is an element of $N_C(v)$.

If $xu \in E(G)$ hold for every $x \in V(D_1)$, then C is a dominating cycle. So we may assume that there exists $x_1 \in V(D_1)$ such that x_1u is not an element of $E(G)$. By the assumption, $|V(D_1)| \geq d_G(x_1) - 1 + 1 \geq (n-4)/3$. Since C is longest, $|V(u^+C^-w^-)|, |V(w^+C^-u^-)| \geq |V(D_1) \cup \{v\}| \geq (n-1)/3$. This implies that $n = |V(D_1)| + |\{v\}| + |V(u^+C^-w^-)| + |V(w^+C^-u^-)| + |\{u, w'\}| \geq (n-4)/3 + 1 + 2(n-1)/3 + 2 = n + 1$, a contradiction.

Case 3. u, v are elements of $V(C)$.

Suppose that $V(C) \cap V(D_1)$ is an empty set. If there exists $x_2 \in V(D_2)$ such that $N_C(x_2)$ is an empty set, then $|V(D_2)| \geq |V(C) - S| + |N_G(x_2)| + |\{x_2\}| \geq 2(n-4)/3 - 2 + (n-4)/3 + 1 = n - 5$, and hence $n = |S| + |V(D_1)| + |V(D_2)| \geq 2 + (n-7)/3 + n - 5 = 4(n-16)/3$,

contradicting $n \geq 65$. Thus, C is a dominating cycle in $G[V(D_2) \cup S]$. Moreover, if $N_S(x)$ is not an empty set for every $x \in V(D_1)$, then C is a dominating cycle. Therefore, we may assume that there exists $x_1 \in V(D_1)$ such that $N_S(x_1)$ is an empty set. By the degree condition, $|V(D_1)| \geq d_G(x_1) + 1 \geq (n-1)/3$ and hence $n \geq |V(D_1)| + |S| + |V(u^+C^-v^-)| + |V(v^+C^-u^-)| \geq (n-1)/3 + 2 + 2(n-1)/3 = n + 1$, a contradiction.

Then we may assume that $V(C) \cap V(D_1)$ is not an empty set. Since D_1 is hamilton-connected, $V(D_1)$ is a subset of $V(C)$. Let $H := G[V(D_2) \cap S]$. Clearly $H - S$ is connected. On the other hand, by (1), $|V(H)| = n - |V(D_1)| \leq (2n + 7)/3$. For every $w \in V(D_2)$, $d_H(w) = d_G(w) \geq (n-4)/3 = \{(2n + 7)/3 - 5\}/2 \geq (|V(H)| - 5)/2$. Moreover, by (4) and $n \geq 52$, $|V(H)| \geq n - (n-2)/2 = n/2 + 1 \geq 27$. Therefore, by Lemma H, every longest uv -path in H is dominating or H is isomorphic to a spanning subgraph of G_1 . But if H is isomorphic to a spanning subgraph of G_1 , then G has a 2-cut set S with $\omega(G-S) = 3$, contradicting Claim 1. Thus, $C \cap H$ is a dominating path and then, C is a dominating cycle.

4 Application for the Algorithm

This chapter, we apply our proposed algorithms to the following sample graph in Fig.3 for the evaluation. This graph is for the demonstrations of the customers paths.

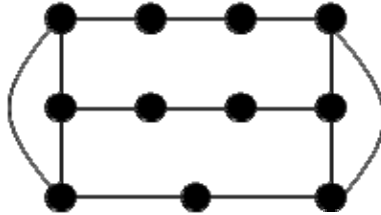


Fig. 3. Example Graphs G_A

The graph G_A has minimum degree $2 \geq (10-4)/3$. Hence we can find a vertex dominating cycle which is also longest cycle in G_B (see Fig. 4).

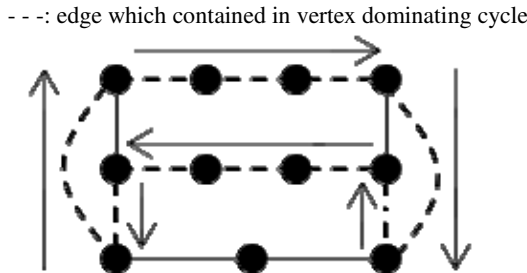


Fig. 4. A longest vertex dominating cycle in the graph G_A

We give the graph G_B (see Fig. 5). The graph G_B has minimum degree $2 < (11-4)/3$. Hence we can not find a vertex dominating cycle in G_B by using above algorithm.

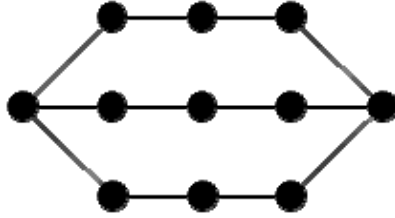


Fig. 5. The graph G_B

However, the above graph can be satisfying the degree condition in Theorem 1 by adding edges. For example, let G'_B be the graph which obtain form G_B by adding some edges (see Fig. 6). The graph G'_B has minimum degree $3 \geq (11-4)/3$. We can obtain a longest vertex dominating cycle in G'_B (see Fig. 7).

-

Fig. 6. The graph G'_B

- - -: edge which contained in vertex dominating cycle

Fig. 7. A longest vertex dominating cycle in the graph G'_B

The above example in Fig. 4 and 7 shows the relations of customers' paths (vertex dominating cycles in graph) and the structure of the store. We regard the edge of the graph as the passage of the store. By using the above algorithm, we can suggest the structure of the store which contains the controlled roots.

5 Conclusion

In this paper, a basic research for the customers' path controlling with the vertex dominating Cycles is described. From the results of the evaluation for the proposed algorithms, finding out the possibilities for controlling the customers passes that be able to optimize the both customers and the store operator needs in the store. As the result of the evaluations, we confirm the possibility for satisfying both of them, with the combinations of the concepts of this proposed algorithms and the customer preferences extraction methods.

This research was with the courtesy support of the local super market in Japan. We express appreciation to those involved.

References

1. Bondy, J.A.: Longest paths and cycles in graphs with high degree, Research Report CORR, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada (1980)
2. Bondy, J.A., Fan, G.: A sufficient condition for dominating cycles. *Discrete Math.* 67, 205–208 (1987)
3. Matsumura, H., Ozeki, K., Sugiyama, T.: A Note on a Longest Cycle which is Vertex Dominating. *AKCE International Journal of Graphs and Combinatorics* 4, 233–243 (2007)
4. Yada, K.: Path Analysis in a Supermarket and String Analysis Technique. *Proc. Of the Institute of Statistical Mathematics* 56(2), 199–213 (2008)
5. Yamada, K., Abe, T., Kimura, H.: Simulation of Planned/Unplanned Purchaser Flow using an Agent Model. In: *The 19th Annual Conference of the Japanese Society for Artificial Intelligence*, 3E3-02 (2005)
6. Larson, J.S., Bradlow, E.T., Fader, P.S.: An Exploratory Look at Supermarket Shopping Paths. *International Journal of Research in Marketing* 22, 395–414 (2005)
7. Diestel, R.: *Graph Theory*, 3rd edn. Springer, Heidelberg (2006)
8. Dirac, G.A.: Some theorems on abstract graphs. *Proc. London Math. Soc.* 2, 69–81 (1952)
9. Ore, O.: Hamilton connected graphs. *J. Math. Pures Appl.* 42, 21–27 (1963)

A Framework for the Quality Assurance of Blended E-Learning Communities

Iraklis Varlamis¹ and Ioannis Apostolakis²

¹ Dept. of Informatics and Telematics
Harokopio University of Athens, Athens, Greece
varlamis@hua.gr

² Dept. of Health Economics
National School of Public Health, Athens, Greece
gapostolakis@esdy.edu.gr

Abstract. E-learning enables learners to decide what to learn, when, how and how fast. In the blended e-learning paradigm, knowledge is delivered using a combination of online and traditional distant education practices. The purpose of this paper is to propose a set of criteria for the evaluation of the educational process in blended e-learning communities. The systematic surveying and evaluation of the various parameters that affect the educational outcome is the primary aim of the quality assurance process. Existing evaluation methods provide general guidelines, which fail to cover the traditional distant education procedures (e.g. educational material, sporadic face-to-face meetings) that accompany e-learning activities. The key reason for the success of a blended e-learning approach is the balance between computer based and face-to-face interactions and the harmonic merge of the two. First, we review the current quality evaluation models for education and focus on the criteria that apply to blended e-learning approaches. Then, we discuss the issues arising from the combination of the two alternatives and propose solutions for improving the quality of the whole process.

Keywords: blended learning, e-learning communities, quality assurance and evaluation.

1 Introduction

A fundamental and recurring action in Quality assurance is to evaluate the system, project or service in order to ensure that it meets the quality standards and achieves the expected outcomes. In this work, we apply the principles of quality assurance and continuous assessment to the educational process inside a blended e-learning community.

Members of e-learning communities, help each other and jointly confront emerging problems. Following the Web 2.0 evolution, a growing number of online-only colleges began to offer several academic degree and certificate programs via the Internet at a wide range of levels and in a wide range of disciplines. Their main aim is to create Virtual Learning Environments (VLEs) for the management of the educational program of the whole institution, through a consistent and standard user interface.

However, they could not avoid offering some campus classes and face-to-face student support services [4], such as registration, advising and counseling. On the other side all other universities begun to offer e-learning services and online support to their students. This new paradigm of “blended e-learning” [5] combines several methods for delivering knowledge and supporting education, such as: traditional learning inside the classroom, distance learning and virtual teaching (e-learning). The selective use of traditional educational tools enforces blended learning strategies and overcomes several obstacles such as: coordination of learners’ activities, absence of the educator, students’ evaluation. It also enforces collaboration of educational institutes and leads to more open, richer and flexible curriculums, thus making the blended model a promising solution for learning communities.

From the quality assurance view, as referred by Mayes [17], quality evaluation must move from atomic to collaborative level. Among the frameworks that have been proposed in the past for evaluating blended learning approaches [10], e-learning solutions [13] or learning communities [19] no one managed to cover all three aspects of learning: pedagogical, technical and social. To the best of our knowledge, this is the first assessment framework which covers pedagogical, technical and social aspects of blended e-learning communities and evaluates the organizational and financial viability of the educational program. The evaluation plan defined in this work can be used as a guideline for evaluating learning programs in different levels and ages and can fit to the specific needs of the educational institute being evaluated.

The following section provides a survey of existing blended learning approaches and introduces the need for an evaluation framework. Section 3 presents the aspects of the educational community that we assess. Section 4 summarizes the criteria of the suggested evaluation plan and section 5 presents our conclusions.

2 Related Work

The first step in defining a framework for the quality assessment of blended e-learning communities is to position these communities precisely among all other learning approaches. Over the years, researchers and educators have introduced and tested a wide range of different teaching and learning methods [15]. Figure 1 presents a classification of these methods.

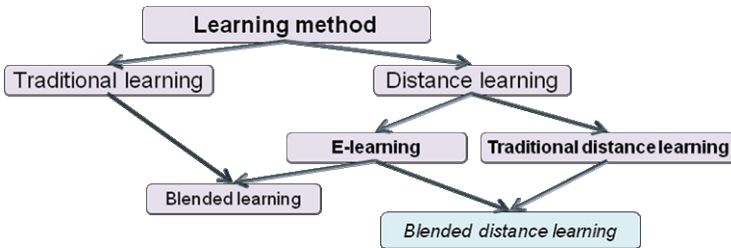


Fig. 1. Taxonomy of Learning approaches

The physical interaction between students and the tutor, through lectures, tutorials, seminars, laboratory and practical classes, is important in traditional learning and thus preferable in primary and secondary education. In the distant approach, both participants work from their own places and follow a more flexible time-schedule. The educational material is recorded, printed or other-ways reproduced and distributed to students who are able to attend courses at their own tempo. Distant learning is preferable for students who have time and space restrictions (e.g. in lifelong learning programs). We further distinguish distant approaches into traditional distant and e-learning ones. In e-learning and computer based learning, the educator is physically absent but always in assistance of the student. The educational material is properly designed to minimize teacher intervention and allow immediate response with the use of ICT technologies.

Although e-learning approaches are gaining the hype, since they provide synchronous means of interaction between students and teachers, blended solutions are preferred when e-learning educational tools are not sufficient to support the learning process. Advances in ICT replace traditional classroom meetings with online sessions and lead to a hybrid approach, which we call *blended distant learning*. For assuring the quality of the learning process, we should consider all possible aspects that affect *learning* as the principal outcome. We must examine: a) what students think and feel about the educational process, b) the resulting increase in students' knowledge and skills, c) the improvement of students' behavior and d) the effects on the students' performance [14].

The methodologies for e-learning activities' evaluation (e.g. Embedding Learning Technologies Institutionally (ELTI) [8], MIT90 [18], Pick&Mix [3] etc) do not apply to blended e-learning communities. Besides, they contain general guidelines for improving the quality of the learning process but do not provide specific evaluation criteria. For example, The E-Learning Maturity Model (eMM) [16] proposes a detailed set of criteria for assessing the learning, development, co-ordination and organization tasks performed by an educational institution, and employs a five-level scale for grading. However, eMM fails to evaluate several distant learning parameters.

As far as it concerns virtual learning communities, the proposed models focus on asynchronous tools (e.g. forums [1, 9]) or on usability issues [7]. In [2], authors propose a detailed set of criteria for virtual learning communities' evaluation and offer a detailed evaluation template, with focus on the knowledge, social and pedagogical aspects of the community. The model does not capitalize on blended e-learning communities and, as a result, neglects the traditional alternatives of "e-" processes.

A first conclusion, from the review of related work, is that some models cover traditional learning and others e-learning or e-learning communities, but none covers both aspects and all possible alternatives. Another conclusion is that all models agree on evaluating: (a) the pedagogical-psychological aspect [6], (b) the technical-functional aspect, (c) the social-cultural aspect, and (d) the organizational-economic aspect [11]. We present in details each of the four aspects, in the following section.

3 Evaluation Aspects

In order to create a concrete framework for the quality assurance of blended e-learning which can apply in a wide range of educational communities, we must

examine all the dimensions of the learning environment. The proposed framework capitalizes on the classification scheme of evaluation criteria introduced by Holst [11] and provides extensions that cover the traditional alternatives of online learning tools and the methodologies that apply in blended environments. According to the metaphor, depicted in Figure 2, a blended e-learning community is like a multilayered sphere, which spins around the organizational axis, and the learning process is a movement towards the sphere's core.

The sphere's core is the pedagogical target of the community: *learning*. The layer covering the core refers to the usability of the community structure, from a technical point of view. The arrows toward the core mean that usability aids learners and educators to achieve their target. The outside layer marks the interest of the members to the community itself. If members are interested and motivated to participate to the community activities, then it is more possible to achieve the community target. The aim of these criteria, which assess the social aspect of the community, is to evaluate whether the community is able to keep members' interest high, to build trust and avoid out-flows (arrows that bounce on the community shell). Finally, evaluation should examine the organization and operation of the learning community, which is important for its long-term existence.

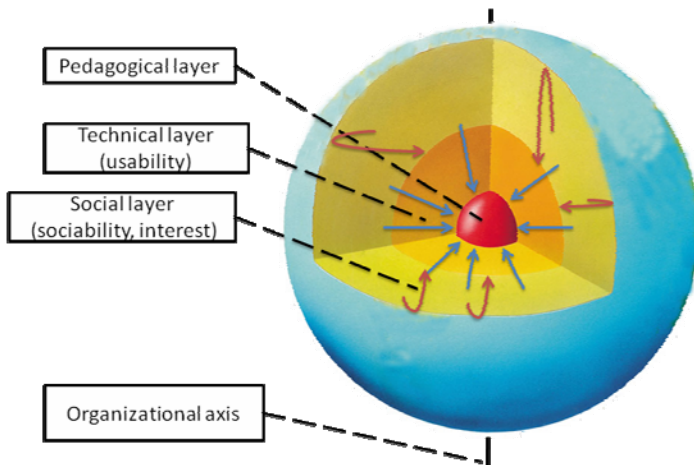


Fig. 2. A metaphor of the learning community

3.1 Pedagogical Aspect

The quality of the learning process, strongly relates to whether it reaches its pedagogical targets, which should be clear for all participants. When the pedagogical targets are clear, members are able to choose the knowledge domain of their interest and make their own schedule. So, fewer members leave the community and members' participation increases due to common interests and common goal.

Tutors are responsible for disseminating the community targets and reminding members about their educational tasks. If visitors know the chosen pedagogical

approach without having to register, it is easier for the tutors to modify it or apply an alternative approach and thus achieve a more effective learning experience. Respectively, if the members take objective comments about their progress, they could immediately try to improve themselves. By progressively increasing the difficulty of the activities, members have enough time to fill their knowledge gaps and broaden their knowledge domain. In addition, with the right scheduling, members have time to respond to their duties inside the community.

3.2 Technical Aspect

The technical infrastructure is the foundation of members' communication. As far as it concerns the virtual part of the learning community, ICT have to be of high quality and members should be familiar with it. According to Keller [12], the usability of the community platform is an intrinsic motivation to learn. Based on the web design and instructional design literature the content offered to the community members must be interactive and the general visual design should facilitate navigation.

The blended environment should be exploited to support members in getting familiar with technology and begin using the online alternatives. In-classroom courses can also complement the virtual environments, when the latter cannot substitute reality, or when it is necessary for the learners to become familiar to the real environment, for example in an anatomy course for doctors.

3.3 Social Aspect

A learning community is above all a community, which means that members reach their targets by communal effort. If the number of members is not sufficient, each member may not be able to get proper support in understanding the knowledge domain and be productive. As the complexity of knowledge domain increases, the community needs more members. Questions should be answered without delay, clarification of complex issues must be ensured and problem solving must be facilitated. Early definition of roles, privileges and responsibilities, is crucial for the operation of the blended learning community. Additionally, the existence of rules about members' behavior prevents members from having problems first in their cooperation and second in communication with others outside the community. Usually, members become more trustful as their participation increases and inspire other members to collaborate with them. Improved interaction between members supports collaborative learning, leading to more effective learning experience. Members' interaction is crucial for the successful operation of the community and it must be supported either it is on-line or on-life.

An important factor for a successful virtual learning community is to support members to evaluate the community. Involving many members in the evaluation process ensures its objectivity. When evaluators are members of the community, they better know its weaknesses and it is easier for them to report most problems. The community, of course, should not only report problems but try to solve them and improve the effectiveness of the learning experience. It should encourage members to

evaluate the community and report specific problems, but also to take into consideration the self-evaluation results and try to solve the reported problems in a legitimate time interval. The virtual activities of the community must be accompanied by scheduled face-to-face meetings that help building trust between community members.

3.4 Organizational Aspect

E-Learning is an efficient and cost-effective way of learning and as a result, it is preferable for companies that offer lifelong learning and training solutions. Although the investment for building and launching an e-learning environment is high, the resulting costs for running e-courses are significantly smaller: travel, accommodation and food expenses are minimized and virtual representations replace costly real-world experiences. Similarly the cost of learning in a virtual community can be low, thanks to simple communication equipments (i.e. a PC with internet connection and a web camera and microphone) and the volunteer contribution of experts.

A restriction for e-learning is the limited availability of Internet and associated resources in some areas of the world. Also, traditional education methods are critical in several cases (e.g. primary education) and are necessary as a complement to e-learning in others (e.g. spontaneous meetings in open-university courses). Savings from e-learning allow the educational institute to spend on traditional forms of education and *vice versa*.

The institutions behind the learning project should evaluate the organizational and financial gains and requirements of the approach and balance between electronic, distant and traditional solutions. They should balance the budget distribution for the wages of content experts, course designers, computer programmers etc, reduce the running costs and increase profits via reusability from partnerships with other e-learning institutions. In the following section we describe the focal points of the evaluation process of the blended e-learning community for all the above aspects.

4 Evaluation Criteria

We divide the evaluation criteria into four main categories matching the aforementioned aspects. Each category is further analyzed in a series of factors that should be evaluated in a continuous basis in order to assure quality. We have defined an extended set of evaluation criteria that cover all aspects and examine their factors and use a 4+1 levels scale for grading. The scale ranges from 'Fully Adequate' -when the solution completely fulfills the criterion- to 'Not Adequate' -when it does not fulfill the criterion at all- and 'Not Measured' -when it is unclear whether the community fulfills the criterion.

The criteria of the Pedagogical category, as depicted in Table 1, examine whether the pedagogical aims and the application field have been made clear for all members and visitors and tutors and learners have agreed on their roles. They check whether tutors are allowed to apply alternative pedagogical practices and the educational solution is efficient in serving the learners' needs and capable to improve their knowledge and skills.

Table 1. Criteria for the evaluation of the pedagogical aspect

Pedagogical approach related
i) Visitors are informed on the pedagogical approach, ii) The application field of the knowledge domain is presented, iii) Visitors are informed on the community goals, iv) Knowledge domain is well known by the tutors, v) Tutors are able to apply alternative pedagogical practices, vi) Instructors and learners have agreed on their roles and on the pedagogical paradigm they employ
Material and activities related
i) Topics in the content area have been divided into suitable for e-learning and suitable for face-to-face instruction, ii) There is a detailed activities' calendar including their purposes and whether they are online or offline, iii) The educational content is accurate, complete and is available in both online and offline formats, iv) Tutors discuss students' progress in person, v) Difficulty of the activities is progressively increased, vi) Activities are fairly scheduled among members, vii) Multimedia components, internet tools and external information sources used in the activities, are listed, viii) Activities are designed to support students to become independent distance learners
Student related
i) Visitors are informed on the prerequisite knowledge, ii) Learners' preferences are recorded in their profiles, iii) Knowledge and skills are available about learners, iv) Attitudinal and motivational information is available about learners, v) A survey on learners' needs is available for visitors, vi) The program goal(s) are approved by appropriate officials within the institution, vii) Student orientation services are available, viii) Students' assessment is performed both using traditional and e-learning methods

Table 2. Criteria for the evaluation of the technical aspect

Requirements
i) Hardware requirements for online or offline e-courses have been defined, ii) Special software is required for online or offline e-courses, iii) Network: technical requirements relating to data transfer (bandwidth, file size, connectivity etc). of the multimedia components, internet tools, and supplementary materials are reasonable, iv) Requirements for specialized equipment
Navigation, Orientation
i) The information is intuitively located and easily accessible, ii) It is easy to understand where I am within the information architecture, iii) Links actually lead to the content they promise to lead (no broken links)
Content & Design
i) International interoperability standards (SCORM) are employed, ii) Multimediaity: Different media are used to convey the information necessary to complete the task, iii) Quality: The quality of the audiovisual and textual content conveys effectively content and interaction capabilities, iv) Graphical interface elements and requirements (browser, plugins, etc), v) The following services are available: Multimedia Archives, Mailing lists and their archives, FAQs, e-books, Webliographies, Reading lists, Experts online.
Communication, collaboration, facilitation
i) Availability of communication tools, ii) Technical support adequacy, iii), Library support adequacy, iv) Learner's guide is available, v) Services for Students with Disabilities.
Educational
i) Flexibility and reusability of the LCMS or LMS components, ii) Teachers' familiarity with the LCMS, iii) The educational material is self-created, iv) Existence and exploitation of tutorial services.

The criteria evaluate the accuracy and completeness of the educational material and activities and if they can be equally performed online or offline. Finally, they examine if members are motivated to discuss with other members in person about their progress, receive facilitation, and participate in activities of progressively increasing difficulty.

The criteria that evaluate the Technical aspect and the quality and usability of the educational solution, as depicted in Table 2, examine: a) the completion of the technical requirements in software, hardware and equipments; b) the accessibility and usability of the online content; c) the compliance to content and design standards; d) the availability of communication, collaboration and facilitation services.

Table 3. Criteria for the evaluation of the social aspect

<i>Collaboration</i>
i) There is sufficient number of members who are encouraged to collaborate, ii) Members assist each other to use ICT, iii) Faculty and staff directories are available and up-to-date.
<i>Feedback</i>
i) Members' feedback concerning activities and their difficulties is directly forwarded to tutors, ii) Members are encouraged to evaluate the community and report specific problems, iii) Each member is rated for its participation and this rating is public, iv) The results of the self-evaluation are taken into consideration and the community tries to solve the reported problems in a legitimate time interval, v) There is a system to accept students' complaints.
<i>Facilitation – Guidance</i>
i) Information on popular courses and course suggestions are provided to students, ii) Course recommendations are made based on collaborative filtering, iii) Instructors respond to learners' inquiries, iv) The instructional staff maintain scheduled office and online hours (synchronous), v) Profile management and career counselling services are available.
<i>Connection to real world</i>
i) Parallel communities are supported (i.e. alumni), ii) Welcome and Graduation ceremony are held online and/or offline, iii) Internship and employment services, iv) The digital divide issue is considered in the design of the e-learning content.
<i>Roles and rules</i>
i) How many of the following roles apply to the faculty members: Instructor, Teaching Assistant, Tutor, Technical Support, Librarian, Counsellor, Graduate Assistant, E-Learning Administrator, Advisor, ii) How many of the following roles apply to the community members: Moderator, Facilitator, Administrator, iii) There is a policy with specific rules about members' behaviour (etiquette and netiquette)

The quality assessment of the Social aspect of the community (Table 3) emphasizes on: a) the degree of support for collaboration; b) the amount and participation of members, c) the exploitation of user feedback; d) the operation of facilitation and guidance services; e) the connection to the real life of members; f) the definition of roles and rules for the community operation.

Finally, with regard to the organizational aspect (Table 4), the operational and economic factors are evaluated. More specifically, we evaluate: a) the strategic and business plans, the structure of the academic calendar and the flexibility of the curriculum; b) the automation of the Registrar's procedures; c) the dissemination plans; d) the sustainability of the financial investment and the viability of the whole educational program.

Table 4. Criteria for the evaluation of the organizational aspect

Operational
i) Mission Statement, Strategic Plan and Business Plan available, ii) Predefined Content Development, Delivery and Maintenance policies, iii) Fix academic calendar or fixed course duration, iv) Curriculum flexibility, v) Clear policies on Intellectual property rights.
Registrar
i) Admission requirements are clear and available, ii) Availability of application forms, iii) Accessibility of students' records, iv) Payment of fees and billing are online, v) Transcript Request Form and official transcripts are both online and offline
Dissemination
i) Newsletter, Community Newspaper, other printed or electronic dissemination material, ii) Organize meetings, conferences, either online or offline, iii) They are able to provide their program and course information completely on-line or in print materials or in a combination, iv) Marketing means (electronic or other)
Investment
i) Allocate budget for e-learning, including wages for content experts, course designers, computer programmers etc., ii) Difference in fees between e-learning and traditional solutions, iii) Reusability of content and resources, iv) Partnerships with other e-learning institutions are encouraged, v) Members can use library and other learning resources from partner institutions, vi) On-line bookstore or partnership with an on-line bookstore.

In order to provide a complete evaluation template that can be applied in a constant basis in the blended e-learning community and reassure the overall quality, we perform a quantification of the evaluation results collected using the aforementioned criteria. The individual marks can be used to calculate separate scores for each sub-category, or category or a final score for the full set of criteria. Supplementary weights can be applied in each category or sub-category depending on the priorities of the community. The resulting formula will have the following form:

$$Score = W_A * \sum_{i=1}^{asize} w_{ai} * m_{ai} + W_B * \sum_{i=1}^{bsize} w_{bi} * m_{bi} + W_C * \sum_{i=1}^{csize} w_{ci} * m_{ci} + W_D * \sum_{i=1}^{dsize} w_{di} * m_{di}$$

where W_A , W_B , W_C , W_D denote the priority of the respective aspect for the community, w_{xi} represents the interest on sub-category i of aspect x and $xsize$ is the number of subcategories that apply in aspect x . Finally, m_{xi} stands for the median of the individual criteria values in the sub-categories of category x . Also:

$$W_A + W_B + W_C + W_D = 1 \text{ and } \sum_{i=1}^{asize} W_{ai} = 1, \sum_{i=1}^{bsize} W_{bi} = 1, \sum_{i=1}^{csize} W_{ci} = 1, \sum_{i=1}^{dsize} W_{di} = 1$$

The detailed presentation of the evaluation form has been omitted due to space limitations. However, it is on our next plans to make the form available in public and use it for the evaluation of a blended e-learning community. Factor analysis and the statistical process of the factors being evaluated will give us a better view on their role in the success of the community and the continuous assessment will allow us to assure the quality of the educational process.

5 Conclusions

The evaluation of learning approaches is a difficult and multi-facet task, which usually results in huge evaluation checklists that cover traditional learning and e-learning. The definition of a strict set of criteria may lead to an inflexible evaluation schema that fails to adapt to the individualities of each approach. In this work, we presented a general evaluation framework that is focused on blended e-learning communities. The framework can easily adapt to personalized education, electronic or in-person teaching, by adjusting the interest to each evaluation factor according to the specific needs of each educational institute that offers Blended E-learning solutions.

References

1. Araújo, L.H.L., Lucena Filho, G.J., Losada, M.: Evaluating Virtual Learning Communities using a Nonlinear Model. In: Proc. of the 8th lasted CATE, Aruba (2005)
2. Athanasiou, G., Maris, N., Apostolakis, I.: Evaluation of virtual learning communities for supporting e-learning in healthcare domain. In: Proc. of 6th ICICTH, Greece (2008)
3. Bacsich, P.: Benchmarks for e-learning in UK HE - adaptation of existing good practice. In: Proc of the 12th ATLC-C 2005, UK (2005)
4. Bajcsy, J.: Basic Information about Engineering Subject for Virtual Education. In: Proceedings of the 5th International Conference on Virtual University, Bratislava (2004)
5. Bielawski, L., Metcalf, D.: Blended eLearning: Integrating Knowledge, Performance Support, and Online Learning. Enterprise-class Edition, 2nd edn. HRD Press (2005)
6. Britain, S., Liber, O.: A Framework for Pedagogical Evaluation of Virtual Learning Environments. Report 41, JISC Technology Application Programme. Wales (1999)
7. Cobb, S.V.G., Neale, H.R., Reynolds, H.: Evaluation of virtual learning environments. In: Proc. of ECDVRAT, Skovde, Sweden, pp. 17–23 (1998)
8. Deepwell, F.: Embedding Quality in e-learning Implementation through Evaluation. *Educational Technology & Society* 10(2), 34–43 (2007)
9. Díaz, L.A., Figaredo, D.: Combined evaluation of on-line learning communities. In: Proc of ICTE 2009: International Conference on Technology and Education, France (2009)
10. Garrison, D., Kanuka, H.: Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education* 7(2), 95–105 (2004)
11. Holst, S.: Evaluation of Collaborative Virtual Learning Environments: The State of the Art. In: Scheuermann, F. (ed.) Proceedings of GMW (2000), ISBN 3-89325-925-2
12. Keller, J.: Motivational design of instruction. In: Reigeluth, C.M. (ed.) *Instructional Design Theories and Models: An overview of their current status*. Erlbaum, Hillsdale (1983)
13. Khan, B.: *E-Learning Quick Checklist*. IGI Publishing (2005), ISBN-10: 1591408121
14. Kirkpatrick, D.L.: *Evaluating Training Programs: The Four Levels*. Berrett-Koehler, San Francisco (1994)
15. de Kock, A., Slegers, P., Voeten, M.: New Learning and the Classification of Learning Environments in Secondary Education. *Review of Educational Research* 74(2) (2004)
16. Marshall, S.J., Mitchell, G.: Benchmarking International E-learning Capability with the E-Learning Maturity Model. In: Proceedings of EDUCAUSE, Australia (2007)
17. Mayes, J.: Quality in the e-university. *Assessment & Evaluation in Higher Education* 26(5), 465–473 (2001)
18. Morton, S., Michael, S.: *The Corporation of the 1990s: Information Technology and Organizational Transformation*. Oxford University Press, New York (1991)
19. Smith, B.L.L., MacGregor, J.: Learning communities and the quest for quality. *Quality Assurance in Education* 17(2), 118–139 (2009)

Quality of Content in Web 2.0 Applications

Iraklis Varlamis

Dept. of Informatics and Telematics,
Harokopio University of Athens, Athens, Greece
varlamis@hua.gr

Abstract. Nowadays, web users are facing a novel and dynamically changing environment, which promotes collaboration in content creation. In the web era, people usually neglected the quality of content, and had minimum authority on reporting good or bad content, whereas, in Web 2.0, content is becoming a priority and people become more critical when they assess its quality. In the former case, each web site had an identity, a single editor or a closed group of editors, who were responsible for content quality management and the stylistic design was the main quality standard for visitors. In the latter case, several applications and a new way of thinking have changed the workplace and created a new approach to quality management. Specifically, the team of editors is open to every web user, content is a collaborative product and the “embedded object” or “inline linking” model is gaining in popularity against hyperlinking.

Keywords: Web 2.0, Content quality, Content control and assurance.

1 Introduction

A study of the Stanford credibility team [9] on 100 web sites showed that the first impression is what counts more for the typical web user. In average, the 2684 users paid far more attention to the look of the site than to its content. Similarly in e-commerce sites the familiarity and credibility of corporate logos strengthens the perceived customer trust [17].

Although the look and feel is critical for the success of a typical web page, “content” is the primary focus of Web 2.0 applications, which capitalize in the simplicity of content instead of dynamic content presentation and scripting. Because of this flexibility in the structure of content, several third-party applications have been developed that aggregate (e.g. in the form of RSS feeds), process plain content (translate, summarize, categorize, rank etc.) and make it available to more web users.

The primary interest of this paper is “content” and mainly the content that is uploaded in a daily basis on Web 2.0 applications. For this reason, we emphasize on the quality of content, present the emerging directions in content quality control and examine the parameters that affect the quality of content such as reliability, accessibility, availability, flexibility. In addition to this, we present the architecture of a system, which can increase the accessibility and availability of content and improve its overall

quality. The system employs web resources to bridge between audiovisual and textual content and increase its availability and accessibility (e.g. text translators, text-to-speech and speech-to-text converters, text and content annotation services). For the quality of content, it adopts the collaborative paradigm of Web 2.0 and several rating and reputation mechanisms to promote high-quality content sources against less credible ones.

We distinguish two main content types: textual and audiovisual. We present solutions that focus on each specific type, but also services that span across the two types and provide flexible solutions for users and applications. As far as it concerns the accessibility of textual content we focus on text translation, text to speech conversion and text annotation. Concerning audiovisual content, we present several speech-to-text and content annotation solutions and focus both on automatic and semi-automatic solutions. In the dimension of content credibility, we emphasize on the collaborative paradigm of Web 2.0 and present the rating and reputation mechanisms that promote the credible sources against less credible ones. Finally, we examine the flexibility, re-usability and availability of Web 2.0 content in contrast to existing Web content.

In the following section we enlist several approaches towards improving content quality in Web 2.0 applications. In section 3 we present the architecture of the suggested system, which combines many of the Web 2.0 novelties, under the prism of intelligent information management, to guarantee content quality. In section 4 we briefly discuss the criticism against the suggested approach and present our counter-arguments. Finally, in section 5 we present the conclusions of this work.

2 Content Quality in Web 2.0: Limits and Solutions

There have been many recent research works on Data Quality in Collaborative Information Systems, which aim at the quality of data in different disciplines, for example in Genomics [15] or Meteorology [19]. All works agree that data quality assurance comprises several tasks, such as duplicate detection, identification of outliers, consistency checking and many more, which are continuously repeated in order to guarantee quality. The preservation of data quality in high standards lies at the convergence of the three aspects, namely organizational, architectural and computational. In a domain specific information system, information lifecycle is usually well defined and the same holds for information providers and structure of data. Thus, it is easier for the system developers to design the information process, to define the quality assessment procedures and to develop a quality control mechanism.

For example, the information process for drug development and disease prevention [15] typically starts with a biological experiment, which is designed based on literature and is performed on the living organism by wet-lab biologists. The result is transformed into digital data format using specific platforms (e.g. Affymetrix) and then data is analyzed by bioinformaticians using computer software. Finally biologists reuse the analyzed information to design more relevant and accurate experiments.

However, the definition of a data quality model in a general purpose information system is not straightforward. Several parameters, such as user needs, computational and/or organizational restrictions affect the importance of each data quality criterion

and modify the resulting data quality assessment model. A Web 2.0 application is an Information System per se, with users, software, hardware and above all: *data*. Quality of data can be of higher or lower importance to users, depending on the application and the criteria for evaluating quality may differ between users depending on their needs. In [5], authors propose a generic model for assessing data quality in social networks, which captures the subjective quality criteria of each user or group of users and the objective criteria of the community and takes into account the collaborative nature of social networks.

In this current work, we emphasize on the quality of content, which is contributed in Web 2.0 applications, we examine the various aspects that affect content quality and present solutions and working schemes for collaboratively improving and evaluating quality of content.

2.1 Content Availability and Accessibility

A first step in increasing the accessibility of textual content is to make it available in many different languages. Despite the improvements in automatic translation services, translating user created content in Web 2.0 applications is usually problematic, due to the informal, conversational style employed. Trying to improve the quality of automatic translation, several Web 2.0 applications have adopted the new collaborative paradigm. Wikipedia, which is now available in more than 250 languages¹, is the most successful crowd-sourcing translation example in which human editors revise and refine machine-translated content.

Apart from translation, metadata and content description is crucial for the quality of nontextual content. Web 2.0 contributed in this direction through social bookmarking and tagging. All social bookmarking engines allow users to provide 'tags' (i.e. keywords) that briefly describe their content, or the content they like. Using these tags, users are able to organize content or search for similar content, even for images, video or audio [10, 18]. Additionally, several applications that extract textual content [11] and technical or semantic information [3] from audiovisual content (e.g. image resolution and format, frame rate and audio compression for videos) can be used to enhance content description. Web 2.0 offers many tools for collaborative video annotation and indexing [21].

A third step towards improving content accessibility is to increase the number of available formats. Speech synthesis and speech recognition software, allow to create several Text to Speech² and Speech to Text³ services and embed them in Web 2.0 applications. The adoption of open standards, such as DAISY⁴ (Digital Accessible Information System) XML, will automate text to speech conversion and facilitate users with "print disabilities". Additional conversion services can be used to convert the original content into multiple resolutions and formats, so that it can be viewed in as many devices as possible (e.g. desktops, laptops, mobile phones, in car devices). Finally, archiving services can be employed for long-term preservation of content [8].

¹ http://meta.wikimedia.org/wiki/List_of_Wikipedias

² <http://vozme.com>, <http://say.expressivo.com>, <http://www.cepstral.com/>

³ Loquendo ASR: <http://www.loquendo.com>

⁴ <http://www.daisy.org/>

2.2 Content Quality Control and Assurance

The main contribution of Web 2.0 in terms of content development is a new crowd-sourcing model, first described by Tim O'Reilly as "the creation of collective intelligence" [20]. Collaborative Content Development refers to a recursive process, where two or more people work together, share knowledge and build consensus outside of professional routines and practices [24].

In contrast to the content of edited Web sites, the quality of user created content is questionable, due to the absence of quality control mechanisms. In order to improve content control, most Web 2.0 applications offer reporting mechanisms for offensive content, and many National and International bodies (e.g. Internet Watch Foundation⁵) attempt to increase web users' awareness against illegal and harmful content [2]. However, few attempts focus on controlling and improving content quality.

Wikipedia is the most appropriate example that demands user-driven content quality control, since the classical expert review model is not feasible, due to the huge amount and the continuous change of content. The Nature Magazine⁶ and Google's Knol⁷ adopt a similar model. Despite its success, the Wiki model still attracts the criticism of academics and encyclopedias and several research works have been focused on rating models that will highlight the most eligible articles [7].

In order to solve several abuse and misuse issues (e.g. wrong or biased reviews, misleading tags), the emerging rating and recommendation schemes [12] employ reputation mechanisms, which prioritize long-living content and long-term contribution [1]. The collaborative rating model, which was designed for content filtering and personalization [13] has recently been employed for ranking content based on popularity and interestingness [16] or for evaluating content quality in general [14]. Digg⁸ is a Web 2.0 application that allows users to report online web content (news, blogs, videos etc) and let other users rate them by voting on them. In the following section, we present a system for the quality control of collaboratively created content.

3 A Collaborative System for Content Quality

According to the European Commission's guidelines for information providers⁹, the quality assurance activities comprise: a) check of text accuracy and proofreading; b) assessment of the presentation clarity; c) copyright clearance; and d) validation of the original content and its translation. The long checklist for quality control¹⁰ can be summarized in the following actions: a) guarantee accessibility of information for everyone (e.g. multilingualism) and for individuals with specific needs (either having physical disabilities or technical restrictions); b) ensure content archiving; c) assure clarity of presentation, style and structure. Having these guidelines in mind, we provide the architecture of a system that allows quality control and assurance of the information contributed in Web 2.0 applications.

⁵ www.iwf.org.uk

⁶ <http://www.nature.com/scitable>

⁷ <http://knol.google.com>

⁸ <http://digg.com>

⁹ http://ec.europa.eu/ipg/quality_control/workflow/content/index_en.htm

¹⁰ http://ec.europa.eu/ipg/quality_control/checklist/index_en.htm

Content is behind every Web 2.0 application either it is a social networking site (e.g. for blogging, media sharing etc) or a community for collaborative knowledge building (e.g. a learning community or a community of practice). The success of a Web 2.0 application equally depends on its ability to appeal users and evoke users' contribution. However, the community assumes that the information provided by all members is of high quality. The challenge in this case, in comparison to web content, is the absence of an editor or a group of editors. Due to the collaborative network of Web 2.0, content is edited, revised and evaluated by the users themselves and consequently, the quality of content is a users' issue. The suggested quality control and assurance approach is generic, is collaborative and strongly connected to users' contribution but also exploits several open and free services and resources in order to facilitate users.

The suggested solution summarizes all the aforementioned quality control activities in four distinct axes, namely: content accessibility, content availability, content clarity and content usefulness. In the first two axes, we propose several technical solutions that can automate the process and improve quality of content in the respective directions. Quality in the last two axes is more subjective, so we count on the contribution of users and propose a rating scheme that allows users to evaluate and consequently highlight content of high quality.

3.1 Automation of Content Accessibility and Availability

In a previous work [23] we introduced an architecture that increases content accessibility and availability in the blogosphere. It exploits the structure of content and additional semantic information, processes, reformats and enriches content using aggregators and mediating services and makes it available to end users. The approach exploits the flexibility of XML, which is the basic format for content in the blogosphere, and assumes a template driven editing model, which allows users to provide content and metadata with minimal effort. With the use of XSL scripts, and intermediate brokering services, the original content can be reformatted and personalized according to users' requirements.

In this work we move one step ahead, and suggest the integration of online services (e.g. automatic translation services¹¹, online text to speech¹² and speech recognition tools, and online media conversion services¹³) for the creation of alternative versions of content. Content is organized in more than one repositories in order to increase availability (see Figure 1) and create a distributed archive of Web 2.0 content.

Content replication and content availability in alternative formats, will allow all overlaying content brokering services to pick the appropriate content format for each user device. The architecture presented above, is very flexible since it allows new content mediation services to be attached, thus providing more and better content alternatives. Moreover, several content brokering services can be created that personalize content to user needs and that exploit the redundancy of content in the distributed repository in order to provide the content in a user-friendly format. Finally, the

¹¹ e.g. <http://babelfish.yahoo.com> or <http://translate.google.com>

¹² e.g. <http://say.expressivo.com>

¹³ e.g. <http://media-convert.com/>

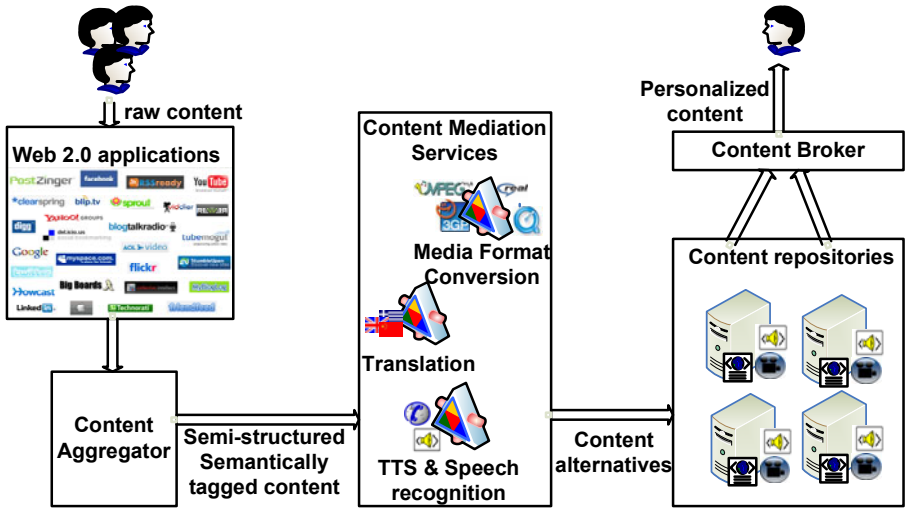


Fig. 1. Information mediation services and repositories

architecture exploits user feedback in order to identify low quality content, as explained in the following subsection, or unreliable content providers. Hence, users may interact with content in the repository and improve its quality and accuracy in a collaborative manner.

3.2 A Rating Mechanism for User-Centric Evaluation of Content Clarity and Usefulness

The idea behind many Web 2.0 applications is to promote user collaboration. Web 2.0 enthusiasts capitalize on this group effort in order to build collective intelligence [20]. The success of this attempt is strongly related to the ability of users to contribute their content and to evaluate the contribution of others. Although a typical quality assessment process is more complex than content rating, successful collaborative applications in Web 2.0 build upon this simple process [14]. Several approaches extend the collaborative rating example to a reward/punish mechanism [6], which promotes high-quality content and demotes spam or to a trust propagation model [22], which learns from the ratings of a user and her trustees.

The proposed rating mechanism, which is based on our previous work [22], exploits users’ feedback on content quality, audits users’ ratings and employs them for building a reputation profile for content contributors. We acknowledge that building a reputation mechanism for social networks is much more complex [25] than simply collecting user rates for content, but we can safely say that our mechanism can be developed on top of user provided feedback, which must be collected for a long period of time.

User feedback on content quality can be collected through simple rating mechanisms, such as a like/dislike flag, or a number in an ordinal scale. Although it is based

on a simple mechanism, the suggested model allows the collective evaluation of a piece of content from many users and provides a useful indication on its quality. User ratings are audited and their analysis provides better clues on the freshness and impact of each piece of content, as well as on the credibility of each user's ratings. For example, a piece of content that receives many positive marks right after its publication is probably of high interest and quality. Similarly, an author who repeatedly publishes content of high interest is probably an influential author. In a similar manner, when a user repetitively assigns bad ratings to contents that other users rate as good, then the credibility of this user decreases.

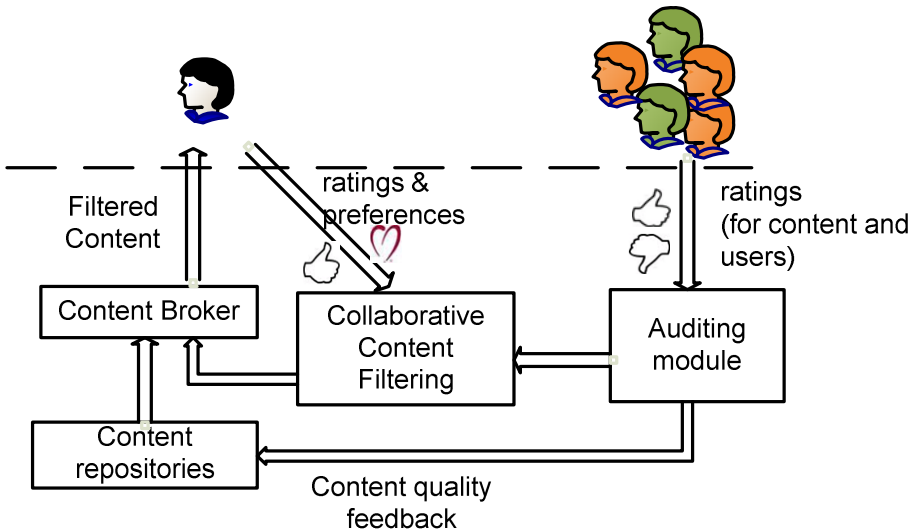


Fig. 2. The content rating mechanism

Figure 2, depicts the application of the proposed mechanism to the Content Brokering module of Figure 1. In addition to the content, which is aggregated, organized and archived in the repositories, users' ratings are audited and processed in order to serve as feedback and help in evaluating quality of content in the repositories. Rating history is processed together with user's preferences in order to provide better recommendations, thus improving the overall quality of the information brokering service.

The detailed modeling of content quality, freshness and impact and of author credibility and trustfulness is strongly depended to the objectives of each Web 2.0 application, so it is outside the scope of a generic mechanism for quality evaluation. It would be of great importance to focus on a specific type of application or on a specific community (e.g. an educational community), analyze the specific needs for content quality and provide a mechanism for content quality control [4] or assurance. However, it is of greater importance to have a generic solution that can be adapted and fine-tuned for each specific application.

4 Criticism and Discussion

The improvement on content accessibility and availability is a technical issue, which also depends on the will of users. The enthusiasts of Web 2.0's simplicity can easily say that adding standards and restrictions to user provided content, will discourage users in providing new content and will lead to a decrease in participation. This claim can be reasonable, if users are requested to know about the standards, learn specific languages for structuring their content and providing useful metadata. However, the trend in Web 2.0 applications shows that the exact opposite holds. In most successful stories, users simply provide their content (e.g. the text of a blog post, the short twitter message, an image on Flickr) and then a lot of metadata, concerning the author, the language, the descriptive tags, the technical details etc., is added automatically or semi-automatically with a few user clicks.

Concerning the rating mechanism, quality experts may argue that quality control is more than a simple rating mechanism for content, which comprises: routine and consistent checks that ensure content correctness and completeness, identification of errors and omissions, content archiving and auditing of all quality control activities. Similarly, quality assurance refers to a planned set system of review procedures, which are usually conducted by independent reviewers. The review procedures usually involve extended questionnaires and huge checklists for the evaluation of every possible aspect that affects content quality. Depending on the significance of the application, quality control and assurance can be of higher or lower importance to the users. For example, a Web 2.0 application that provides medical consultation to patients or a collaborative educational platform capitalize on the quality and correctness of content, and thus may require additional control mechanisms and a more profound content evaluation. On the other side, a social bookmarking application can afford a temporary decrease in content quality (e.g. due to spam bookmarks), but will reside on users' feedback in order to identify and eliminate low quality content or malicious content contributors.

5 Conclusions

The current study, addressed the various aspects of quality of content in collaborative applications. The survey of web translation and conversion applications revealed a large number of open and free solutions that can increase content availability and improve accessibility. In addition to this, the joint efforts of web users can further amend machine translated and converted content, in favor of the users. Users can also guarantee the reliability and correctness of content, as happens in wiki applications.

The proposed system combines the power of Web 2.0 users and applications in order to improve content quality in every dimension and reassure that content always reaches a high quality level. As far as it concerns availability and accessibility, the system exploits the modular structure of Web 2.0 content, the user provided metadata and freely accessible web services in order to enrich content and provide content alternatives. Concerning content quality assessment, the system collects user-provided feedback (preferences and ratings for content and users) and processes it in order to

create useful report on content quality. Low quality content will be demoted and useful and interesting content will be favored. The suggested rating system can be used to create global evaluations of content quality, but also can be employed for collaborative content filtering, thus providing users with content of interest. Finally, the auditing module, allows the monitoring of the aggregated content in a continuous basis, thus achieving maintenance of content quality. The next step of our work is to implement the proposed system and test its performance in a real scenario, for a specific community of web users.

References

1. Adler, B.T., de Alfaro, L.: A content-driven reputation system for the Wikipedia. In: Proc. of the 16th Intl. World Wide Web Conf. (WWW 2007), ACM Press, New York (2007)
2. Akdeniz, Y.: Controlling Illegal and Harmful Content on the Internet. In: Wall, D.S. (ed.) *Crime and the Internet*, pp. 113–140. Routledge, London (November 2001)
3. Athanasiadis, T., Avrithis, Y.: Adding Semantics to Audiovisual Content: The FAETHON Project. In: Enser, P.G.B., Kompatsiaris, Y., O'Connor, N.E., Smeaton, A., Smeulders, A.W.M. (eds.) CIVR 2004. LNCS, vol. 3115, pp. 665–673. Springer, Heidelberg (2004)
4. Avouris, N., Solomos, K.: Social mechanisms for content quality control in web-based learning: An agent approach. In: Jacko, J., Stephanidis, C. (eds.) *Human-Computer Interaction*, vol. 1, pp. 891–895. Lawrence Erlbaum Assoc., Mahwah (2003)
5. Caballero, I., Verbo, E., Serrano, M., Calero, C., Piattini, M.: Tailoring Data Quality Models Using Social Network Preferences. In: Chen, L., Liu, C., Liu, Q., Deng, K. (eds.) DASFAA 2009. LNCS, vol. 5667, pp. 152–166. Springer, Heidelberg (2009)
6. Cheng, R., Vassileva, J.: Adaptive Reward Mechanism for Sustainable Online Learning Community. In: *Int. Conf. on Artificial Intelligence in Education*, pp. 152–159 (2005)
7. Cusinato, A., Della Mea, V., Di Salvatore, F., Mizzaro, S.: QuWi: quality control in Wikipedia. In: 3rd workshop on Information credibility on the web, WICOW 2009 (2009)
8. Day, M.: The Long-Term Preservation of Web Content. In: *Web Archiving*, pp. 177–199. Springer, Heidelberg (2006), doi:10.1007/978-3-540-46332-0_8
9. Fogg, B., Soohoo, C., Danielson, D., Marable, L., Stanford, J., Tauber, E.: How Do People Evaluate a Web Site's Credibility? A Consumer WebWatch research report. Stanford Persuasive Technology Lab, Cordura Hall 226, Stanford University, Stanford (2003)
10. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social Bookmarking Tools (I): A General Review. *D-Lib Magazine* 11(4) (April 2005), ISSN 1082-9873
11. Jung, K., Kim, K.I., Jain, A.K.: Text information extraction in images and video: a survey. *Pattern Recognition* 37(5), 977–997 (2004)
12. Kolbitsch, J., Maurer, H.: The Transformation of the Web: How Emerging Communities Shape the Information we Consume. *Journal for Universal Computer Science* 12(2) (2006)
13. Lee, C., Kim, Y., Rhee, P.: Web personalization expert with combining collaborative filtering and association rule mining technique. *Expert Systems with Applications* 21, 131–137 (2001)
14. Lerman, K.: Dynamics of a Collaborative Rating System. In: Zhang, H., Spiliopoulou, M., Mobasher, B., Giles, C.L., McCallum, A., Nasraoui, O., Srivastava, J., Yen, J. (eds.) *WebKDD 2007*. LNCS, vol. 5439, pp. 77–96. Springer, Heidelberg (2009)
15. Liu, Q., Lin, X.: Genomic Information Quality. In: *Proceedings of DASFAA Workshop on Managing Data Quality in Collaborative Information Systems*, New Delhi, India (2008)

16. Louta, M., Varlamis, I.: Blog rating as an iterative collaborative process. In: *Semantics in Adaptive and Personalised Services: Methods, Tools and Applications*. Springer series on Studies in Computational Intelligence (2010)
17. Lowry, P.B., Roberts, T.L., Higbee, T.: First Impressions with Websites: The Effect of the Familiarity and Credibility of Corporate Logos on Perceived Consumer Swift Trust of Websites. In: Jacko, J.A. (ed.) *HCI 2007*. LNCS, vol. 4553, pp. 77–85. Springer, Heidelberg (2007)
18. Lund, D., Hammond, T., Flack, M., Hannay, T.: Social Bookmarking Tools (II): A Case Study – Connotea. *D-Lib Magazine* 11(4) (April 2005), ISSN 1082-9873
19. Mateo, M.A., Leung, C.: CHARIOT: A Comprehensive Data Integration and Quality Assurance Model for Agro-Meteorological Data. In: *Proceedings of DASFAA Workshop on Managing Data Quality in Collaborative Information Systems*, New Delhi, India (2008)
20. O’Reilly, T.: *What Is Web 2.0*. O’Reilly Network (2005),
<http://oreilly.com/web2/archive/what-is-web-20.html>
(Retrieved 2010-02-02)
21. Schroeter, R., Hunter, J., Kosovic, D.: FilmEd - Collaborative Video Indexing, Annotation and Discussion Tools Over Broadband Networks. In: *International Conference on Multi-Media Modeling*, Brisbane, Australia (2004)
22. Varlamis, I., Louta, M.: Towards a Personalized Blog Site Recommendation System: a Collaborative Rating Approach. In: *Proceedings of the 4th International Workshop on Semantic Media Adaptation and Personalization*, San Sebastian, Spain (December 2009)
23. Varlamis, I., Giannakouloupoulos, A., Gouscos, D.: Increased content accessibility for wikis and blogs. In: *Proceeding of the 4th Mediterranean Conference on Information Systems MCIS 2009*, Athens, Greece (2009)
24. Vickery, G., Wunsch-Vincent, S.: Participative Web and User-created Content: Web 2.0, Wikis and Social Networking. In: *OECD 2007* (2007)
25. Yu, B., Singh, M.P.: A social mechanism of reputation management in electronic communities. In: Klusch, M., Kerschberg, L. (eds.) *CIA 2000*. LNCS (LNAI), vol. 1860, pp. 154–165. Springer, Heidelberg (2000)

Telepediatrics Education on the Semantic Web

Sofia Sidirokastriti and Anastasia N. Kastania

Department of Informatics,
Athens University of Economics and Business, Greece
sidirokastr@aueb.gr, ank@aueb.gr

Abstract. An e-learning web application for telepediatrics education is presented. Semantic web technology is used in the design and development of the platform. The platform offers personalized e-learning services using various developed agents. A Web Service was also developed to monitor a child's electrocardiograph from distance. Quality assurance aspects fulfilled are reliability of information included, understandability and accuracy of the content, disclosure related to the purpose of the smart website, inclusion of links so that the user can compare the information and make sure that is trustworthy, user friendly design, inclusion of feedback and alerts.

Keywords: telepediatrics, e-learning, ontology, agents, web service.

1 Introduction

Telemedicine is a rapidly evolving practice of clinical medicine from distance. Technology is used in service of medicine. High-quality medical services can be offered. This makes it possible for people in remote places to be examined by doctors who can monitor patients through the computer using computer networks. Patients especially the ones with chronic diseases can be treated in their homes, rather than spending their lives in hospitals. This is ideal for ill children. Every child must have the opportunity to be a child, to relax and have fun. However, there are many ill children who spend countless hours in hospitals. Even in the best pediatric hospitals, these children see depressed, sick and sad people all the time and therefore are forced to grow up sooner than necessary.

Although telemedicine has much to offer to both patients and doctors, some people are cautious due to the fact that the information handled by the telemedicine systems is private and moreover the smallest mistake could cost a life. However, the whole purpose is to save lives that are currently mistreated or even not able to be treated at all. Education helps people understand the advantages of telemedicine. Furthermore, education can also be useful for people who contribute in the development of training material and platforms who have different backgrounds (doctors, engineers, computer scientists or even people and children) and are not all familiarized with technology or telemedicine.

The article presents a web application that tries to contribute to telepediatrics education, a valuable branch of telemedicine. The current knowledge in telepediatrics is presented in a functional training interactive environment. To achieve that semantic

web technology is being used [4,11,12]. An ontology represents the knowledge while intelligent agents interact with it in order to assist the user's educational needs. The whole purpose is to develop an e-learning platform for the telepediatrics and telemedicine training needs. The software also includes an implementation of a telepediatrics' program to monitor a child's electrocardiograph from distance.

2 Design

Based on semantic web technologies a Telepediatrics web application is developed. The system consists of a telepediatrics ontology which contains the telepediatrics knowledge. It is supported by intelligent agents that implement the interaction between the web interface and the ontology. A web service is developed in order to transfer the medical information produced by an electrocardiograph to the ontology. The web service also interacts with the agents.

The use of semantic web technologies to build this system ensures the following properties: automation, adaptability to user demands, re-usable information.

2.1 System Description – Application Architecture

The system's design is presented in Fig. 1. An Owl – RDF Ontology constitutes the conceptual model for the basic terms of telepediatrics. It is a multi-agent system consisting by two types of intelligent software agents, JAVA and PHP agents. JAVA agents are developed using JADE and JENA libraries. They communicate with each other exchanging ACL messages. The interaction with the ontology is succeeded using both JENA methods and SPARQL queries.

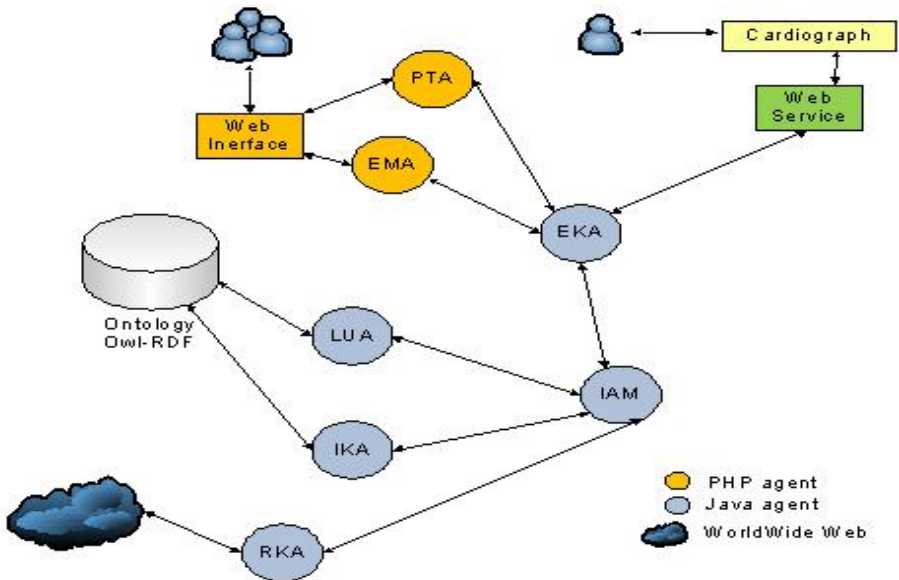


Fig. 1. Telepediatrics Application Architecture

PHP agents are responsible for establishing the connection between the web interface and the Java agents, using xml-rpc calls. They are also in charge of personalizing the outgoing information based on each user's needs.

The web interface is mainly developed in php. HTML, AJAX, Javascript were also used to build a friendly environment and to establish an adequate speed for the application. The system supports multiple users.

The electrocardiograph is used in order to assess the patient, whereas a web service is developed to transfer those exams to the application so that the authorized doctor is able to see them. The web service interacts with the java agents using xml-rpc calls.

2.2 Ontology

The telepediatrics ontology is designed to meet the educational purpose of the current application. Therefore, it contains all the essential terms of the knowledge domain as also their relations. The basic telepediatric classification is presented in Fig. 2.

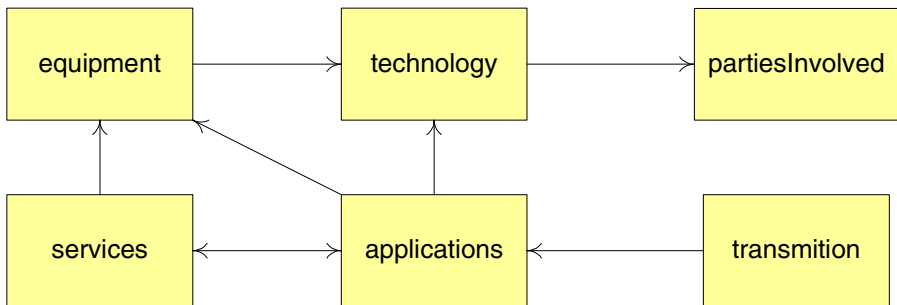


Fig. 2. Basic Telepediatrics' ontology classification

This includes categories whose entities are related according to one or more general criteria. The total number of the entities is prohibitory in order to make a full analysis. There are six main categories in the Telepediatrics schema presented below:

- **Applications:** The class contains the main telepediatrics applications. It constitutes the basic entity of the current ontology. The applications' hierarchy is based on different criteria, so that most probable relationships between them are supported. Initially, they are separated based on the connection model that is required for every one of them. We can have real time or asynchronous applications, whereas some of them are supported by both types of services. Additionally, they are classified based on their conceptual meaning and the telepediatrics' subject to which they are related. The goal of the current hierarchy is to help relate the terms, so they can be easily searched. In Fig. 3., the main applications organization is presented.
- **Technology:** This class contains all the technology (connection, communication types, computer infrastructures, network, data technologies are only some of the subclasses) that is needed to the above applications.

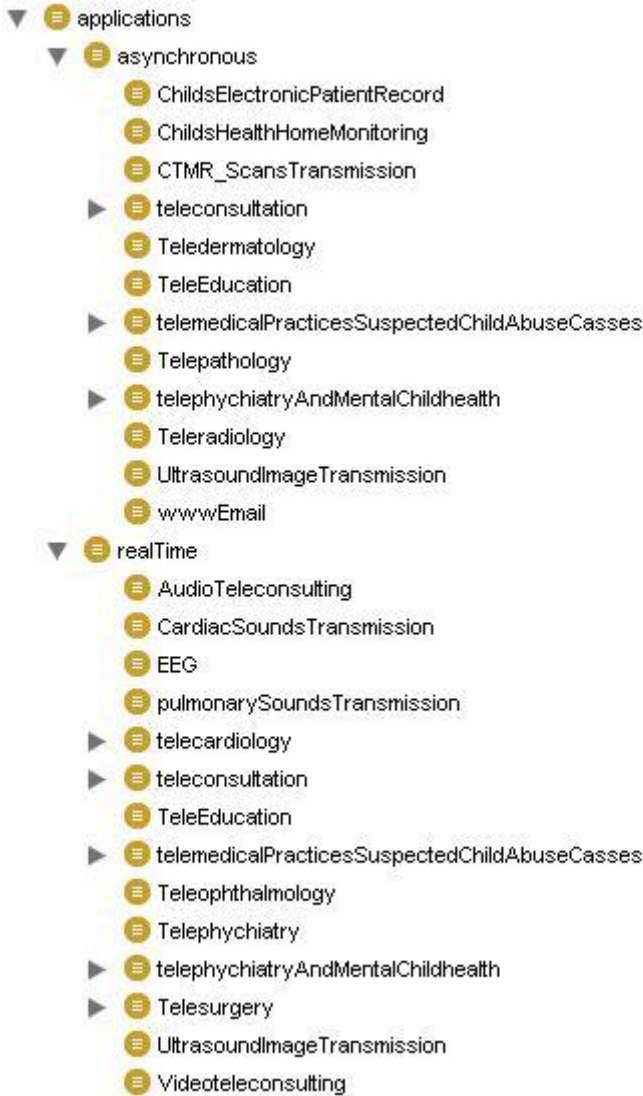


Fig. 3. Basic Applications' hierarchy

- **Equipment:** The current class consists of the equipment (medical devices, computer infrastructures and other electronic equipment) that is used for applications.
- **Services:** All kind of services are concluded both the ones used by the applications and some applications themselves that are services such as teleconsultation.
- **Transmission:** It contains applications and other classes referred to transmission of information.
- **Parties Involved:** The people and departments on which the use of these applications has consequences.

2.3 Agents

The list of the designed agents for our software and their functions can be studied in the following table (Table 1).

Table 1. Designed agents with their functions [8]

BUILDDED AGENTS	
NAME	FUNTIIONS
EKA (Exchange Knowledge Agent) Java agent	Establishes a connection in order to accomplish exchanges of information, search and insertion orders, between PTA, EMA and LUA, IKA agents.
IAM (InterAgent Manager) - Java agent	It is in charge to coordinate the activity and message interchange between the JAVA agents (EKA, LUA,IKA, RKA).
LUA (Look Up Agent) Java agent	Find the data into the ontology according to semantic definitions, search syntax and filter rules stated by PTA and EMA .
IKA (Insert Knowledge Agent) - Java agent	Insert the data specified by PTA into the ontology.
RKA (Retrieve Knowledge Agent) Java agent	Retrieves more information from the internet according to users demands and updates with it the current ontology
EMA (Event Manager Agent) - Php agent	Creates and sends search orders to EKA according to users demands
PTA (Personal Training Agent) - Php agent	Responsible to help the user during his visit by proposing topics that apply to his profile. Also creates and updates each user's profile.

The Personal Training Agent (PTA) in general can find users' personal interests without bothering the users. Therefore, it is particularly suitable for personalized e-learning by recommending learning materials. The major functionality of personal training agent in e-learning is to analyze users' browsing profile and propose related study material to each of them. Most agents of e-learning perform this analysis using the best learning rules. They are set up with the method of analyzing the browsing order in advance. These agents assume that the order of the content that users have browsed has some form of relations and because of that their results are only suitable for some fields. A lot of researchers [5,6,9,10] have analyzed the relations between knowledge fields and courses, instead of browsing order relations.

The PTA tries to analyze the users' profile, which is recorded during their browsing sequence, without assuming the relationship of browsing sequence and without adopting any form of rules in advance. The proposed algorithm that PTA uses is based on

Naïve Bayesian Classifier (NBC) [1,3] which assumes that all input items are independent and performs fast analysis. These properties make it particularly suitable for our intelligent agent.

Based on the recorded browsing sequence PTA calculates a probability, which indicates how interesting each topic is considered for a specific user. This probability is influenced by the following criteria:

1. Time: the time the user has spent to every topic and it's subtopics
2. Visits: how many times the user has browsed the specific topic or it's subtopics
3. Topic's length: The size of every topic impacts the time needed for it to be read. The larger the topic, the more time is needed.

According to the above criteria the probability is formed as shown below:

$$\max \{w_{visits} \times P_{visits} + w_{timeSpent} \times P_{timeSpent}\} \quad \text{with} \quad w_{visits} + w_{timeSpent} = 1$$

and

$$P_{timeSpent} = \frac{t_n(x_n/x) + (m \times p)}{t + m}, \quad P_{visits} = \frac{v_n + (m \times p)}{v + m}$$

where:

Pvisits : The probability the topic is interesting based on the visit's number.

PtimeSpent : The probability the topic is interesting based on the time. The current probability also takes into consideration the topic's length for more accuracy.

Wvisits : The importance (weight) of P_{visits} .

WtimeSpent : The importance (weight) of $P_{timeSpent}$.

V_n : The number of the visits to a specific topic and its' subtopics.

V : The total visits to all the topics of the current user.

t_n : The time that the user spends on a specific topic and its' subtopics.

X_n : The topic's length in bytes, words or pages.

t : The total time the current user has spent reading.

X : The total length of all the learning material.

m : arbitrary number.

p : the general probability of showing interest in a topic.

Then the system calculates the issue with the highest probability and using that option recommends new teaching materials, by displaying them on the web site at the following user login.

2.4 Web Service

The web service is required to be used in every patient's computer in order to transfer the recorded electrocardiographs. Every four minutes it checks for new electrocardiographs and transfers them to the central computer where the software is located and inserts them into the ontology. In order to achieve this, the service communicates with EKA agent by sending him an insertion order.

3 Application Development – Demonstration

The Web application consists of two basic parts: i) personalized e-learning environment and ii) electrocardiograph transfer application.

i) Personalized e-learning Environment

The parts described above are combined to provide a friendly e-learning environment (Fig. 4). The system dedicated to helping the user in every possible way. The menu is modified in such hierarchy so that most topics' relations are demonstrated, just by looking at it. Furthermore, every time a user chooses a topic an agent proposes learning material that is related to his choice, so that he is fully informed on it by the end of his reading. Search option is also offered in the site, in case the user needs more information than the ontology contains. In addition, every time a user logs in the system proposes topics based on his profile created during his previous visits using the algorithm proposed above.

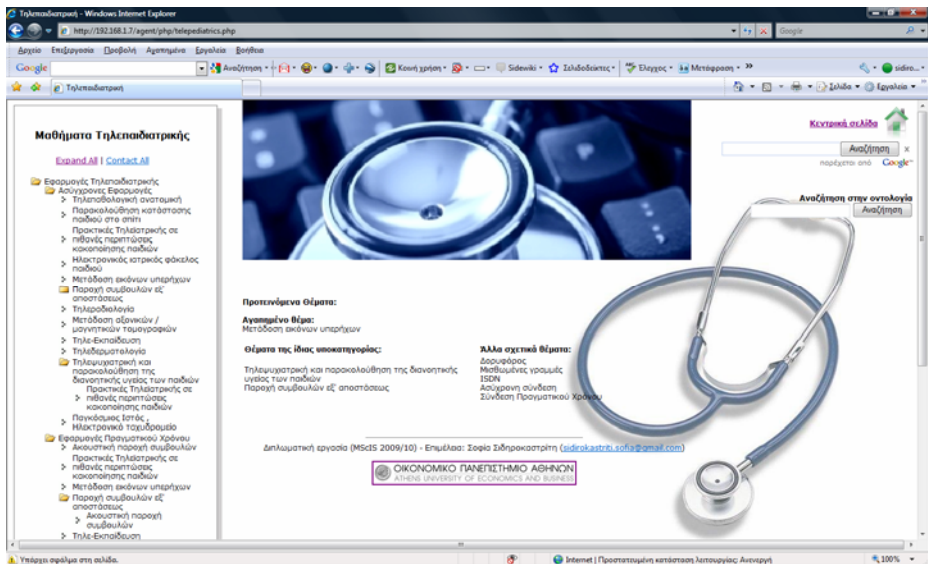


Fig. 4. Personalized e-learning Environment

ii) Electrocardiograph Transfer Application

The e-learning web application demonstrates the benefits of telepediatrics and telemedicine and can be easily used in real life. A web service is installed on patient network and the patients are equipped with electrocardiographs. The parent has to place the electrodes of the electrocardiograph in the child's body and start the electrocardiograph operation. Then the service transfers the electrocardiograms in the central Web application through which each doctor monitors his patients (Fig. 5). Doctors access rights are determined to allow each doctor to see only his patients records, so that patient's confidentiality is assured.



Fig. 5. Electrocardiograph Web Environment

4 Quality Assurance

The system that was described above refers to telepediatrics education and the quality assurance aspect is important.

The quality factors for medical websites are [2] :

- **Reliability** of the information included. The information that the website contains should come from reliable scientific sources of the current field. This point has been considered during the development. All the educational information in the proposed telepediatrics e-learning platform comes from reputable sources, is updated by the latest researches [7] and every user is allowed to explore at the same time the web for even more information.
- **Content** of the website. All content should be presented to the readers and be sorted in such way that is understandable and precise. No faulty promises should be given to the reader. As far as the application is a result of academic research there was no purpose to hide or twist given information. Furthermore, the way the content is being sorted and grouped serves even better the current aspect.
- **Disclosure.** The purpose of each smart website should be clearly clarified. Also, the users should be informed about which personal information is monitored and for what reason. The current implementation uses personal information such as user's browsing profile, which is sensitive, but only to help the user through his reading. Therefore, when the user subscribes the system informs that he/she will be monitored and also for the reason that this is necessary. User's permission is required for subscribing.

- **Links** must be included so that the user can compare the information and make sure that is trustworthy. Also, the links should be easy to explore and connected with one another. The current implementation offers the opportunity for web search, which allows the user to compare the information included with any link he wants. Every link included in the system proposes relative links allowing the user to connect to all the included knowledge.
- **Design** should be simple and friendly to the user which was the goal from the beginning for the current system.
- **Interactivity – Caveats.** Feedback and alerts are necessary to improve the website. Alerts are already included. The website also contains information about the creators where feedback can be sent by every user.

5 Conclusions and Outlook

Herein, we have succeeded in designing and implementing an e-learning platform for telepediatrics education on the Semantic Web. A telepediatrics ontology and various proposed e-learning agents have been developed and extensively tested. A quality assurance plan was adopted during platform evaluation. Furthermore, we have experimented successfully with developing a web service that acquires and transfers electrocardiograms through the Internet and inserts them into the ontology. The transformation of the presented e-learning platform to allow for teleworking, tediagnosis and the full range of telemedicine services for children, assuring quality and patient safety, is currently under investigation.

References

- [1] Moore, A.W.: Naïve Bayes Classifiers. University Lectures. School of Computer Science. Carnegie Mellon University (2004), <http://www.autonlab.org/tutorials/>
- [2] Gountava, E.: Quality of health information on the web. TEI of Epirus/ Ioannina. Greece
- [3] Meisner, E.: Naive Bayes Classifier example (November 2003), <http://www.inf.u-szeged.hu/~ormandi/teaching/mi2/06-naiveBayes-example.pdf>
- [4] Kappel, G., Pröll, B., Reich, S., Retschitzegger, W.: Web Engineering - The Discipline of Systematic Development of Web Applications, ch. 14.1. Wiley, Chichester (May 2006)
- [5] Lin, J.-L., Chen, M.-H.: An Intelligent Agent for Personalized E-Learning. ACM, Shih-Hsin University (2008)
- [6] Oriana, L., Semeraro, G.: Student Profiles to Improve Searching in E-Learning Systems. International Journal of Continuing Engineering Education and Life Long Learning 17(4), 392–401 (2007)
- [7] PEDITOP: Promotion of advanced educational innovations for training in Paediatrics, <http://www.peditop.com>
- [8] Ferrer-Roca, O.: Adaptive and Adaptable distant Telemedicine training. Faculty Medicine University of La Laguna (2003)

- [9] Ahu, S., Mobasher, B., Burke, R.: Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search. *IEEE Intelligent Informatics Bulletin* 8(1), 7–18 (2007)
- [10] Bo, S., Chen, A.: An Examination of Learning Profiles in Physical Education. *Journal of Teaching in Physical Education* 26(2), 145–160 (2007)
- [11] Berners-Lee, T., Miller, E.: Semantic Web. *ERCIM NEWS* number 51. W3C (October 2002)
- [12] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* (May 2001)

Web Applications and Public Diplomacy

Antigoni Koffa and Anastasia N. Kastania

Department of Informatics,
Athens University of Economics and Business, Greece
koffa@aueb.gr, ank@aueb.gr

Abstract. This article exploits the discipline of Public Diplomacy taking into account the contribution of Web applications to practice it. A web application for public websites assessment is implemented. This application is a tool designed to help the government and public institutions to find their omissions and their by any chance errors in their websites, so that they can adjust and enrich them, aiming at providing excellent service to the citizens.

Keywords: public diplomacy, public web sites, Certification Frameworks, e-government, web application, smart assessment.

1 Introduction

This study examines how Web technologies and applications can contribute to the establishment of properly structured websites designed for Public institutions, aiming at the better utilization from citizens and simultaneously the appropriate web presence of a country. A web assessment tool for Public institutions' web sites was designed and developed. The objective of this assessment tool is the promotion of the nation's "brand name" via provision of sufficient, effective and easily accessible information from foreign citizens.

As reported by the Center of Public diplomacy, "Public Diplomacy intends to influence and inform public opinion in other countries using cultural exchanges, publications, television and radio" [26]. Public Diplomacy is essential for the nation in order to disseminate its views and practice politics [1, 10, 13, 16, 20, 22, 23, 26, 29]. A direct means, which contributes to the creation of public opinion, is the Internet. Thus, the contribution of citizens is strengthened and virtual communities are created. These are virtual societies of citizens, with participants in international level, who exchange opinions and interact socially. Web applications are complex, interactive software systems that deliver personalized services accessible via different devices, offer the option of user transactions and usually store data in a database [2, 9, 12, 25, 28, 31]. Web applications can also be used to the profit of Public Diplomacy.

The designed web application uses various online web analysis tools which are freely available on the Internet. The choice of the tools was made conducting extensive Internet research. An interface for the insertion of the URLs of Public websites to

be evaluated was implemented and then an automatic evaluation platform through the web application's implementation. The assessment is giving an overall score for each website depending on whether it meets or not the criteria examined.

1.1 Public Diplomacy

Public diplomacy is about the impact of social attitudes in the development and implementation of external government actions. It covers aspects of international relations beyond traditional diplomacy, such as influencing public opinion in foreign countries, the interaction of different groups and interests in one country with those in other countries, recording the impact of international relations in politics, communication among diplomats and foreign correspondents in a process of intercultural contact [1, 10, 13]. Central concept in public diplomacy is the cross-border flow of information and ideas.

States implement the strategy of nation branding (creating "brand name" for a state) [5], using the marketing and Public Diplomacy to achieve their goal. The difference between Public Diplomacy and nation branding is that the first is about creating relationships, and the latter about promoting a nation's name. Public diplomacy is the process of creating the message that a country wants to promote abroad and then it has to apply appropriate techniques of persuasion - depending on the intended audience - and appropriate tools to interpret and analyse the reactions of the public towards the messages it receives.

Nowadays, the main instrument of Public Diplomacy is the Internet. The lack of national boundaries and restrictions on the content, the easily accessible information and the access speed to the content, make the Internet a powerful tool for shaping public opinion. Nations have the opportunity to develop and strengthen their diplomatic positions, by giving arguments and directly affect the public worldwide. As long as all countries use the Internet as a way to promote themselves, it is essential that the information is provided by a reliable service. Typical examples are services that use modern technology (i.e. use of satellites) to disseminate the views of the countries to the global Internet audience.

1.2 E-Governance

E-Governance is the utilization of information and communication technology in public administrations, in connection with essential organizational changes and new skills of existing human resources, in order to improve the provided services, the internal processes, as well as the wider management of public sector [4, 6, 18, 24].

E-Governance allows the development of governmental effectiveness while at the same time it ensures access without discriminations in citizens and enterprises. E-Governance provides the following benefits:

- Increases the productivity of Public Administration
- Reduces the service cost for citizens and businesses
- Reduces the communication and service time with the public (call centers, front desk)

- Provides better coordination between organizations - common standards
- Additional benefits are the reorganization of processes, which are gradually optimized using the information and communication technologies
- Offers the possibility of new services provision and working methods (e.g. teleworking, forums, consultations, tele-education)
- Develops better services for citizens and businesses
- Improves transaction security and data integrity
- Services are provided continuously (e.g. all day)
- Services do not discriminate citizens based on gender, color, age
- Offers the potential for new services provision (e.g. e-Democracy [21])

1.3 Virtual Communities, Internet and Practice of Public Diplomacy

Virtual communities are groups of mutual support, information, communication and general social interaction that emerged through Internet development [3, 8, 17, 27, 33]. Public diplomacy should examine the trend of public opinion on the Internet. New technology combined with the available software facilitates the simultaneous and automated monitoring of multiple blogs in real time helping to find and deal with an "unsatisfactory" opinion dissemination on the Internet. Public diplomacy is not limited to promoting a national image. In the era of new technologies and a myriad of information channels, the ultimate reliability of information, in particular the Internet, invaluable when it comes from people who belong to same potential group whose members share understandings. Therefore, effective public diplomacy aims to recruit some members of virtual communities rather than simply providing information and other data defining an appropriate set of ideas that might affect members of the community. Public diplomacy should establish a connection to an international audience, addressing global issues and not to be maintained to promote the national image.

1.4 Certification Frameworks for Public Websites

The purpose of a Certification Framework for Public Web sites is to improve different design parameters and operation of public websites such as naming, organization and layout of content, navigation, the search capabilities, content accessibility, user authentication and the protection of personal data. This will better support the citizens and companies in the discovery and exploitation of public information, by setting rules, standards and specifications for the design, development and maintenance of public websites. It also aims at encouraging the public administration to provide citizens, businesses and other entities interactive e-Government services, defining the rules for providing and supporting these services. The adoption of a Certification Framework offers opportunities for improving public service transactions, increase the performance quality, reduce the execution time and improve electronic services reliability. Usability, culture and communication, graphical user interfaces, quality, performance, security are included in a model for the creation of Certification Frameworks and quality seals [11].

2 Features of Intelligent Websites for Public Diplomacy

The main feature of intelligent applications is behaviour change based on users input. Intelligent information processing (Fig. 1) can be achieved through algorithms for search, recommendations, groupings, classification, and the combination of classifiers [15].

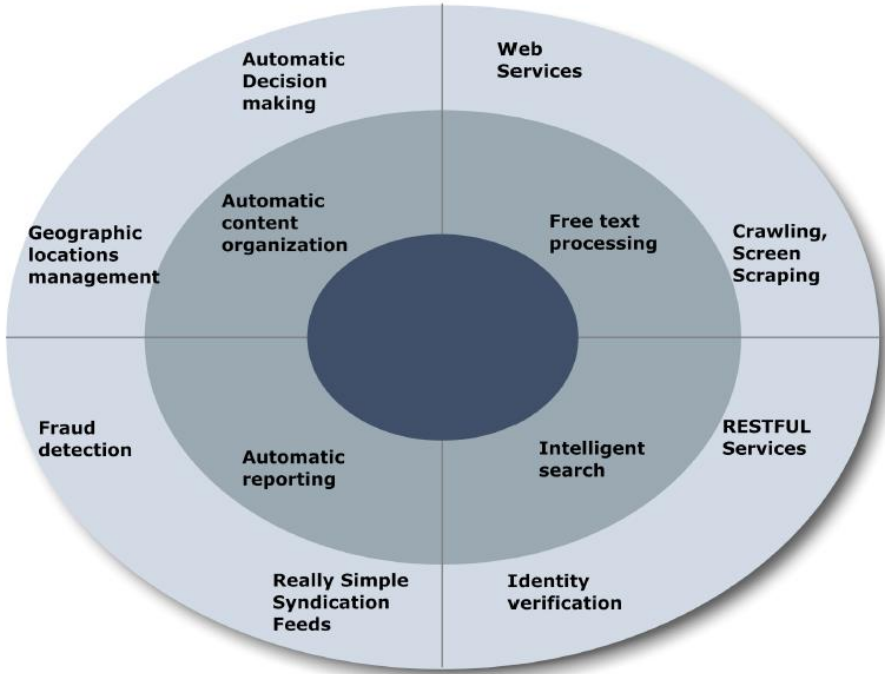


Fig. 1. Aspects involved in shaping a smart web application

Public diplomacy applications that can benefit from intelligence are social networking sites, mashups, portals, wikis and media sharing sites. In social networking sites techniques available for social network analysis (including the dynamics of network and behaviour) and for cultural analysis (consensus analysis, standard multivariate tools).

Artificial intelligence and machine learning (decision trees, rule-based methods, neural networks, support vector machines etc.) can be incorporated in an intelligent web application [15, 30]. Reference structures are dictionaries, knowledge bases and ontologies. The application fields are smart searching, recommendation engines, automatic content organization, matching discovery on social-networking sites, news group discussions organization, shared bookmarks management, spam filtering and e-mails categorization based on content [15].

3 Quality Assurance Frameworks for E-government Websites

Some of the most frequent usability problems in public institutions' web sites are the difficulty in finding the needed information or service, the difficult operation of e-services, the language understandability and the need for better service regarding the e-service provided on the website. Additionally there are issues like back office efficiency and system reliability that substantiate the necessity of a quality perspective in the development and provision of e-government services. There are four main categories that summarize the key factors that ensure quality in e-government portals [7, 14, 19]:

- Back office system performance: it includes factors typically found in quality models for traditional government services.
- Site technical performance: it includes factors concerning the technical performance of the site such as security.
- Site Quality: it includes factors concerning the site usability, information resources and interface.
- Citizens' overall satisfaction: it refers to the overall level of quality perceived by each citizen against their expectations.

Other quality factors [7, 14, 19] are reliability, linkage, content, ease of use, self-service, personalization and citizen service.

4 Website On-Line Assessment Application

Several website assessment tools are available online (Fig. 2). The proposed php Web application imports the URLs of the Public Web sites that someone might want to evaluate. The criteria selected for the evaluation accompanied with the related tools (Table 1, Fig.3) are stored in an XML file for subsequent access by the application. To find and understand the structure of the POST request for each tool Wireshark was used [32].

The interface operates as follows: First, there is a form in which the user fills the number with the websites to be evaluated. Depending on this number, the software generates input fields for the URLs of the websites. Having entered the URLs, the analysis is performed sending the URLs to the online tools. The URL is automatically inserted in the corresponding entry form of the online tools and is also submitted automatically. All these functions are not visible to the application user. The outputs of each tool, which are essentially html pages, are stored in txt files. Each file is then parsed and a certain character sequence is located, which differs for each tool-criterion, and determines whether this criterion is satisfied or not (Fig. 4). Depending on whether or not a criterion is met, this is scored with 1 or 0, respectively. The total score for each URL is kept in a table.

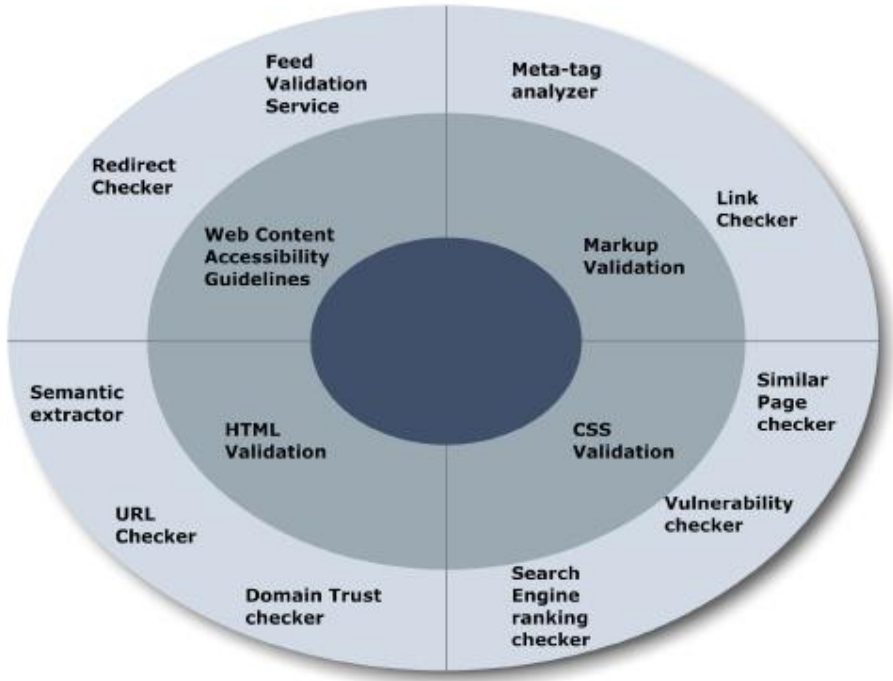


Fig. 2. Available commercial and free tool categories for on-line website assessment

Table 1. Free tool categories for on-line website assessment

Functionality	Tool URL
W3C Link Checker	http://validator.w3.org/checklink
W3C Markup Validation Service	http://validator.w3.org/
W3C CSS Validation Service	http://jigsaw.w3.org/css-validator/
Semantic extractor	http://www.w3.org/2003/12/semantic-extractor.html
W3C Feed Validation Service	http://validator.w3.org/feed/

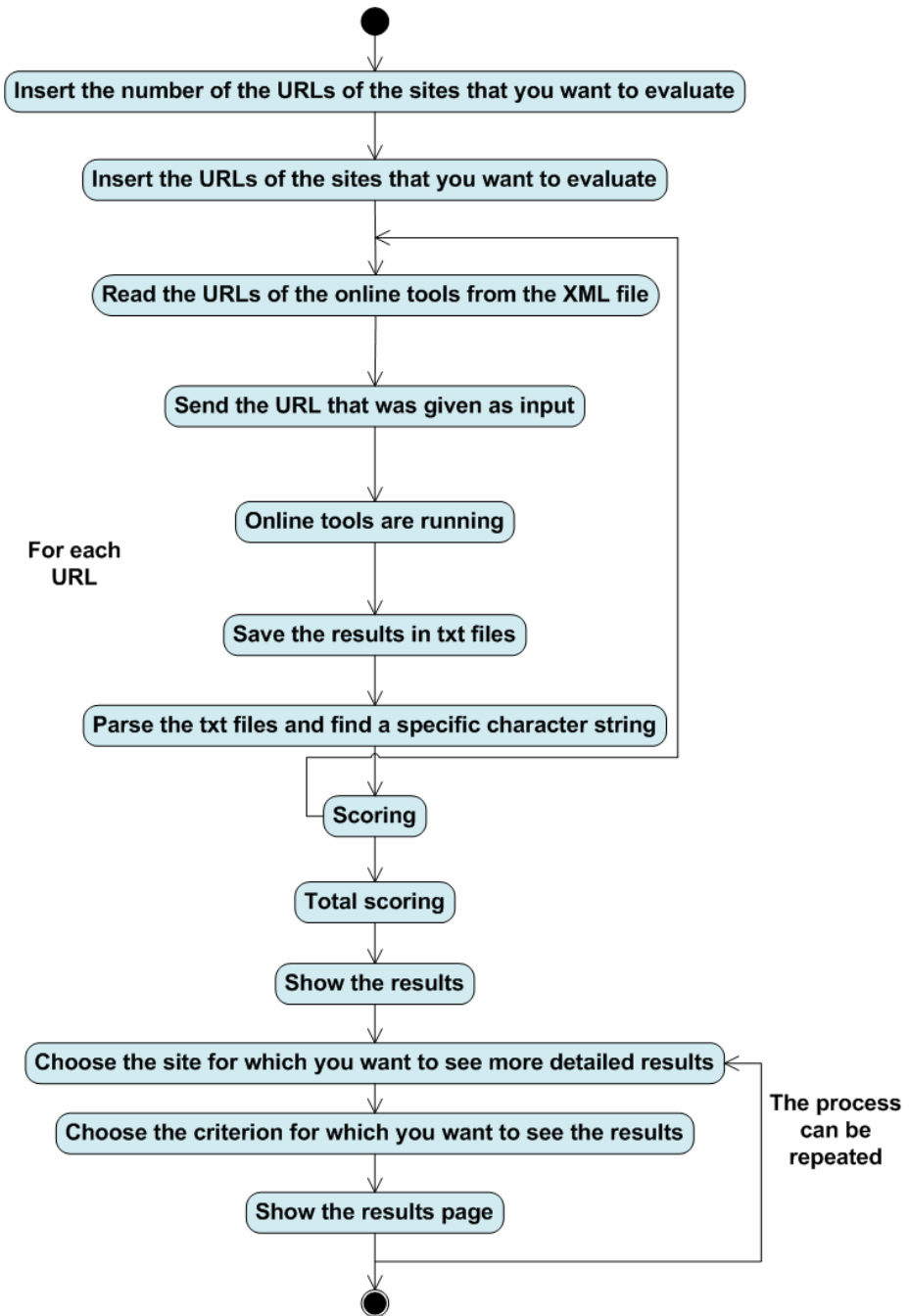


Fig. 3. Activity diagram of the developed assessment tool

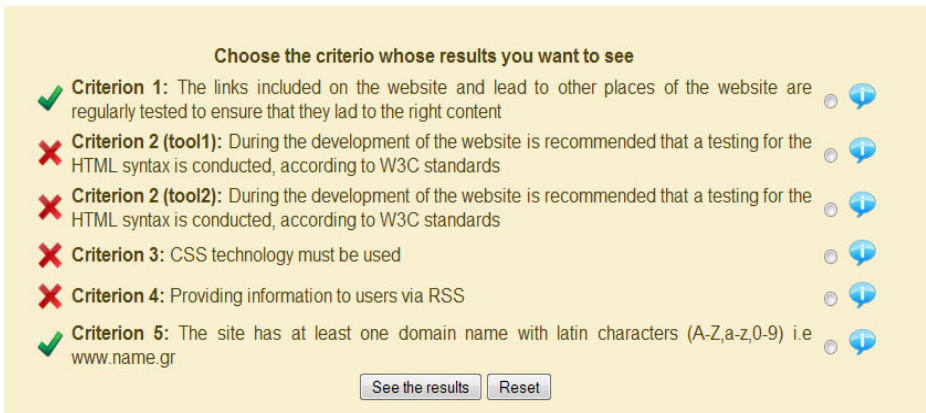


Fig. 4. Overall results for a specific site using the developed tool

5 Conclusions and Outlook

We have succeeded in developing an on-line web based assessment tool for public websites evaluation. The proposed tool can be used to create an assessment list for various web sites that provide information and services to the public. The tool can propose public website improvements through massive processing of the outputs of various existing tools for on-line website assessment. Future work will focus on transforming this tool into an open source intelligent web based platform incorporating both social network analysis and cultural domain analysis techniques with the ultimate goal to serve as an on-line instrument for improving the practice of Public Diplomacy in the cyberspace.

References

1. Batora, J.: Foreign Ministries and the Information Revolution: Going Virtual? (Diplomatic Studies). Martinus Nijhoff Publishers (2008)
2. Brajnik, G.: Quality Models based on Automatic Webtesting. In: CHI 2002 Workshop: Automatically Evaluating Usability of Web Sites, Minneapolis, USA (2002)
3. Brown, G.: Social Media 100 Success Secrets: Social Media, Web 2.0 User Generated Content and Virtual Communities – 100 Most Asked Mass Collaboration Questions. Emereo Pty Ltd. (2009)
4. Budd, L.: e-Governance: Managing or Governing. Routledge, Taylor & Francis Group (2009)
5. Dinnie, K.: Nation branding: Concepts, Issues, Practice. Butterworth-Heinemann, Butterworths (2007)
6. Garson, G.D.: Public Information Technology and e-Governance: Managing the Virtual State. Jones and Bartlett, USA (2006)
7. Halaris, C., Magoutas, B., Papadomichelaki, X., Mentzas, G.: Classification and Synthesis of Quality Approaches in E-government Services. In: Internet Research-Westport then Bradford, vol. 17(4), pp. 378–401. Emerald Group Publishing Limited (2007)

8. Hinds, D., Lee, R.: Social Network Structure as a Critical Success Condition for Virtual Communities. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (2008)
9. Hitz, M., Leitner, G., Melcher, R.: Usability of Web Applications. In: Kappel, G., Proll, B., Reich, S., Retschitzegger (eds.) *Web Engineering*. John Wiley & Sons Ltd., Chichester (2006)
10. Hofstede, G.: *Culture's consequences: international differences in work-related values*. Sage, Beverly Hills (1980)
11. Kastania, A.N., Zimeras, S.: Book Web-based Applications in Health Care and Biomedicine. In: *Evaluation for Web-Based Applications*. *Annals of Information Systems*, vol. 7. Springer, Heidelberg (2009)
12. Kotsis, G.: Performance of Web applications. In: Kappel, G., Proll, B., Reich, S., Retschitzegger, W. (eds.) *Web Engineering*. John Wiley & Sons Ltd., Chichester (2006)
13. Leonard, M., Stead, C., Smewing, C.: *Public Diplomacy*. Foreign Policy Centre, London (2002)
14. Magoutas, B., Halaris, C., Mentzas, G.: An Ontology for the Multi-perspective Evaluation of Quality in E-Government Services. LNCS, pp. 318–329. Springer, Heidelberg (2007)
15. Marmanis, H., Babenko, D.: *Algorithms of the Intelligent Web*. Manning Publications Co. (2009)
16. Melissen, J., Lee, D., Sharp, P.: *The New Public Diplomacy: Soft Power in International Relations (Studies in Diplomacy)*. Palgrave Macmillan, Oxford (2007)
17. Nastos, E.: An open distance learning model in Virtual Communities for Public Diplomacy. National School of Public Administration, Greece (2008)
18. OECD, OECD e-Government Studies *The e-Government Imperative* (2003)
19. Papadomichelaki, X., Magoutas, B., Halaris, C., Apostolou, D., Mentzas, G.: A Review of Quality Dimensions in e-Government Services. In: Wimmer, M.A., Scholl, H.J., Grönlund, Å., Andersen, K.V. (eds.) *EGOV 2006*. LNCS, vol. 4084, pp. 128–138. Springer, Heidelberg (2006)
20. Richmond, Y.: *Practicing Public Diplomacy: A Cold War Odyssey (Explorations in Culture and International History)*. Berghahn Books (2008)
21. Shane, M.P.: *Democracy Online: The Prospects for Political Renewal Through the Internet*. Routledge, Taylor & Francis Group (2004)
22. Signitzer, B.H., Coombs, T.: Public relations and public diplomacy: Conceptual coverages. *Public Relations Review* 18(2), 137–147 (1992)
23. Snow, N., Taylor, P.M.: *Routledge Handbook of Public Diplomacy*. Routledge, Taylor & Francis Group (2008)
24. Soliman, K.S., Affisco, J.F.: *E-Government*. Emerald Group Publishing Limited (2006)
25. Steindl, C., Ramler, R., Altmann, J.: Testing Web Applications. In: Kappel, G., Proll, B., Reich, S., Retschitzegger, W. (eds.) *Web Engineering*. John Wiley & Sons Ltd., Chichester (2006)
26. USC Center on Public Diplomacy, <http://www.publicdiplomacy.org>
27. Vassileva, J., Grassmann, W.: A system dynamics approach to study virtual communities. In: Proceedings of the 40th Hawaii International Conference on System Sciences (2007)
28. W3C Quality Assurance Activity Statement, <http://www.w3.org/QA/Activity.html>
29. Waller, J.M.: *The Public Diplomacy Reader*. The Institute of World Politics Press, Washington (2007)

30. Web Intelligence Consortium, <http://wi-consortium.org/>
31. Wimmer, M., Kemper, A., Seltsam, S.: Security for Web Applications. In: Kappel, G., Proll, B., Reich, S., Retschitzegger, W. (eds.) Web Engineering. John Wiley & Sons Ltd., Chichester (2006)
32. Wireshark, Stanford University, <http://www.wireshark.org/download.html>
33. Wu Song, F.: Virtual Communities: Bowling Alone, Online Together (Digital Formations). Peter Lang Publishing (2009)

A Hybrid Face Recognition System for Managing Time of Going to Work and Getting away from Office

Yoshinori Adachi, Zeng Yunfei, Masahiro Ozaki, and Yuji Iwahori

Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
adachiy@isc.chubu.ac.jp

Abstract. A handy method of individual identification based on the face recognition was examined aiming at the construction of time management system of going to work and getting away from office. The foreground was easily extracted by the background difference method to acquire the facial image in the fixed place, and the face region was able to be specified by a flesh-colored extraction with the filter. Individual identification was tried by specifying the positions of the face parts, generating the characteristic vector of ten dimensions, and using fuzzy reasoning from the position in the principal ingredient score space. As a result, it turned out that a difficult group of individual identification existed. For such a group, it was confirmed that the support vector machine was effective to identification. Then, an hybrid identification method was proposed.

1 Introduction

Along with advances in globalization, the situation that various nationalities and races' people work for the same office has appeared.

In this research, in such social circumstances, the face recognition system that was able to deal also with the person with a different flesh-color and different frame aimed to manage time of going to work and getting away from office was examined. For face recognition in a work management system, because the direction of the face and the brightness of the background are fixed, facial characteristics can be decided comparatively easy. Therefore, it is thought that face recognition in a simple method is possible.

In this research, it is examined whether face recognition is possible by the principal component analysis.

2 Image Extraction

In this research, individual identification by the face image for the time management system of going to work and getting away from office is examined. Therefore, image processing is carried out under a fixed background (hue, saturation, and brightness).

Therefore, it is thought that an individual image can be easily obtained by the background subtraction [1].

In this research, when the distance of each correspondent element of the background image and the original image was bigger than that of the threshold in RGB color space, it was made to be as the foreground.

3 Extraction of Flesh-Colored Region

Appropriately cutting out the face region from the original image leads to the accuracy of the face part extraction that is processing afterwards and shortening the processing time.

In this research, the extraction of the face region was tried by using the HSV color space in which brightness and saturation were separated. Because a flesh-colored feature was different depending on the race, it was difficult to extract the face region with one filter. Therefore three filters (for yellow race, for black race, and for white race) were prepared. Table 1 shows the numerical value of each filter. And the extraction results are shown in Fig. 1.

Table 1. Filters in HSV color space according to races

Filter	H(hue)	S(saturation)
Yellow race	0~40, 160~180	50~128
Black race	5~30, 100~150	0~50
White race	0~20, 180~200	50~200

To select a filter was done as follows. (1) Select the horizontal line around cheek, which is easily selected from the top of the subtracted foreground. (2) Project pixels on the line to HS space. (3) Calculate including rate in each filtered region. (4) Select the filter that gives highest rate as the proper one.

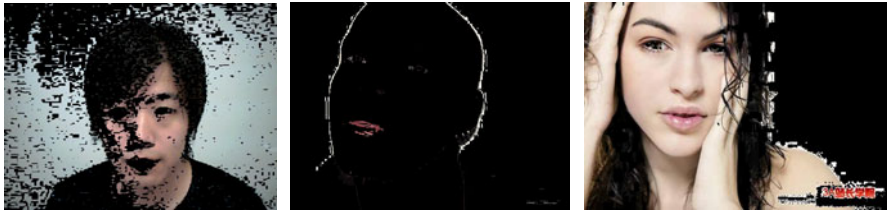
4 Detection and Feature of Face Parts

The facial characteristics are shown well by the three parts (eyes, nose, and mouth) and the externals [2, 3]. The searching range by the object detector of OpenCV [4] can be narrowed by using arrangement information on average face. The detected face part area was edging detected by the Sobel method. The positions of the face parts were specified by the processing of the Hough transform etc. An example of eyes detection is depicted in Fig. 2.

A relative position was assumed to be facial characteristics. Here, ten values shown in Table 2 were assumed to be facial characteristics.



a. Yellow race's filter



b. White race's filter



c. Black race's filter

Fig. 1. Example of extraction result of each filter

5 Identification by Principal Ingredient Score

Five face images of 11 subjects were collected at a different date. One image of each subject is depicted in Fig. 3. And the features have been extracted. Four images were arbitrarily chosen and the mean value and the standard deviation were calculated. And the principal component analysis was applied to the mean values. In that case, they were standardized because the sizes of the values were different and were analyzed.

Up to the third principal ingredient were used to adjust the contribution rate to 85% or more. The principal ingredient score [5] was calculated from data of the subject and was projected in three dimension space, and the distances between subjects in this space were calculated. Those distances are listed in Table 3.

Face recognition was preceded by fuzzy reasoning by using the membership function shown in Fig. 4. Where σ is standardized standard deviation.

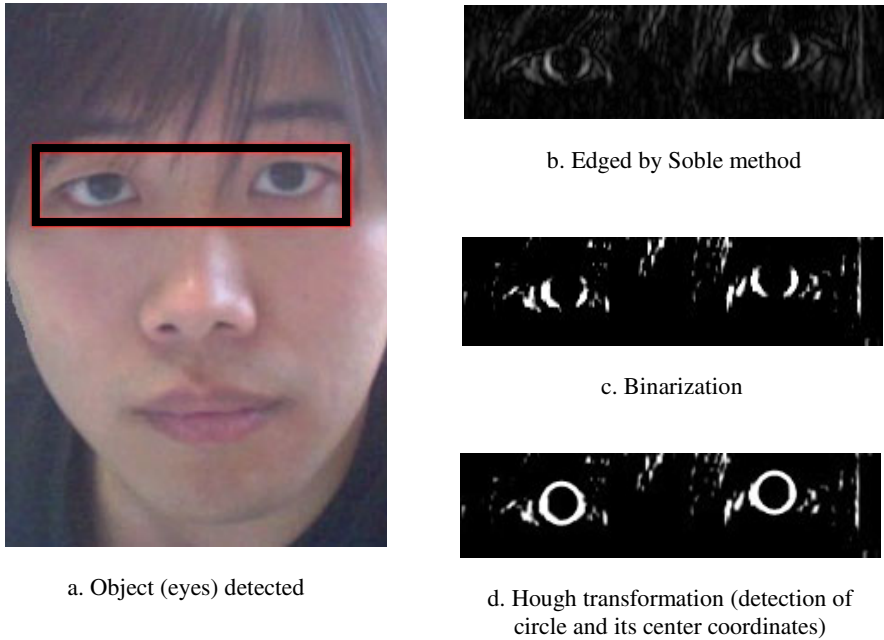


Fig. 2. An example of face parts extraction and location detection of eyes

Table 2. Features of face

No.	Feature	No.	Feature
1	Distance of nose left hole-left eye	6	Angle of right eye-left eye-nose left hole
2	Distances of midway point of both eyes-mouth	7	Angle of left eye-right eye-nose right hole
3	Distance of mouth-left eye	8	Distances of midway point of both eyes-mandible
4	Distance of mouth-right eye	9	Distance of midway point of both nose holes-mandible
5	Distance between both nose holes	10	Distance of midway point of both nose holes-mouth

When the distance between subjects is smaller than 1.0, an identification is difficult. Therefore the subject No.5, No.9 and No.10 seem to be hard to distinguish. This situation is shown in Fig. 5.

Actually, by using the fifth image, it was not possible to identify individual. However, as shown in Fig. 5, it is possible to classify the neighboring group and to restrict the searching members in the farther identification process.



Fig. 3. Images of Subjects

Table 3. Distance in three dimension principal ingredient score space

Subject No.	1	2	3	4	5	6	7	8	9	10	11
11	1.06	5.05	4.14	4.21	3.86	2.75	2.50	2.00	4.39	4.15	0
10	4.81	4.84	1.85	6.44	0.38	3.29	2.16	5.68	0.38	0	
9	5.03	4.69	2.06	6.61	0.75	3.63	2.33	5.84	0		
8	1.88	4.72	5.57	4.57	5.46	4.72	3.77	0			
7	3.38	3.81	1.95	5.90	1.98	3.18	0				
6	3.09	6.65	3.54	4.57	2.93	0					
5	4.53	4.96	1.70	6.25	0						
4	3.26	7.31	7.34	0							
3	4.98	5.18	0								
2	5.44	0									
1	0										

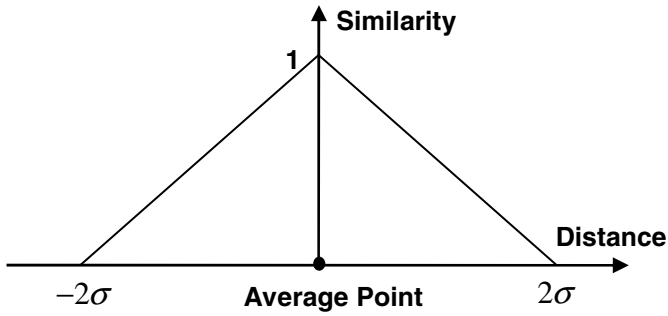


Fig. 4. Fuzzy membership function used for identification

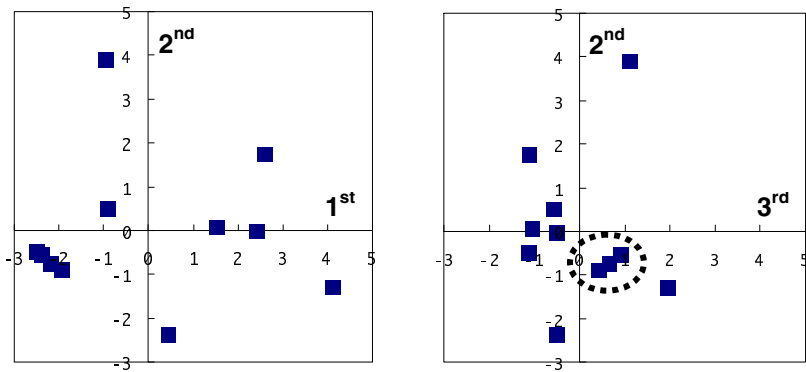


Fig. 5. Positions of subjects in three dimensional principal ingredient score space. Dotted circle indicates short distant group, Subjects No.5, No.9, and No.10.

6 Identification by SVM

The identification machine with support vector machine (SVM) is made for the group to which the distance between members is short in the principal ingredient score space. By specifying the group, the data that should be studied with SVM can be squeezed beforehand, and the identification study becomes easily, and the identification accuracy can be improved.

Three identification machines are made with SVM for Subject No.5, No.9, and No.10 respectively. They are close to each other in the principal ingredient score space and hard to be identified. The identification machine that can identify only No.5 was made by specifying two data sets. One is No.5's four feature data and the other is other subject's eight feature data. Similarly, the identification machines for No.9 and No.10 were made respectively.

The identification results using identification machines were all correct as shown in Table 4.

In this research, the number of subjects was small, only 11, SVM could distinguish all of them. However in the case of large group, the learning process must be time consuming and the identification accuracy must be not so good. Therefore we make a new identification method by using both strong points of the principal ingredient score space and the SVM.

Table 4. Identification result by SVM

Subject	Identification machine 5	Identification machine 9	Identification machine 10
No.5	○	×	×
No.9	×	○	×
No.10	×	×	○

○: Properly recognized, ×: Denied

7 Proposed Hybrid Face Recognition System

From above-mentioned examination, the hybrid face recognition system shown in Fig. 6 was proposed in this research.

This system first makes the feature vector, and projects this in the principal ingredient score space, and the identification is carried out by the fuzzy inference in this space.

In the case that the distance between subjects in the principal ingredient score space is not large enough, the identification machine of the subject who belongs to the short distance group is made with SVM and the subject is tried to be identified based on this identification machine.

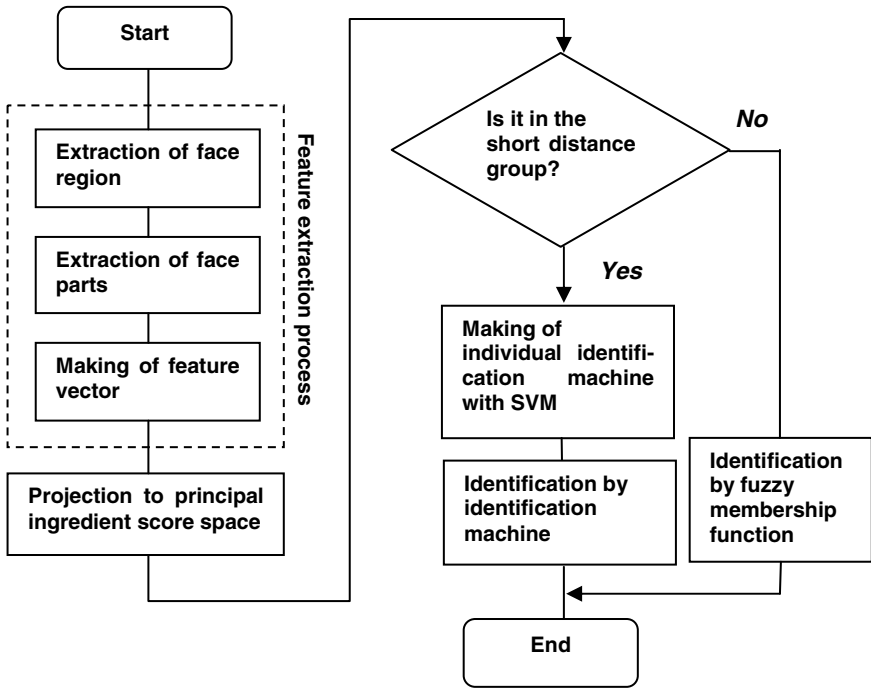


Fig. 6. Proposed hybrid face recognition system

8 Conclusion

In this research, the face recognition system that aimed to manage time of going to work and getting away from office was studied.

The following results were obtained in the process.

- i) Three filters (for yellow race and for white race and for black race) are needed for the extraction of the face region.
- ii) The relative distances of the face parts were used as the individual characteristics, and identification was carried out by the principal component analysis. However, there is a group that the distance between subjects is too short and the personal identification is difficult.
- iii) By making the identification machine with SVM after specification the group, it has been understood to be able to identify the individual by considerable accuracy.
- iv) The hybrid identification system that was combined simple principal component analysis and SVM that need study was proposed.

In this research, an experiment is done by the data of 11 subjects at most, and it is necessary to do the proof test by a practicable number of subjects.

Acknowledgment

This research is supported by JSPS Grant-in-Aid for Scientific Research, Scientific Research (C) (19500830) and Chubu University Grant.

References

- [1] Tanaka, T., et al.: Object extraction that uses non-parametric dynamic background model. In: 13th Image Sensing Symposium (SSII 2007), IN-20 (2007)
- [2] Echigo, T., et al.: Person image processing. Ohm Company (2007)
- [3] Kamiya, T., Adachi, Y.: Research on direction recognition of face. Master's thesis of graduate school of Business Administration and Information Science, Chubu University (2002)
- [4] OpenCV programming book production team: OpenCV Programming Book. Mainichi communications (2007)
- [5] Tanaka, Y., Wakimoto, K.: Multivariate statistics and analysis methods. Gendaisugakusha (1987)

Multi-Relational Pattern Mining System for General Database Systems

Nobuhiro Inuzuka* and Toshiyuki Makino

Nagoya Institute of Technology,
Gokiso-cho Showa, Nagoya 466-8555, Japan
inuzuka@nitech.ac.jp, makino@nous.nitech.ac.jp

Abstract. Multi-relational data mining (MRDM) is to enumerate frequently appeared patterns in data, the patterns which are appeared not only in a relational table but over a collection of tables. Although a database usually consists of many relational tables, most of data mining approaches treat patterns only on a table. An approach based on ILP (inductive logic programming) is a promising approach and it treats patterns on many tables. Pattern miners based on the ILP approach produce expressive patterns and are wide-applicative but computationally expensive. MAPIX [2] has an advantage that it constructs patterns by combining atomic properties extracted from sampled examples. By restricting patterns into combinations of the atomic properties it gained efficiency compared with other algorithms. In order to scale MAPIX to treat large dataset on standard relational database systems, this paper studies implementation issues.

1 Introduction

Relational pattern mining has been discussed in the framework of multi-relational data mining (MRDM) and it is suitable to use the technique of inductive logic programming (ILP). WARMR [1] is a representative algorithm. For efficiency it uses a prune method which is similar to the method used in Apriori [4]. In spite of the cut-down procedure it has limitation, because of the exponentially growing space of patterns with respect to the length of patterns and the number of relations. MAPIX acquired much efficiency at the sacrifice of the variety of patterns. It only finds patterns as combination of attributes (extended attributes in relational form), which are dynamically constructed as a set of first-order literals from examples. It is bottom-up in the sense that attributes are not given in advance but are constructed from data. It first constructs all attributes, called property items, which are appeared in examples. Then it applies an Apriori-like procedure for the property items. It succeeded to prohibit duplication of patterns in the sense of logical equivalence.

In this paper we study on implementation of MAPIX with close connection to relational database management system (RDBMS). Usual setting of MRDM

* This research is partially supported by JSPS, Grant-in-Aid for Scientific Research (C) (20500132).

bases implementation on Prolog system and data are assumed to be manipulated on main memory. In order to apply MRDM methods for wider application areas, we try to implement MAPIX on RDBMS.

MAPIX has another advantage that it needs few additional knowledge to data to be mined compared with other ILP-based system. Usual systems need to give data a language bias, which is a specification of pattern forms. Our previous paper [7] proposed a basic method to combine MAPIX algorithm with RDBMS but left some limitation. The limitation was about the form of tables.

2 Multi-Relational Pattern Mining

Consider a database R_{train} including four relational tables as shown in Fig. 1. A relation $\text{train}(\cdot)$ keeps trains and $\text{has-car}(\cdot, \cdot)$ shows cars to which each train connects. Other two tables $\text{triangle}(\cdot)$ and $\text{circle}(\cdot)$ show attributes of loads kept in cars. For multi-relational pattern mining we choose a table (a *key* in WARMR's term) in the database and try to find patterns which are appeared in many objects, more objects than a prescribed threshold, in the chosen table. Such patterns are called frequent patterns. For example we may see a pattern that many trains have at least two cars of which a car keeps a triangle shaped load and the other keeps a circle shaped load, described by,

$$\text{train}(A) \wedge \text{has-car}(A, B) \wedge \text{has-car}(A, C) \wedge \text{triangle}(B) \wedge \text{circle}(C).$$

A conventional successful algorithm WARMR finds frequent patterns in a top-down way. That is, it generates and tests patterns from simple to complex. If a simple pattern is found infrequent WARMR does not try the pattern grow longer. MAPIX has a different strategy for finding patterns. It restricts patterns into combination of basic patterns, called property items. Property items can be seen as a natural extension of attributes in first order logic and we see it in later paragraphs. MAPIX also restricts only property items appeared in sampled objects. By the restriction MAPIX does not have completeness, that is, it does not enumerate all frequent patterns. MAPIX gained much efficiency at the sacrifice of the restriction, although the successor of MAPIX has approached to complete enumeration [3].

The reason of inefficiency of WARMR is not only by the top-down method. It comes from duplication of patterns. Two differently appeared patterns may be

train	has-car	triangle	circle
t_1	t_1 c_1	c_1	c_2
t_2	t_1 c_2	c_3	c_5
t_3	t_2 c_3	c_4	
	t_3 c_4		
	t_3 c_5		

Fig. 1. Database R_{train} with four tables including key table train

equivalent logically and it is difficult to cut down all equivalent patterns. MAPIX avoids all logical duplication in searching patterns[2,3].

2.1 Property Items

We assume readers familiar with terms of logic programming. Arguments of predicates are given execution mode. Input mode is denoted by $+$ and output is denoted by $-$. We assume modes of the predicates in R_{train} as $\text{has-car}(+, -)$, $\text{triangle}(+)$ and $\text{circle}(+)$. We give output modes to all arguments of key predicate because of a technical reason.

In MAPIX, predicates are classified into two types. Predicates with only input mode arguments are called check predicates. Predicates including output mode are called path predicates. $\text{has-car}(+, -)$ is a path predicate and $\text{triangle}(+)$ and $\text{circle}(+)$ are check predicates. We call a literal of check (path) predicate a check (path) literal.

A path literal has a function like a mapping, a path literal derives a term as an output from input terms. A check literal has a function like an attribute. It takes some terms and describes a character, such as its shape. For example, for a train t_1 , a literal $\text{train}(t_1)$ recorded in the key table, let us imagine a set of literals, $\{\text{has-car}(t_1, c_1), \text{triangle}(c_1)\}$. The first literal $\text{has-car}(t_1, c_1)$ has a function deriving a car c_1 from t_1 and then the second literal describes an attribute for c_1 as it is a triangle.

The set of literals can be seen as an extended attribute. An extended attribute has two parts. One consists of path literals and they derive from a term to a term and the other consists of a check literal and describes a fact referring the derived terms. We call such an extended attribute a property item.

A property item is generalized by replacing terms by variables and represented as follows.

$$\text{train}(A) \leftarrow \text{has-car}(A, B) \wedge \text{triangle}(B).$$

We used a clausal formula for a property item, where a key literal is given as a head and path and check literals are combined by conjunction in its body part. For detailed terminology and semantics are given in [2].

2.2 MAPIX Algorithm

An outline of MAPIX algorithm is as follows.

1. It samples a set of examples (tuples) from a key table.
2. For each example it collects all relevant literals from database.
3. For the set of relevant literals of each example it extracts and generates all property items.
4. Extracted property items are gathered together and logically duplicated property items are eliminated.
5. It measures frequencies of property items in databases and eliminates infrequent property items.
6. By using an analogous method to Apriori[4] it enumerates all frequent patterns made by conjunction of property items.

Relevant literals of an example are literals in databases which have connection to the example as in,

- The example itself is relevant literal.
- If every input mode argument of a literal in the database is appeared in an output mode arguments of relevant literal, the literal is also a relevant literal.

The relevant literals keep all information of the example. For example relevant literals of the example $\text{train}(t_1)$ are,

$\text{train}(t_1), \text{has-car}(t_1, c_1), \text{has-car}(t_1, c_2), \text{triangle}(c_1), \text{circle}(c_2).$

The idea of relevant literal is related to the idea of saturation clauses in ILP literature [5].

3 An Implementation Combining Relational Database Management Systems

Pattern mining algorithms based on ILP approach usually treat data on main memory but not on database in storage systems, and assume an environment of logic programming, such as Prolog system. The environment has advantages that logical manipulations are easier and extracted logical patterns can be combined with other knowledge-base immediately. However, usability of such algorithms is limited because they require transformation of data into logic programming environment and large scale data can not be manageable. Furthermore, in order to control patterns to be mined, most of systems require additional information including mode and many other language bias. MAPIX is austere. It needs only table schemes and type and mode information.

3.1 An Implementation SQLMAPIX

We assume all data are kept in a database in RDBMS. Our implementation uses Prolog and ODBC (open database connectivity) interface in order to manipulate database from a Prolog program. We also assume that the Prolog program keeps database scheme and input/output-mode information of data tables. We also give the Prolog program a name of key table and a frequency threshold. Our MAPIX system, which we call SQLMAPIX, is summarized as follows.

Input in RDBMS: A database.

Input of SQLMAPIX in Prolog: A frequency threshold, schemes and attribute types and input/output modes of tables in the database, and a name of key table.

Output of SQLMAPIX: All frequent patterns appeared in DB the pattern which are derived by combining property items.

We carefully divide manipulations of MAPIX algorithm into ones in a Prolog program and ones on RDBMS. For dividing manipulation we assume that a set of relevant literals of an example is enough small to manipulate on main memory and assume that any relation table may be too large to bring on main memory. Hence we strictly prohibit to take manipulations over whole data of a table with Prolog.

We have the following procedure by these consideration. Lines headed by **RDBMS** is manipulation on RDBMS and ones headed by **Prolog** is manipulation by a Prolog program.

RDBMS: It samples and stores examples (tuples) from a key table on RDBMS.

RDBMS: It generates relevant literals of the sampled examples.

Prolog: For each sampled example the set of relevant literals of the example are transmitted to a Prolog process and transformed in a logical formula.

Prolog: It processes the relevant literals of every example in a logical formula and generates property items.

Prolog: All property items are gathered and logical duplications are eliminated.

Prolog: For each property item it generates an SQL query to test if each example satisfies the property items. The queries are issued for RDBMS.

RDBMS: It generates a transaction table in which property items that an example satisfies are recorded. It is similar to a market basket database which keeps items bought by each custom.

RDBMS: By processing using Apriori-like procedure, it generates all frequent combination of property items in the transaction database.

That is, operations to sample examples, to generate relevant literals, to generate a transaction database, and to process the Apriori-like procedure, are operated on RDBMS.

Sampling is operated on RDBMS by a standard selection operation using random ordering. Generating relevant literals is explained later. Transaction tables are generated from extracted property items. Each property item is transformed to a query to select tuples that satisfy the property item in the target table. The results are gathered into a table like market basket database. From transaction database frequent patterns combining property items are enumerated by Apriori-like method. Implementation of the method on RDBMS is investigated in [6].

3.2 Details of Generating Relevant Literals

To collect relevant literals, the following operation, can be considered,

$$\begin{aligned} \text{add-relevant-literals}(S, c, R) = & S \cup \{ R(t_1, \dots, t_k) \mid \\ & (\forall i \in \{1, \dots, k\} \text{ if } i\text{-th arg. is } +- \text{mode then} \\ & t_i \text{ is occurred in } S \text{ as } -- \text{mode and in the same type)} \\ & \wedge (\exists j \in \{1 \dots k\} t_j \text{ is appeared in the index } c) \}, \end{aligned}$$

where S is a set of partially collected relevant literals, c is an index of an argument (column) of a literal in S of $--$ mode, and R is a table of which tuples are added as

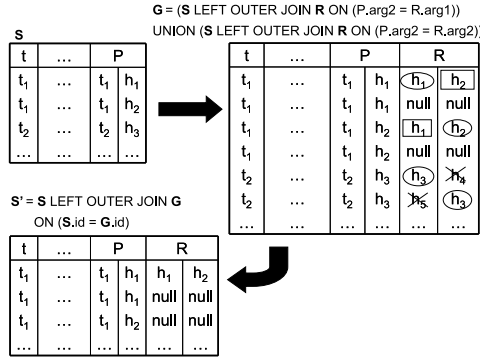


Fig. 2. A procedure to join operation

relevant. This operation collects relevant literals which are connected to a literal in S at the column indexed c . When we operate this iteratively using every $--$ mode index of every literal in S and every table as R , all relevant literals are collected.

In the above a relevant literal is connected at its j -th arg. to c . When we divide by arguments connected to c ,

$$\text{add-relevant-literals}(S, c, R) = S \cup \bigcup_{j \in \{1, \dots, k\}} \text{add-relevant-literals-at}(S, c, R, j),$$

is yield, where

$$\begin{aligned} & \text{add-relevant-literals-at}(S, c, R, j) \\ &= \{ R(t_1, \dots, t_k) \mid (\forall i \in \{1, \dots, k\} \text{ if } i\text{-th arg. is } +- \text{ mode then} \\ & \quad t_i \text{ is occurred in } S \text{ as } -- \text{ mode and in the same type}) \\ & \quad \wedge (t_j \text{ is appeared in the index } c) \}. \end{aligned}$$

The procedure is given in Table [11](#). The main procedure `generate_relevant_literals(B, E)` collects all relevant literals. It controls the index c using `Open` list. It keeps index of $--$ args. which are not processed yet. `add_relevant_literals` has a function as explained above. The loop of lines 2 to 4 repeats the function of `add-relevant-literals-at` for each $+-$ mode args. of R . `occur_check` confirms the condition that all $+-$ args. in R must be connected to somewhere in S . Although a row in S is replicated by the join in line 4 of `add_relevant_literals`, `occur_check` may delete all rows for a row in S and loses information. Line 5 of `occur_check` brings back the information lost from the original S using the **OUTER JOIN** operation.

[Fig. 2](#) illustrates to collect relevant literals. S is a table which partially collects literals. In S , t is a target column. Table P is a path predicate that involves a $+-$ mode arg. and a $--$ mode arg., and is joined in S . Also, it is assumed that all args. have the same type. Let us consider to join table R . It involves two

Table 1. A procedure to generate relevant literals

```

generate_relevant_literals( $B, E$ )
input  the set  $B$  of table names with their schemes, column types and
        modes; the key table  $E$  in  $B$  with its scheme and column types;
output the table  $S$  of relevant literals;
1. Let  $S \subseteq E$  the table of sampled examples from key table  $E$ ;
2.  $\text{Open} :=$  the set of all columns of  $E$ ; % all columns are --mode.
3. While  $\text{Open} \neq \emptyset$  do
4.   Choose a column  $c$  from  $\text{Open}$ ;
5.   For each table  $R$  in  $B$  do
6.      $S := \text{add\_relevant\_literals}(S, c, R)$ ;
7.     If some rows of the new column
8.       for  $R$  have values that are not null then
9.       Add all new --mode column for  $R$  to  $\text{Open}$ ;
10.     $\text{Open} := \text{Open} - \{c\}$ ;

add_relevant_literals( $S, c, R$ )
input  the partial table  $S$  of relevant literals, a column  $c$  from  $\text{Open}$ ,
        a table  $R$  in  $B$ ;
output the updated partial table  $S$  of relevant literals;
1.  $G := \emptyset$ ;
2. For each +-mode column  $d$  of  $R$  do
3.   If  $c$  and  $d$  are the same type then
4.      $G := G \cup S$  LEFT OUTER JOIN  $R$  ON  $S.c = R.d$ ;
5. If  $R$  has more than one +-mode column then
6.    $S := \text{occur\_check}(G, R, S)$ ;
7. else  $S := G$ ;

occur_check( $G, R, S$ )
input  the table  $G$  of relevant literals, a table  $R$  in  $B$ ,
        the partial table  $S$  of relevant literals;
output the table  $S$  of relevant literals;
1. For each +-mode column  $e$  of  $R$  do
2.    $S_e := \bigcup_{\substack{i \in \text{col. in } S \text{ of the} \\ \text{same type with } e \\ \text{and --mode}}} \text{projection of columns } t \text{ and } i \text{ in } S$ 
3.   %  $t$  is a column originated in key table
4.    $G :=$  selection of  $G$  WHERE  $G.t = S_e.t$  AND  $G.e = S_e.i$ 
5.  $S := S$  LEFT OUTER JOIN  $G$  ON  $S.id = G.id$ ;

```

columns of +-mode. Therefore, G is generated by union of two results of join on the two columns. Finally, it deletes the rows where term of each columns in table R do not occur in S and the result is generated in S' . The lost rows are recovered with null values.

Operations above are executed by SQL manipulation issued in a Prolog program. From the resulted table, rows for each examples are transmitted to the Prolog process and transformed into logical formulae. Then property items are extracted from the formulae using the procedure in [2].

4 Experiments

We examined SQLMAPIX using two datasets. The first is for grammar structure of English. The experiment is due to compare runtime with original MAPIX. The dataset were prepared in [2], and includes information of 3369 English sentences from Wall Street Journal. Sentences are analyzed and grammatical structure are expressed in tables, which are similar to *has-car*. POS-tag information is also used.

Using this dataset we examined to generate a transaction database and measured runtime sampling 100, 1000 or 3369 examples. Note that sampling is used for extraction of property items but a transaction database is generated for all examples. Table. 2 is the result. The runtime in Table. 2 is the average of 10 trials, but 5 trials for 1000 and 3369 examples in MAPIX and 3369 in SQLMAPIX. SQLMAPIX was found to process faster than the original MAPIX.

Table 2. Runtime of MAPIX and SQLMAPIX using English sentences

Examples sampled for extracting pr. items	runtime (sec.)	
	MAPIX	SQLMAPIX
100	726	277
1000	7,283	1,348
3369	24,853	4,231

Another experiment uses the dataset named Mutagenesis which has been used for ILP problems. This examines SQLMAPIX for complex table schemes. It has predicates including two or more +-mode args. The datasets includes a predicates **atom** and **bond** which represent molecules' structure. We used 230 examples. To generate transaction database SQLMAPIX took about 15 minutes while original MAPIX executed in 2 minutes. By the experiment we can confirm the correctness of SQLMAPIX but found that SQLMAPIX has less feasibility in complex domains.

5 Conclusion

We described an implementation of Multi-relational pattern mining combining RDBMS. A difficult point is to combine information across many tables on

database. We proposed a method to extract combined information using SQL operations transmitted from a Prolog program. Multi-Relational data mining is powerful but costly in general. The proposing method showed a direction that Multi-relational mining method can be applied with large scale databases.

Acknowledgment

This work is partially supported by Grant-in-Aid for Scientific Research (c), No.20500132 from MEXT, Japan and by a Research for Promoting Technological Seeds, from JST, Japan.

References

1. Dehaspe, L., De Raedt, L.: Mining Association Rules in Multiple Relations. In: Džeroski, S., Lavrač, N. (eds.) ILP 1997. LNCS (LNAI), vol. 1297, pp. 125–132. Springer, Heidelberg (1997)
2. Motoyama, J., Urazawa, S., Nakano, T., Inuzuka, N.: A Mining Algorithm Using Property Items Extracted from Sampled Examples. In: Muggleton, S.H., Otero, R., Tamaddoni-Nezhad, A. (eds.) ILP 2006. LNCS (LNAI), vol. 4455, pp. 335–350. Springer, Heidelberg (2007)
3. Inuzuka, N., Motoyama, J., Urazawa, S., Nakano, T.: Relational pattern mining based on equivalent classes of properties extracted from samples. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 582–591. Springer, Heidelberg (2008)
4. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: VLDB, pp. 487–499 (1994)
5. Rouveirol, C.: Extensions of Inversion of Resolution Applied to Theory Completion. In: Inductive Logic Programming, pp. 63–92. Academic Press, London (1992)
6. Sarawagi, S., Thomas, S., Agrawal, R.: Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. In: SIGMOD, pp. 343–354 (1998)
7. Inuzuka, N., Makino, T.: Implementing Multi-relational Mining with Relational Database Systems. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009. LNCS, vol. 5712, pp. 123–128. Springer, Heidelberg (2009)

Recovering 3-D Shape Based on Light Fall-Off Stereo under Point Light Source Illumination and Perspective Projection

Yuji Iwahori¹, Claire Rouveyrol², Robert J. Woodham³,
Yoshinori Adachi¹, and Kunio Kasugai⁴

¹ Faculty of Engineering, 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 Engineering Student in Computer Science, ENSEIRB
B.P. 99, 33402 Talence CEDEX, France
galiannee@gmail.com

³ Department of Computer Science, University of British Columbia
Vancouver, B.C. Canada V6T 1Z4
woodham@cs.ubc.ca

⁴ Department of Gastroenterology, Aichi Medical University
Nagakute-cho, Aichi-gun, Aichi 480-1195, Japan
kuku3487@aichi-med-u.ac.jp

Abstract. In the medical imaging applications, endoscope image is used to observe the human body. As a method to recover 3-D shape from shading images, light fall-off stereo has been proposed using the inverse square law for illuminance with point light source illumination. This paper extends the principle of light fall-off stereo and proposes a new approach under the assumption of both point light source illumination and perspective projection using two images with the different size of target object. To solve the problem that target has the specular reflectance in general, removing the points with specular components and interpolating the surface can handle the target with specular reflectance. The proposed approach is demonstrated through computer simulation and real experiment.

Keywords: Light Fall-off Stereo, Endoscope Images, Point Light Source Illumination, Perspective Projection, Shape from Shading.

1 Introduction

The purpose of the medical applications using the endoscope image includes the easier observation and judgment through recovering 3-D shape from the observed image. The application sometimes includes the pathological condition of the polyp with the geometrical shape. These applications are important in the research of computer vision technology. The special endoscope with the laser

light beam head [4] or the endoscope where two cameras are mounted in the head [2] have been developed. However, these approaches are difficult to be applied to the general endoscope which is widely used in the medical institution.

Here, the general endoscope is assumed to recover the 3-D shape for the medical application. Stereo based endoscope has been proposed to recover the shape [5] which determines the corresponding point between two images based on the time difference with the change of the shape of internal organs. The approach uses the geometrical calculation for the corresponding points and uses the interpolation for remained points of the object. While Light Fall-off Stereo (LFS) approach [8] uses the inverse square law for illuminance to compute the depth of each point of the target image. This approach still uses the condition that only the light source moves while the camera is fixed. However, the actual endoscope environment uses the conditions that lighting and the observation are point light source illumination and perspective projection.

Another approach [7] has been proposed under the assumption of perspective projection. The height is updated with the modification of equation but the lighting condition still uses the parallel light source.

This paper extends LFS approach with two images to the condition that both point light source and camera are moved under the perspective projection. The approach is applied to the endoscope image environment. Results are demonstrated through computer simulation and experiments.

2 Previous Light Fall-Off Stereo

2.1 Principle

The Light Fall-off Stereo is an approach that recovers an image's depth-map from the light source using two images. The inverse square law for illuminance is used to determine the distance between the light source and the object. This law is defined as

$$I(p) = \frac{k_p}{r_p^2} \quad (1)$$

where k_p represents a constant related to the intensity of the light source as well as the reflectance and orientation of the surface point p , while r_p represents the distance between the light source and surface point p .

Two images are taken under the condition that the light source is moved from the object and the incident angle (i.e., the angle between light source direction vector and surface normal vector) at the point p stays unchanged as shown in Fig. 1.

As the angles ω stay unchanged and only take part in the determination of the coefficient k_p from equation (1), k_p is considered unchanged between the two images taken. This results in the following Eq. (2).

$$\frac{I'_p}{I_p} = \frac{r_p^2}{r'^2_p} \quad (2)$$

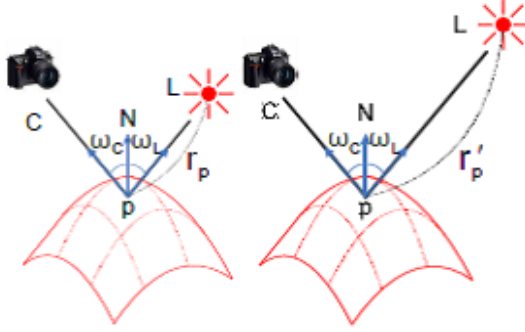


Fig. 1. LFS camera and light source settings

Then the depth r_p is computed using the following equation.

$$r_p = \frac{\Delta r}{\sqrt{\frac{I_p}{I'_p} - 1}} \quad (3)$$

where r_p represents the distance between the light source and the point p of the surface, $\Delta r = r'_p - r_p$ represents the distance between two light source positions, and I and I' represent the intensities of at any point p of the two images.

2.2 Using More Images

There exists an improvement to this approach using more than two images. The improvement searches the depth values which minimize the following energy function.

$$E = (1 - \lambda) \sum_{x,y} \sum_{i=0}^N (K_i - \bar{K})^2 * \lambda \sum_{x,y} (u_{x,y}^2 + v_{x,y}^2) \quad (4)$$

where N represents the number of images used, and K_i represents

$$K_i = \sqrt{I_{x,y}^i} (r_{x,y} + \Delta r_i)$$

$u_{x,y}$ and $v_{x,y}$ are defined as

$$u_{x,y} = r_{x+1,y} - 2r_{x,y} + r_{x-1,y}$$

$$v_{x,y} = r_{x,y+1} - 2r_{x,y} + r_{x,y-1}$$

These $u_{x,y}$ and $v_{x,y}$ are the symmetric second derivatives of $r_{x,y}$ used for the smoothness constraints.

3 Adaptation for Endoscope Images

3.1 Conditions in Endoscope Images

In the case of endoscope images, several changes to the original Light Fall-off Stereo settings need to be taken into account.

1. The light source and the camera are fixed as the same position and moved together.
2. As the endoscope progresses in a human body, the distance between the camera and the object of interest is restricted and near.

Two main problems arise from the appearance of the above constraints.

1. The intensity changes between two images are small due to the space restrictions.
2. A zoom-out effect appears between two images, that is, the object size varies between two images and it is not necessarily present in the same part of the image.

The first problem can be overcome by using an appropriate precision for the image encoding. To overcome the second problem, the use of a bilinear interpolation is applied to bring two images to the same scale.

3.2 Extension of LFS to Specular Reflection Points

To remove the aberrations obtained so far because of the specular reflectance model of the sphere, yet another improvement to the approach is included. In this approach, instead of computing the depth map using the Light Fall-off Stereo on the specular reflection points, a prior treatment is applied to the images. This treatment consists of creating a mask which will correspond to the specular reflection points of the object. The detection of the specular reflection points in the image is achieved by combining the use of a threshold based on the observation of the intensity evolution in each line/column of the image as shown in Fig. 2.

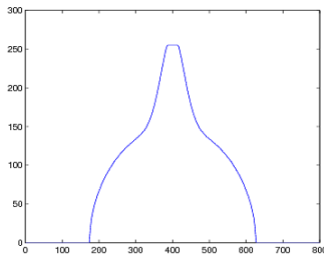


Fig. 2. Example of the intensity evolution for a line of a specular sphere

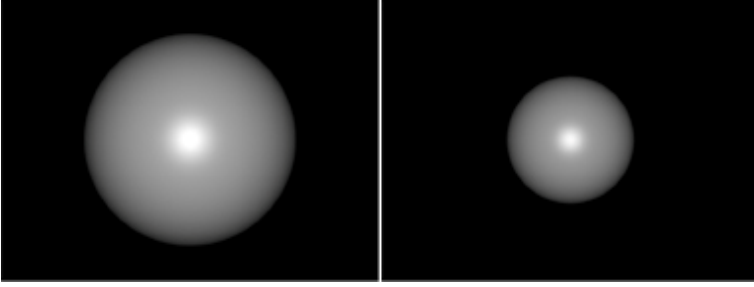


Fig. 3. Input Images of Synthesized Specular Sphere

Upon entering the specular reflection, the function gradient increases abruptly. This characteristic added to the use of a threshold leads to the construction of the mask for the specular reflection points.

Once this mask is computed, the determination of the depth is done using the Light Fall-off Stereo principle, the only difference coming from the use of the new factor which corresponds to the value of the mask for a chosen pixel. In case of the pixel being part of the specular reflection point, the value of the mask for this pixel is equal to 0; otherwise it is equal to 1. A 3-D shape is thus obtained for the object. However, because of the use of a mask, a hole is present in the 3-D shape. It is now necessary to fill this gap.

Input images of synthesized specular sphere are shown in Fig 3. Two images are transformed into almost the same size of sphere, then light fall-off stereo approach is applied to those images.

The presence of a specular reflection points is due to the fact that the light is almost completely reflected towards the camera direction. The light source and the camera being positioned at the exact same location, it means that the surface reflects the light right back to its source. Taking this fact into account, it can be considered that the specular point consists of an almost plane surface which is perpendicular to the light source direction. To fill the gap present in the 3-D shape, the creation of a practically plane patch which will be placed on top of this gap would constitute an acceptable approximation. This patch is a uniformly slightly curved patch centered on the specular point. The base of the patch is fixed at a value equal to the mean depth of the boarding points of the patch.

An illustration of the gap and the corresponding result with the patch applied is as follows.

4 Experiments

The final implementation of the Light Fall-off Stereo for endoscope images assumes the following hypothesis:

1. Two images are taken with the same camera and light source parameters,
2. The images viewpoints are similar to one another,
3. The second image corresponds to the zoomed-out image.

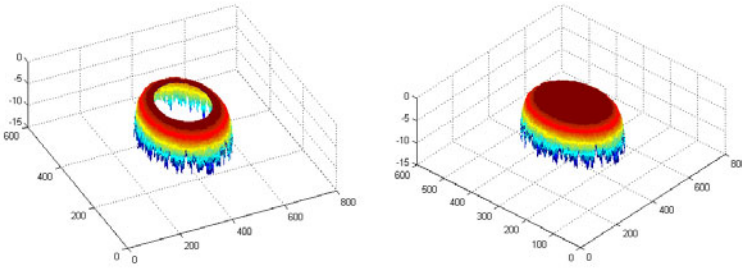


Fig. 4. Left: result before patching, Right: result after patching

Contrarily to the previous cases where the object is easily extracted from the black background, in endoscope images, the object cannot be automatically extracted without the use of a previous learning phase, or the use of image features in order to determine the matching areas of the two images. In this implementation, the matching areas between two images were selected manually.

Two images are displayed as a first step, the user is asked to select the corresponding areas in each image by clicking on the top-left and bottom-right corners of these areas in each image. Then, based on the areas specified by the user, a bilinear interpolation is applied to bring the two areas to the same scale. These two new images are then used as initial images for the different steps of the main algorithm:

1. Computing of the mask for the images,
2. Computing of the centers for the specular points present in the images,
3. Computing of the depth-map (containing gap(s) corresponding to the specular points),
4. Computing and application of the patches.

Input images obtained with the actual medical endoscope are shown in Fig. 5. In the case of computer synthesized images or photos following the Light Fall-off Stereo constraints, it is easy to have exploitable images. However, in the

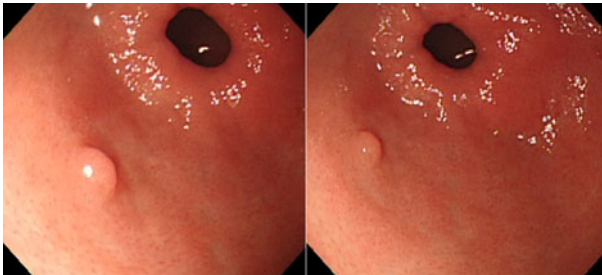


Fig. 5. Input images for the final implementation

case of endoscope images, it is not easy to obtain two images of a same area with the same camera orientation, as well as a zoom-out effect between the two images. This means that the approach is applied to the images that would correspond to each other in endoscope video sequences. The image preparation step gives some effect to the recovered results. It is also important that the matching areas are given with high correspondence in the procedure of the recovery of the 3-D shape. Obtained result for this final implementation of the proposed approach is shown in Fig. 6.

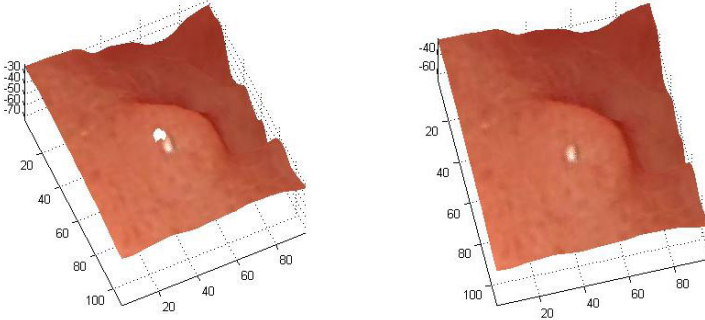


Fig. 6. Results – Left: Result before Application of Patch, Right: Final Result

The final result after the application of patch gives acceptable 3-D shape by this proposed approach.

The Light Fall-off Stereo approach uses several approximations right from the start. Further, the original conditions of this approach are different from the ones generally encountered in endoscope images, which leads to new approximations. The proposed approach extended the original Light Fall-off Stereo to the condition of both point light source illumination and perspective projection. The approach was applied to the actual endoscope image with the specular points which were recovered by the appropriate interpolation processing.

5 Conclusion

This paper proposed a new approach to extend the previous Light Fall-off Stereo approach to the condition of both point light source illumination and perspective projection. The proposed approach gives the robust result for the endoscope environment where the distance between light source and object surface is near. The effectiveness of the proposed approach is demonstrated through computer simulation and experiments with actual endoscope image.

A final great improvement would be the automatic extraction of the initial images from an endoscope video sequence, or at least a program which scans a video sequence and selects images that could potentially be candidates for a pairing so that user can make the last decision.

Acknowledgment

This research was done while Claire Rouveyrol visited Iwahori Lab. of Chubu University from ENSEIRB as the research internship. Iwahori's research is supported by JSPS Grant-in-Aid for Scientific Research (C) (20500618) and Chubu University Grant. Woodham's research is supported by the Natural Sciences and Engineering Research Council (NSERC). Kasugai's research is supported by the Japanese Foundation for Research and Promotion of Endoscopy.

References

1. Hasegawa, K., Sato, Y.: High Speed Endoscope Type 3-D Measurement System. In: Proceedings of the Asian Conference on Computer Vision (ACCV), vol. 2, pp. 887–892 (2000)
2. Mourgues, F., Devernavy, F., Coste-Manière, È.: 3D Reconstruction of the Operating Field for Image Overlay in 3D-endoscopic Surgery. In: Proceedings of the IEEE and ACM International Symposium on Augmented Reality (ISAR), pp. 191–192 (2001)
3. Hakamata, S., Miyoshi, D., Nakaguchi, T., Tsumura, N., Miyake, Y.: Reconstruction of 3D Organ Image Using Endoscope with Magneto-position-sensor. IEICE Technical Report 106(145), 13–18 (2006) (in Japanese)
4. Nakatani, H., Abe, K., Miyakawa, A., Terakawa, S.: Three-Dimensional Measurement Endoscope System with Virtual Rulers. *Journal of Biomedical Optics* 12(5), 051803 (2007)
5. Thormaehlen, T., Broszio, H., Meier, P.N.: Three-Dimensional Endoscopy. In: Falk Symposium, pp. 199–212 (2001)
6. Iwahori, Y., Sugie, H., Ishii, N.: Reconstructing Shape from Shading Images under Point Light Source Illumination. In: Proceedings of IEEE 10th International Conference on Pattern Recognition (ICPR 1990), vol. 1, pp. 83–87 (1990)
7. Yuen, S.Y., Tsui, Y.Y., Chow, C.K.: A Fast Marching Formulation of Perspective Shape from Shading under Frontal Illumination. *Pattern Recognition Letters* 28(7), 806–824 (2007)
8. Miao, L., Liang, W., Ruigang, Y., Minglun, G.: Light Fall-off Stereo. In: Proceedings of IEEE Computer Vision and Pattern Recognition, CVPR (2007)

Shadow Detection Method Based on Dirichlet Process Mixture Model

Wataru Kurahashi¹, Shinji Fukui², Yuji Iwahori¹, and Robert J. Woodham³

¹ Dept. of Computer Science, Chubu University
1200 Matsumoto-cho, Kasugai 487-8501, Japan
kurahasi@cvt1.cs.chubu.ac.jp, iwahori@cs.chubu.ac.jp

² Faculty of Education, Aichi University of Education
1 Hirosawa, Igaya-cho, Kariya 448-8542, Japan
sfukui@aecc.aichi-edu.ac.jp

³ Dept. of Computer Science, University of British Columbia
Vancouver, B.C. Canada V6T 1Z4
woodham@cs.ubc.ca

Abstract. In this paper, a new method for shadow detection is proposed. The proposed method models shadows by the Gaussian mixture model with the Dirichlet process mixture model. The parameters of the shadow model with the Dirichlet process mixture model are estimated by the Dirichlet Process EM algorithm. The proper number of the distribution can be determined through the process estimating the parameters without calculating the probability densities with two or more different number of distributions. Shadows are detected by the probability density calculated with the shadow model. The method for improving the accuracy of the result is also proposed.

1 Introduction

In many fields of computer vision, a detection method of moving objects is used as a preprocessing. Many methods for object detection have a problem that shadows are detected as moving objects. Shadows often have a harmful effect on the result. Here, the shadow is defined as cast shadow of moving object in this paper. The method for object detection requires the methods for detecting shadows and for removing them.

Methods which convert color space to remove shadows are proposed [1][2]. It is difficult for them to remove all shadows stably. Shadow detection methods modeling shadows are also proposed [3][4]. Detection methods based on a shadow model often use a probability density distribution, such as the Gaussian distribution. They model shadows by mixture distribution models and the number of distribution should be determined. It is difficult to determine the proper number of distributions.

This paper proposes a new approach for shadow detection. The proposed method models shadows by the Gaussian mixture distribution. The Dirichlet process mixture (DPM) model, which is a typical nonparametric Bayesian model, is

used for the prior distribution of the Gaussian mixture distribution. The parameters of the shadow model are estimated by the Dirichlet Process EM (DPEM) algorithm [5]. The DPEM algorithm can obtain the appropriate number of distributions without calculating probability densities with two or more different number of distributions. The proposed shadow model makes it possible to detect shadows with high performance. The method for improving the result of shadow detection is also proposed in this paper. Results are demonstrated by the experiments using real images.

2 Construction of Shadow Model

The proposed method models shadows after converting RGB color space into YUV color space. The method uses a background image. The differences of U and V components of each pixel in shadow regions between the background image and a target image become small but the difference of Y component becomes large. On the other hand, the differences of all components in regions of moving objects become large. The proposed method uses the differences of YUV components as the observed data. The shadow model of the proposed method is explained in the following.

First, the frequency distribution of the data in shadow regions is obtained. Next, the probability of occurrence for the frequency distribution is approximated by the Gaussian mixture distribution. The EM algorithm is well known as a method for estimating parameters of a mixture distribution. The algorithm needs to determine the number of the distribution *a priori*. However, it is difficult to determine the proper number of Gaussian distributions for the shadow model when the EM algorithm is used. The proposed method uses the DPM model for the prior distribution of the Gaussian mixture model and estimates the parameters of the mixture model by the DPEM algorithm.

The Gaussian mixture model with the DPM model is a very flexible model because it can define a probability model with countably infinite distributions. In fact, the distributions are truncated at a sufficiently large number of distributions because it is difficult to treat the infinite distributions. The effect of truncation to the approximation is very small because many distributions hardly contribute to the mixture distribution in many cases.

The implementation methods for estimating parameters of the mixture distribution with the DPM model have been proposed [5][7]. The proposed method uses the DPEM algorithm to estimate the parameters. It is based on the EM algorithm and can be implemented easily. The DPEM algorithm used by the proposed method is the extended EM algorithm with observing the prior probability distribution for the mixture ratio of the distribution as Stick Breaking Process (SBP)[6], which is an intuitive approach for the Dirichlet process. It can calculate not only the parameters of the mixture distribution but also the mixture rates.

The proposed method gives manually the data in the shadow regions to learn the shadow model. The frequency distribution of the data is approximated by the Gaussian mixture model with the parameters estimated by the DPEM algorithm. The DPEM algorithm uses a sufficiently large number of distributions when it estimates the parameters and the mixture rates of the mixture distribution. After estimating the mixture rates, the number of the distribution for the shadow model is determined according to the mixture rates. The obtained mixture model is treated as the shadow model.

3 Shadow Detection

After constructing the shadow model, shadows are detected by using the model. The detection process is as follows: At first, the probability that the pixel in the regions extracted by the background subtraction exists in the shadow region is calculated. Next, the shadow regions and the object regions are separated through the threshold processing. At last, the result is refined by using the detected object regions. The shadow detection process and the process for refining the result are explained in the following.

Let the number of the distribution be K . The probability that the pixel \mathbf{x} belongs to the shadow region is calculated by Eq. (1).

$$p_{EM}(\mathbf{s}(\mathbf{x})|\alpha_{EM}, \boldsymbol{\theta}_{EM}) = \sum_{k=1}^K \alpha_k \mathcal{N}(\mathbf{s}(\mathbf{x}); \boldsymbol{\theta}_k) \quad (1)$$

$$\alpha_{EM} = \{\alpha_1, \dots, \alpha_K\}, \quad \boldsymbol{\theta}_{EM} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$$

where $\mathbf{s}(\mathbf{x})$ means the data at \mathbf{x} , $\mathcal{N}(\cdot; \boldsymbol{\theta}_k)$ means k -th Gaussian distribution, α_k means the mixture ratio for k -th distribution and $\boldsymbol{\theta}_k$ means the parameters of k -th distribution. α_k and $\boldsymbol{\theta}_k$ are estimated by the DPEM algorithm.

$p_{EM}(\mathbf{s}(\mathbf{x})|\alpha_{EM}, \boldsymbol{\theta}_{EM})$ becomes large when the pixel exists in a shadow region. \mathbf{x} is regarded as the pixel in the shadow region when $p_{EM}(\mathbf{s}(\mathbf{x})|\alpha_{EM}, \boldsymbol{\theta}_{EM})$ is larger than a threshold. Otherwise, \mathbf{x} is regarded as the pixel in the object region.

After the threshold processing, the object regions are used to refine the result. Let the object region be $R_{(i)}$ ($i = 1, \dots, I$), where I means the number of the object region¹ and let the center position of $R_{(i)}$ and the variance-covariance matrix of the pixel coordinates in $R_{(i)}$ be $\boldsymbol{\mu}_{R_{(i)}}$ and $\boldsymbol{\Sigma}_{R_{(i)}}$ respectively. The probability that a point does not belong to the object region is defined by Eq. (2).

$$p_R(\mathbf{x}_{(i)}|\boldsymbol{\theta}_{R_{(i)}}) = 1 - \exp(-c \overline{\mathbf{x}_{(i)}}^\top \boldsymbol{\Sigma}_{R_{(i)}}^{-1} \overline{\mathbf{x}_{(i)}}) \quad (2)$$

where $\boldsymbol{\theta}_{R_{(i)}} = \{\boldsymbol{\mu}_{R_{(i)}}, \boldsymbol{\Sigma}_{R_{(i)}}\}$, $\overline{\mathbf{x}_{(i)}} = \mathbf{x}_{(i)} - \boldsymbol{\mu}_{R_{(i)}}$, $\mathbf{x}_{(i)}$ means the point to which the closest point in the center points of the object regions is $\boldsymbol{\mu}_{R_{(i)}}$, and c means a constant.

¹ The small object regions are treated as noises, and they are not included in $R_{(1)}, \dots$ and $R_{(I)}$.

The shadow regions are obtained again by the threshold processing using the values calculated by Eq. (3).

$$p(\mathbf{s}(\mathbf{x}_{(i)}), \mathbf{x}_{(i)} | \alpha_{EM}, \boldsymbol{\theta}_{EM}, \boldsymbol{\theta}_{R_{(i)}}) = (1 - \lambda) p_R(\mathbf{x}_{(i)} | \boldsymbol{\theta}_{R_{(i)}}) + \lambda p_{EM}(\mathbf{s}(\mathbf{x}_{(i)}) | \alpha_{EM}, \boldsymbol{\theta}_{EM}) p_R(\mathbf{x}_{(i)} | \boldsymbol{\theta}_{R_{(i)}}) \quad (3)$$

where λ ($0 \leq \lambda \leq 1$) means the parameter for the mixture ratio. $p_R(\mathbf{x}_{(i)} | \boldsymbol{\theta}_{R_{(i)}})$ is multiplied by $p_{EM}(\mathbf{s}(\mathbf{x}_{(i)}) | \alpha_{EM}, \boldsymbol{\theta}_{EM})$ to reduce the points in the object region misjudged as the pixel in the shadow region. $p_R(\mathbf{x}_{(i)} | \boldsymbol{\theta}_{R_{(i)}})$ is added to reduce the points in the shadow region misjudged as the pixel in the object region.

4 Experiments

The experiments using the real images were done to confirm the effectiveness of the proposed method. The size of each frame is 720×480 pixels and each pixel has 8bit color value for each RGB component. A PC with Core i7 and 4G Bytes main memory was used for the experiments. The threshold for the shadow detection was set to 0.01. c in Eq. (2) was set to $1/8$. λ in Eq. (3) was set to 0.99.

Four different scenes (Scene 1, Scene 2, Scene 3 and Scene 4) were used in the experiments. The input images used for the experiments are shown in Fig. 1 and the background images are shown in Fig. 2.

The results of shadow detection are shown in Fig. 3. The blue region in those figures means the object region and the red one means the shadow region. The median filter and the morphology process are applied to the results to refine them. The results show that the proposed method can detect most shadows.

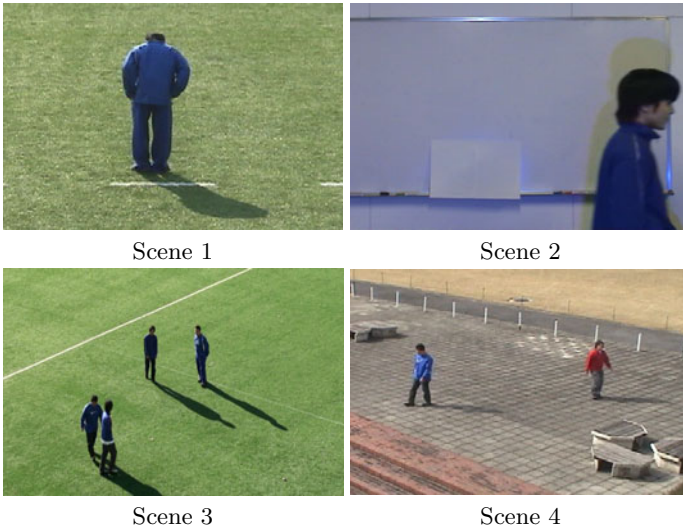


Fig. 1. Input Images

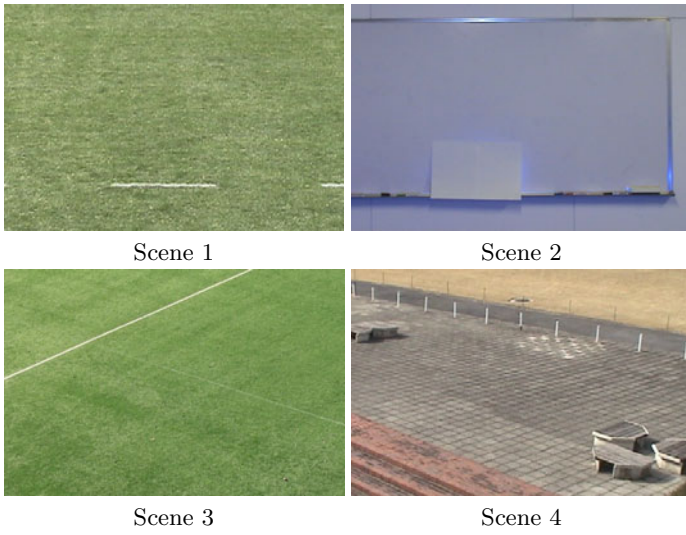


Fig. 2. Background Images

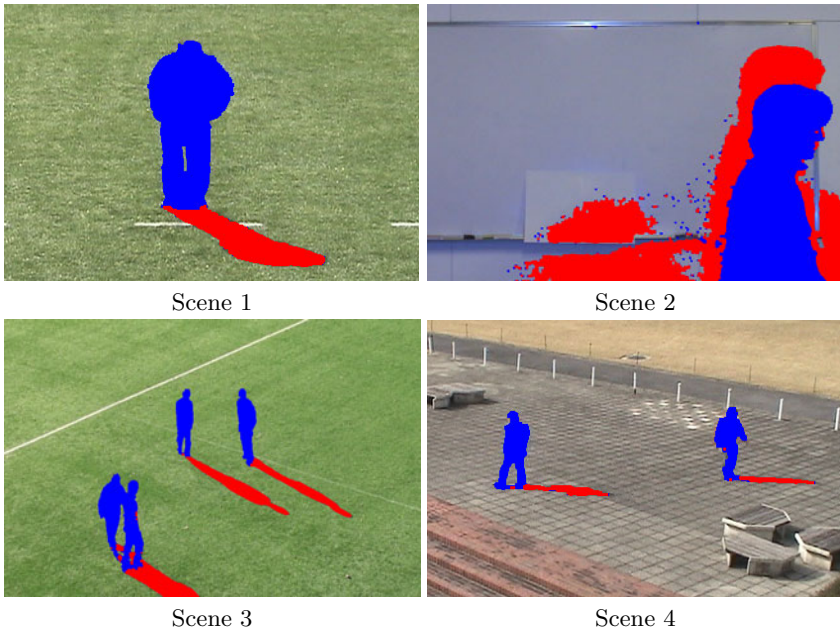


Fig. 3. Experimental Results of Proposed Method

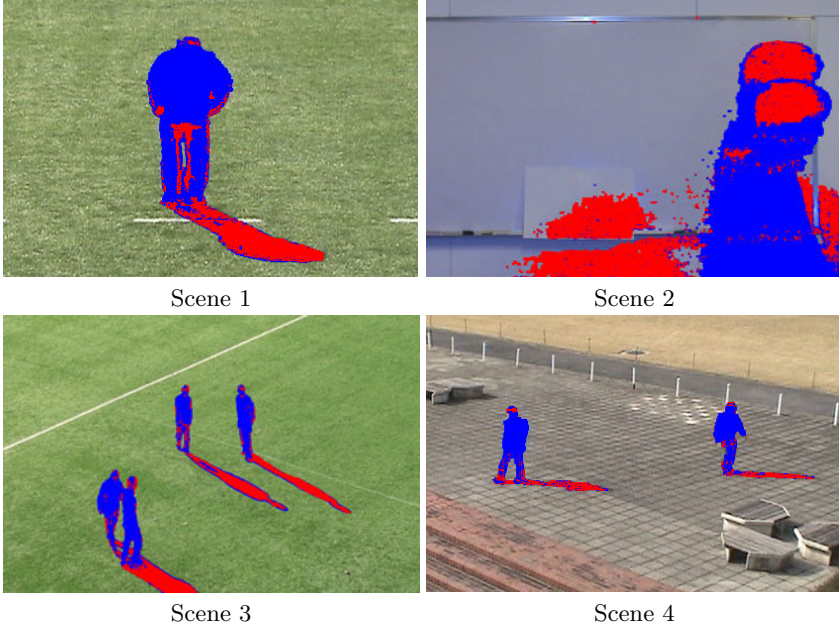


Fig. 4. Experimental Results of [3]

Two metrics proposed in [8] are introduced to evaluate the results. One is the shadow detection rate η , and the other is the shadow discrimination rate ξ . η and ξ are calculated by Eq. (4) and Eq. (5)

$$\eta = \frac{TP_s}{TP_s + FN_s} \quad (4)$$

$$\xi = \frac{\overline{TP}_f}{TP_f + FN_f} \quad (5)$$

where TP means the number of true positives, FN means the number of false negatives, the subscription s means shadow, subscription f means foreground and \overline{TP}_f means the number subtracting the number of points misjudged as shadows on foreground objects from the correct number of points of foreground objects. Table 1 shows η , ξ and K_i of each scene for the proposed method. $K_i (i = 1, \dots, 4)$ is the number of distribution estimated for each scene. η and ξ of all scenes are more than 90 %. It is shown that the proposed method detects the shadows with high performance.

The experimental results of the method [3] are shown in Figure 4. η and ξ of each scene by [3] are shown in Table 2. They show that the proposed method can get the better results than the method [3]. The method [3] uses the data obtained by threshold processing to construct the shadow model. It tends to extract the dark region as the shadow region and to misjudge in such a region.

Table 1. Shadow Detection Rates (η), Shadow Discrimination Rates (ξ) and Number of Distributions(K_i)

scene	1	2	3	4
η [%]	98.04	93.43	97.01	92.94
ξ [%]	97.46	98.73	97.40	93.64
K_i	2	3	2	2

Table 2. Shadow Detection Rates (η) and Shadow Discrimination Rates (ξ) of [3]

scene	1	2	3	4
η [%]	85.95	65.12	81.40	83.93
ξ [%]	71.78	66.17	72.41	73.23

Table 3. Processing Times of Proposed Method (ms)

Scene	1	2	3	4
Time	276.82	727.43	355.92	216.89

Table 4. Processing times for the case that all Gaussian distributions are used (ms)

Scene	1	2	3	4
Time	830.60	3360.61	1263.16	360.47

The processing time of the proposed method for each scene is measured. Table [3] shows the processing times for all scenes. The table shows that it takes more than 200 ms for all scenes to obtain the shadow regions by the proposed method. However, the processing time can be reduced by using GPU for handling the process of the proposed method.

The proposed method selects the Gaussian distributions used for the shadow model according to the mixture rates. The processing times of the method without the selection of the Gaussian distributions are also measured to confirm the effectiveness of the selection. Table [4] shows the processing times for all scenes. The processing times shown in Table [4] increase greatly in comparison with those shown in Table [3]. η and ξ for this case becomes the same values shown in Table [1]. These results show that the selection of the Gaussian distributions used for the shadow model causes the reduction of the processing time without reducing the accuracy of the shadow detection.

5 Conclusion

This paper proposed a new shadow detection method. The proposed method models shadows by the Gaussian mixture model with the Dirichlet process mixture model. The parameters of the model are estimated by the Dirichlet process

EM algorithm. The Dirichlet process EM algorithm can determine the proper number of distributions through the process of estimating the parameters.

After modeling the shadow by the Gaussian mixture model, the probability that a pixel belongs to the shadow region is calculated. After the shadow regions are detected through the threshold processing using the probability, the probability that the point belongs to the shadow region and does not belong to the object region is calculated at each pixel and final results are obtained. The results show that the proposed method can detect shadows with high performance.

Future work includes determining the parameters automatically and increasing the accuracy of the result.

Acknowledgment

Iwahori's research is supported by JSPS Grant-in-Aid for Scientific Research (C)(20500168) and Chubu University Grant. Woodham's research is supported by the Natural Sciences and Engineering Research Council (NSERC).

References

1. Cucchiara, R., Grana, C., Piccardi, M., Sirotti, S., Prati, A.: Improving shadow suppression in moving object detection with hsvcolor information. In: Proc. IEEE Intelligent Transportation Systems Conf., pp. 334–339 (2001)
2. Blauensteiner, P., Wildenauer, H., Hanbury, A., Kampel, M.: On colour spaces for change detection and shadow suppression. In: Proc. 11th Computer Vision Winter Workshop, Telc, Czech Republic, pp. 87–92 (2006)
3. Tanaka, T., Shimada, A., Taniguchi, R., Daisaku, A.: Object detection based on non-parametric adaptive background and shadow modeling. *IPSI SIG Notes. CVIM* 42(12), 105–112 (2007) (in Japanese)
4. Wang, Y., Cheng, H.D., Shan, J.: Detecting shadows of moving vehicles based on hmm. In: *ICPR 2008*, pp. 1–4 (2008)
5. Kimura, T., Nakada, Y., Doucet, A., Matsumoto, T.: Semi-supervised learning scheme using dirichlet process em-algorithm. *IEICE Tech. Rep.* 108(484), 77–82 (2009)
6. Sethuraman, J.: A constructive definition of dirichlet priors. *Statistica Sinica* 4, 639–650 (1994)
7. Neal, R.M.: Markov chain sampling methods for Dirichlet Process mixture models. *Journal of Computational and Graphical Statistics* 9, 249–265 (2000)
8. Prati, A., Mikic, I., Trivedi, M., Cucchiara, R.: Detecting Moving Shadows: Algorithms and Evaluation. *IEEE Trans. on PAMI* 25(7), 918–923 (2003)

Vowel Sound Recognition Using a Spectrum Envelope Feature Detection Method and Neural Network

Masashi Kawaguchi¹, Naohiro Yonekura¹, Takashi Jimbo², and Naohiro Ishii³

¹ Department of Electrical & Electronic Engineering, Suzuka National College of Technology, Shiroko, Suzuka Mie 510-0294, Japan
masashi@elec.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 Science, Aichi Institute of Technology, Yachigusa, Yagusa-cho, Toyota, 470-0392 Japan
ishii@aitech.ac.jp

Abstract. We proposed vowel sound recognition using a spectrum envelope feature detection method. At first, we collected sound with a headset microphone and Windows sound recorder. I used disintegration Fourier transform and cepstrum analysis to create sound and extracted the characteristic point in the frequency axis of each vowel. I took the characteristic point from the first to the second. This is sample data of each vowel sound. With a neural network, I made a program to perform vowel sound recognition. As a result, I was able to finish the basic vowel recognition program. With this method two undulate mountains in the spectrum envelope graph of each voice data are selected. Each vowel sound can be categorized clearly in spite of the environment of the speaker indefiniteness. It is possible to improve the voice recognition in the speaker indefiniteness in the simple algorithm.

Keywords: sound recognition, neural network, spectrum envelope, Fourier transform.

1 Introduction

In the field of speech recognition, many methods have been proposed by researchers, such as the Hidden Markov Model (HMM), which is one of the stochastic models, and the MFCC model, which calculates and distinguishes the cepstrum parameter.

On the other hand, in the field of voice recognition concerning speaker indefiniteness, papers have described practical speaker independent voice recognition using segmental features, the recognition of spoken sequences of Japanese isolated vowels based on structural representation of speech, and the recognition of Japanese vowels using a neural network.

The first of these used segmental features, the HMM and spectrum motion information, however, clear categorization was not perfect. [1] With the second, tree diagrams of 5 vowels of the speaker were used. This model also requires the learning process. [2] Finally, the third used a normal power spectrum which was analyzed by FFT. [3]

In this paper, we propose a Spectrum Envelope Feature Detection Method in which two undulate mountains in the spectrum envelope graph of each voice data are selected. Each vowel sound can be categorized clearly in spite of the environment of the speaker indefiniteness. It is possible to improve the voice recognition in the speaker indefiniteness in the simple algorithm.

Each vowel sound can be categorized clearly in spite of the environment of the speaker indefiniteness. It is possible to improve the voice recognition in the speaker indefiniteness in the simple algorithm.

2 Basic Theory for Voice Recognition

First, we show the basic theory for voice recognition. It includes the Structure of WAV file, Fourier Transform, Windows Function, Cepstrum Analysis and Neural Network.[4][5]

2.1 The Structure of WAV File

The WAV type format is the Microsoft Windows standard type in Voice data. It is one kind of RIFF format. The structure of RIFF is shown in Table 1 under the heading

Table 1. WAV File Head Information

Byte	Head Information
4 Byte	RIFF
4 Byte	File size
4 Byte	It is shown that the type of RIFF is WAVE
4 Byte	The definition of the format
4 Byte	The byte number
2 Byte	Format ID
2 Byte	The channel number
4 Byte	The sampling frequency.
4 Byte	The data rate.
2 Byte	The block size.
2 Byte	The bit number per sample.
4 Byte	Data chunk
4 Byte	Byte number of waveform data.
n Byte	Waveform data.

“Head Information”. Thus, when we acquire the wave form data, the first 44 bytes are disregarded, leaving data only after 44 bytes in the file.

The WAV file is recorded as a discrete data array by transforming the sampling theorem from a continuous signal speech waveform. A sampling frequency of 8kHz, results in 8000 times per one second.

2.2 Fourier Transform

Fourier Transform means “Transform from the time domain $x(t)$ to frequency domain $X(f)$ ”, denoted by the following equation.

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \\ &= \int_{-\infty}^{\infty} x(t) (\cos 2\pi ft - j \sin 2\pi ft) dt \end{aligned} \quad (1)$$

On the computer system, wave form data has a discrete value. So we changed the discrete equation, $k\Delta f$ is the frequency, f and $n\Delta t$ is the time, t .

$$X(k) = \sum_{k=0}^{N-1} x(n) e^{-j2\pi(k\Delta f)(n\Delta f) \Delta t} \quad (2)$$

We used the DFT(Discrete Fourier transform) and IDFT(Inverse discrete Fourier transform) equations as follows.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi k}{N}n} \quad (3)$$

2.3 The Window Function

When we make DFT from a speech signal, we have to cut out the waveform for analysis. We used the windows function for cutting out the waveform.

2.4 Cepstrum Analysis

The power spectrum of the speech signal is convoluted by the articulation filter on the sound source signal. The cepstrum analysis can separate the speech sound and the articulation filter. In particular, it is important that the articulation filter have a smoothing value compared to the power spectrum of the speech signal.

2.5 Neural Network

We constructed a neural network system consisting of three layers as shown in Fig. 3. There are 2 units in the first layer. 2 units mean the first and second feature quantity as shown in chapter 3. There are another five units in the second and third layer. Five units in the third layer mean each five vowel sound, "a", "i", "u", "e" and "o".[6]

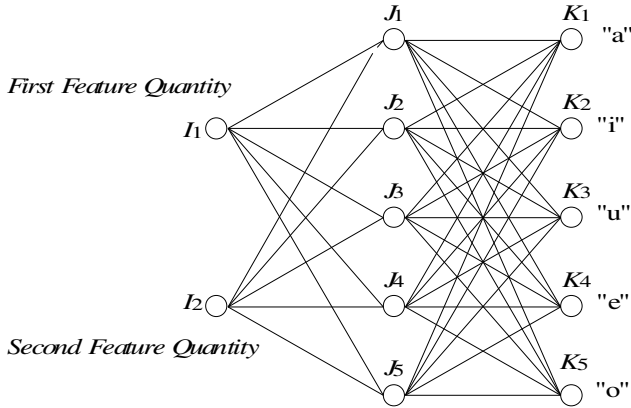


Fig. 1. Structure of Neural Network of this system

3 Experimental Result

At first, we record the vowel sound using the sound recorder on a Windows OS. We show the performance of the microphone and recording type of audio file in Table 2 and Table 3.

3.1 The Analysis Procedure

We undertook the following procedure. After this procedure, the real number component is taken out. It is shown in Fig. 2.

1. The data is read by the binary type. The numbers of data are 256 after cutting out.
2. The framing processing removed the uncontinuous waving form.
3. DFT(Discrete Fourier transform), each frequency independent intensity is converted into the amplitude spectrum
4. The logarithmic conversion, the amplitude spectrum is converted into logarithm amplitude spectrum.
5. IDFT(Inverse discrete Fourier transform), the logarithm amplitude spectrum is converted into the logarithm cepstrum.

6. The liftering processing, cutting off the coefficient of the logarithm cepstrum in the low-level term
7. DFT, Logarithm amplitude spectrum component of the articulation filter appears in the real number column.
8. Taking out the real number component.

Table 2. The Performance of Microphone

Subjects	
The Intensity of Input	-58dBV/ μ Bar, -39dBV/Pa+/-4dB
Frequency characteristics	100~16,000Hz
Noise Filtering	-30dB 250Hz, Avarage -16dB 20Hz~5kHz

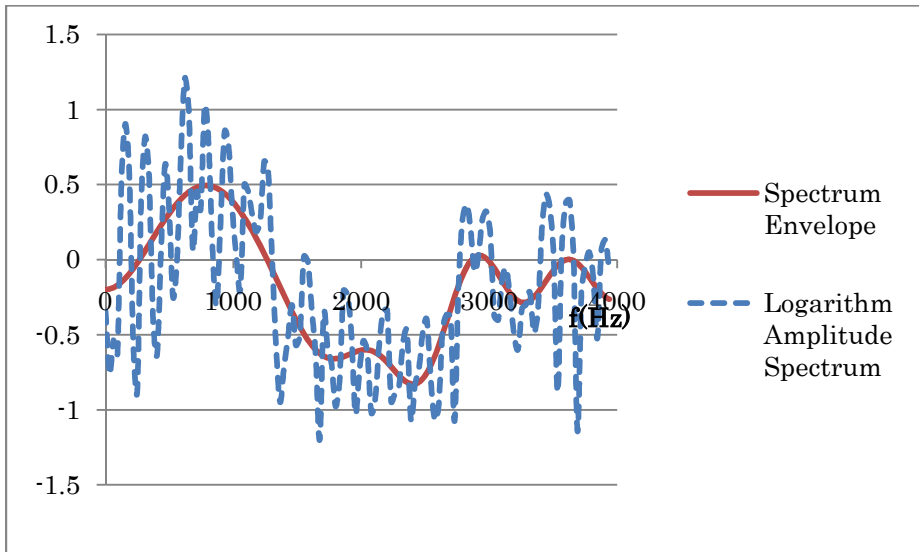


Fig. 2. The result of analysis for the vowel sound “a”

In Fig. 2, the dotted line shows the logarithm amplitude spectrum, the continuous line shows the spectrum envelope, final procedure.

3.2 Getting the Feature Quantity

For Getting the Feature Points, we measured the spectrum envelope. We selected the two peak points in the amplitude within 3500Hz. We measured the frequency in proportion to two peak points. These data were acquired in every vowel sound, shown in Fig. 3.

Table 3. Recording Type of Audio File

Subjects	
Saving Type	Wave
Sampling Frequency	8000Hz
Quantization.bit	8bit
Cannel	Monophony

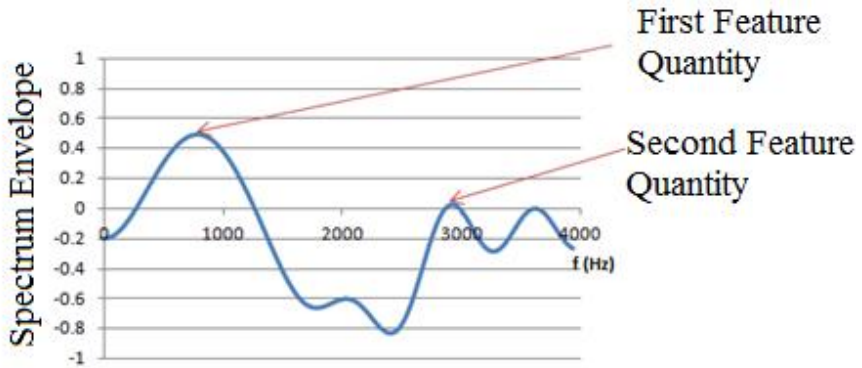


Fig. 3. First feature point and second feature point

3.3 Categorized Each Vowel Sound

After getting two feature quantities, we analyze the distribution of these data. We get sound data from three persons, A,B and C. Each person sounds two times in each vowel sound. We described the distribution of the feature quantity in Fig. 4. Their data can be categorized clearly compared with other methods. Furthermore, there is little information quantity under the speaker indefiniteness.

3.4 Recognition Experiment Using Neural Network

We converted the feature quantity into a neural network. The neural network consists of 2 input units and 5 output units. Five output units mean each vowel sound, "a", "i", "u", "e" and "o".

The speaker environment is specified, the output value of the neural network is in the range of 0.874 and 0.998. On the other hand, with speaker environment indefiniteness, the output value is in the range of 0.746 and 0.999. These values were satisfactory for recognizing the sounds.

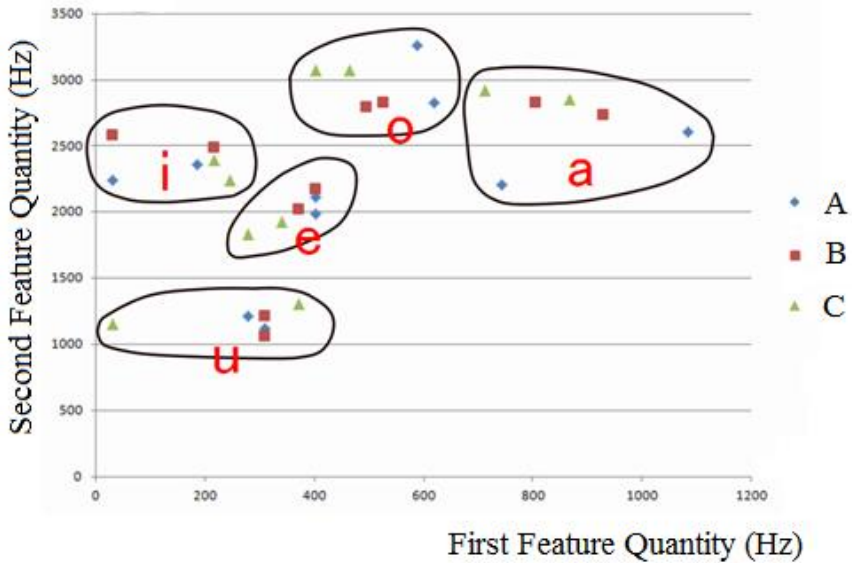


Fig. 4. The distribution of the feature quantity

4 Conclusion

We propose the Spectrum Envelope Feature Detection Method. This method includes selecting two undulate mountains in the spectrum envelope graph of each voice data.

Each vowel sound can be categorized clearly in spite of the speaker indefiniteness environment. Because the minimum output value is 0.746. It is possible to improve the voice recognition in the speaker indefiniteness in the simple input data.

This model uses two feature quantities. In the future it would be useful to increase the number of feature quantities, getting more peak points from the spectrum envelope. This will improve the recognized ratio and expand the consonant recognition under the speaker indefiniteness environment.

References

1. Kimura, T., Ishida, A., Niyada, K.: Practical Speaker Independent Voice Recognition Using Segmental Features. T. IEICE Japan J85, D-II(3), 398–405 (2002)
2. Murakami, T., Minematsu, N., Hirose, K.: Recognition of Spoken Sequences of Japanese Isolated Vowels Based on Structural Representation of Speech. T. IEICE Japan J91, A(2), 181–191 (2008)
3. Tsubota, A., Iijima, N., Sone, M., Mitsui, H., Yoshida, Y.: Phonemes Recognition of Japanese Vowels Using Neural Network. The Institute of Image Information and Television Engineers VAI91-37, pp. 29–36 (1991)
4. Kumazawa, I.: Learning and Neural Network, Morikita (1998), ISBN 978-4-627-70291-2
5. Shikano, K.: IT Text, Sound Recognition System, Ohmsha (2001), ISBN: 4-274-13228-5
6. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel distributed processing, pp. 318–362. MIT Press, Cambridge (1986)

Information Extraction Using XPath

Masashi Okada, Naohiro Ishii, and Ippei Torii

Aichi Institute of Technology

1247 Yachigusa, Yakusacho, Toyota, Japan 470-0392, Japan

kapio@jt9.so-net.ne.jp, {ishii, mac}@aitech.ac.jp

Abstract. To improve the classification accuracy of documents, it will be important to characterize not only words but also their relations among words. The classification method from this point of view will need another approach for the analysis of documents. In this paper, first, how to find the pattern tree in the XML data tree as the embedded sub-tree is developed simply by applying XPath technique. This problem is applicable to the search of the characterized words and their relations in the XML documents. Second, next problem is what kind of words and their relations exist in the XML documents. This problem is how to find the most frequent patterns in the documents, which is called often the most frequent sub-trees in the XML domain. The second problem finding the most frequent sub-trees is solved simply here by applying XPath technique.

1 Introduction

Broad band networks have made a great progress in information technology for much data communication and much information transactions in the computer networks. Though the data is becoming greatly large in the volume, the machine classification of the data classification as text data, is not easy under these computing circumstances. Most conventional methods to classify the documents are based on the occurrences of words, which characterize their documents[1,2]. To improve the classification accuracy of documents, it will be important to characterize not only words but also their relations among words. The classification method from this point of view will need another approach for the analysis of documents. Extensible Markup Language(XML) is a hierarchical markup language used to describe semi-structured data in a way that is independent of its appearance. XML has emerged as a standard for data representation and exchange on the World Wide Web. An XML document is modeled as an ordered labeled tree, where nodes represent elements, attributes, and text data and edges represent element inclusion as parent child relationship. An XPath is also a language for matching paths and, more generally, patterns in tree-structured data and XML documents[3,4,5]. XPath allows navigation in XML trees and returning a set of matching nodes. XPath expressions look like directory navigation paths. For example, the XPath “/paper/section/figure” navigates from the root of a document through the top-level “paper” element to its “section” child elements and on its “figure” child elements. The results of the expression give the set of all the “figure” child elements in the order they occurred. At each step in the navigation, the selected nodes for this

step can be filtered using qualifiers. A qualifier is a Boolean expression between brackets that can test path existence. There are fundamental problems how to find the objective sub-tree in the document data tree, which is called here pattern tree.

These problems are studied without using XPath techniques[6,7,8,9,10].

In this paper, first, how to find the pattern tree in the data tree as the embedded sub-tree is developed simply by applying XPath technique. This problem is applicable to the search of the characterized words and their relations in the documents. Second, next problem is what kind of words and their relations exist in the documents. This problem is how to find the most frequent patterns in the documents, which is called often the most frequent sub-trees in the XML domain. The problem constructing new information will be developed from the most frequent sub-trees by applying XPath technique, in which the developed technique will have to be iterated.

2 Tree Structure of XML Document and XPath

The XML document and data have the layered structure that consists of the nest of the element (inclusion relation). Then, the tree structure model is regarded as a logical model of the XML document and the XML data. The element can be shown by labeled tree with only one root node[3,4]. Fig. 1 shows an example of showing the XML document by the tree structure. The information extraction from these XML documents is basically important. The extracted information will be often useful to make the integrated new information in several documents. To solve these problems, some techniques are expected. The XPath is a language for finding information in an XML document. XPath uses path expression to select nodes or nodes sets in an XML document.

2.1 Location Paths

XPath represent a pathway to the node of destination data, which is called a location path. There are two kinds of location path: relative location paths and absolute location paths. A relative location path consists of one or more location steps separated by /. The steps in a relative location path are composed together from left to right from the context node, which is focused at the present time. An axis is introduced to specify the tree relationship between nodes selected by the location step and the context node. In Fig. 1, an example of XPath is shown. A node F in Fig. 1 is the context node, which is located on self axis.

An absolute location path consists of / optionally followed by a relative location path.

A / by itself selects the root node of the document containing the context node. The syntax for a location step is the axis name and the node test separated by a double colon, followed by zero or more expressions each in square brackets. For example, in `child::para[position()=1]`, `child` is the name of the axis, `para` is the node test and `[position()=1]` is a predicate[3,4].

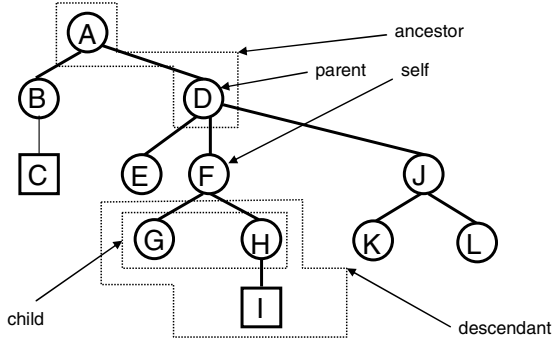


Fig. 1. Example of XPath axes

2.2 Extraction of Embedded Sub-tree Pattern

A retrieval method is developed here by giving the objective sub-tree pattern, which is called pattern tree. As an example, a pattern tree of nodes A, C, B, G, and D as shown in Fig. 2. The embedded pattern tree is searched as follows.

- (1) Node A is searched as the root node of the pattern tree. The document root (parent node as root node) is regarded as the context node. By applying XPath formula as descendant::A , an element node A is found in the data tree.(Fig. 3)
- (2) The first child node C of the node A in the pattern tree, is searched. Since C belongs to descendant of the parent node A in the data tree, XPath formula descendant::C is applied and an element node C in the data tree is found. Then, the node C becomes a context node.(Fig. 4)

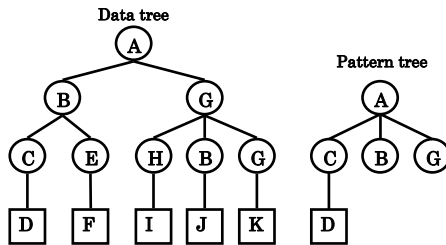


Fig. 2. Extraction of embedded pattern tree in data tree

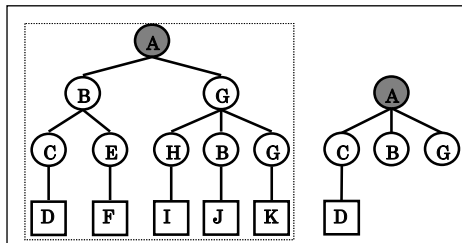


Fig. 3. Extraction of node A at Step(1)

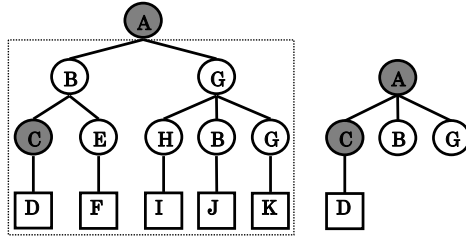


Fig. 4. Extraction of nodes A and C at Step (2)

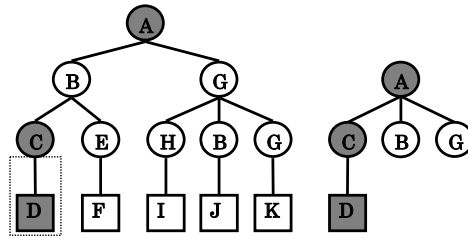


Fig. 5. Extraction of nodes A, C, and D at Step (3)

- (3) The text node D is searched as the first child node of C in the pattern tree. Since node D is a descendant of parent node C in the data tree, XPath formula `descendant::text()[string(.)='D']` is applied by replacing text node as C (Fig. 5)
- (4) The second child B of parent A in pattern tree is searched. XPath formula `descendant::B` is applied by assigning A as the context node. Then, a set X (both Bs in Fig. 6) is obtained. Next, by assigning B as the context node, XPath formula `ancestor::B|descendant::B` ("|" shows set union) makes a set Y (left side B in Fig. 6). Thus, $X - Y$ ("-" shows a set difference) becomes a node B in data tree. (Fig. 6)
- (5) The third child G of parent A in pattern tree is searched. Under the condition that G is a descendant of parent A and also is not an ancestor or descendant of C and B, XPath formula `descendant::G` is applied by assigning A as context node. Then, a set X (both Gs in Fig. 6) is obtained. Next, XPath `ancestor::G|descendant::G` is applied by assigning C as the context node. Then, a set Y (null) is obtained. Finally, XPath `ancestor::G|descendant::G` is applied by assigning B as the context node. Then, a set Z (G on the discovered B) is obtained. By the set operation $(X - Y) \cap (X - Z)$, node G is obtained in data tree. (Fig. 7)

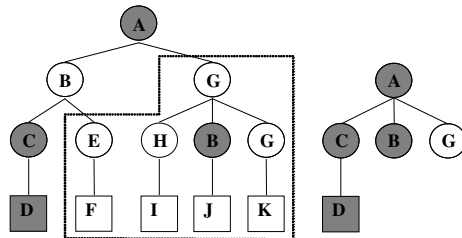


Fig. 6. Extraction of nodes A, C, D and B at Step (4)

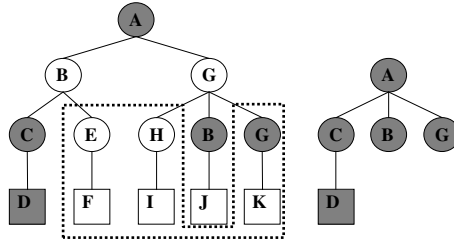


Fig. 7. Extraction of nodes A, C, D, B and G at Step (5)

3 Extraction of Frequent Patterns

Frequent patterns are cues for the understanding and characterizing documents. To characterize the documents, frequent words are often used as conventional method. To improve the classification accuracy of documents, relation of frequent words will be useful.

3.1 Extraction Algorithm

Extraction algorithm of the frequent pattern is described as follows.

(Step 1) The most frequent node is searched in data tree. The node is regarded as the root node in pattern tree. The criteria of the selection are described as follows.

- (1.1) Pattern tree consists of root node, which is used most frequently in data tree.
- (1.2) Among nodes with the same number of the occurrence in data tree, the node in the upper row is chosen as pattern tree.
- (1.3) When nodes are also on the same row, the respective patterns are made.

(Step 2) Embedded sub-tree pattern, made in Step(1) is searched in data tree. When terminal node(node without child) in pattern tree is found, a node with most occurrences followed by the leaf node is chosen.

- (2.1) Node with most occurrences is chosen. The node is added as a child of leaf node in pattern tree. Thus, a new pattern tree is made.
- (2.2) When more than two same nodes are found, the node with the most occurrences is chosen as a child of leaf node. Thus, a new pattern is made.
- (2.3) Among nodes with the same number of the occurrence in data tree, the node in the upper row is chosen as pattern tree. Thus, a new pattern is made.
- (2.4) When nodes are also on the same row, respective patterns are made. These Step(1) and Step(2) are iterated for the final new pattern.

3.2 Extraction Process

The algorithm described in the above is realized in the following example in Fig. 8.

Fig. 8 shows a given data tree. The frequent sub-tree as a pattern tree is made successively in the following steps.

- 1) In the first, the context node is assigned to the document root in data tree. By applying XPath formula descendant::*, all nodes are selected. By Step (1) in algorithm, the most frequent element node in data tree is searched in Fig. 8. Nodes B,C, and

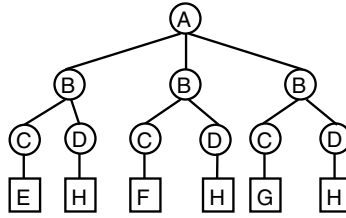


Fig. 8. Data tree in the experiment

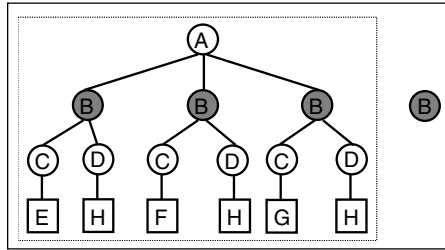


Fig. 9. Node B is extracted and assigned as root node in pattern tree

D appears 3 times in data tree, respectively. Since node B is on the most upper row, node B is chosen as root node in pattern tree. (Fig. 9)

2) Pattern tree is searched in data tree in Fig. 8, which is embedded in data tree. After XPath formula `descendant::B` is applied, 3 times of B appear in data tree. Since node B is a leaf node, searched node is assigned as context node in data tree and XPath formula `descendant::*` is applied to find the most frequent node by Step (2). Then nodes C and D appear 3 times on the same row. Thus, nodes C and D are followed by node B as child nodes. This makes a new pattern tree. (Fig. 10)

3) New pattern tree is searched as the embedded tree in data tree. This is done by XPath formula `descendant::B`. By assigning node B as context node, XPath formula `descendant::C` is applied. Since C is a leaf node in pattern tree, XPath formula `descendant::*` is applied by Step (2). Then, text nodes E, F and H are found. Here, node E is regarded as the first child of C. Next, the second child of B, D is checked. By assigning B as the context node, `descendant::D` makes a set X. By assigning C as context node, `ancestor::D|descendant::D` makes a set Y. Then, X-Y searches node D. By assigning node in X-Y as context node, XPath formula `descendant::*` is applied by Step (2). Thus, text node H is chosen as the child node of D. (Fig. 11)

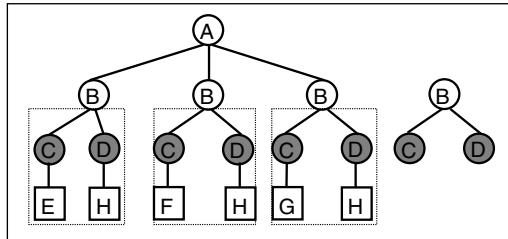


Fig. 10. Nodes C and D are followed by B

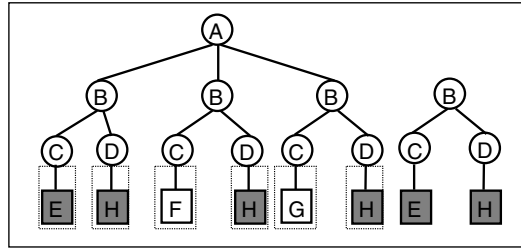


Fig. 11. Nodes E and H are attached

4) Final pattern tree is made by the iteration of the above algorithm, which is shown as the most frequent sub-tree in the right side of Fig. 11.

4 Experiments of Detection of Most Frequent Pattern

The search of the most frequent trees developed here, was applied to practical data on the Web, which is Reed University XML data[11]. This data contains the educational campus data as course name, credit unit, room, time, instructor, building et al. Under some conditions, the objective search was carried out. A part of the XML document is shown in Fig. 12.

```

<root>
<course>
  <reg_num>10577</reg_num>
  <subj>ANTH</subj>
  <crse>211</crse>
  <sect>F01</sect>
  <title>Introduction to Anthropology</title>
  <units>1.0</units>
  <instructor>Brightman</instructor>
  <days>M-W</days>
  <time>
    <start_time>03:10PM</start_time>
    <end_time>04:30</end_time>
  </time>
  <place>
    <building>ELIOT</building>
    <room>414</room>
  </place>
</course>

<course>
  . . . . .
</root>

```

Fig. 12. University classes data in XML

```

<course>
  <reg_num>10624</reg_num>
  <subj>ANTH</subj>
  <crse>431</crse>
  <sect>S01</sect>
  <title>Field Biology of Amphibians</title>
  <units>0.5</units>
  <instructor>Kaplan</instructor>
  <days>T</days>
  <time>
    <start_time>10:00</start_time>
    <end_time>11:50</end_time>
  </time>
  <place>
    <building>ELIOT</building>
    <room>240A</room>
  </place>
</course>

```

Fig. 13. Extraction of most frequent tree

The most frequent tree are searched in data trees as shown in Fig. 12. The searched results are shown in Fig. 13 by the proposed method here. Most frequent trees, next frequent trees, and so on will characterize the documents. From using frequent words by the conventional classification methods of documents, frequent trees including frequent words will be useful for the classification of documents. These frequent trees technique for the classification is the next study from the developed method here.

5 Conclusion

The XML tree structure model was adopted for finding the embedded sub-tree by giving the pattern tree, which is typical pattern to be searched in the document. The embedded tree is searched by the XPath language, which is a query language introduced by the W3C for retrieving information in XML documents. Then, the most frequent pattern trees was also made dynamically by the XPath. Since these methods developed here consider words and relations among words, they will be applicable for the improvement of the classification accuracy of documents than the conventional methods. This is remained as the future problem.

References

1. Ishii, N., Bao, Y., Hoki, Y., Tanaka, H.: Rough Set Reduct Based Classification. In: *New Advances in Intelligent Decision Technologies (IDT 2009)*. Studies in Computational Intelligence, vol. 199, pp. 373–382. Springer, Heidelberg (2009)
2. Bao, Y., Tsuchiya, E., Ishii, N., Du, X.: Classification by Instance-Based Learning Algorithm. In: *Gallagher, M., Hogan, J.P., Maire, F. (eds.) IDEAL 2005*. LNCS, vol. 3578, pp. 133–140. Springer, Heidelberg (2005)

3. Geneves, P., Layaida, N.: A System for the Static Analysis of XPath. *ACM transactions on Information Systems* 24(4), 475–502 (2006)
4. Benedikt, M., Koch, C.: XPath Leashed. *ACM Computing Surveys* 41(1), Article 3, 3:1–3:52 (2008)
5. Yang, L.H., Lee, M.L., Hsu, W.: Efficient Mining of XML Query Patterns for Caching. In: *Proc. of the 29th VLDB Conference*, vol. 29, pp. 69–80 (2003)
6. Zaki, M.J.: Efficiently Mining Frequent Trees in a Forest: Algorithms and Applications. *IEEE Trans. on Knowledge and Data Engineering* 17(8), 1021–1035 (2005)
7. Zaki, M.J.: Efficiently Mining Frequent Embedded Unordered Trees. *Fundamenta Informatica* 65, 1–20 (2005)
8. Zaki, M.J., Aggarwal, C.C.: XRules, An Effective Structural Classifier for XML Data. In: Whang, K.-Y., Jeon, J., Shim, K., Srivastava, J. (eds.) *PAKDD 2003. LNCS (LNAI)*, vol. 2637, pp. 316–325. Springer, Heidelberg (2003)
9. Asai, T., Arimura, T., Uno, T., Nakano, S.: Discovering Frequent Substructures in Large Unordered Trees. In: *Proc. Sixth Int. Conf. Discovery Science*, pp. 47–61 (October 2003)
10. Chi, Y., Yang, Y., Munz, R.R.: Indexing and Mining Free Trees. In: *Proc. Third IEEE Int. Conf. Data Mining*, pp. 509–512 (2003)
11. <http://www.cs.washington.edu/research/xmldatasets/www/repository.html>

Information Visualization System for Activation of Shopping Streets

Ippei Torii, Yousuke Okada, Takahito Niwa, Manabu Onogi, and Naohiro Ishii

Dept. of Information Science
Aichi Institute of Technology
Yachigusa 1247, Yakusa-cho, Toyota-shi, Aichi, Japan
{mac,x07232xx,ishii}@aitech.ac.jp
<http://www.ait.ac.jp/>

Abstract. This paper attempts to activate a large scale shopping streets (shotengai) using Internet technique. A new web architecture is made for activation of shotengai, which is a typical Japanese shopping streets in the city and towns. Recently, decline of shotengai is a serious problem by development of shopping centers. “Osu” is one of the most famous shotengai in Nagoya, Japan, which includes about 400 stores. We developed Osu shotengai official web site, called “AT Osu”. First, the information of 400 stores is collected at Osu shotengai. Next, for a new web site, “Information Visualization System” is proposed here to express information freshness of present shotengai situation. It includes “Message Upload System” which administers the information from stores using systematic output method by RS1.0 effectively. Further, we introduce a new approach for competing for a store owner with another one by their prominence and attractive design, which is described in text information. This is useful for the web composition from image base to text base to increase motivation of store owners. By the new developed web site of At Osu, number of visitors of “At Osu” has increased rapidly.

Keywords: Activation of Shopping Streets, Information Visualization.

1 Introduction

Shotengai is a typical Japanese commercial district stores along a certain street. Recently, decline in shotengai is a serious problem by development of large scale shopping centers. According to the investigation by Ministry of Economy, Trade and Industry in Japan, 70.3% shotengai stores say “decline” or “stagnation, but likely to decline” (Fig. 1) [4]. Other studies have concluded from sociology and economics [17, 12]. But the effect is not clear by these researches and methods. So, we present a method of activation of shotengai using Internet technique and show the effect by our method clearly.

Osu shotengai (Naka-ku, Nagoya, Aichi, Japan) (Fig. 2) is one of the most famous shotengai, which is Japanese temple town called Osu Kannon. Osu has

600 meters × 400 meters area, and across some roads. Osu shotengai consists of 8 groups which are called Banshoji-dorishotengai, Osu-shintenchi-dorishotengai, Nagoya-osu-Higashi nioumon-dori shotengai, Osu-nioumon-dori shotengai, Osu-kannon-dori shotengai, Osu monzenmachi shotengai, Osu hon-dori shotengai and Akamon-myouou-dori shotengai. It includes about 400 shops in Osu.

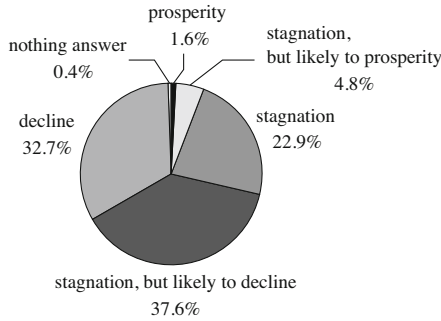


Fig. 1. Shotengai investigation of actual conditions



Fig. 2. Shotengai investigation of actual conditions

After the postwar, the town of Osu has declined gradually. But, shotengai have planed some events for example “Osu street performer festival” since 1978. Then, young people gather at Osu to wear with unique fashion. Now, the stores of old-clothes, accessory, and miscellaneous goods stand 70% of all. Osu was crowded and also was a gathering town around young person who are from teens to twenties. But, consumption got depressed along with worsening economic conditions in Japan at 2008. It’s influence was seen also at Osu. Under these conditions, we renew a web site, “At Osu” [\[1\]](#) to activate Osu shotengai. In this paper, we have developed a method of web technique and showed result of a large scale shotengai activation project using Internet. We hope our instance for the activation of shotengai will be applicable to another shoutengai projects in Japan.

2 Existing States of Internet and Shotengai

Digital Communication made rapid progress as popularization of personal computer [5,6]. People can communicate with world at real-time. They can send and receive a lot of contents for example characters, images, sounds, movies, and so on. Then, they build the meta-universe which can economically and socially acts in it so that they can obtain, edit and send information at liberty. High quality web site of a large scale shopping mall are developed for their advertise and Internet shopping mall. But, it is difficult for shotengai to develop a high quality web site, because the financial deficit and aging, etc. influence. We aim at the web site architecture of the university initiation to solve these problems.

3 As Educational and Research Project between Academia and Industry

In cooperation between industry and the academia, the multiplicative effects are born from exchange with related persons and their ideas. Both of them is activated from their results. We believe academic-industrial alliance can help each other for the objective project. The research at academia can get more from industry, and on the other hand they may bring something to industry. It is important and proper for usability and findability of the web site to design “At Osu” [10,14,11], since we have to develop easy, usable and findable user interface to show a lot of information of shotengai. And, the project will develop university students’ ability of problem solving and communication by practical problems. They can try to make their efforts not to think cost. We analyze the results of visitors accesses, and evolve it spending a year or more.

4 Information Visualization System

Our research has 4 characteristic features, which are different from conventional methods [16,15].

- First is the information administer and unification of about 400 stores using Contents Management System (CMS) and a lot of RDF Site Summary (RSS) 1.0.
- Second is the system of expressing information freshness.
- Third is the method for competing for a store owner with another one by prominency to increase motivation of store owners.
- Fourth is the attractive design which is web composition shifting from image base to text base.

We call the system which includes 4 features in above “Information Visualization System”. Fig. 3 shows an explanation of the sub-section. Fig. 9 in Appendix shows a workflow of the proposed system in this study.

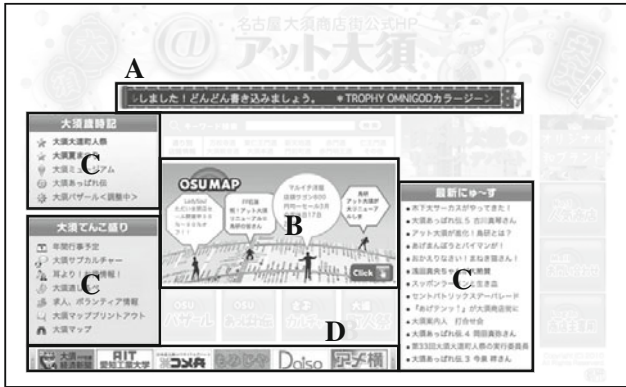


Fig. 3. A is “Osu Marquee” which we explain in 4.3”. B is “Main Window” which we explain in 4.3. C means web composition shifting from image base to text base. D means screen with movement using Ajax.

4.1 Information Administration by CMS and RSS1.0

“Message Upload System” to connect store owners with users, is constructed firstly. In this system, store owner can upload messages directly on Internet and users can obtain messages through some user interface. We have to develop a large scale web system which provide about 400 username and password, and administer to unify every information. Recently, a large scale web system uses CMS written by Hypertext Preprocessor (PHP) to generate dynamic contents. In this research, “Extensible Object Oriented Portal System” (XOOPS, GNU general public license) which is one of the general purpose CMS to administer a lot of information [7]. We register unique ID, username and password each store using user registration function. After that, we provide about 400 independent web log system in XOOPS. The system output message which is defined RSS1.0 file format. RSS1.0 is written by Extensible Markup language (XML) [13]. And it provides more systematic format environment because Resource Description Framework (RDF) restricts it. RDF is a framework which divide information into subject, predicate and object, and express text file as directed graph. XML is a language for meta-data which is structure data effectively.

4.2 “Osu Map” Visualizes Freshness of Information

“Osu Map” system in Fig. 4 shows all the stores location information which is RSS1.0 generated by CMS. The RSS1.0 can not only broadcast information but also sort data automatic by computer, because RSS1.0 is defined by XML. “Osu Map” system is controlled by Flash (ActionScript), because Flash is proper platform for the method of visualizing freshness. If users clicked the marker, the system called RSS1.0 extraction function. Then the system load the date, store name and up-loaded message. The date is used by deciding the kind of balloon. The color and shape of balloon express freshness of information. If

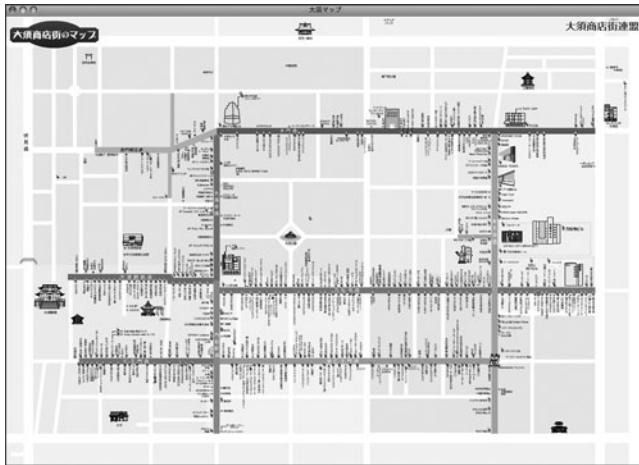


Fig. 4. “Osu map” which is the main item of Information Visualization System

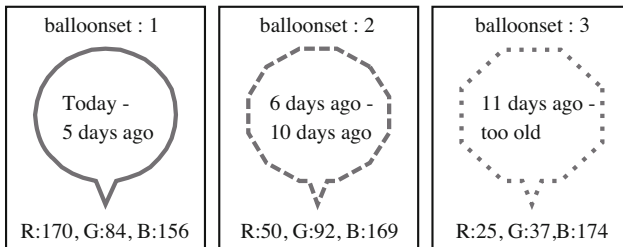


Fig. 5. The balloon with shape and color which express freshness

the information has become old it, the color and shape are dark and angular, respectively. We show Fig. 5 to explain the meaning of the balloon.

4.3 Competing for a Store Owner with Another One

First of all, we should raise the motivation of the store owners who is the subject person, because the amount of latest information concerning the website is proportional to the number of accesses by users. Therefore, we provide “Osu marquee” which displays seven of the latest information up-loaded by owners on top page. Marquee is one of the display styles of characters that scroll right and left. And the system input up-loaded date by the RSS1.0 format, and are controls by Flash with unique ID. Generally, top page is the most accessed page by users in the website. So, the advertising potential is also very high. Store owners compete for a store owner with another one. Moreover, we provide “Main Window” to compete for a store owner with another one. It displays 4 up-loaded message from the latest information. “Marquee” has variable text area, but “Main Window” is fixed the text area, because “Main Window” attaches importance to visual expression.

4.4 Attractive Design for Web Site

Externals in web site are one of the important factors [8]. We referred to the Chinese Internet shopping mall (Taobao “http://www.taobao.com/”, dangdang “http://www.dangdang.com/”, joya (amazon) “http://www.amazon.cn/”, Paipai “http://www.paipai.com/”, and so on) for activation of shotengai in Japan. In the Internet shopping mall supported by the rapid growth of GDP in China, there is energy that fuels the purchase desire of 384 million users who are Chinese. We consider common features of web sites in China to tell energy, and injected energies into “At Osu”. We noticed the regular arrangement of colors. All arrangement of colors obey the psychology [9]. Every web sites include color of orange and blue. We show color’s nature of psychology in Table 1. We introduce it in “At Osu” in Fig. 6. “Screen with movement using Ajax and Flash” and “web composition shifting from image base to text base” produce lively on “At Osu”. We show what is the newest “At Osu” in Fig. 6.

Table 1. Color’s nature of psychology

Orange	Energy, Active, kind, healthy, lightly, warmth
Blue	Clean, cool, reason, trust, quiet, reliability



Fig. 6. The newest top page design on March 26, 2010

5 Relationship between Improvement Points and Results of Visitors in Site

We show all the results of web results in Fig. 7, Fig. 8 and Table 2. In Table 2, we explain the relationship between improvement points of web and results of visit. Visits [3] represent the number of individual sessions initiated by all the visitors to your site. If a user is inactive on your site for 30 minutes or more, any future activity will be attributed to a new session. Users that leave your site and return

within 30 minutes will be counted as part of the original session. Bounce rate [2] is the percentage of single-page visits or visits in which the person left your site from the entrance (landing) page. Use this metric to measure visit quality - a high bounce rate generally indicates that site entrance pages aren't relevant to your visitors.

6 Conclusions

In this paper, we developed Information Visualization System which includes database processing, mainly focusing on activation of shotengai using Internet technique. In the Message Upload System for the store owner, the up-loaded message is output to three RSS1.0 (Osu map, Marquee, and direct output) and a XHTML (message up-loaded list). Moreover, we obtain a TEXT data from the database directly and use it for the system supplement of "Osu map". A choice of in total five output format is generated. The information is delivered systematically using RSS1.0. In addition, even though the administrator who is inexperienced in a computer, he or she can perform maintenance and administration of web site, easily. We studied much in the design which pursued findability of web site. We made the top page which has attractive color, text links and movement. We succeeded in compete for a store owner with another one by their prominency, to get the space of the advertisement.

References

1. At osu, <http://www.osu.co.jp/>
2. What does bounce rate mean? - analytics help, <http://www.google.com/support/analytics/bin/answer.py?hl=en&answer=81986>
3. What's the difference between clicks, visits, visitors, pageviews, and unique pageviews? - analytics help, <http://www.google.com/support/analytics/bin/answer.py?hl=en&answer=57164>
4. Investigation of Actual Conditions in Shotengai. In: The Small and Medium Enterprise Agency, Ministry of Economy Trade and Industry in Japan (2007)
5. Anderson, C.: Free: The future of a radical price. Hyperion (2009)
6. Graham, P.: Hackers and painters: big ideas from the computer age. O'Reilly, Sebastopol (2004)
7. Kondo, M., Goto, M., Hattori, A., Yasuda, T., Yokoi, S.: Practical use and future issues of open source cms for regional portal site. IPSJ SIG Notes 2008(16), 29–34 (2008)
8. Kuroda, H., Ozawa, T., Kameda, H.: Implementation and evaluation of a web page design system based on impression words. Technical report of IEICE. Thought and language 107(387), 19–24 (2007)
9. Lindsay, P., Norman, D.: An introduction to psychology. Academic Press, London (1977)
10. Morville, P.: Ambient findability: what we find changes who we become. O'Reilly Media, Inc., Sebastopol (2005)
11. Morville, P., Rosenfeld, L.: Information architecture for the world wide web. O'Reilly Media, Inc., Sebastopol (2006)

12. Otake, A., Nakayama, T., Inoue, Y.: 7018 a study on the ideal method of the activation of shopping street from the activation of the downtown in nara city (48), 433-436 (2008)
13. Ray, E., Maden, C.: Learning XML. O'Reilly and Associates, Inc., Sebastopol (2001)
14. Shneiderman, B.: Designing the user interface: strategies for effective human-computer interaction. Addison-Wesley Longman Publishing Co., Inc., Boston (1997)
15. Yonemoto, K.: Selection lesson of the good practice by the ministry education and science in the fiscal year 2005- management the virtual shopping district in the internet by industry-university cooperation: Education by the collaboration of the shopping town and college in the country. IEICE technical report 107(201), 1-2 (2007)
16. Yoshimura, K., Yamaga, K.: 7106 a study on the use of information technology for the activation of shopping streets: part1. a preparatory research on the possession of internet web sites. Summaries of technical papers of Annual Meeting Architectural Institute of Japan. F-1, Urban planning, building economics and housing problems 2004, 267-268 (2004)
17. Yoshimura, K., Yamaga, K.: 7243 a study on the activation of shopping streets in yokohama city: From view point of the information technology use and the community facilities introduction. Summaries of technical papers of Annual Meeting Architectural Institute of Japan. F-1, Urban planning, building economics and housing problems 2005, 547-548 (2005)

Appendix



Fig. 7. Design transitions from May 2009 to Mar 2010

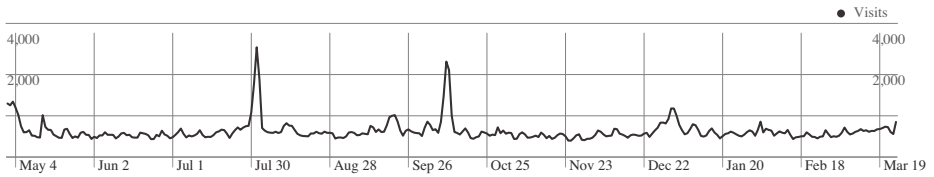


Fig. 8. The number of visits from May 2009 to Mar 2010

Table 2. Relationship between improvement points and results

Date	Improvement Points	Results
May 2009	Be born new “At Osu”	Average time on site became the current twice. The number of visits exceeded 1,300.
Jun 2009	Put the banner by Flash for advertisement on top page, verify the visualization.	Bounce rate has decreased.
Jul 2009	Introduce the latest information in shotengai on top page.	Returning visitor has increased.
Aug 2009	Correct the arrangement of top page, and users can collect the information easier.	Differences of the number of visits between pages have decreased.
Sep 2009	Add “Marquee”, and be able to display the event in shotengai	Average time on site has increased.
Oct 2009	Change the “Main Window” from the link of Osu Map to the news of season.	Average time on site has increased, and bounce rate has decreased.
Nov 2009	Put the banner by Ajax for advertisement on top page.	Users have visited site from the enterprise again.
Dec 2009	Change “Marquee” from displaying the news in shotengai to the message which is the uploaded it by store owners. Store owners frequent upload message.	The number of visits exceeded 1500. Pageviews and average time on site has increased.
Mar 2010	Renewal “At Osu”. Change the design greatly. Especially we provide “Main Window”. It displays 4 up-loaded message from the latest information.	The number of visits exceeded 1800.

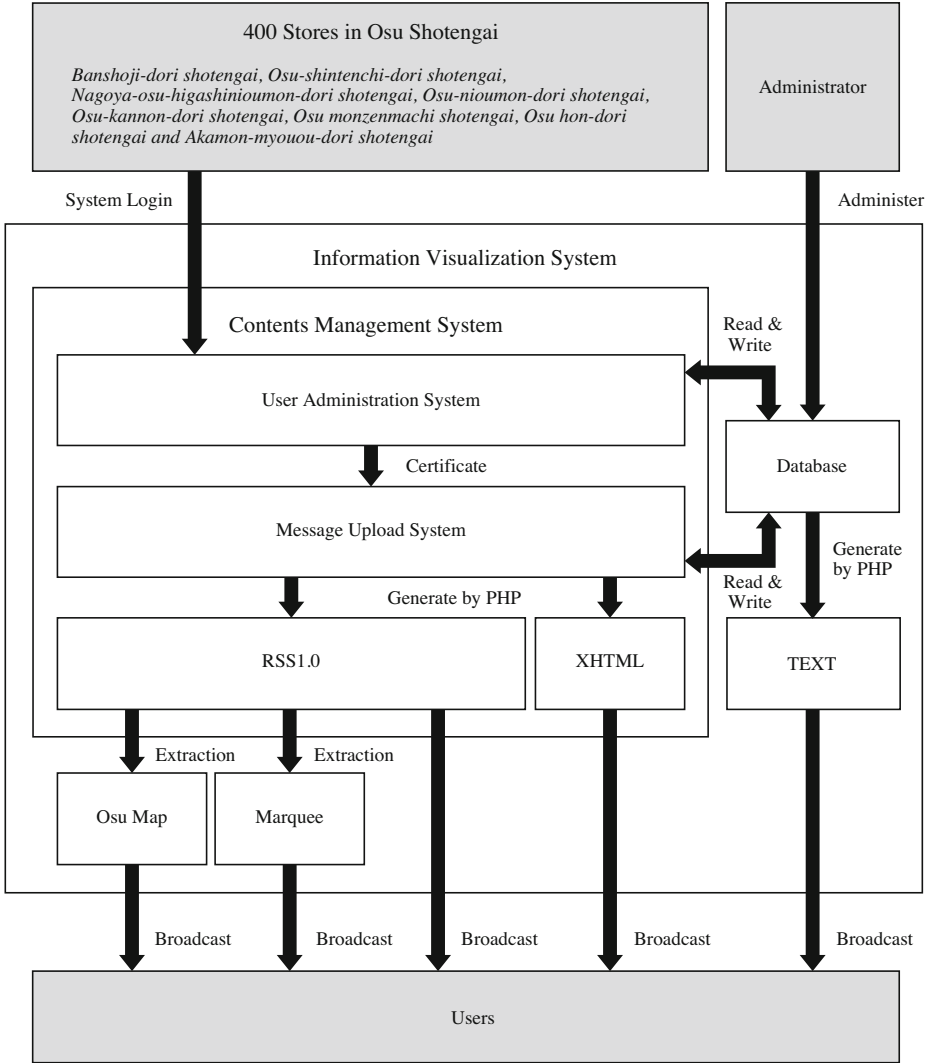


Fig. 9. The workflow of Information Visualization System

Introduction to Intelligent Network Routing Based on EVALPSN

Kazumi Nakamatsu¹, Jair Minoro Abe², and Takashi Watanabe³

¹ School of H.S.E., University of Hyogo, Himeji, Japan
nakamatu@shse.u-hyogo.ac.jp

² Paulista University, University of Sao Paulo, Sao Paulo, Brazil
jairabe@uol.com.br

³ Faculty of Information, Shizuoka University, Hamamatsu, Japan
watanabe@inf.shizuoka.ac.jp

Abstract. The goal of our work is to make up an EVALPSN based multi-path ad-hoc network routing protocol system in which various kinds of protocols can be dealt with uniformly. In this paper, as the first step to our goal we introduce a single-path ad-hoc mobile network routing protocol based on a paraconsistent annotated logic program EVALPSN with a simple example of ad-hoc network routing, which is a translation of the DSR(Dynamic Source Routing) protocol into EVALPSN. Mainly route discovery is focused on in this paper.

Keywords: ad-hoc network, multi-path network routing protocol, paraconsistent annotated logic program, EVALPSN.

1 Introduction

We have already developed a paraconsistent annotated logic program called Extended Vector Annotated Logic Program with Strong Negation(abbr. EVALPSN) in [3,11] that has been applied to various real-time intelligent control and safety verification systems such as pipeline process control in [4]. Moreover, EVALPSN can deal with before-after relation between processes(time intervals) and has been applied to process order control in [8,9]. EVALPSN based intelligent control /verification have the following features: (1) since EVALPSN can deal with deontic notions such as forbiddance, control properties in deontic expression can be easily and directly formalized in EVALPSN; (2) logical verification of operation control can be easily carried out as logic programming; (3) since some restricted fragment of EVALPSN can be implemented on microchips as electronic circuits [5], EVALPSN is a suitable tool providing real-time control. Therefore, EVALPSN can provide a useful platform for all kinds of intelligent information systems in both practical and theoretical senses.

Recently, wireless ad-hoc communication network systems in which nodes can communicate directly each other without via sites have been focused on. In such ad-hoc networks each distributed mobile node has to control communication between nodes autonomously and some ad-hoc network control methods have been

proposed. Among them routing protocol is a control method that deal with two mechanism, route discovery and route maintenance. If we consider data transmission efficiency in ad-hoc mobile networks, multi-path routing protocols that deal with more than two routes for communication should be more appropriate than single-path routing protocols. Various kinds of multi-path routing protocols that are object-orientedly improved versions of the basic one have been proposed and each multi-path routing protocol has different advantages. Therefore, if we utilize various kinds of multi-path routing protocols efficiently, an intelligent platform on which they can be implemented is indispensable. Our near future work is to make up an intelligent multi-path routing system based on EVALPSN in which various kinds of multi-path routing protocols can be treated uniformly. As the first step to our goal, we have implemented the single-path DSR(Dynamic Source Routing) Protocol in EVALPSN, which is introduced in this paper.

The Dynamic Source Routing (DSR) protocol is a simple and efficient routing protocol designed for multi-hop wireless ad-hoc networks of mobile nodes [12]. The DSR protocol consists of two mechanisms, route discovery and route maintenance, and three kinds of messages, route request message, route reply message for route discovery and route error message for route maintenance are used in the DSR. The operations for the three messages at each node can be expressed in rule sentences, and they are easily translated into EVALPSN.

This paper is organized in the following manner: first, EVALPSN is reviewed and the basic ad-hoc network routing protocol DSR is introduced and its node operations are described in if-then rule form; next the node operations are formalized in EVALPSN and a simple example of the EVALPSN based network routing is introduced; last, this paper is concluded with our future work.

We omit the explanation and definition of EVALPSN due to space restriction. The details of EVALPSN can be found in [11].

2 Ad-Hoc Network Routing Protocol DSR

As preliminary, we provide some definitions. A *route* is defined as an ordered list of the finite number of nodes such as $\langle n_1, n_2, \dots, n_k \rangle$, where each $n_i (1 \leq i \leq k)$ represents a node id; the *origin* node is as a node that has sent the message initially; the *object* node is as a node that has just received the message; and the *source* node is as a node that has sent the message immediately before the object node receives the message.

First of all, we introduce a well known single-path routing protocol, DSR(Dynamic Source Routing) protocol [12]. The DSR protocol deals with three kinds of network messages called *route request* message to discover routes, *route reply* message to reply for route request if the requested route has been made up, and *route error* message to report route error for the origin node if the route has been cut. The DSR protocol contains the node operations for the three kinds of network messages, which are summarized in the following seven operation rules. Those operation rules will be translated into EVALPSN in the following section.

[Node Operations in DSR Protocol]

- 1. route request message discarding.** If the object node receives a route request message that has been already received, it should be discarded.
- 2. route request message replying.** If the object node is the destination of the route request message that should not be discarded, the route request message should not be flooded and the route reply message for the route request should be sent back to the origin node according to the reverse order of the route node list in the route request message;
- 3. route request flooding.** otherwise, the route request message should be flooded with adding the object node id to the route node list in the route request message.
- 4. route reply message transferring.** If the object node receives a route reply message whose destination node is not the object node, the route reply message should be transferred according to the reverse order of the route node list.
- 5. route reply message termination.** If the object node is the destination of a route reply message, the route reply message should be terminated with completing the route discovery.
- 6. route error message transferring.** If the object node detects a route error directly or receives a route error message from other nodes, and the object node is not the destination of the route error message, the route error message should be transferred according to the reverse order of the route node list;
- 7. route error message termination.** otherwise, the route error message should be terminated with starting another route discovery.

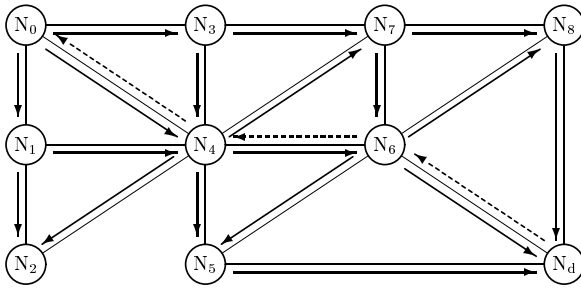


Fig. 1. Network Route Message transferring

We describe a simple example of the DSR protocol operations with taking the network in Figure 1.

[Example]

Suppose that a route discovery request to node N_d has been issued at N_0 ; then node N_0 floods the route request message to all the neighbor nodes N_1 , N_3 and N_4 ; if node N_4 receives the route request message from node N_0 , node N_4 should check whether the same route request message has been already received or not; then since the route request message has not been received by node N_4 before, it should be flooded to nodes N_5 , N_6 and N_7 with adding node N_4 id n_4 to the end

of the route node list $\langle n_0, (n_4) \rangle$ in the route request message; on the other hand, if the same route request message arrived at node N_4 via node N_3 later than the previous flooding of the message, it should be discarded; suppose that node N_d has received the same route request message whose destination is node N_d from node N_6 , which has the route node list $\langle n_0, n_4, n_6 \rangle$, then node N_d should send the route reply message with the route node list $\langle n_0, n_4, n_6, n_d \rangle$ back to the origin node N_0 via nodes N_6 and N_4 ; if node N_6 has received the route reply message from node N_d , it should be transferred to node N_4 according to the reverse order of the route node list.

The route request message flooding is shown by the solid line arrow symbol \longrightarrow and the route reply path is by the chained line arrow \dashrightarrow in Figure III.

3 DSR(Dynamic Source Routing) in EVALPSN

In order to translate the DSR protocol into EVALPSN, three kinds of messages, route request, route reply and route error should be formalized in EVALP clauses as follows.

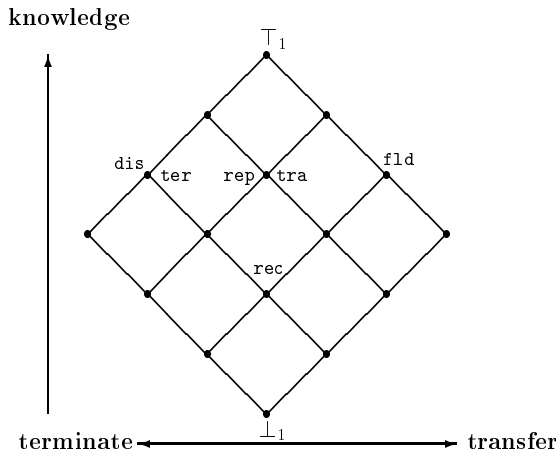


Fig. 2. The Complete Lattice of Route Message Operations

Route request message contains route request message id $reqid$, the object node obn where the message has been received, the source node son where the message has been flooded, the destination node den of the route request message, and the ordered list $nseq$ of nodes representing the route from the origin node to the source node. Node operations for route request messages are receiving (represented by annotation rec in EVALPSN), flooding (by annotation fld), which means sending the route request message to anonymous nodes, replying (by annotation rep), which means sending the route reply message to the origin node according to the reverse order of the route node list in the route request

message, and discarding (by annotation **dis**), which means discarding the route request message if the same one has been already received.

Here we analyze the node operations represented by the annotations **rec**, **fld**, **rep** and **dis** with the independent bi-criteria, message transferability and knowledge amount. If we take the viewpoint of message transferability, we may have the order between the node operations, flooding, transferring(replying), discarding, and operation receiving is neutral for message transferability. On the other hand, if we take the viewpoint of knowledge amount, since four operations, flooding, transferring(replying) and discarding, should be executed after operation receiving, the four operations imply knowledge that the message has been already received. Therefore, if we distribute those annotations representing the node operations in the complete lattice of vector annotations, they are located as shown in Figure 2. The complete lattice is a bi-lattice in which its vertical criterion is knowledge amount and its horizontal one is message transferability. For example, we may regard that operation flooding has the highest message transferability and more knowledge than operation receiving in terms of message treatment. On the other hand, operation discarding may be regarded that it has the lowest message transferability.

The route request message is formalized in the EVALP literal,

$$rreq(req_{id}, obn, son, den, nseq, t) : [\mu_1, \mu],$$

where $\mu_1 \in \{\perp(0, 0), \dots, \mathbf{rec}(1, 1), \dots, \mathbf{dis}(1, 3), \dots, \mathbf{rep}(2, 2), \dots, \mathbf{fld}(3, 1), \dots, \top(3, 3)\}$, and $\mu \in \mathcal{T}_e$. For example, the EVALP clause $rreq(req_{id}, obn, son, den, nseq, t) : [\mathbf{fld}, \beta]$ can be intuitively interpreted that the route request message req_{id} flooded from the source node son must be flooded from the object node obn with node list $nseq$ and the destination node den at time t .

Route reply message contains route reply message id rep_{id} , which is the same as the corresponding route request message id, the object node obn , the source node son , the destination node des , which is the origin node of the corresponding route request message, the ordered list $nseq$ of nodes representing the route. Node operations for route reply messages are receiving (represented by annotation **rec** in EVALPSN as well as route request messages), transferring (by annotation **tra**), which means sending the route reply message directly to the next node in the ordered list $nseq$ of nodes, and terminating (by annotation **ter**), which means the termination operation for route reply messages if the object node is the destination of the route reply message. The complete lattice structure of annotations **rec**, **tra** and **ter** for route reply messages are also shown in Figure 2 as well as those for route request messages. Therefore, route reply message is formalized in the EVALP literal,

$$rrep(rep_{id}, obn, son, den, nseq, t) : [\mu_2, \mu],$$

where $\mu_2 \in \{\perp(0, 0), \dots, \mathbf{rec}(1, 1), \dots, \mathbf{ter}(1, 3), \dots, \mathbf{tra}(2, 2), \dots, \top(3, 3)\}$, and $\mu \in \mathcal{T}_e$. For example, the EVALP clause $rrep(rep_{id}, obn, son, den, nseq, t) : [\mathbf{rec}, \alpha]$ can be intuitively interpreted that the route reply rep_{id} transferred from the

source node *son* has been received with the node list *nseq* and the destination node *den* by the object node *obn* at time *t*.

Route error message contains route error message id err_{id} , the object node *obn* where the message has been received, the source node *son* where the message has been transferred, the destination node *den* of the route error message, and the ordered list *nseq* of nodes representing the route. Node operations for route error messages are receiving (represented by annotation **rec** in EVALPSN), transferring (by annotation **tra**) and terminating (by annotation **ter**) as well as those for route reply message. Therefore, route error message is formalized in the EVALP literals,

$$rerr(err_{id}, obn, son, den, eseq, t) : [\mu_2, \mu],$$

where $\mu_2 \in \{\perp(0, 0), \dots, \mathbf{rec}(1, 1), \dots, \mathbf{ter}(1, 3), \dots, \mathbf{tra}(2, 2), \dots, \top_1(3, 3)\}$, and $\mu \in \mathcal{T}_e$. For example, the EVALP clause $rerr(err_{id}, obn, son, den, eseq, t) : [\mathbf{s}, \beta]$ can be intuitively interpreted that the route error message err_{id} transferred from the node *son* must be sent with the node list *eseq* and the destination node *den* from the object node *obn* at time *t*.

We need to define another EVALP literal to formalize the DSR operations.

Relation between nodes in ordered list of nodes is formalized in the EVALP literal,

$$rel(n_i, n_j, nseq) : [\mu_3, \mu],$$

where $\mu_3 \in \{\perp(0, 0), \dots, \mathbf{pr}(2, 0), \mathbf{eq}(1, 1), \mathbf{su}(0, 2), \dots, \top(1, 1)\}$, and $\mu \in \mathcal{T}_e$, and annotations **pr**, **eq** and **su** declare that one node precedes/is equal to/succeeds another one. For example, the EVALP literal, $rel(n_4, n_0, <n_0, n_4>) : [\mathbf{su}, \alpha]$ can be intuitively interpreted that node $N_4(n_4)$ succeeds node $N_0(n_0)$ in the ordered list *nseq* of nodes.

The node operations described in the previous section are translated into EVALPSN.

1. route request message discarding

$$\begin{aligned} rreq(req_{id}, obn, son, den, nseq, t) &: [\mathbf{rec}, \alpha] \wedge \\ rreq(req_{id}, obn, son, den, nseq, t') &: [\mathbf{rec}, \alpha] \wedge \\ bf(t', t) &: [\mathbf{t}, \alpha] \rightarrow \\ rreq(req_{id}, obn, son, den, nseq, t) &: [\mathbf{dis}, \beta] \end{aligned} \quad (1)$$

where $bf(t', t) : [\mathbf{t}, \alpha]$ declares that time t' is before time t .

2. route request message replying

$$\begin{aligned} rreq(req_{id}, obn, son, obn, nseq, t) &: [\mathbf{rec}, \alpha] \wedge \\ \sim rreq(req_{id}, obn, son, obn, nseq, t) &: [\mathbf{dis}, \beta] \rightarrow \\ rrep(req_{id}, son, obn, hn, nseq \cup <obn>, t) &: [\mathbf{tra}, \beta], \end{aligned} \quad (2)$$

where $nseq \cup <obn>$ represents the list that the object node *obn* is added to the node list *nseq* as its last member, and *hn* represents the origin node of the route request in the node list *nseq*.

3. route request message flooding

$$\begin{aligned} rcr(req_{id}, obn, den, t) : [\mathbf{is}, \alpha] \rightarrow \\ rreq(req_{id}, obn, des, \langle n_o \rangle, t) : [\mathbf{fld}, \beta] \end{aligned} \quad (3)$$

$$\begin{aligned} rreq(req_{id}, obn, son, den, nseq, t) : [\mathbf{rec}, \alpha] \wedge \\ \sim rreq(req_{id}, obn, son, den, nseq, t) : [\mathbf{dis}, \beta] \wedge \\ \sim rrep(req_{id}, son, obn, hn, nseq \cup \langle obn \rangle, t) : [\mathbf{tra}, \beta] \rightarrow \\ rreq(req_{id}, ano, obn, den, nseq \cup \langle obn \rangle, t) : [\mathbf{fld}, \beta], \end{aligned} \quad (4)$$

where $rcr(req_{id}, n_o, den, t) : [\mathbf{is}, \alpha]$ declares that the network user of node n_o has issued the route request req_{id} .

4. route reply message transferring

$$\begin{aligned} rrep(req_{id}, obn, son, den, nseq, t) : [\mathbf{rec}, \alpha] \wedge \\ \sim rrep(req_{id}, obn, son, den, nseq, t) : [\mathbf{ter}, \beta] \wedge \\ \wedge rel(nen, obn, nseq) : [\mathbf{prec}, \alpha] \rightarrow rrep(req_{id}, nen, obn, den, nseq, t) : [\mathbf{tra}, \beta] \end{aligned} \quad (5)$$

5. route reply message termination

$$\begin{aligned} rrep(req_{id}, obn, son, obn, nseq, t) : [\mathbf{rec}, \alpha] \rightarrow \\ rrep(req_{id}, obn, son, obn, nseq, t) : [\mathbf{ter}, \beta]. \end{aligned} \quad (6)$$

6. route error message transferring

$$\begin{aligned} error(obn, ern, t) : [\mathbf{d}, \alpha] \wedge rrep(req_{id}, obn, ern, den, nseq, t') : [\mathbf{tra}, \alpha] \wedge \\ bf(t', t) : [\mathbf{t}, \alpha] \wedge rel(nen, obn, nseq) : [\mathbf{prec}, \alpha] \rightarrow \\ rerr(req_{id}, nen, obn, den, nseq, t) : [\mathbf{tra}, \beta], \end{aligned} \quad (7)$$

where $error(obn, ern, t) : [\mathbf{d}, \alpha]$ declares that a route error has been detected between the object node obn and the error node ern at time t ,

$$\begin{aligned} rerr(req_{id}, obn, son, den, nseq, t) : [\mathbf{rec}, \alpha] \wedge \\ rrep(req_{id}, obn, ern, den, nseq, t') : [\mathbf{tra}, \alpha] \wedge \\ bf(t', t) : [\mathbf{t}, \alpha] \wedge rel(nen, obn, nseq) : [\mathbf{prec}, \alpha] \rightarrow \\ rerr(req_{id}, nen, obn, den, nseq, t) : [\mathbf{tra}, \beta]. \end{aligned} \quad (8)$$

7. route error message termination

$$\begin{aligned} rerr(req_{id}, obn, son, obn, nseq, t) : [\mathbf{rec}, \alpha] \rightarrow \\ rerr(req_{id}, obn, son, obn, nseq, t) : [\mathbf{ter}, \beta] \\ (rcr(req'_{id}, obn, den, t) : [\mathbf{is}, \beta]), \end{aligned} \quad (9)$$

where the route error message termination may be changed for new route discovery.

4 Example

Here we take the same DSR protocol example shown in Figure 1 and show the EVALPSN network routing based on the DSR protocol.

Stage 1 (time t_1) at node $N_0(n_0)$, suppose that a network user issued the route request req_1 with the destination node $N_d(n_d)$, then we have the fact EVALP clause,

$$rcr(req_1, n_0, n_d, t_1):[\mathbf{is}, \alpha],$$

and the obligatory EVALP clause,

$$rreq(req_1, ano, n_0, n_d, < n_0 >, t_1):[\mathbf{fld}, \beta]$$

for flooding the route request message req_1 is derived by the EVALP clause (3).

Stage 2 (time t_2) at node $N_4(n_4)$, suppose that the flooded route request message req_1 from node N_0 has been received, then we have the fact EVALP clause,

$$rreq(req_1, n_4, n_0, n_d, < n_0 >, t_2):[\mathbf{rec}, \alpha]. \quad (10)$$

Moreover, since the obligatory EVALP clauses,

$$\begin{aligned} rreq(req_1, n_4, n_0, n_d, < n_0 >, t_2):[\mathbf{dis}, \beta] \quad \text{and} \\ rrep(req_1, n_0, n_4, n_0, < n_0, n_4 >, t_2):[\mathbf{tra}, \beta] \end{aligned}$$

cannot be derived by the EVALP clauses (1) and (2), we obtain only the obligatory EVALP clause,

$$rreq(req_1, ano, n_0, n_d, < n_0 >, t_1):[\mathbf{fld}, \beta]$$

for flooding the route request message req_1 by the EVALP clause (3).

Stage 3 (time t_3) at node $N_4(n_4)$, suppose that the flooded route request message from node N_3 has been received, then we have the fact EVALP clause,

$$rreq(req_1, n_4, n_3, n_d, < n_0, n_3 >, t_3):[\mathbf{rec}, \alpha],$$

and we already have the fact EVALP clause (10), which can be regarded as the record of route request messages received at node N_4 before, then the obligatory EVALP clause,

$$rreq(req_1, n_4, n_3, n_d, < n_0, n_3 >, t_3):[\mathbf{dis}, \beta]$$

for discarding the route request with the same request id req_1 is derived by the EVALP clause (1).

Stage 4 (time t_4) at node $N_6(n_6)$, suppose that the flooded route request message from node N_4 has been received, then the same operations at node N_4 are carried out and the obligatory EVALP clause,

$$rreq(req_1, ano, n_4, n_d, < n_0, n_4, n_6 >, t_4):[\mathbf{fld}, \beta],$$

for flooding the route request message req_1 is derived by the EVALP clause (3).

Stage 5 (time t_5) at node $N_d(n_d)$, suppose that the flooded route request message req_1 from node N_6 has been received, then we have the fact EVALP clause,

$$rreq(req_1, n_d, n_6, n_d, < n_0, n_4, n_6 >, t_5): [\mathbf{rec}, \alpha].$$

Since the obligatory EVALP clause,

$$rreq(req_1, n_d, n_6, n_d, < n_0, n_4, n_6 >, t_5): [\mathbf{dis}, \beta],$$

cannot be derived by the EVALP clause (II), the obligatory EVALP clause,

$$rrep(req_1, n_d, n_6, n_0, < n_0, n_4, n_6, n_d >, t_5): [\mathbf{tra}, \beta],$$

for transferring the route reply message for the route request message req_1 is derived.

Stage 6 (time t_6) at node $N_6(n_6)$, suppose that the route reply message req_1 transferred from node N_d has been received, then we have the fact EVALP clause,

$$rrep(req_1, n_6, n_d, n_0, < n_0, n_4, n_6, n_d >, t_6): [\mathbf{rec}, \alpha].$$

However, the obligatory EVALP clause,

$$rrep(req_1, n_6, n_d, n_0, < n_0, n_4, n_6, n_d >, t_6): [\mathbf{ter}, \beta]$$

cannot be derived by the EVALP clause (6). Therefore, we obtain the obligatory EVALP clause,


$$rrep(req_1, n_4, n_6, n_0, < n_0, n_4, n_6, n_d >, t_6): [\mathbf{tra}, \beta],$$

for transferring the route reply message req_1 is derived by the EVALP clause (5).

5 Future Work and Conclusion

Our goal of this work is to make up an EVALPSN based intelligent routing system, which can deal with various kinds of routing protocols on the same EVALPSN platform uniformly. As the first step to the goal, we have introduced the most basic EVALPSN single-path routing protocol based on the DSR protocol in this paper, as the next step, we are planning to make up an EVALPSN multi-path routing system in which various kinds of multi-path routing protocols can be dealt with on the same EVALPSN platform uniformly.

As the advantage of the EVALPSN multi-path routing system, we are considering to apply defeasible/plausible reasoning in EVALPSN [7,10] to two kinds of decision making, one is route-discovery in multi-path routing and another one is protocol-selection in the EVALPSN multi-path ad-hoc routing system. It is preferable to discover more than two mutually independent routes as possible in multi-path routing, and EVALPSN defeasible/plausible reasoning can be applied to discovering such routes. Furthermore, it is also preferable to select the most suitable routing protocol under the network environment such as

data traffic congestion in real-time, and the defeasible/plausible reasoning can be also applied to selecting such a suitable protocol. In practice we have already applied EVALPSN defeasible reasoning to traffic signal control aiming at traffic jam reduction in [6], and the research results could be applied to ad-hoc network routing with intelligent data traffic control. 

References

1. Johnson, D.B., Maltz, D.A.: Dynamic Source Routing in Ad-Hoc Wireless Networks. In: *Mobile Computing*, ch. 5, pp. 153–181. Kluwer Academic Publishers, Dordrecht (1996)
2. Johnson, D.B., Maltz, D.A., Hu, A.Y., Jetcheva, J.G.: The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks. Internet draft, draft-ietfmanet-dsr-0.9.txt (2003)
3. Nakamatsu, K., Abe, J.M., Suzuki, A.: Annotated Semantics for Defeasible Deontic Reasoning. In: Ziarko, W.P., Yao, Y. (eds.) *RSCTC 2000*. LNCS (LNAI), vol. 2005, pp. 432–440. Springer, Heidelberg (2001)
4. Nakamatsu, K.: Pipeline Valve Control Based on EVALPSN Safety Verification. *J. Advanced Computational Intelligence and Intelligent Informatics* 10, 647–656 (2006)
5. Nakamatsu, K., Mita, Y., Shibata, T.: An Intelligent Action Control System Based on Extended Vector Annotated Logic Program and its Hardware Implementation. *J. Intelligent Automation and Soft Computing* 13, 289–304 (2007)
6. Nakamatsu, K., Abe, J.M., Akama, S.: An Intelligent Coordinated Traffic Signal Control Based on EVALPSN. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) *KES 2007, Part II*. LNCS (LNAI), vol. 4693, pp. 869–876. Springer, Heidelberg (2007)
7. Nakamatsu, K.: The Paraconsistent Annotated Logic Program EVALPSN and its Application. In: *Computational Intelligence: A Compendium*. Studies in Computational Intelligence, vol. 115, pp. 233–306. Springer, Heidelberg (2008)
8. Nakamatsu, K., Abe, J.M., Akama, S.: Paraconsistent Before-after Relation Reasoning Based on EVALPSN. In: *New Directions in Intelligent Interactive Multimedia*. Studies in Computational Intelligence, vol. 142, pp. 265–274. Springer, Heidelberg (2008)
9. Nakamatsu, K., Akama, S., Abe, J.M.: Transitive Reasoning of Before-After Relation Based on Bf-EVALPSN. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) *KES 2008, Part II*. LNCS (LNAI), vol. 5178, pp. 474–482. Springer, Heidelberg (2008)
10. Nakamatsu, K., Imai, T., Abe, J.M., Akama, S.: An Introduction to Plausible Reasoning Based on EVALPSN. In: *New Advances in Intelligent Decision Technologies*. Studies in Computational Intelligence, vol. 199, pp. 363–372. Springer, Heidelberg (2009)
11. Nakamatsu, K., Abe, J.M.: The development of Paraconsistent Annotated Logic Program. *Int'l J. Reasoning-based Intelligent Systems* 1, 92–112 (2009)

¹ This work is financially supported by Japanese Scientific Research Grant (C) No. 20500074.

Introduction to Intelligent Elevator Control Based on EVALPSN

Kazumi Nakamatsu¹, Jair Minoro Abe²,
Seiki Akama³, and Roumen Kountchev⁴

¹ University of Hyogo, Himeji, Japan
nakamatu@shse.u-hyogo.ac.jp

² Paulista University, University of Sao Paulo, Sao Paulo, Brazil
jairabe@uol.com.br

³ University of Tsukuba, Tsukuba, Japan
akama@jcom.home.ne.jp

⁴ Technical University of Sofia, Sofia, Bulgaria
rkountch@tu-sofia.ac.bg

Abstract. In this paper, we introduce the basic idea of single-car/single-shaft elevator control based on EVALPSN. Car operation in the elevator system is logically verified by the EVALPSN representing the car operation properties.

Keywords: elevator control, defeasible deontic reasoning, logical verification, annotated logic program, EVALPSN.

1 Introduction

We have already developed a paraconsistent annotated logic program called Extended Vector Annotated Logic Program with Strong Negation (abbr. EVALPSN) [3,7] which can be applied to various real-time intelligent control and safety verification systems such as pipeline process control [4,5]. The features and advantages of EVALPSN based control are: (1) since EVALPSN can deal with deontic notions such as forbiddance, control policy in deontic expression can be easily translated into EVALPSN; (2) logical verification of operation control can be easily carried out as logic programming; (3) since some restricted fragment of EVALPSN can be implemented on microchips as electronic circuits, real, suitable for real-time control.

In this paper, we apply EVALPSN to elevator control and introduce the basic idea of EVALSN based elevator control. Recently, multi-car elevator systems in which more than two cars are operated in the same shaft have been implemented practically and started their services in some countries, however their control should be required verifying the safety for operation. Our work is aiming to provide an EVALPSN based control system for multi-car elevator systems. As the first step to our goal, we introduce the basic idea of EVALPSN based elevator control with taking a single-car/single-shaft elevator system as an object in this paper. The car operation in the elevator system is divided into two kinds

of operation, one is preservice operation for picking up passengers waiting at floors and another one is passenger service operation for carrying passengers to their destinations. For each operation we suppose operation properties to be assured for safe operation. The properties are translated into EVALPSN and floor call and passenger request are verified for their safety based on the EVALPSN. We introduce the formalization of verifying the safety for elevator operation in EVALPSN and show a basic and essential example of the verification.

This paper is organized in the following manner: first, EVALPSN is reviewed briefly, and the outline of the single-car/single-shaft elevator and its service properties are introduced; next, the service properties are translated into an EVALPSN and a simple example of the EVALPSN based elevator control is provided; last, the future development of EVALPSN based elevator control to single-car/multi-shaft elevator systems and multi-car/multi-shaft elevator systems are introduced as the future work.

2 EVALPSN

We review EVALPSN briefly [3]. Generally, a truth value called an *annotation* is explicitly attached to each literal in annotated logic programs [1]. For example, let p be a literal, μ an annotation, then $p:\mu$ is called an *annotated literal*. The set of annotations constitutes a complete lattice. An annotation in EVALPSN has a form of $[(i, j), \mu]$ called an *extended vector annotation*. The first component (i, j) is called a *vector annotation* and the set of vector annotations constitutes the complete lattice,

$$\mathcal{T}_v(n) = \{ (x, y) | 0 \leq x \leq n, 0 \leq y \leq n, x, y \text{ and } n \text{ are integers} \}$$

in Figure 1. The ordering (\preceq_v) of $\mathcal{T}_v(n)$ is defined as: let $(x_1, y_1), (x_2, y_2) \in \mathcal{T}_v(n)$,

$$(x_1, y_1) \preceq_v (x_2, y_2) \text{ iff } x_1 \leq x_2 \text{ and } y_1 \leq y_2.$$

For each extended vector annotated literal $p:[(i, j), \mu]$, the integer i denotes the amount of positive information to support the literal p and the integer j denotes that of negative one. The second component μ is an index of fact and deontic

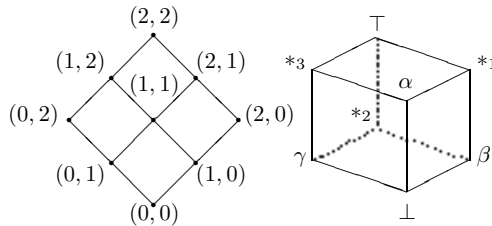


Fig. 1. Lattice $\mathcal{T}_v(2)$ and Lattice \mathcal{T}_d

notions such as obligation, and the set of the second components constitutes the complete lattice,

$$\mathcal{T}_d = \{\perp, \alpha, \beta, \gamma, *_1, *_2, *_3, \top\}.$$

The ordering(\preceq_d) of \mathcal{T}_d is described by the Hasse's diagram in Figure 1. The intuitive meaning of each member of \mathcal{T}_d is \perp (unknown), α (fact), β (obligation), γ (non-obligation), $*_1$ (fact and obligation), $*_2$ (obligation and non-obligation), $*_3$ (fact and non-obligation), and \top (inconsistency). Then the complete lattice $\mathcal{T}_e(n)$ of extended vector annotations is defined as the product $\mathcal{T}_v(n) \times \mathcal{T}_d$. The ordering(\preceq_e) of $\mathcal{T}_e(n)$ is defined as : let $[(i_1, j_1), \mu_1]$ and $[(i_2, j_2), \mu_2] \in \mathcal{T}_e$,

$$[(i_1, j_1), \mu_1] \preceq_e [(i_2, j_2), \mu_2] \text{ iff } (i_1, j_1) \preceq_v (i_2, j_2) \text{ and } \mu_1 \preceq_d \mu_2.$$

There are two kinds of *epistemic negation* (\neg_1 and \neg_2) in EVALPSN, both of which are defined as mappings over $\mathcal{T}_v(n)$ and \mathcal{T}_d , respectively.

Definition 1(epistemic negations \neg_1 and \neg_2 in EVALPSN)

$$\begin{aligned} \neg_1([(i, j), \mu]) &= [(j, i), \mu], & \forall \mu \in \mathcal{T}_d, \\ \neg_2([(i, j), \perp]) &= [(i, j), \perp], & \neg_2([(i, j), \alpha]) &= [(i, j), \alpha], \\ \neg_2([(i, j), \beta]) &= [(i, j), \gamma], & \neg_2([(i, j), \gamma]) &= [(i, j), \beta], \\ \neg_2([(i, j), *_1]) &= [(i, j), *_3], & \neg_2([(i, j), *_2]) &= [(i, j), *_2], \\ \neg_2([(i, j), *_3]) &= [(i, j), *_1], & \neg_2([(i, j), \top]) &= [(i, j), \top]. \end{aligned}$$

If we regard the epistemic negations as syntactical operations, the epistemic negations followed by literals can be eliminated by the syntactical operations. For example, $\neg_1(p: [(2, 0), \alpha]) = p: [(0, 2), \alpha]$ and $\neg_2(q: [(1, 0), \beta]) = p: [(1, 0), \gamma]$.

There is another negation called *strong negation* (\sim) in EVALPSN, and it is treated as well as classical negation.

Definition 2(strong negation \sim) [2]. Let F be any formula and \neg be \neg_1 or \neg_2 .

$$\sim F =_{def} F \rightarrow ((F \rightarrow F) \wedge \neg(F \rightarrow F)).$$

Definition 3 (well extended vector annotated literal). Let p be a literal.

$$p: [(i, 0), \mu] \quad \text{and} \quad p: [(0, j), \mu]$$

are called *weva(well extended vector annotated)-literals*, where $i, j \in \{1, 2, \dots, n\}$, and $\mu \in \{\alpha, \beta, \gamma\}$.

Definition 4 (EVALPSN). If L_0, \dots, L_n are weva-literals,

$$L_1 \wedge \dots \wedge L_i \wedge \sim L_{i+1} \wedge \dots \wedge \sim L_n \rightarrow L_0$$

is called an *EVALPSN clause*. An *EVALPSN* is a finite set of EVALPSN clauses.

We note that if the annotations α and β represent fact and obligation, notions “fact”, “obligation”, “forbiddance” and “permission” can be represented by extended vector annotations, $[(m, 0), \alpha]$, $[(m, 0), \beta]$, $[(0, m), \beta]$, and $[(0, m), \gamma]$, respectively in EVALPSN, where m is a positive integer.

3 Single-car Elevator Control Based on EVALPSN

For simplicity, we suppose that (refer to Figure 2): the elevator system physically consists of single car/single shaft with ten floors, and each floor has its floor id; generally, there are two kinds of requests from passengers; one is a *floor call* issued by passengers waiting at floors, which contains two components, the request direction (up or down) and the floor id where the call has been issued; and another one is a *passenger request* issued by passengers in the car, which contains one factor, the destination floor id; moreover, a floor where a floor call is issued is called a *passenger floor*, a floor where the car is located is called the *current floor*, a *route* is logically defined as a series of floors, the *current route* is a route in which the car is serving at present, the terminal floor of the current route is called the *destination floor*, and a route from the destination floor to a passenger floor is called a *preservice route*, which is a service route for picking up passengers at the passenger floor.

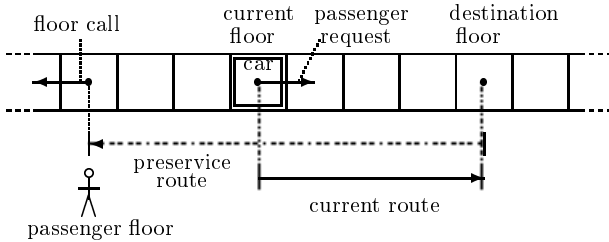


Fig. 2. Single Car Elevator System

Now we outline the elevator control with referring to Figure 3; if a floor call represented as an arrow symbol in Figure 3 has been issued, in order to respond to the floor call its preservice route represented as a dotted arrow in Figure 3 should be created, and whether the preservice route can be processed by the car or not is logically verified. We classified the relations between the current route and preservice routes into six cases shown in Figure 3:

- (floor call 1) the passenger floor is in the current route, and the request direction and the current route direction are the same, then its preservice route should be the route from the current floor to the passenger floor, and the current route direction and the preservice route direction should be the same;
- (floor call 2) the passenger floor is forward of the destination floor, and the request direction and the current route direction are the same, then the preservice route should be the route from the destination floor to the passenger floor, and the current route direction and the preservice route direction are the same;

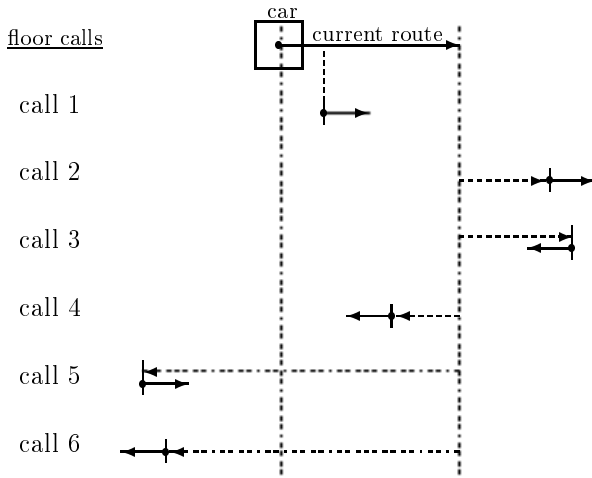


Fig. 3. Elevator Control Principles

- (floor call 3) the passenger floor is forward of the destination floor and the request direction and the current route direction are not the same, the preservice route should be the route from the destination floor to the passenger floor, and the current route direction and the preservice route direction are the same;
- (floor call 4) the passenger floor is in the current route, and the request direction and the current route direction are not the same, then the preservice route should be the route from the destination floor to the passenger floor, and the current route direction and the preservice route direction are not the same;
- (floor call 5) the passenger floor is backward of the current floor and the request direction and the current route direction are the same, then the preservice route should be the route from the destination floor to the passenger floor, and the current route direction and the preservice route direction are not the same;
- (floor call 6) the passenger floor is backward of the current floor and the request direction and the current route direction are not the same, then the preservice route should be the route from the destination floor to the passenger floor, and the current route direction and the preservice route direction are not the same.

We suppose operation properties for the single-car elevator control as follows.

SP-1. Floor calls whose preservice route has the same direction as the current route should be processed prior to other floor calls.

Property **SP-1** can be interpreted deontically that if the preservice route direction is not the same as the current route direction, the preservice route is not permitted to be migrated into the current route. Therefore, the preservice route direction of floor calls 4, 5 and 6 is not the same as the current route direction, and they must wait for being processed until the current route direction has

turned. On the other hand, floor calls 1, 2 and 3 should be sequentially processed from the nearest passenger floor of the current floor.

Another car operation policy is for passenger requests.

SP-2. Passenger requests having the same direction as the current route should be served prior to other passenger requests.

Property **SP-2** can be interpreted deontically that if the passenger request direction is not the same as the current route direction, the passenger request route is not permitted to be migrated into the current route.

We will translate the operation properties into EVALPSN to construct an EVALPSN based elevator control system in the following section.

4 EVALPSN for Singel-Car Elevator Control

First of all, We define some EVALP literals used in the knowledge representation of the elevator control system.

Current Floor is formalized in the EVALP literal, $fl(fl_{id}, c_{id}, t) : [\mu_1, \mu]$, where

$$\mu_1 \in \{\perp(0, 0), \circ(1, 0), \vee(0, 1), \top(1, 1)\}, \quad \text{and} \quad \mu \in \mathcal{T}_e,$$

and annotations \circ and \vee show that floor fl_{id} is occupied by car c_{id} or not(vacant). For example, EVALP clause $fl(fl_{id}, c_{id}, t) : [\circ, \alpha]$ can be intuitively interpreted that car c_{id} has occupied floor fl_{id} since time t .

Notes: if the elevator system has more than one shaft, floor id may be a pair of integers such as (n, m) or $n - m$ in which n is the shaft id and m is the floor number, if the elevator system is single-car/single-shaft, the term c_{id} representing car id in the EVALP literal is not necessary.

Route Direction is formalized in the EVALP clause, $dir(fl_i, fl_j) : [\mu_2, \mu]$, where

$$\mu_2 \in \{\perp(0, 0), \text{up}(2, 0), \text{no}(1, 1), \text{dw}(0, 2), \top(2, 2)\}, \quad \text{and} \quad \mu \in \mathcal{T}_e,$$

and annotations **up**, **dw** and **no** show that the direction of the route from floor fl_i to floor fl_j is upward, downward or no direction. For example, EVALP clause $dir(fl_i, fl_j) : [\text{up}, \alpha]$ can be intuitively interpreted that the direction of the route from floor fl_i to floor fl_j is upward.

Floor Call is formalized in the EVALP literal, $fc(fl_r, dir, t) : [\mu_3, \mu]$, where

$$\mu_3 \in \{\perp(0, 0), \text{is}(1, 0), \text{ui}(0, 1), \top(1, 1)\}, \quad \text{and} \quad \mu \in \mathcal{T}_e,$$

and annotations **is** and **ui** show that the floor call has been issued or not, $dir \in \{\text{up}, \text{dw}\}$, and **up** and **dw** express ‘upward’ and ‘downward’ respectively. For example, EVALP clause $fc(fl_r, dir, t) : [\text{is}, \alpha]$ is intuitively interpreted that the floor call with direction dir at floor fl_r has been issued at time t .

Passenger Request is formalized in an EVALP literal, $pr(fl_r, c_{id}, t) : [\mu_4, \mu]$, where

$$\mu_4 \in \{\perp(0, 0), \text{is}(1, 0), \text{ui}(0, 1), \top(1, 1)\}, \quad \text{and} \quad \mu \in \mathcal{T}_e,$$

and annotations **is** and **ui** show that the passenger request has been issued or not, For example, EVALP clause $pr(fl_r, c_{id}, t) : [\mathbf{is}, \alpha]$ is intuitively interpreted that the passenger request bound for floor fl_r has been issued at time t .

Route is formalized in the EVALP literal, $ro(fl_c, fl_d, c_{id}, t) : [\mu_5, \mu]$, where

$$\mu_5 \in \{\perp(0, 0), \mathbf{s}(1, 0), \mathbf{x}(0, 1), \top(1, 1)\}, \quad \text{and} \quad \mu \in \mathcal{T}_e,$$

and annotations **s** and **x** show that the route from floor fl_c to floor fl_d is set for car c_{id} or not. Here "a route is set" means the route is ready to be served by the car. For example, EVALP clause $ro(fl_c, fl_d, c_{id}, t) : [\mathbf{x}, \beta]$ is intuitively interpreted that the route from floor fl_c to floor fl_d must not be set(**s**) for car c_{id} at time t .

Note: car id c_{id} is not necessary if the elevator system is single-car/single-shaft.

Car Stop is formalized in the EVALP literal, $st(fl, c_{id}, dir, t) : [\mu_6, \mu]$, where

$$\mu_6 \in \{\perp(0, 0), \mathbf{s}(1, 0), \mathbf{x}(0, 1), \top(1, 1)\}, \quad \text{and} \quad \mu \in \mathcal{T}_e,$$

and annotations **s** and **x** show that a car stop bound for dir at floor fl is set for car c_{id} or not, $dir \in \{up, te, dw\}$, and te expresses that the car stop floor is the terminal floor of the route. For example, EVALP clause $st(fl, c_{id}, dir, t) : [\mathbf{x}, \gamma]$ is intuitively interpreted that the car stop for car c_{id} with direction dir is permitted to be set(**s**) at floor fl at time t .

Then, properties **SP-1** and **SP-2** are translated into the EVALPSN clauses,

SP-1

$$fc(fl_r, up/dw, t) : [\mathbf{is}, \alpha] \wedge ro(fl_c, fl_o, c_{id}, t) : [\mathbf{s}, \alpha] \wedge fl(fl_c, c_{id}, t) : [\mathbf{o}, \alpha] \wedge dir(fl_c, fl_o) : [\mathbf{dw/up}, \alpha] \wedge dir(fl_o, fl_r) : [\mathbf{up/dw}, \alpha] \rightarrow \quad (1)$$

$$ro(fl_c, fl_r, c_{id}, t) : [\mathbf{x}, \beta], \quad (2)$$

where the conjunction part (1) represents the condition that the current route direction and the preservice route direction are not the same, and the consequent EVALP clause (2) represents the forbiddance that the route from floor fl_c to floor fl_r must not be set, that is to say, the floor call must not be processed at time t .

SP-2

$$pr(fl_r, c_{id}, t) : [\mathbf{is}, \alpha] \wedge ro(fl_c, fl_d, c_{id}, t) : [\mathbf{s}, \alpha] \wedge fl(fl_c, c_{id}, t) : [\mathbf{o}, \alpha] \wedge dir(fl_c, fl_d) : [\mathbf{up/dw}, \alpha] \wedge dir(fl_c, fl_r) : [\mathbf{dw/up}, \alpha] \rightarrow \quad (3)$$

$$ro(fl_c, fl_r, c_{id}, t) : [\mathbf{x}, \beta], \quad (4)$$

where the conjunction part (3) represents the condition that the current route direction and the passenger request route direction are not the same, and the consequent EVALP clause (4) represents the forbiddance that the route from floor fl_c to floor fl_r must not be set, that is to say, the passenger request must be ignored at time t .

On the other hand, if forbiddance from setting routes cannot be derived, permission for setting routes should be derived. Therefore, we need the following EVALPSN clause **DP** for deriving permission.

$$\mathbf{DP} \quad \sim ro(fl_c, fl_r, c_{id}, t):[\mathbf{x}, \beta] \rightarrow ro(fl_c, fl_r, c_{id}, t):[\mathbf{x}, \gamma]. \quad (5)$$

Car stops should be set according to route set permission. If the route is generated by floor call, the corresponding car stop has direction ‘up’ or ‘down(dw)’, and if the route is generated by passenger request, the corresponding one has direction ‘terminal(te)’, which can be formalized by the EVALP clauses,

$$\mathbf{CS} \quad fc(fl_r, up/dw, t):[\mathbf{is}, \alpha] \wedge ro(fl_c, fl_r, c_{id}, t):[\mathbf{x}, \gamma] \rightarrow st(fl_r, c_{id}, up/dw, t):[\mathbf{s}, \beta], \quad (6)$$

$$pr(fl_r, c_{id}, t):[\mathbf{is}, \alpha] \wedge ro(fl_c, fl_r, c_{id}, t):[\mathbf{x}, \gamma] \rightarrow st(fl_r, c_{id}, te, t):[\mathbf{s}, \beta]. \quad (7)$$

Moreover, we need EVALP clauses for updating the current route, that is to say, for example, if the current route from 2nd floor to 5th floor has already been set, then if a route from the current floor 2nd floor to 7th floor is permitted to be set, the new current route from 2nd floor to 7th floor should be set. Such current route update can be formalized in the EVALP clauses,

CRU

$$ro(fl_c, fl_d, c_{id}, t):[\mathbf{s}, \alpha] \wedge ro(fl_c, fl_n, c_{id}, t):[\mathbf{x}, \gamma] \wedge fl(c_{id}, fl_c, t):[\mathbf{o}, \alpha] \wedge \sim dir(fl_c, fl_d):[\mathbf{up/dw}, \alpha] \wedge dir(fl_d, fl_n):[\mathbf{dw/up}, \alpha] \rightarrow ro(fl_c, fl_n, c_{id}, t):[\mathbf{s}, \beta]. \quad (8)$$

In practice, although release conditions for routes and car stops after they have been processed or migrated into other routes should be considered here, it is not essential, therefore we omit the release conditions for routes and stops.

Here we present a simple example for EVALPSN elevator control.

Example

We suppose a single-car/single-shaft elevator system with 10 floors. For simplicity, we represent a floor call as an ordered pair, (call issued floor, request direction), a root as (origin floor \implies destination floor) and a passenger request (in the car) as (request issued floor \implies destination floor). We consider the elevator control scenario, **stage 1**, **stage 2**, \dots , **stage 6**, with two floor calls call A and B and one passenger request req B by the passenger of call B, which is summarized in Table [II](#).

stage 1 suppose that call A (4,dw) is issued when the car is at 2nd floor with no request processing, EVALP **SP-1** cannot derive the forbiddance from setting a new route (2 \implies 4), thus route (2 \implies 4) is permitted to be set by EVALPSN **DP**, and must be set by EVALPSN **CRU**, moreover downward car stop 4 \downarrow is set at 4th floor by EVALPSN **CS**;

Table 1. Scenario for Example

Stage (floor)	Request \ EVALPSN	SP-1, DP	SP-2, DP	CS (Car Stop)	CRU (Current Route)
Stage 1 (2F)	call A (4,dw)	$2 \Rightarrow 4$	—	$4 \downarrow$	$2 \Rightarrow 4$
Stage 2 (3F)	call A	—	—	—	—
	call B (5,up)	$3 \Rightarrow 5$	—	$5 \uparrow$	$3 \Rightarrow 5$
Stage 3 (4F)	call A	—	—	—	—
	call B	$4 \Rightarrow 5$	—	$5 \uparrow$	$4 \Rightarrow 5$
Stage 4 (5F)	call A	—	—	—	—
	req B ($5 \Rightarrow 6$)	—	$5 \Rightarrow 6$	6	$5 \Rightarrow 6$
Stage 5 (6F)	call A	$6 \Rightarrow 4$	—	$4 \downarrow$	$6 \Rightarrow 4$
Stage 6 (5F)	call A	$5 \Rightarrow 4$	—	$4 \downarrow$	$5 \Rightarrow 4$

- stage 2** suppose that call B (5,up) is issued when 3rd floor is the current floor and the current route is ($3 \Rightarrow 4$), EVALP **SP-1** cannot derive the forbiddance from setting the new route ($3 \Rightarrow 5$), thus route ($3 \Rightarrow 5$) is permitted to be set by EVALPSN **DP**, and it must be set by EVALPSN **CRU**, moreover car stop $5 \uparrow$ is set at 5th floor by EVALPSN **CS**, on the other hand EVALP **SP-1** derives the forbiddance from setting the new preservice route ($3 \Rightarrow 4$) for call A (4,dw), therefore car stop $4 \downarrow$ has been canceled(unset);
- stage 3** suppose that both call A (4,dw) and call B (5,up) are tried to be processed again when 4th floor is the current floor and the current floor is ($4 \Rightarrow 5$), EVALP **SP-1** still derives the forbiddance from setting the new preservice route ($5 \Rightarrow 4$), for call A, on the other hand the forbiddance from setting the new preservice route ($4 \Rightarrow 5$) for call B cannot be derived, therefore route ($4 \Rightarrow 5$) and car stop $5 \uparrow$ must be set;
- stage 4** suppose that req B ($5 \Rightarrow 6$) is issued when the passenger gets on the car at the 5th floor and call A is tried again, call B and car stop $5 \uparrow$ are released at 5th floor, [\[1\]](#) then the current route is ($5 \Rightarrow 5$), EVALP **SP-1** still derives the forbiddance from setting the new preservice route ($5 \Rightarrow 4$), for call A, on the other hand EVALP **SP-2** cannot derive the forbiddance from setting the new route ($5 \Rightarrow 6$) for req B, therefore rout ($5 \Rightarrow 6$) and car stop 6 (terminal stop) are set;
- stage 5** suppose that call A (4,dw) is tried when the car has arrived at 6th floor and the passenger has got off the car, then req B and car stop 6 are released, and the current route is ($6 \Rightarrow 6$), EVALP **SP-1** cannot derive the forbiddance from setting the new preservice route ($6 \Rightarrow 4$) for call A, therefore route ($6 \Rightarrow 4$) and car stop $4 \downarrow$ are set;
- stage 6** suppose that the car is moving downward at 5th floor, the current route is ($5 \Rightarrow 4$), call A is tried again, EVALP **SP-1** cannot derive the forbiddance from setting the new preservice route ($5 \Rightarrow 4$) for call A, therefore route ($5 \Rightarrow 4$) and car stop $4 \downarrow$ are set again.

¹ We do not address such release conditions in details.

5 Conclusion and Future Work

In this paper, we have introduced the basic ideas of an EVALPSN based elevator control system having single-car/single-shaft, which is based on the logical verification of elevator operation policies though, our goal is to extend and apply the idea to multi-car/multi-shaft elevator control systems and to obtain a flexible and logical verification based multi-car elevator control. Then the ideas of the safety verification for railway interlocking or pipeline valve control seem to be applicable. Either way, we believe that the obtained elevator control system has high extendability such as two-car elevator system to three-car one because it is based on EVALPSN.

As another future work, we need to consider optimization of the EVALPSN elevator control, then efficient shaft assignment may be required if the elevator system has more than one shaft. Therefore, we are planning to apply plausible reasoning based on EVALPSN [6] to the shaft assignment problem. ²

References

1. Blair, H.A., Subrahmanian, V.S.: Paraconsistent Logic Programming. *Theoretical Computer Science* 68, 135–154 (1989)
2. da Costa, N.C.A., Subrahmanian, V.S., Vago, C.: The Paraconsistent Logics *PT*. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 37, 139–148 (1989)
3. Nakamatsu, K., Abe, J.M., Suzuki, A.: Annotated Semantics for Defeasible Deontic Reasoning. In: Ziarko, W.P., Yao, Y. (eds.) *RSTC 2000. LNCS (LNAI)*, vol. 2005, pp. 432–440. Springer, Heidelberg (2001)
4. Nakamatsu, K.: Pipeline Valve Control Based on EVALPSN Safety Verification. *J. Advanced Computational Intelligence and Intelligent Informatics* 10, 647–656 (2006)
5. Nakamatsu, K., Mita, Y., Shibata, T.: An Intelligent Action Control System Based on Extended Vector Annotated Logic Program and its Hardware Implementation. *J. Intelligent Automation and Soft Computing* 13, 289–304 (2007)
6. Nakamatsu, K., Imai, T., Abe, J.M., Akama, S.: An Introduction to Plausible Reasoning Based on EVALPSN. In: *New Advances in Intelligent Decision Technologies. SCI*, vol. 199, pp. 363–372. Springer, Heidelberg (2009)
7. Nakamatsu, K., Abe, J.M.: The development of Paraconsistent Annotated Logic Program. *Int'l J. Reasoning-based Intelligent Systems* 1, 92–112 (2009)

² This work is financially supported by Japanese Scientific Research Grant (C) No. 20500074.

Monadic Curry System N_1^*

Jair Minoro Abe^{1,2}, Kazumi Nakamatsu³, and Seiki Akama⁴

¹ Graduate Program in Production Engineering, ICET - Paulista University
R. Dr. Bacelar, 1212, CEP 04026-002 São Paulo – SP – Brazil

² Institute For Advanced Studies – University of São Paulo, Brazil
jairabe@uol.com.br

³ School of Human Science and Environment/H.S.E. – University of Hyogo – Japan
nakamatu@shse.u-hyogo.ac.jp

⁴ C-Republic, Tokyo, Japan
akama@jcom.home.ne.jp

Abstract. This paper is a sequel to [5], [6]. We present the Curry monadic system N_1^* which has as extensions the Curry monadic algebras C_1^* and P_1^* . All those systems are extensions of the classical monadic algebras introduced by Halmos [13]. Also the Curry monadic system N_1 constitutes an algebraic version of the non-alethic predicate logic N_1^* .

Keywords: Curry algebra, algebraic logic, paraconsistent logic, paracomplete logic, non-alethic logic.

1 Introduction

The concept of Curry System was introduced in a systematic way in [7], although some issues had been published by Da Costa previously. Curry systems can systematize a general theory of algebraization of logical systems. Actually, all mathematical treatment of logical notions can be viewed as Curry systems. More than this, enriching or modifying the concepts of Curry system, we can obtain as particular cases, logical matrix, Kripke structures, theory of models, which are not directly coped with problem of algebraization. In a certain sense, we can say that logic reduces to the study of Curry systems [7]. In [5] it was obtained a Curry monadic version of the predicate calculi C_n ($1 \leq n \leq \omega$) introduced by Da Costa. In [6], a Curry monadic version of the paracomplete predicate logics P_n ($1 \leq n \leq \omega$) was obtained. This paper is a sequel to these studies. We present the Curry monadic systems N_n ($1 \leq n \leq \omega$) which has as extensions the Curry monadic algebras C_n^* and P_n^* and can be viewed as an algebraic version of the non-alethic predicate logic N_n^* . All these systems constitute a generalization of the monadic algebras introduced by Halmos [13].

2 Background

We begin with some basic concepts. For a detailed account see [7].

A system $\langle A, \equiv, \leq \rangle$ is called a pre-ordered system if for all $x \in A$, $x \leq x$; for all $x, y, z \in A$, $x \leq y$ and $y \leq z$ imply $x \leq z$; for all $x, y, x', y' \in A$, $x \leq y$, $x \equiv x'$, and $y \equiv y'$ imply $x' \leq y'$.

A pre-ordered system $\langle A, \equiv, \leq \rangle$ is called a partially-ordered system if for all $x, y \in A$, $x \leq y$ and $y \leq x$ imply $x \equiv y$;

A partially-ordered system $\langle A, \equiv, \leq \rangle$ is called a pre-lattice system if for all $x, y \in A$, the set of $\sup\{x, y\} \neq \emptyset$ and the set of $\inf\{x, y\} \neq \emptyset$. We denote by $x \vee y$ one element of $\sup\{x, y\}$ and by $x \wedge y$ one element of $\inf\{x, y\}$.

A system $\langle A, \equiv, \leq, \rightarrow \rangle$ is called an implicative pre-lattice if $\langle A, \equiv, \leq \rangle$ is a pre-lattice, and for all $x, y, z \in A$, $x \wedge (x \rightarrow y) \leq y$ and $x \wedge y \leq z$ iff $x \leq y \rightarrow z$.

$\langle A, \equiv, \leq, \rightarrow \rangle$ is called classical implicative pre-lattice if it is an implicative pre-lattice and $(x \rightarrow y) \rightarrow x \leq x$ (Peirce's law).

With this definition we can extend the majority of algebraic systems to pre-algebraic systems considering an equivalence relation \equiv instead of equality relation. In this way we can obtain, v.g., the concepts of Boolean pre-algebras, pre-filters, pre-lattices, etc.

3 The Curry Algebra C_1

Definition 3.1. A Curry algebra C_1 (or a C_1 -algebra) is an implicative pre-lattice $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ with an maximum element 1 and operators \wedge, \vee , and $'$ satisfying the conditions below, where $x^0 =_{\text{Def.}} (x \wedge x')$:

- | | |
|--|--------------------------------------|
| (1) $x \vee x' \equiv 1$ | (5) $x \wedge y \leq (x \wedge y)^0$ |
| (2) $x'' \leq x$ | (6) $x \wedge y \leq (x \vee y)^0$ |
| (3) $y \leq (x \rightarrow y) \rightarrow ((x \rightarrow y') \rightarrow x')$ | (7) $x \leq (x')^0$ |
| (4) $x \wedge y \leq (x \rightarrow y)^0$ | |

Example 3.2. Let's consider the calculus C_1 [9]. A is the set of all formulas of C_1 . Let's consider as operations, the logical connectives of conjunction, disjunction, implication, and negation. Let's define the relation on A .

$x \equiv y$ iff $\vdash x \leftrightarrow y$. It is easy to check that \equiv is an equivalence relation on A .

$x \leq y$ iff $x \equiv x \wedge y$ and $y \leq x$ iff $y \equiv x \wedge y$. Also we take as 1 any fixed axiom instance.

The structure composed $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ is a C_1 -algebra.

Theorem 3.3. Let's $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a C_1 -algebra. Then the operator $'$ is non-monotone relatively \equiv .

Proof. See [15].

The earlier theorem says that every quotient algebra in the calculus C_1 is trivial. It's worthwhile to observe that [5] mention the existence of such non-monotone operators, but didn't give any concrete example.

Theorem 3.4. A C_1 -algebra is distributive and has a greatest element, as well as a first element.

Definition 3.5. Let x be an element of a C_1 -algebra. We put $x^* = x' \wedge x^0$.

Theorem 3.6. In a C_1 -algebra, x^* is a Boolean complement of x ; so $x \vee x^* \equiv 1$ and $x \wedge x^* \equiv 0$. Moreover, in a C_1 -algebra, the structure composed by the underlying set and by operations \wedge , \vee , and $*$ is a (pre) Boolean algebra. If we pass to the quotient by the basic relation \equiv , we obtain a Boolean algebra in the usual sense.

Definition 3.7. Let $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a C_1 -algebra and $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, * \rangle$ the Boolean algebra obtained as in the above theorem. Any Boolean algebra that is isomorphic to the quotient algebra of $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, * \rangle$ by \equiv is called Boolean algebra *associated with the* C_1 -algebra.

Hence, we have the following representation theorems for C_1 -algebras.

Theorem 3.8. Any C_1 -algebra is associated with a field of sets. Moreover, any C_1 -algebra is associated with the field of sets simultaneously open and closed of a totally disconnected compact Hausdorff space.

One open problem concerning C_1 -algebras remains. How many non-isomorphic associated with the C_1 -algebra are there?

4 The Curry Algebras C_n ($1 < n < \omega$)

Now we show a chain of Curry algebras beginning with the C_1 -algebra.

Let $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a C_1 -algebra. If $x \in A$, x^1 abbreviates x . x^n ($1 < n < \omega$) abbreviates $\overset{0, \dots, 0}{x^{n-\text{times}}}$. Also, $x^{(1)}$ abbreviates x^1 . $x^{(n+1)}$ ($1 < n < \omega$) abbreviates $x^{(n)} \wedge x^{n-1}$.

Definition 4.1. A C_n -algebra ($1 < n < \omega$) is an implicative pre-lattice $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ with a first element 1 and operators \wedge , \vee , and $'$ satisfying the conditions below:

- | | |
|--|--|
| (1) $x \vee x' \equiv 1$ | (5) $x^{(n)} \wedge y^{(n)} \leq (x \wedge y)^{(n)}$ |
| (2) $x'' \leq x$ | (6) $x^{(n)} \wedge y^{(n)} \leq (x \vee y)^{(n)}$ |
| (3) $y^{(n)} \leq (x \rightarrow y) \rightarrow ((x \rightarrow y') \rightarrow x')$ | (7) $x^{(n)} \leq (x')^{(n)}$ |
| (4) $x^{(n)} \wedge y^{(n)} \leq (x \rightarrow y)^{(n)}$ | |

Usual algebraic structural concepts like homomorphism, monomorphism, etc. can be introduced for Curry algebras without extensive comments.

Theorem 4.2. Every C_n -algebra is embedded in any C_{n-1} -algebra ($1 < n < \omega$).

If we indicate a C_n -algebra by C_n , the embedding hierarchy can be represented as $C_1 > C_2 > \dots > C_n > \dots$

5 The Monadic Curry Algebras C_1^*

In this section we introduce the monadic Curry algebra C_1^* .

Definition 5.1. Let A be a C_1 -algebra. Let \exists (existential quantifier) and \forall (universal quantifier) be operators on A . (\exists, \forall) is called a quantifier on A if

- | | |
|---|---|
| (1) $\exists 0 \equiv 0$ | (7) $\forall 1 \equiv 1$ |
| (2) $x \leq \exists x$ | (8) $\forall x \leq x$ |
| (3) $\exists(x \vee y) \equiv \exists x \vee \exists y$ | (9) $\forall(x \vee y) \equiv \forall x \vee \forall y$ |
| (4) $\exists \exists x \equiv \exists x$ | (10) $\forall \forall x \equiv \forall x$ |
| (5) $\exists(\exists x)^* \equiv (\exists x)^*$ | (11) $\forall(\forall x)^* \equiv (\forall x)^*$ |
| (6) $\exists(x \wedge \exists y) \equiv \exists x \wedge \exists y$ | |

We suppose in the above definition that, if $x \equiv y$, then $\exists x \equiv \exists y$ and $\forall x \equiv \forall y$. \exists is called existential quantifier on A and \forall is called universal quantifier on A . The pair $\langle A, (\exists, \forall) \rangle$ is called a monadic Curry algebra C_1^* or a C_1^* -monadic algebra (or C_1^* -algebra).

Given a Curry algebra C_1 , let's assume that there is an universal quantifier defined on it, i.e., a structure $\langle A, \forall \rangle$ such that conditions 7-11 above are satisfied. If we define $\exists_{1,x} =_{\text{Def.}} (\forall x^*)^*$, then \exists_1 is an existential quantifier (i.e. satisfying 1-6) and the structure composed by $\langle A, (\exists_1, \forall) \rangle$ is a C_1^* -monadic algebra. Also, we can get a monadic algebra considering an existential quantifier \exists on a Curry algebra C_1 satisfying conditions 1-6 of above definition and defining a universal quantifier (i.e, satisfying 7-11) as $\forall_{1,x} =_{\text{Def.}} (\exists x^*)^*$. Then the structure composed by $\langle A, (\exists, \forall_1) \rangle$ is a C_1^* -monadic algebra. Given a Curry algebra C_1 , in general, the algebras obtained $\langle A, (\exists_1, \forall) \rangle$ and $\langle A, (\exists, \forall_1) \rangle$ are not isomorphic. Also, given a C_1^* -monadic algebra $\langle A, (\exists, \forall) \rangle$, define the new quantifiers \exists_1 and \forall_1 . In general, we have $\exists_1 \neq \exists$ and $\forall_1 \neq \forall$.

Theorem 5.2. In a C_1^* -monadic algebra $\langle A, (\exists, \forall) \rangle$, the structure composed by the underlying set and by operations $\wedge, \vee, *, \exists$, and \forall is a (pre) monadic algebra. If we pass to the quotient by the basic relation \equiv , we obtain a monadic algebra in the usual sense [4].

Definition 5.3. Let $\langle A, (\exists, \forall) \rangle$ be a C_1^* -monadic algebra, and $\langle A, \equiv, \leq, \rightarrow, *, \exists, \forall \rangle$ the monadic algebra obtained as in the above theorem. Any monadic algebra that is isomorphic to the quotient algebra of $\langle A, \equiv, \leq, \rightarrow, *, \exists, \forall \rangle$ by \equiv is called monadic algebra *associated with the C_1^* -monadic algebra*.

Hence, we can establish the following representation theorems for C_1^* -monadic algebras.

Theorem 5.4. If C is a C_1^* -monadic algebra, then for its associated monadic algebra A , there exists a set X and there exists a Boolean algebra B , such that (i) A is isomorphic to a B -valued functional algebra A' with domain X , and (ii) for every element p of A' there exists a point x in X with $p(x) = \exists p(x)$.

Theorems 5.3 and 5.4 show us that C_1^* -monadic algebras constitute interesting generalization of the concept of monadic algebras. Here, there is an open problem. How many non-isomorphic monadic algebras associated with a C_1^* -monadic algebra are there?

6 The Curry Algebra P_1

Definition 6.1. A Curry algebra P_1 (or a P_1 -algebra) is a classical implicative pre-lattice $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ with a greatest element 1 and operators \wedge, \vee , and $'$ satisfying the conditions below, where $x^\# =_{\text{Def.}} x \vee x'$:

- | | |
|--|---|
| (1) $x \leq x''$ | (5) $x^\# \wedge y^\# \leq (x \vee y)^\#$; |
| (2) $(x \wedge x') \equiv 1$ | (6) $x^\# \leq (x')^\#$ |
| (3) $x^\# \wedge y^\# \leq (x \rightarrow y)^\#$; | (7) $x^\# \leq (x \rightarrow y) \rightarrow ((x \rightarrow y') \rightarrow x')$ |
| (4) $x^\# \wedge y^\# \leq (x \wedge y)^\#$; | (8) $x \leq (x' \rightarrow y)$ |

Example 6.2. Let's consider the calculus P_1 . A is the set of all formulas of P_1 . Let's consider as operations, the logical connectives of conjunction, disjunction, implication, and negation. Let's define the relation on A :

$x \equiv y$ iff $\vdash x \leftrightarrow y$. It is easy to check that \equiv is an equivalence relation on A .

$x \leq y$ iff $x \equiv x \wedge y$ and $y \leq x$ iff $y \equiv x \wedge y$. Also we take as 1 any fixed axiom instance.

The structure composed $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ is a P_1 -algebra.

Theorem 6.3. Let's $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a P_1 -algebra. Then the operator $'$ is non-monotone relatively \equiv .

Theorem 6.4. A P_1 -algebra is distributive and has a greatest element, as well as a first element.

Definition 6.5. Let x be an element of a P_1 -algebra. We put $x^* = x \rightarrow (y \wedge y')$, where y is a fixed element.

Theorem 6.6. In a P_1 -algebra, x^* is a Boolean complement of x ; so $x \vee x^* \equiv 1$ and $x \wedge x^* \equiv 0$. Moreover, in a P_1 -algebra, the structure composed by the underlying set and by operations \wedge, \vee , and $*$ is a (pre) Boolean algebra. If we pass to the quotient by the basic relation \equiv , we obtain a Boolean algebra in the usual sense.

Definition 6.7. Let $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a P_1 -algebra and $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, * \rangle$ the Boolean algebra obtained as in the above theorem. Any Boolean algebra that is isomorphic to the quotient algebra of $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, * \rangle$ by \equiv is called Boolean algebra associated with the P_1 -algebra.

Hence, we have the following representation theorems for P_1 -algebras.

Theorem 6.8. Any P_1 -algebra is associated with a field of sets. Moreover, any P_1 -algebra is associated with the field of sets simultaneously open and closed of a totally disconnected compact Hausdorff space.

One open problem concerning P_1 -algebras remains. How many non-isomorphic associated with the P_1 -algebra are there?

7 The Curry Algebras P_n

Now we show a chain of Curry algebras beginning with the P_1 -algebra.

Let $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a P_1 -algebra. If $x \in A$, x^1 abbreviates $x^\#$. x^n ($1 < n < \omega$) abbreviates $x^\# \wedge x^{\#\#} \wedge \dots \wedge x^{\#\#\dots\#}$, where the symbol $\#$ occurs n times. Also, $x^{(n)}$ abbreviates $x^1 \wedge x^2 \wedge \dots \wedge x^n$.

Definition 7.1. A P_n -algebra ($1 < n < \omega$) is an implicative pre-lattice $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ with a first element 1 and operators \wedge, \vee , and $'$ satisfying the conditions:

- | | |
|---|--|
| (1) $x \leq x''$ | (5) $x^{(n)} \wedge y^{(n)} \leq (x \vee y)^{(n)}$; |
| (2) $(x \wedge x') \equiv 1$ | (6) $x^{(n)} \leq (x')^{(n)}$ |
| (3) $x^{(n)} \wedge y^{(n)} \leq (x \rightarrow y)^{(n)}$; | (7) $x^{(n)} \leq (x \rightarrow y) \rightarrow ((x \rightarrow y') \rightarrow x')$ |
| (4) $x^{(n)} \wedge y^{(n)} \leq (x \wedge y)^{(n)}$; | (8) $x \leq (x' \rightarrow y)$ |

Usual algebraic structural concepts like homomorphism, monomorphism, etc. can be introduced for Curry algebras without extensive comments.

Theorem 7.2. Every P_n -algebra is embedded in any P_{n-1} -algebra ($1 < n < \omega$).

If we indicate a P_n -algebra by P_n , the embedding hierarchy can be represented as $P_1 > P_2 > \dots > P_n > \dots$

Definition 7.4. A P_ω -algebra is an implicative pre-lattice $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ with a first element 1 and operators \wedge, \vee , and $'$ satisfying the conditions below:

- | | |
|---------------------------------|------------------|
| (1) $(x \wedge x')' \equiv 1$ | (3) $x \leq x''$ |
| (2) $x \leq (x' \rightarrow y)$ | |

We propose in the sequence some extensions of the P_1 -algebras.

8 The Monadic Curry Algebras P_1^*

In this section we present some monadic Curry algebras P_1^* .

Definition 8.1. Let A be a P_1 -algebra. Let \exists and \forall be operators on A . (\exists, \forall) is called a quantifier on A if

- | | |
|---|---|
| (1) $\exists 0 \equiv 0$; | (7) $\forall 1 \equiv 1$; |
| (2) $x \leq \exists x$; | (8) $\forall x \leq x$; |
| (3) $\exists(x \vee y) \equiv \exists x \vee \exists y$; | (9) $\forall(x \vee y) \equiv \forall x \vee \forall y$; |
| (4) $\exists \exists x \equiv \exists x$; | (10) $\forall \forall x \equiv \forall x$; |
| (5) $\exists(\exists x)^* \equiv (\exists x)^*$; | (11) $\forall(\forall x)^* \equiv (\forall x)^*$ |
| (6) $\exists(x \wedge \exists y) \equiv \exists x \wedge \exists y$; | |

We suppose in the above definition that, if $x \equiv y$, then $\exists x \equiv \exists y$ and $\forall x \equiv \forall y$. \exists is called existential quantifier on A and \forall is called universal quantifier on A . The pair $\langle A, (\exists, \forall) \rangle$ is called a monadic Curry algebra P_1^* or a P_1^* -monadic algebra (or P_1^* -algebra).

Given a Curry algebra P_1 , let's assume that there is an universal quantifier defined on it, i.e., a structure $\langle A, \forall \rangle$ such that conditions (7) - (11) above are satisfied. If we define $\exists_1 x =_{\text{Def.}} (\forall x^*)^*$, then \exists_1 is an existential quantifier (i.e. satisfying (1) - (6) and the structure composed by $\langle A, (\exists_1, \forall) \rangle$ is a P_1^* -monadic algebra. Also, we can get a monadic algebra considering an existential quantifier \exists on a Curry algebra P_1 satisfying conditions (1) - (6) of above definition and defining a universal quantifier (i.e., satisfying (7) - (11) as $\forall_1 x =_{\text{Def.}} (\exists x^*)^*$. Then the structure composed by $\langle A, (\exists, \forall_1) \rangle$ is a P_1^* -monadic algebra. Given a Curry algebra P_1 , in general, the algebras obtained $\langle A, (\exists_1, \forall) \rangle$ and $\langle A, (\exists, \forall_1) \rangle$ are not isomorphic. Also, given a P_1^* -monadic algebra $\langle A, (\exists, \forall) \rangle$, define the new quantifiers \exists_1 and \forall_1 . In general, we have $\exists_1 \neq \exists$ and $\forall_1 \neq \forall$.

Theorem 8.2. In a P_1^* -monadic algebra $\langle A, (\exists, \forall) \rangle$, the structure composed by the underlying set and by operations $\wedge, \vee, *, \exists$, and \forall is a (pre) monadic algebra. If we pass to the quotient by the basic relation \equiv , we obtain a monadic algebra in the usual sense [10].

Definition 8.3. Let $\langle A, (\exists, \forall) \rangle$ be a P_1^* -monadic algebra, and $\langle A, \equiv, \leq, \rightarrow, *, \exists, \forall \rangle$ the monadic algebra obtained as in the above theorem. Any monadic algebra that is isomorphic to the quotient algebra of $\langle A, \equiv, \leq, \rightarrow, *, \exists, \forall \rangle$ by \equiv is called monadic algebra *associated with the P_1^* -monadic algebra*.

Hence, we can establish the following representation theorems for P_1^* -monadic algebras.

Theorem 8.4. If P is a P_1^* -monadic algebra, then for its associated monadic algebra A , there exists a set X and there exists a Boolean algebra B , such that (i) A is isomorphic to a B -valued functional algebra A' with domain X , and (ii) for every element p of A' there exists a point x in X with $p(x) = \exists p(x)$.

Theorems 8.3 and 8.4 show us that P_1^* -monadic algebras constitute interesting generalization of the concept of monadic algebras. Here, there is an open problem. How many non-isomorphic monadic algebras associated with a P_1^* -monadic algebra are there?

9 The Curry Algebra N_1

Definition 9.1. A Curry algebra N_1 (or a N_1 -algebra) is a classical implicative pre-lattice $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ with a greatest element 1 and operators $\wedge, \vee,$ and $'$ satisfying the conditions below, where $x^0 =_{\text{Def.}} (x \wedge x')$ and $x^\# =_{\text{Def.}} x \vee x'$:

- (1) $x^0 \wedge y^\# \leq ((x \rightarrow y) \rightarrow ((x \rightarrow y') \rightarrow x'))$
- (2) $x^0 \wedge y^0 \leq (x \rightarrow y)^0 \wedge (x \wedge y)^0 \wedge (x \vee y)^0 \wedge (x')^0$
- (3) $x^\# \wedge y^\# \leq (x \rightarrow y)^\# \wedge (x \wedge y)^\# \wedge (x \vee y)^\# \wedge (x')^\#$
- (4) $x^0 \leq (x \rightarrow x'') \wedge (x \rightarrow (x' \rightarrow y))$
- (5) $x^\# \leq x'' \rightarrow x$
- (6) $x^0 \vee x^\# \equiv 1$

Example 9.2. Let's consider the calculus N_1 [12]. A is the set of all formulas of N_1 . Let's consider as operations, the logical connectives of conjunction, disjunction, implication, and negation. Let's define the relation on A :

$x \equiv y$ iff $\vdash x \leftrightarrow y$. It is easy to check that \equiv is an equivalence relation on A .

$x \leq y$ iff $x \equiv x \wedge y$ and $y \leq x$ iff $y \equiv x \wedge y$. Also we take as 1 any fixed axiom instance.

The structure composed $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ is a N_1 -algebra.

Theorem 9.3. Adding the postulate $(x \wedge x')' \equiv 1$ we obtain a P_1 -algebra.

Theorem 9.4. Adding the postulate $x \vee x' \equiv 1$ we obtain a C_1 -algebra.

Theorem 9.5. Adding the postulates $(x \wedge x')' \equiv 1$ and $x \vee x' \equiv 1$ we obtain a (pre) Boolean algebra.

Theorem 9.6. Let's $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a N_1 -algebra. Then the operator $'$ is non-monotone relatively \equiv .

Theorem 9.7. N_1 -algebra is distributive and has a greatest element, as well as a first element.

Definition 9.8 Let x be an element of a N_1 -algebra. We put $x^* =_{\text{Def.}} x' \wedge x^0$.

Theorem 9.9. In a N_1 -algebra, x^* is a Boolean complement of x ; so $x \vee x^* \equiv 1$ and $x \wedge x^* \equiv 0$. Moreover, in a N_1 -algebra, the structure composed by the underlying set and by operations $\wedge, \vee,$ and $*$ is a (pre) Boolean algebra. If we pass to the quotient by the basic relation \equiv , we obtain a Boolean algebra in the usual sense.

Definition 9.10. Let $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a N_1 -algebra and $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, * \rangle$ the Boolean algebra obtained as in the above theorem. Any Boolean algebra that is isomorphic to the quotient algebra of $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, * \rangle$ by \equiv is called Boolean algebra *associated with the N_1 -algebra*.

Hence, we have the following representation theorems for N_1 -algebras.

Theorem 9.11. Any N_1 -algebra is associated with a field of sets. Moreover, any N_1 -algebra is associated with the field of sets simultaneously open and closed of a totally disconnected compact Hausdorff space.

One open problem concerning N_1 -algebras remains. How many non-isomorphic associated with the N_1 -algebra are there?

10 The Curry Algebras N_n^*

Now we show a chain of Curry algebras beginning with the N_1 -algebra.

Let $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ be a N_1 -algebra. If $x \in A$, remember that $x^0 =_{\text{Def.}} (x \wedge x')$. x^1 abbreviates x^0 . x^n ($1 < n < \omega$) abbreviates $(x^{n-1})^0$. Also, $x^{(1)}$ abbreviates x^1 . $x^{(n)}$ ($1 < n < \omega$) abbreviates $x^{(n-1)} \wedge x^n$. Also if x^1 abbreviates $x^\#$ and x^n ($1 < n < \omega$) abbreviates $x^\# \wedge x^{\#\#} \wedge \dots \wedge x^{\#\#\dots\#}$ (where the symbol $\#$ occurs n times), $x^{[n]}$ abbreviates $x^1 \wedge x^2 \wedge \dots \wedge x^n$.

Definition 10.1. A N_n -algebra ($1 < n < \omega$) is an implicative pre-lattice $\langle A, \equiv, \leq, \wedge, \vee, \rightarrow, ' \rangle$ with a first element 1 and operators \wedge, \vee , and $'$ satisfying the conditions:

- (1) $x^{(n)} \wedge y^{[n]} \leq ((x \rightarrow y) \rightarrow ((x \rightarrow y') \rightarrow x'))$
- (2) $x^{(n)} \wedge y^{(n)} \leq (x \rightarrow y)^{(n)} \wedge (x \wedge y)^{(n)} \wedge (x \vee y)^{(n)} \wedge (x')^{(n)}$
- (3) $x^{[n]} \wedge y^{[n]} \leq (x \rightarrow y)^{[n]} \wedge (x \wedge y)^{[n]} \wedge (x \vee y)^{[n]} \wedge (x')^{[n]}$
- (4) $x^{(n)} \leq (x \rightarrow x') \wedge (x \rightarrow (x' \rightarrow y))$
- (5) $x^{[n]} \leq x' \rightarrow x$
- (6) $x^{(n)} \vee x^{[n]} \equiv 1$

Theorem 10.2. Adding the postulate $x^{[n]} \equiv 1$ we obtain a C_n -algebra.

Theorem 10.3. Adding the postulate $x^{(n)} \equiv 1$ we obtain a P_n -algebra.

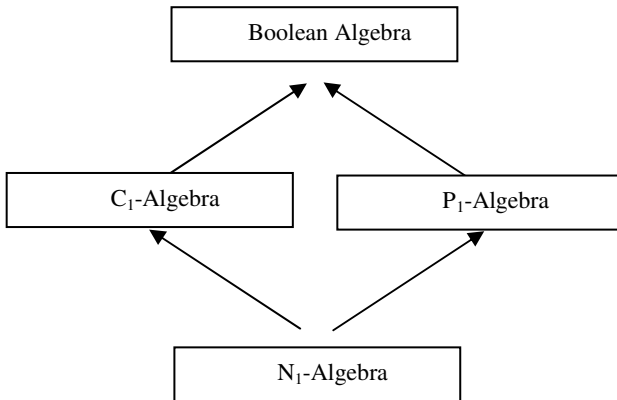


Fig. 1. The relationship among algebras

Theorem 10.4. Every N_n -algebra is embedded in any N_{n-1} -algebra ($1 < n < \omega$).

If we indicate a N_n -algebra by N_n , the embedding hierarchy can be represented as $N_1 > N_2 > \dots > N_n > \dots$

We propose in the sequence some extensions of the N_1 -algebras.

11 The Monadic Curry Algebras N_1^*

In this section we present some monadic Curry algebras N_1^* .

Definition 11.1. Let A be a N_1 -algebra. Let \exists and \forall be operators on A . (\exists, \forall) is called a quantifier on A if

- | | |
|---|---|
| (1) $\exists 0 \equiv 0$; | (7) $\forall 1 \equiv 1$; |
| (2) $x \leq \exists x$; | (8) $\forall x \leq x$; |
| (3) $\exists(x \vee y) \equiv \exists x \vee \exists y$; | (9) $\forall(x \vee y) \equiv \forall x \vee \forall y$; |
| (4) $\exists \exists x \equiv \exists x$; | (10) $\forall \forall x \equiv \forall x$; |
| (5) $\exists(\exists x)^* \equiv (\exists x)^*$; | (11) $\forall(\forall x)^* \equiv (\forall x)^*$ |
| (6) $\exists(x \wedge \exists y) \equiv \exists x \wedge \exists y$; | |

We suppose in the above definition that, if $x \equiv y$, then $\exists x \equiv \exists y$ and $\forall x \equiv \forall y$. \exists is called existential quantifier on A and \forall is called universal quantifier on A . The pair $\langle A, (\exists, \forall) \rangle$ is called a monadic Curry algebra N_1^* or a N_1^* -monadic algebra (or N_1^* -algebra).

Given a Curry algebra P_1 , let's assume that there is an universal quantifier defined on it, i.e., a structure $\langle A, \forall \rangle$ such that conditions (7) - (11) above are satisfied. If we define $\exists_1 x =_{\text{Def}} (\forall x^*)^*$, then \exists_1 is an existential quantifier (i.e. satisfying (1) - (6) and the structure composed by $\langle A, (\exists_1, \forall) \rangle$ is a N_1^* -monadic algebra. Also, we can get a monadic algebra considering an existential quantifier \exists on a Curry algebra P_1 satisfying conditions (1) - (6) of above definition and defining a universal quantifier (i.e., satisfying (7) - (11) as $\forall_1 x =_{\text{Def}} (\exists x^*)^*$. Then the structure composed by $\langle A, (\exists, \forall_1) \rangle$ is a N_1^* -monadic algebra. Given a Curry algebra N_1 , in general, the algebras obtained $\langle A, (\exists_1, \forall) \rangle$ and $\langle A, (\exists, \forall_1) \rangle$ are not isomorphic. Also, given a N_1^* -monadic algebra $\langle A, (\exists, \forall) \rangle$, define the new quantifiers \exists_1 and \forall_1 . In general, we have $\exists_1 \neq \exists$ and $\forall_1 \neq \forall$.

Theorem 11.2. In a N_1^* -monadic algebra $\langle A, (\exists, \forall) \rangle$, the structure composed by the underlying set and by operations $\wedge, \vee, *, \exists$, and \forall is a (pre) monadic algebra. If we pass to the quotient by the basic relation \equiv , we obtain a monadic algebra in the usual sense [10].

Definition 11.3. Let $\langle A, (\exists, \forall) \rangle$ be a N_1^* -monadic algebra, and $\langle A, \equiv, \leq, \rightarrow, *, \exists, \forall \rangle$ the monadic algebra obtained as in the above theorem. Any monadic algebra that is isomorphic to the quotient algebra of $\langle A, \equiv, \leq, \rightarrow, *, \exists, \forall \rangle$ by \equiv is called monadic algebra *associated with the N_1^* -monadic algebra*.

Hence, we can establish the following representation theorems for N_1^* -monadic algebras.

Theorem 11.4. If N is a N_1^* -monadic algebra, then for its associated monadic algebra A , there exists a set X and there exists a Boolean algebra B , such that (i) A is isomorphic to a B -valued functional algebra A' with domain X , and (ii) for every element p of A' there exists a point x in X with $p(x) = \exists p(x)$.

Theorems 11.2 and 11.3 show us that N_1^* -monadic algebras constitute interesting generalization of the concept of monadic algebras. Here, there is an open problem. How many non-isomorphic monadic algebras associated with a N_1^* -monadic algebra are there?

12 Conclusions

In this paper we've applied the concept of Curry algebra in order to obtain an algebraic version of the monadic predicate logic N_1^* . Also such algebra has as extension the monadic Curry algebras C_1^* and P_1^* . All such algebras are generalizations of the monadic algebras introduced by Halmos [13]. We hope to say more in forthcoming papers.

References

1. Abe, J.M.: A note on Curry algebras, Bulletin of the Section of Logic. Polish Academy of Sciences 16(4), 151–158 (1987)
2. Abe, J.M.: Curry Algebras N_1 . Atti Acc. Lincei Rend. Fis. 7(9), 125–128 (1996)
3. Abe, J.M.: Curry algebras Pr. Logique et Analyse 161-162-163, 5–15 (1998)
4. Abe, J.M., Akama, S., Nakamatsu, K.: Monadic Curry Algebras Qr. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 893–900. Springer, Heidelberg (2007)
5. Abe, J.M., Nakamatsu, K., Akama, S.: An Algebraic Version of the Monadic System C_1 . In: New Advances in Intelligent Decision Technologies. Studies in Computational Intelligence, vol. 199, pp. 341–349. Springer, Heidelberg (2009)
6. Abe, J.M., Nakamatsu, K., Akama, S.: A Note on Monadic Curry System P_1 . In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009. LNCS, vol. 5712, pp. 388–394. Springer, Heidelberg (2009)
7. Barros, C.M., da Costa, N.C.A., Abe, J.M.: Tópico de teoria dos sistemas ordenados: vol. II, sistemas de Curry, Coleção Documentos, Série Lógica e Teoria da Ciência, IEA-USP, 20, 132p (1995)
8. Curry, H.B.: Foundations of Mathematical Logic. Dover, New York (1977)
9. Da Costa, N.C.A.: On the theory of inconsistent formal systems. Notre Dame J. of Formal Logic 15, 497–510 (1974)
10. Da Costa, N.C.A., Marconi, D.: A note on paracomplete logic. Atti. Acc. Lincei Rend. Fis. 80(8), 504–509 (1986)
11. Da Costa, N.C.A.: Logics that are both paraconsistent and paracomplete. Atti. Acc. Lincei. Rend. Fis. 83, 29–32 (1990)
12. Eytan, M.: Tableaux of Hintikka et Tout ça: un Point de Vue Algebrique. Math. Sci. Humaines 48, 21–27 (1975)
13. Halmos, P.R.: Algebraic Logic. Chelsea Publishing Co., New York (1962)
14. Kleene, S.C.: Introduction to Metamathematics. Van Nostrand, Princeton (1952)
15. Mortensen, C.: Every quotient algebra for C_1 is trivial. Notre Dame J. of Formal Logic 21, 694–700 (1977)

A Sensing System for an Autonomous Mobile Robot Based on the Paraconsistent Artificial Neural Network

Claudio Rodrigo Torres^{1,4}, Jair Minoro Abe², Germano Lambert-Torres¹,
João Inácio Da Silva Filho³, and Helga Gonzaga Martins¹

¹ Artificial Intelligence Application Group – GAIA, Federal University of Itajubá
Av. BPS 1303 – 37.500-903 Itajubá – MG

² Paulista University, UNIP. Dr. Bacelar, 1.212,
São Paulo, SP, CEP 04026-002

³ Universidade Santa Cecília – UNISANTA, Santos, SP

⁴ Universidade Metodista de São Paulo, São Bernardo do Campo, SP
c . r . t @ u o l . c o m . b r

Abstract. This paper shows a sensing system for an autonomous mobile robot. The Sensing System is based on the Paraconsistent Neural Network. The type of artificial neural network used in this work is based on the Paraconsistent Evidential Logic – Et. The objective of the Sensing System is to inform the other robot components the position where there is an obstacle. The reached results have been satisfactory.

Keywords: Paraconsistent Neural Network, Paraconsistent Logic, Autoumous Mobile Robot, Sensing System.

1 Introduction

This article describes a sensing system for an autonomous mobile robot which is able to achieve a predetermined point in an environment divided into coordinates.

In this work, an autonomous mobile robot is considered as a system divided in three other subsystems: Planning Subsystem, Sensing Subsystem and Mechanical Subsystem. The Planning Subsystem is responsible for generating the sequence of movements the robot must perform to achieve a predetermined point. The Sensing Subsystem has the objective of informing the Planning Subsystem the position where there are obstacles. And the Mechanical Subsystem is the robot itself, it means, the mobile mechanical platform that carries all the devices from the other subsystems and perform the sequence of movements determined by the Planning Subsystem.

The Planning Subsystem and the Sensing Subsystem have been already implemented. But, the Mechanical Subsystem has not been implemented yet.

The Sensing Subsystem uses the Paraconsistent Artificial Neural Network [1]. This type of artificial neural network is based on the Paraconsistent Evidential Logic – Et. Thus, we describe some concepts about the Paraconsistent Evidential Logic – Et in the next section.

2 Paraconsistent Evidential Logic – Eτ

The Paraconsistent Evidential Logic – Eτ [2] is a type of paraconsistent logic in which there may be a Favorable Evidence Degree - μ and a Contrary Evidence Degree - λ for each analyzed sentence.

The Favorable Evidence Degree – μ is a value between 0 and 1 that represents the favorable evidence that the sentence is true.

The Contrary Evidence Degree - λ is a value between 0 and 1 that represents the contrary evidence that the sentence is true.

Through the Favorable and Contrary Degrees it is possible to represent four extremes logic states as shown in the figure 1.

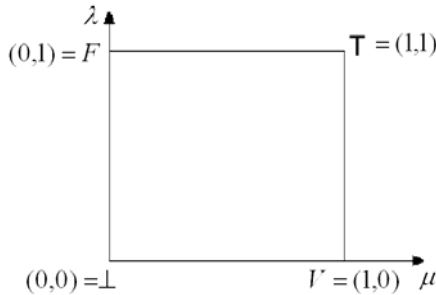


Fig. 1. Extreme Logic States

The four extreme logic states are:

- True (V)
- False (F)
- Paracomplete (⊥)
- Inconsistent (T)

In [3] is proposed the Paralyzer Algorithm. By this algorithm is also possible to represent non-extreme logic state. The figure 2 shows this.

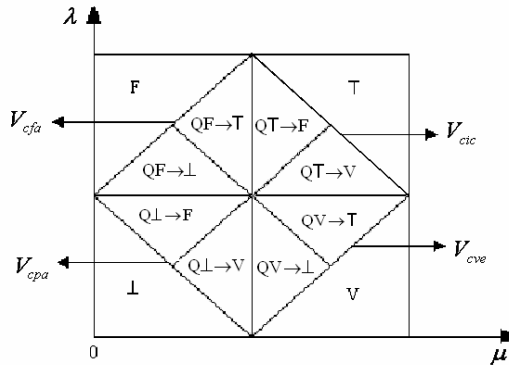


Fig. 2. Non-extreme logic states

The eight non-extreme logic states are:

- Quasi-true tending to Inconsistent - $QV \rightarrow T$
- Quasi-true tending to Paracomplete - $QV \rightarrow \perp$
- Quasi-false tending to Inconsistent - $QF \rightarrow T$
- 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 - $Q\perp \rightarrow F$

We also introduce the Uncertainty Degree: $G_{un}(\mu, \lambda) = \mu + \lambda - 1$ and the Certainty Degree: $G_c(\mu, \lambda) = \mu - \lambda$ ($0 \leq \mu, \lambda \leq 1$)

Some additional control values are:

- V_{cic} = maximum value of uncertainty control
- V_{cve} = maximum value of certainty control
- V_{cpa} = minimum value of uncertainty control
- V_{cfa} = minimum value of certainty control

In the next section we describe the proposed sensing system.

3 Sensing System

In [4] is presented a method of robot perception and world modeling that uses a probabilistic tessellated representation of spatial information called the Occupancy Grid. We propose in this work a similar method but instead of using probabilistic representation, we use the Paraconsistent Evidential Logic Et.

The sensing system we propose is prepared to receive data from ultrasonic sensors. Its aim is to generate a Favorable Evidence Degree for each environment position. The Favorable Evidence Degree is related to the sentence: there is obstacle in the position. And the sensing system is divided into two parts. The first part is responsible for receiving the data from the sensors and sending information to the second part of the system. And the second part is the Paraconsistent Artificial Neural Network itself. The figure 3 shows this idea.

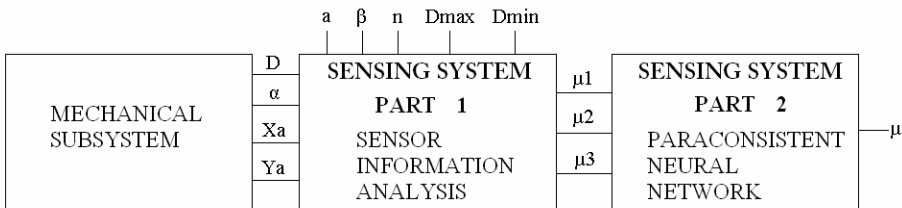


Fig. 3. Representation of the Sensing System

The robot sensors are on the Mechanical Subsystem. So, this subsystem must treat the data generated by the sensors and send information to the first part of the Sensing Subsystem. The data the Mechanical Subsystem must send to the first part of the Sensing Subsystem are: D , α , X_a and Y_a .

- a) The distance between the sensor and the obstacle (D).
- b) The angle between the horizontal axis of the environment and the direction to the front of the sensor (α). The figure 4 shows the angle α .

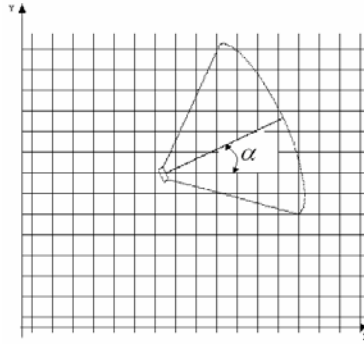


Fig. 4. Angle α

- c) The coordinate where the robot is (X_a , Y_a).

In the first part of the Sensing Subsystem there are also some configuration parameters. They are:

- a) The distance between the environment coordinates (a). The figure 5 shows this.

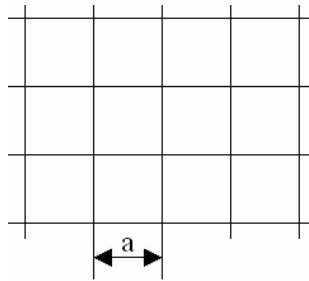


Fig. 5. Distance between coordinates

- b) The angle of the ultrasonic sensor conical field of view (β). The figure 6 shows this.
- c) The number of positions on the arch BC, shown in the figure 6, that the system measures (n).
- d) The maximum distance measured by the sensor that the system considers (D_{max})
- e) The minimum distance measured by the sensor that the system considers (D_{min})

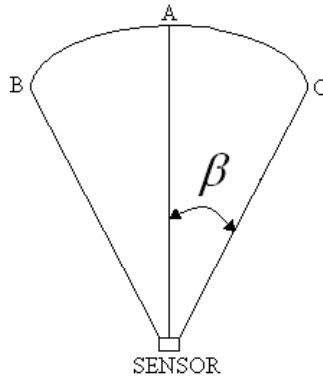


Fig. 6. Ultrasonic sensor conical field of view (β)

The first part of the Sensing System generates three Favorable Evidence Degree, μ_1 , μ_2 and μ_3 .

The Favorable Evidence Degree μ_1 is related to the distance between the sensor and the obstacle. The nearer the obstacle is to the sensor the bigger μ_1 value is.

The Favorable Evidence Degree μ_2 is related to the coordinate position on the arch AC shown in the figure 6. As the analyzed coordinate is near to the point A as bigger is μ_2 value. And as the analyzed coordinate is near to the points A and C, as smaller is the μ_2 value. The inspiration of this idea comes from [5]. It says that the probability of the obstacle be near to the point A is high. And this probability decreases as we analyze the region near to the points B and C.

Eventually, the Favorable Evidence Degree μ_3 is the previous value of the coordinate Favorable Evidence Degree.

4 Paraconsistent Artificial Neural Network

The Sensing Subsystem neural network is composed of two types of cells: Analytic Paraconsistent Artificial Neural Cell – CNAPa and Passage Paraconsistent Artificial Neural Cell - CNAPpa. In the follow we describe the cells.

4.1 Analytic Paraconsistent Artificial Neural Cell – CNAPa

This cell has two in-puts (μ_{RA} and μ_{RB}) and two out-puts (S1 and S2). Also there are two configuration parameter in-puts (Ftct and Ftc). The figure 7 shows the graphic representation of this cell.

The in-put evidence degrees are:

$$\mu_{RA}, \text{ such as: } 0 \leq \mu_{RA} \leq 1$$

$$\mu_{RB}, \text{ such as: } 0 \leq \mu_{RB} \leq 1$$

There are also two control values:

$$\text{Contradiction Tolerance Factor – Ft}_{ct}, \text{ such as: } 0 \leq Ft_{ct} \leq 1$$

$$\text{Certainty Tolerance Factor – Ft}_{c}, \text{ such as: } 0 \leq Ft_c \leq 1$$

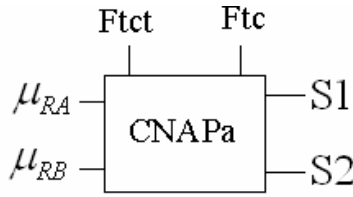


Fig. 7. Graphic representation of the Analytic Paraconsistent Artificial Neural Cell

The Analytic Paraconsistent Artificial Neural Cell – CNAPa has two out-puts. The out-put 1 (S1) is the Resultant Evidence Degree - μ_E .

$$\mu_E, \text{ such as: } 0 \leq \mu_E \leq 1$$

The out-put 2 (S2) is the Resultant Evidence Interval – φ_E .

$$\varphi_E, \text{ such as: } 0 \leq \varphi_E \leq 1$$

The Analytic Paraconsistent Artificial Neural Cell calculates the maximum value of certainty - V_{cve} , the minimum value of certainty control - V_{cfa} , the maximum value of uncertainty control - V_{cic} and the minimum value of uncertainty control - V_{cpa} by this way:

$$V_{cve} = \frac{1 + Ft_c}{2} \tag{1}$$

$$V_{cfa} = \frac{1 - Ft_c}{2} \tag{2}$$

$$V_{cic} = \frac{1 + Ft_{ct}}{2} \tag{3}$$

$$V_{cpa} = \frac{1 - Ft_{ct}}{2} \tag{4}$$

The Resultant Evidence Degree – μ_E , is determined in this way:

$$\mu_E = \frac{G_c + 1}{2} \tag{5}$$

As $G_c = \mu - \lambda$, we can say that:

$$\mu_E = \frac{\mu - \lambda + 1}{2} \tag{6}$$

We call as Certainty Interval (φ) the Certainty Degree interval that can be modified without changing the Uncertainty Degree value. This value is determined in this way.

$$\varphi = 1 - |G_{ct}| \tag{7}$$

4.2 Passage Paraconsistent Artificial Neural Cell – CNAPpa

The Passage Paraconsistent Artificial Neural Cell – CNAPpa has one in-put (μ), one out-put (S1) and one parameter control in-put (Ftc). The figure 8 shows the graphic representation of CNAPpa.

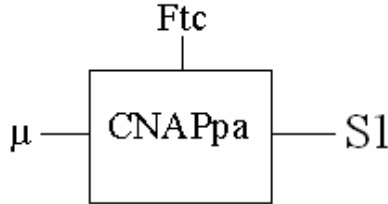


Fig. 8. Graphic representation of the Passage Paraconsistent Artificial Neural Cell

The in-put is the Favorable Evidence Degree (μ).

μ , such as: $0 \leq \mu \leq 1$

The value of the out-put S1 is the same as the in-put μ . But, the out-put value may be limited through the parameter control in-put Ftc.

The Passage Paraconsistent Artificial Neural Cell calculates the maximum value of certainty - V_{cve} and the minimum value of certainty control - V_{cfa} by this way:

$$V_{cve} = \frac{1 + Ft_c}{2}$$

$$V_{cfa} = \frac{1 - Ft_c}{2}$$

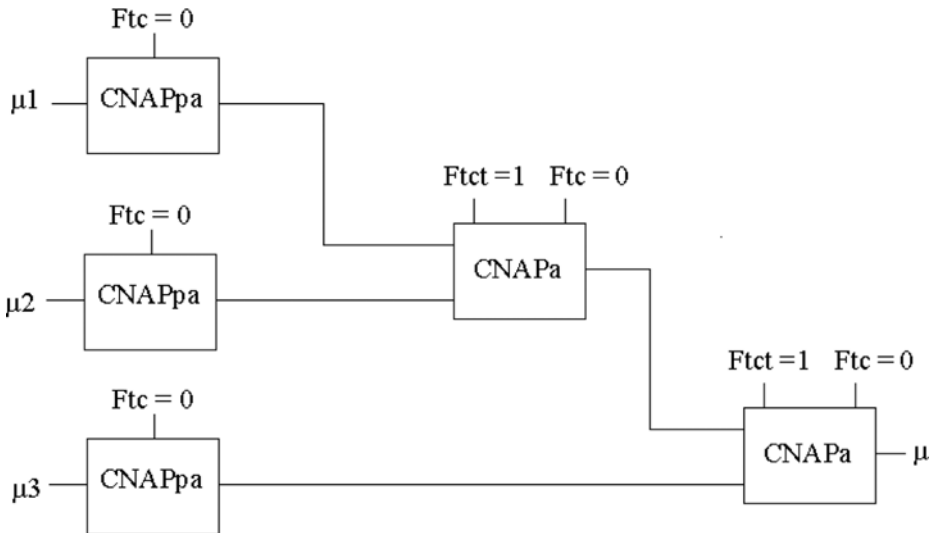


Fig. 9. The chosen Paraconsistent Neural Network Architecture for the Sensing System

It also determines the Resultant Evidence Degree - μ_E by this way:

$$\mu_E = \frac{\mu - \lambda + 1}{2}$$

$$\lambda = 1 - \mu$$

The out-put S1 assumes the same value as in the in-put μ when the follow situation is true:

$$[(V_{cve} \leq \mu_E) \text{ or } (\mu_E \leq V_{cfa})]$$

Otherwise, S1 is 0,5.

4.3 Paraconsistent Artificial Neural Architecture

It is possible to see in the figure 9 the chosen Paraconsistent Neural Network Architecture for the Sensing Subsystem.

5 Reached Results

The Sensing System has been tested. The tests consisted in simulating the Sensing System in-puts and analyzing the database generated by the Sensing System. This database stores the Favorable Evidence Degree for each environment position.

We show here the result of just one test composed of the information from only one ultrasonic sensor. The configuration parameters of the test were the following. The distance between the environment coordinates (α): 10. The angle of the ultrasonic sensor conical field of view (β): 30. The number of positions on the arch of the sensor conical field of view that the system measures (n): 10. The maximum distance measured by the sensor, that the system considers (D_{max}): 800. The minimum distance measured by the sensor, that the system considers (D_{min}): 8.

The Mechanical Subsystem treats the data from the sensors and generates the Sensing Subsystem in-puts. We needed to simulate the Sensing Subsystem in-puts because the Mechanical Subsystem has not been implemented yet.

Thus, the simulated Sensing Subsystem data were the ones we describe in the follow. The distance between the sensor and the obstacle (D): 200. The angle between the horizontal axis of the environment and the direction to the front of the sensor (α): 30. The coordinate where the robot is (X_a, Y_a): (0, 0).

We simulated the first measuring of the sensor, then, μ_3 was initially 0.

It is possible to see in the figure 10 the representation of the coordinates in which the Sensing System considered to have obstacles in. Summarizing, the figure 10 is a graphical representation of the database generated by the Sensing Subsystem.

The analyzed coordinates and their Favorable Evidence Degree are: A (18,10): $\mu = 0.438$. B (17,11): $\mu = 0.413$. C (17,12): $\mu = 0.388$. D (16,13): $\mu = 0.363$. E (15,14): $\mu = 0.338$. F (15,15): $\mu = 0.313$. G (14,15): $\mu = 0.288$. H (13,16): $\mu = 0.263$. I (12,17): $\mu = 0.238$. J (11,17): $\mu = 0.213$. K (10,18): $\mu = 0.188$. L (18,10): $\mu = 0.413$. M (19,9): $\mu = 0.388$. N (19,8): $\mu = 0.363$. O (20,7): $\mu = 0.338$. P (20,6): $\mu = 0.313$. Q (20,5): $\mu = 0.288$. R (20,4): $\mu = 0.263$. S (20,3): $\mu = 0.238$. T (20,2): $\mu = 0.213$. U (20,0): $\mu = 0.188$.

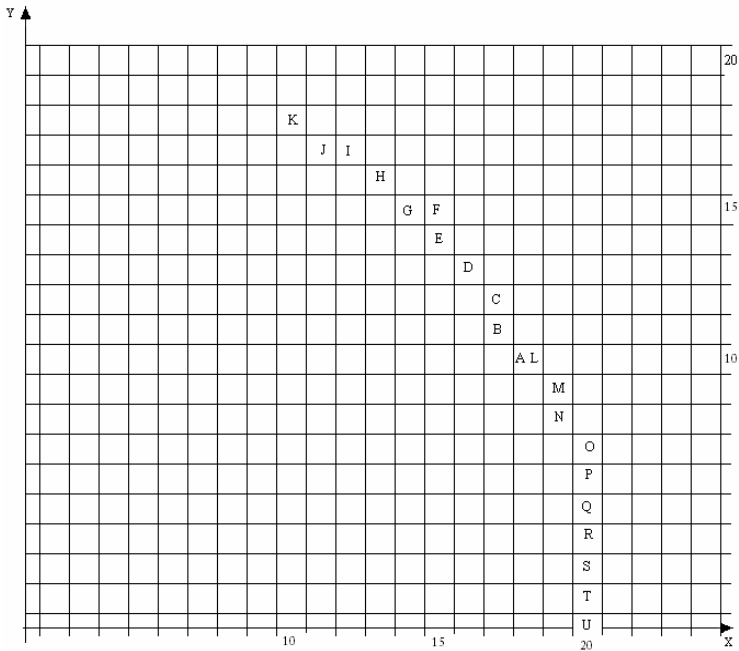


Fig. 10. The graphical representation of the database generated by the Sensing Subsystem

If we consider the sequence of positions from K to U as an arch, we perceive that the Favorable Evidence Degree (μ) decreases as the coordinate is farther distant from the center of the arch. It means that the system is working as desired.

6 Conclusions

This paper presents a proposal of an autonomous mobile robot composed of three modules: Sensing Subsystem, Planning Subsystem and Mechanical Subsystem. The Mechanical Subsystem has not been implemented yet. It is emphasized here the Sensing Subsystem.

The aim of the Sensing Subsystem is to inform the Planning Subsystem the positions in which may have obstacles in. It considers the environment divided into coordinates.

The Sensing Subsystem is based on the Paraconsistent Artificial Neural Network. The Sensing Subsystem neural network is composed of two types of cells: Analytic Paraconsistent Artificial Neural Cell – CNAPa and Passage Paraconsistent Artificial Neural Cell - CNAPpa.

The out-put of the Sensing Subsystem is the Favorable Evidence Degree related to the sentence: there is obstacle in the position. In fact, the Sensing Subsystem generates a database with the Favorable Evidence Degree for each analyzed coordinate.

Some tests were made with the Sensing Subsystem. The reached results were satisfactory.

The next step is the implementation of the Mechanical Subsystem and the connection of the three subsystems.

References

1. Da Silva Filho, J.I., Abe, J.M., Lambert-Torres, G.: *Inteligência Artificial com Redes de Análises Paraconsistentes: Teoria e Aplicação*. Rio de Janeiro: LTC (2008)
2. Abe, J.M.: *Fundamentos da Lógica Anotada (Foundations of Annotated Logics)*, in Portuguese, Ph. D. Thesis, University of São Paulo, São Paulo (1992)
3. Da Silva Filho, J.I.: *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, University of São Paulo, São Paulo (1999)
4. Elfes, A.: Using occupancy grids for mobile robot perception and navigation. *Comp. Mag.* 22(6), 46–57 (1989)
5. Borenstein, J., Koren, Y.: The Vector field Histogram: Fast Obstacle Avoidance for Mobile Robots. *IEEE Journal of Robotics and Automation* 7, 278–288 (1991)

Paraconsistent Artificial Neural Networks and EEG Analysis

Jair Minoro Abe^{1,2}, Helder F.S. Lopes², Kazumi Nakamatsu³, and Seiki Akama⁴

¹ Graduate Program in Production Engineering, ICET - Paulista University

R. Dr. Bacelar, 1212, CEP 04026-002 São Paulo – SP – Brazil

² Institute For Advanced Studies – University of São Paulo, Brazil

jairabe@uol.com.br, helder.mobile@gmail.com

³ School of Human Science and Environment/H.S.E. – University of Hyogo – Japan

nakamatu@shse.u-hyogo.ac.jp

⁴ C-Republic, Tokyo, Japan

akama@jcom.home.ne.jp

Abstract. The aim of this paper is to present a study of brain EEG waves through a new ANN based on Paraconsistent Annotated Evidential Logic Et which is capable of manipulating concepts like impreciseness, inconsistency, and paracompleteness in a nontrivial manner. As application, the Paraconsistent Artificial Neural Network – PANN showed capable of recognizing children with Dyslexia with Kappa index at a rate of 80%.

Keywords: artificial neural network, paraconsistent logics, annotated logics, pattern recognition, Dyslexia.

1 Introduction

In this paper we employ a new kind of ANN based on paraconsistent annotated evidential logic E_t , which is capable of manipulating imprecise, inconsistent and para-complete data in order to make a first study of the recognition of EEG standards.

The EEG is a brain electric signal activity register, resultant of the space-time representation of synchronic postsynaptic potentials. The most probable is that the main generating sources of these electric fields are perpendicularly guided regarding to the cortical surface, as the cortical pyramidal neurons.

The graphic registration of the sign of EEG can be interpreted as voltage fluctuation with mixture of rhythms, being frequently sinusoidal, ranging from 1 to 70 Hz. In the clinical-physiological practice, such frequencies are grouped in frequency bands: delta (0.5 to 4.0 Hz), theta (4.1 to 8.0 Hz), alpha (8.1 to 12.5 Hz), and beta (> 13.0 Hz). During the relaxed awake, normal EEG in adults is predominantly composed by alpha band frequency, which is generated by interactions of the slum-cortical and thalamocortical systems.

EEG analysis, as well as any other measurements devices, is limited and subjected to the inherent imprecision of the several sources involved: equipment, movement of the patient, electric registers and individual variability of physician visual analysis.

Such imprecision can often include conflicting information or paracomplete data. The majority of theories and techniques available are based on classical logic and so they cannot handle adequately such set of information, at least directly. Although several theories have been developed in order to overcome such limitations, v.g. Fuzzy set theory, Rough set theory, non-monotonic reasoning, among others, they cannot manipulate inconsistencies and paracompleteness directly. So, we need a new kind of logic to deal with uncertainty, inconsistent and paracomplete data, namely the paraconsistent annotated evidential logic $E\tau$.

2 Background

Paraconsistent Artificial Neural Network – PANN is a new artificial neural network introduced in [8]. Its basis leans on paraconsistent annotated logic $E\tau$ [1]. Let us present it briefly.

The atomic formulas of the logic $E\tau$ are 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 (all considerations are taken with $0 \leq \mu, \lambda \leq 1$): Uncertainty degree: $G_{un}(\mu, \lambda) = \mu + \lambda - 1$; Certainty degree: $G_{ce}(\mu, \lambda) = \mu - \lambda$; 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*, 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 = $-Ft_{cc}$

All states are represented in the next figure (Fig. 1).

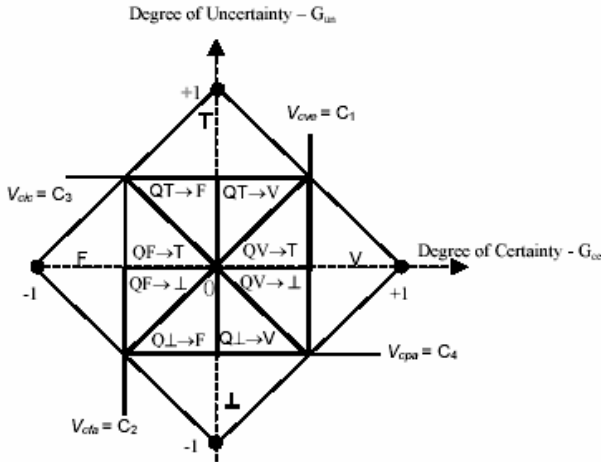


Fig. 1. Extreme and Non-extreme states

3 The Main Artificial Neural Cells

In the PANN, the certainty degree G_{cc} indicates the ‘measure’ falsity or truth degree. 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.

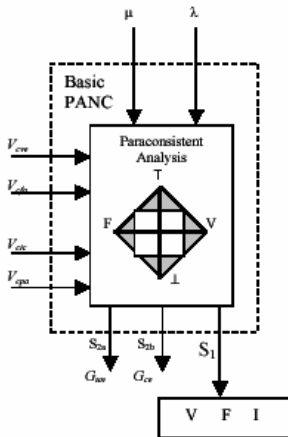


Fig. 2. Basic cell of PANN

The resulting certainty degree G_{cc} is obtained as follows:

If: $V_{cfa} \leq G_{un} \leq V_{cve}$ or $V_{cic} \leq G_{un} \leq V_{cpa}$
 $\Rightarrow G_{cc} = \text{Indefinition}$

For: $V_{cpa} \leq G_{un} \leq V_{cic}$

If: $G_{un} \leq V_{cfa} \Rightarrow G_{cc} = \text{False}$ with degree G_{un}
 $V_{cic} \leq G_{un} \Rightarrow G_{cc} = \text{True}$ 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 uncertainty degree, G_{cc} = resulting certainty degree, and X = constant of Indefinision.

Using the concepts of *basic* Paraconsistent Artificial Neural Cell – 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	$\mu, \lambda,$ $F_{t_{ct}}, F_{t_{ce}}$	$\lambda_c = 1 - \lambda$ $G_{un} G_{ce},$ $\mu_r = (G_{ce} + 1)/2$	If $ G_{ce} > F_{t_{ce}}$ then $S_1 = \mu_r$ and $S_2 = 0$ If $ G_{un} > F_{t_{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 $S_1 = \lambda$

4 Experimental Procedures

Recent researches reveal that 10% of the world population in school age suffer of learning and/or behavioral disorders caused by neurological problems, such as ADHD, dyslexia, and dyscalculia, with predictable consequences in those students' insufficient performance in the school [5], [6], [10], [11], [21], [22].

Concisely, a child without intellectual lowering is characterized as bearer of Attention-deficit/hyperactivity disorder (ADHD) when it presents signs of:

Inattention: difficulty in maintaining attention in tasks or games; the child seems not to hear what is spoken; difficulty in organizing tasks or activities; the child loses things; the child becomes distracted with any incentive, etc.

Hyperactivity: frequently the child leaves the class room; the child is always inconveniencing friends; the child runs and climbs in trees, pieces of furniture, etc; the child speaks a lot, etc.

Impulsiveness: the child interrupts the activities of colleagues; the child doesn't wait his time; aggressiveness crises; etc.

Dyslexia: when the child begins to present difficulties to recognize letters or to read them and to write them, although the child has not a disturbed intelligence, that is, a normal IQ;

Dyscalculia: when the child presents difficulties to recognize amounts or numbers and/or to figure out arithmetic calculations.

A child can present any combination among the disturbances above. All those disturbances have their origin in a cerebral dysfunction that can have multiple causes, many times showing a hereditary tendency. Since from the first discoveries made by [8], those disturbances have been associated to cortical diffuse lesions and/or more specific, temporal-parietal areas lesions in the case of dyslexia and dyscalculia [5], [11], [22]. The disturbances of ADHD disorder seem to be associated to an alteration of the dopaminergic system, that is, it is involved with mechanisms of attention and they seem to involve a frontal-lobe dysfunction and basal ganglia areas [6], [22].

EEG alterations seem to be associated those disturbances. Thus, some authors have proposed that there is an increase of the delta activity in EEG in those tasks that demand a larger attention to the internal processes. Other authors [16], [20] have

described alterations of the delta activity in dyslexia and dyscalculia children sufferers. [12] has proposed that a phase of the EEG component would be associated to the action of the memory work. More recently, [14] has showed delta activity is reduced in occipitals areas, but not in frontals, when dyslexic's children were compared with normal ones.

In this way, the study of the delta and theta bands becomes important in the context of the analysis of learning disturbances.

So, in this paper we've studied two types of waves, specifically delta and theta waves band, where the size of frequency established clinically ranges from 1.0 Hz to 3.5 Hz and 4.0 Hz to 7.5 Hz respectively.

The process of wave analysis by PANN consists previously of data capturing, adaptation of the values for screen examination, elimination of the negative cycle and normalization of the values for PANN analysis. It is worth to observe such process does not allow loss of any wave essential characteristics for our analysis.

As the actual EEG examination values can vary highly, in module, something 10 μV to 1500 μV , we make a normalization of the values between 100 μV and -100 μV by a simple linear conversion, to facilitate the manipulation the data:

$$x = \frac{100 \cdot a}{m}$$

Where: m is the maximum value of the exam.
 a is the current value of the exam.
 x is the current normalized value.

The minimum value of the exam is taken as zero value and the remaining values are translated proportionally.

5 PANN for Morphological Analysis

This method is used primarily for PANN (Fig. 4) to balance the data received from expert systems. After this process uses a decision-making lattice to determine the soundness of the recognition (Fig. 3).

The wave that has the highest favorable evidence and lowest contrary evidence is chosen as the more similar wave to the analyzed wave.

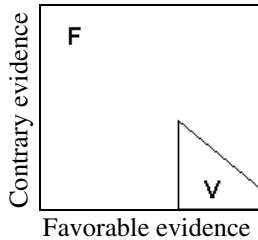


Fig. 3. Lattice for decision-making used in morphological analysis used after making PANN; F: logical state false (it is interpreted as wave not similar); V: logical state true (it is interpreted as wave similar).

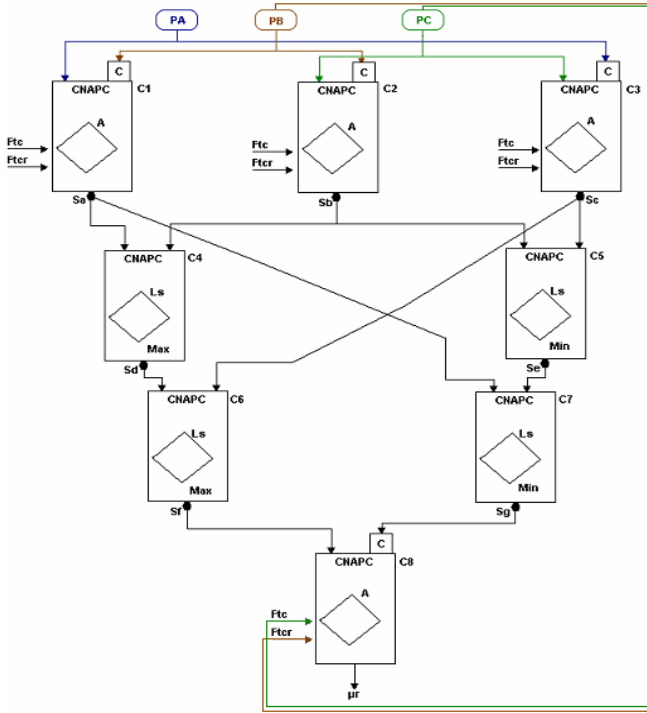


Fig. 4. The architecture for morphological analysis. Three expert systems operate: PA, for check the number of wave peaks; PB, for checking similar points, and PC, for checking different points:

C1–PANC which processes input data of PA and PB

C2–PANC which processes input data of PB and PC

C3–PANC which processes input data of PC and PA

C1, C2, and C3 constitute the 1st layer of the architecture

C4–PANC which calculates the maximum evidence value between cells C1 and C2

C5–PANC which calculates the minimum evidence value between cells C2 and C3

C4 and C5 constitute the 2nd layer of the architecture

C6–PANC which calculates the maximum evidence value between cells C4 and C3

C7–PANC which calculates the minimum evidence value between cells C1 and C5

C6 and C7 constitute the 3rd layer of the architecture

C8 analyzes the experts PA, PB, and PC and gives the resulting decision value

PANC A = Paraconsistent artificial neural cell of analytic connection

PANCLs_{Max} = Paraconsistent artificial neural cell of simple logic connection of maximization

PANCLs_{Min} = Paraconsistent artificial neural cell of simple logic connection of minimization

Ft_{ce} = Certainty tolerance factor; Ft_{ct} = Contradiction tolerance factor

S_a = Output of C1 cell; S_b = Output of C2 cell; S_c = Output of C3 cell; S_d = Output of C4 cell

S_e = Output of C5 cell; S_f = Output of C6 cell; S_g = Output of C7 cell

C = Complemented value of input

μ = Value of output of PANN

Table 3. Lattice for decision-making used in the morphological analysis (Fig. 3)

Logical states of the lattice		
True	Ef > 0,61	Ec < 0,40 G _c > 0,22
False	Ef < 0,61	Ec > 0,40 G _c <= 0,23

Ec: contrary evidence; Ef: favorable evidence; G_{ce}: certainty degree;

For an adequate PANN wave analysis, it is necessary that each input of PANN is properly calculated. These input variables are called expert systems, as they are specific routines for extracting information.

5.1 Expert System 1 – Checking the Number of Wave Peaks

The aim of the *expert system 1* is to compare the waves and analyze their differences regarding the number of peaks.

$$Se_1 = 1 - \left(\frac{|bd - vt|}{bd + vt} \right)$$

Where: *vt* is the number of peaks of the wave.
bd is the number of peaks of the wave stored in the database.
Se₁ is the value resulting from the calculation.

5.2 Expert System 2 – Checking Similar Points

The aim of the *expert system 2* is to compare the waves and analyze their differences regarding of similar points.

$$Se_2 = \frac{\sum_{j=1}^n (x_j)}{n}$$

Where: *n* is the total number of elements.
x is the element of the current position.
j is the current position.
Se₂ is the value resulting from the calculation.

5.3 Expert System 3 – Checking Different Points

The aim of the *expert system 3* is to compare the waves and analyze their differences regarding of different points.

$$Se_3 = 1 - \left(\frac{\sum_{j=1}^n \left(\frac{|x_j - y_j|}{a} \right)}{n} \right)$$

Where: *n* is the total number of elements.
a is the maximum amount allowed.
j is the current position.
x is the value of wave 1.
y is the value of wave 2.
Se₃ is the value resulting from the calculation.

6 Tests

Seven exams of different EEG were analyzed, being two exams belonging to adults without any learning disturbance and five exams belonging to children with learning disturbances (exams and respective diagnoses given by ENSCER - Teaching the Brain, EINA - Studies in Natural Intelligence and Artificial Ltda).

Each analysis was divided in three rehearsals, each rehearsal consisted of 10 seconds of the analyzed, free from visual analysis of spikes and artifacts regarding to the channels T3 and T4.

In the first battery it was used of a filter for recognition of waves belonging to the Delta band. In the second battery it was used a filter for recognition of waves belonging to the Theta band. In the third battery it was not used any filters for recognition, i.e., the system was free to recognize any wave type. The total number of exams is 180.

Table 4. Contingency table

		Visual Analysis					Total
		Delta	Theta	Alpha	Beta	Unrecognized	
RNAP Analysis	Delta	31	3	0	0	0	34
	Theta	15	88	1	1	0	105
	Alpha	0	5	22	0	0	27
	Beta	0	0	1	3	0	4
	N/D	7	2	1	0	0	10
Total		53	98	25	4	0	180

Index Kappa = 0.80

Table 5. Statistical results - sensitivity and specificity: Delta waves

		Visual analysis		
		Delta	Not Delta	Total
RNAP Analysis	True	31	124	155
	False	22	3	25
	Total	53	127	180

Sensitivity = 58%; Specificity = 97%

Table 6. Statistical results - sensitivity and specificity: Theta waves

		Visual analysis		
		Theta	Not Theta	Total
RNAP Analysis	True	88	65	153
	False	10	17	27
	Total	98	82	180

Sensitivity = 89%; Specificity = 79%

Table 7. Statistical results - sensitivity and specificity: Alpha waves

		Visual analysis		
		Alpha	Not Alpha	Total
RNAP Analysis	True	22	150	172
	False	3	5	8
	Total	25	155	180

Sensitivity = 88%; Specificity = 96%

Table 8. Statistical results - sensitivity and specificity: Beta waves

		Visual analysis		
		Beta	Not Beta	Total
RNAP Analysis	True	3	175	178
	False	1	1	2
	Total	4	176	180

Sensitivity = 75%; Specificity = 99%

Table 9. Statistical results - sensitivity and specificity: Unrecognized waves

		Visual analysis		
		Unrecognized	Recognized	Total
RNAP Analysis	True	0	180	180
	False	0	0	0
	Total	0	180	180

Sensitivity = 100%; Specificity = 100%

7 Conclusions

These findings suggest that the sensitivity with respect to the Delta waves is 58%. This is an indication that there must be improvements in the detection of peaks in the band Delta. We believe that such improvements are possible to be made in this direction. The sensitivities of the theta, alpha and beta waves are reasonable, but that improvements can be tried.

Regarding the specificity, the method showed more reliable results. Taking into account an overall assessment in the sense we take the arithmetic mean of sensitivity (75.50%) and specificity (92.75%), we find reasonable results that encourage us to seek improvements in this study.

Even finding a low sensitivity in the recognition of delta waves, the methodology of pattern recognition using morphological analysis showed to be effective, achieving recognize patterns of waves similar to patterns stored in the database, allowing quantifications and qualifications of the examination of EEG data to be used by PANN in their process analysis of examination.

References

1. Abe, J.M.: Foundations of Annotated Logics, PhD thesis. University of São Paulo, Brazil (1992) (in Portuguese)
2. Abe, J.M.: Some Aspects of Paraconsistent Systems and Applications. *Logique et Analyse* 157, 83–96 (1997)
3. Abe, J.M., Lopes, H.F.S., Anghinah, R.: Paraconsistent Artificial Neural Network and Alzheimer Disease: A Preliminary Study. *Dementia & Neuropsychologia* 3, 241–247 (2007)
4. Anghinah, R.: Estudo da densidade espectral e da coerência do eletrencefalograma em indivíduos adultos normais e com doença de Alzheimer provável, PhD thesis, Faculdade de Medicina da Universidade de São Paulo, São Paulo (2003) (in Portuguese)

5. Ansari, D., Karmiloff-Smith, A.: Atypical trajectories of number development: a neuro-constructivist perspective. *Trends In Cognitive Sciences* 12, 511–516 (2002)
6. Blonds, T.A.: Attention-Deficit Disorders and Hyperactivity. In *Developmental Disabilities in Infancy and Ramus, F., Developmental dyslexia: specific phonological deficit or general sensorimotor dysfunction? Current Opinion in Neurobiology* 13, 1–7 (2003)
7. Da Silva Filho, J.I.: Métodos de interpretação da Lógica Paraconsistente Anotada com anotação com dois valores LPA2v com construção de Algoritmo e implementação de Circuitos Eletrônicos, EPUSP, PhD thesis, São Paulo (1999) (in Portuguese)
8. Da Silva Filho, J.I., Abe, J.M., Torres, G.L.: Inteligência Artificial com as Redes de Análises Paraconsistentes. LTC-Livros Técnicos e Científicos Editora S.A, São Paulo,, 313 (2008) (in Portuguese)
9. Gallaburda, A.M., Sherman, G.F., Rosen, G.G., Aboitiz, F., Genschiwind, N.: Developmental dyslexia: four consecutive patients with cortical anomalies. *Ann. Neurology* 18, 2122–2333 (1985)
10. Hynd, G.W., Hooper, R., Takahashi, T.: Dyslexia and Language-Based disabilities. In: Coffey, Brumbak (eds.) *Text Book of Pediatric Neuropsychiatry*, pp. 691–718. American Psychiatric Press, Washington (1985)
11. Lindsay, R.L.: Dyscalculia. In: Capute, Accardo (eds.) *Developmental Disabilities in Infancy and Childhood*, pp. 405–415. Paul Brookes Publishing Co., Baltimore (1996)
12. Lopes, H.F.S.: Aplicação de redes neurais artificiais paraconsistentes como método de auxílio no diagnóstico da doença de Alzheimer, MSc Dissertation, Faculdade de Medicina- USP, São Paulo, 473p. (2009) (in Portuguese)
13. Klimeshc, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Ver.* 29, 169–195 (1999)
14. Klimesch, W., Doppelmayr, H., Wimmer, J., Schwaiger, D., Rôhm, D., Bruber, W., Hutzler, F.: Theta band power changes in normal and dyslexic children. *Clinical Neurophysiology* 113, 1174–1185 (2001)
15. Kocuyigit, Y., Alkan, A., Erol, H.: Classification of EEG Recordings by Using Fast Independent Component Analysis and Artificial Neural Network. *Journal of Medical Systems* 32(1), 17–20 (2008)
16. Niedermeyer, E., da Silva, F.L.: *Electroencephalography*, 5th edn. Lippincott Williams & Wilkins (2005)
17. Rocha, A.F., Massad, E.: How the human brain is endowed for mathematical reasoning. *Mathematics Today* 39, 81–84 (2003)
18. Rocha, A.F., Massad, E., Rocha, F.T.: Arithmetic reasoning: Experiments and modeling (Submitted)
19. Rocha, A.F., Massad, E., Leite, C.: Brain plasticity and arithmetic learning in congenitally injured brains (2003) (Submitted)
20. Rocha, A.F., Massad, E., Pereira, Jr., A.: *The Brain: From Fuzzy Arithmetic to Quantum Computing*. Springer, Heidelberg (2004) (Submitted)
21. Temple, E.: Brain mechanisms in normal and dyslexic readers. *Current Opinion in Neurobiology* 12, 178–183 (2002)
22. Voeller, K.K.S.: Attention-Deficit / Hyperactivity: Neurobiological and clinical aspects of attention and disorders of attention. In: Coffey, Brumbak (eds.) *Text Book of Pediatric Neuropsychiatry*, pp. 691–718. American Psychiatric Press, Washington (1998)

A Reasoning-Based Strategy for Exploring the Synergy among Alternative Crops

Hércules Antonio do Prado^{1,2}, Edilson Ferneda¹, and Ricardo Coelho de Faria³

¹ Graduate Program on Knowledge and IT Management, Catholic University of Brasilia, Brasília, DF, Brazil

hercules@{ucb.br, embrapa.br}, eferneda@pos.ucb.br

² Embrapa – Management and Strategy Secretariat, Brasília, DF, Brazil

³ Undergraduate Program on Economy, Catholic University of Brasilia, Brasília, DF, Brazil
ricardoc@ucb.br

Abstract. This paper focuses on the problem of choosing one among many alternatives, each one expressed as a combination of factors. The problem is approached with a reasoning-based strategy that takes into account the relations among the alternatives discovered by means of a data mining technique. The problem of choosing a combination of different vegetables based on the synergistic effects of the combination and considering some socioeconomic variables is taken as a case study. The idea is to identify combinations that lead to gains in productivity, profitability, and lower costs. Experts recognize that some combinations of cultures can generate synergistic effects that can lead to profits or losses, depending on the production variables involved. However, the analysis of the results of the cultivation of multiple varieties involves a space of possibilities whose treatment is not trivial. Among the alternatives available, this study explored the Combinatorial Neural Model (CNM) that assures the hypotheses generation for all possible combinations, within limits defined by the model parameters. The study was carried out on data collected from farms in the Brazilian Federal District. Two approaches to the problem are presented: (i) the first one based on univariate and bi-dimensional data analysis and (ii) a multidimensional analysis based on the CNM results.

Keywords: Reasoning-based systems, Data mining, Farming systems.

1 Introduction

Many variables influence the results of cultivations in agricultural establishments. The synergy existing among alternative crops combinations are a common sense knowledge, mainly because each combination shares a set of inputs like defensives, adubation and labor. For a given combination, one could benefit from scale gains of shared inputs. This way, a producer may combine many vegetables (e.g., carrots, cucumber, beet, and cassava, or carrot, lettuce, and tomato).

This work is a result from a research project that aimed at creating management strategies for the agricultural production taking into account market, technical, and economic parameters. The approach adopted takes advantage from the associative

nature of Combinatorial Neural Model (CNM), a hybrid (symbolic-connectionist) neural network able to map a high dimension input space into a smaller set of classes, preserving the meaning of the relation. It presents the well-known learning ability of the neural networks while keeping a symbolic knowledge representation. In order to evaluate the model, a field research was carried out along with a set of producers from Brasília, the Brazilian government headquarter, collecting information about the vegetables planted, the crops cost, productivity, and the profitability, as well other social-economic information that could help to build the model.

The next sections include: (i) the methodology adopted, including the model building with the CNM and data preprocessing, (ii) the findings of the study, and (iii) a discussion on the results, along with some ideas on future works.

2 Methodology

2.1 Building a Combinatorial Neural Model

The Combinatorial Neural Model (CNM) is a hybrid model for intelligent systems that adopt the better of two worlds: the learning ability from neural networks and the expressivity of symbolic knowledge representation. It was proposed by Machado and Freitas (1989 & 1991) as an alternative to overcome the black box limitation of the Multilayer Perceptron (Haykin, 2001), by applying the neural network structure along with a symbolic processing. CNM can identify regularities among input vectors and output values, performing a symbolic mapping. It uses supervised learning and a feedforward topology with three layers (see Figure 1): (i) the input layer, in which each node corresponds to a triple object-attribute-value that describes some dimension (here called *evidence* and denoted by e) in the domain; (ii) an intermediary or combinatorial layer with neurons connected to one or more neurons in the input layer, representing the logical conjunction AND; and (iii) the output layer, with one neuron for each possible class (here called *hypothesis* and denoted by h), that is connected to one or more neurons in the combinatorial layer by the logical disjunction OR.

The synapses may be inhibitory or excitatory and have assigned a weight between zero (non connected) and one (fully connected). The network is created as follows: (i)

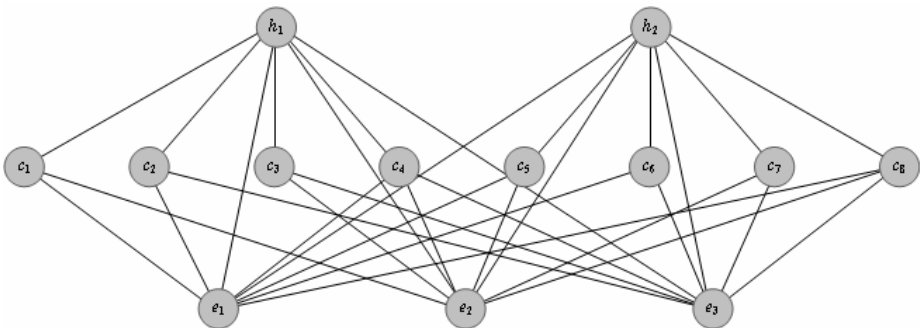


Fig. 1. An example of combinatorial neural network

one neuron in the input layer for each evidence in the training set; (ii) a neuron in the output layer for each class in the training set; and (iii) one neuron in the combinatorial layer for each possible combination of evidences (with order ≥ 2) from the input layer. Combinations of order = 1 are connected directly to the output layer. CNM is trained according to the following learning algorithm in Figure 2.

PUNISHMENT_AND_REWARD_LEARNING_RULE

(i) **Set** the initial value of the accumulator in each arc to zero;

(ii) **For** each example from the training set, **do**:
 Propagate the evidences from input nodes to the output layer;

(iii) **For** each arc arriving to a neuron in the output layer, **do**:
 If the output neuron corresponds to a correct class
 Then backpropagate from this neuron to the input nodes increasing the accumulator of each traversed arc by its evidential flow (reward)
 Else backpropagate decreasing the accumulators (punishment).

Fig. 2. Learning algorithm for CNM

For the sake of simplicity, without quality losses, we used a constant value 1 for the evidential flow. For the CNM training, the examples are presented to the input layer, triggering a signal from each neuron matched to the combinatorial layer, having their weights increased. Otherwise, their weights are weakened. After the training, the accumulators associated to each arc in the output layer will belong to the interval $[-c, c]$, where c is the number of cases in the training set. After the training process, the weights network is pruned as follows: (i) remove all arcs arriving to the output layer with accumulators below a threshold specified by the user and (ii) remove all neurons and arcs from the input and combinatorial layers disconnected after the first step.

The relations that remained after the pruning are considered rules in the application domain. Two basic learning metrics are applied during the model generation that allow the evaluation of the resulting rules (Feldens, 1997):

- *Rule confidence* is the percentages that express the number E of examples for whose the rule R is valid with respect to the total cases of class C . For example, if the rule $R = \text{"IF } beet \text{ and } pod \text{ THEN Item_cost} = high\text{"}$ holds for 15 out of 20 cases with $Item_cost = high$, so $Confidence_R = 75\%$.
- *Rule support* is the percentages that express the number of examples E in which the rule R holds with respect to the total amount of cases (T). If R holds for 15 out of 1000 cases from the database, $Support_R = 1,5\%$.

The main problem with CNM is its well-know low performance due to the exponential growing of the combinatorial layer. However, it received many improvements (Feldens and Castilho, 1997; Machado et al., 1998; Beckenkamp et al., 1998; Prado et al., 1998; Prado et al., 1999; Prado et al., 2000; Noivo et al., 2004) that has turned it feasible as a useful approach for building classifiers.

2.2 Data Description

The data came from a survey applied to a population of 148 agricultural producers from the Brazilian city of Brasília, with data collected from 2007. After a data quality inspection, 118 producers remained, corresponding to near 80% of the whole population. Each producer may explore different combinations of vegetables (e.g., carrot, beet, cassava, and cucumber; carrot, lettuce, and tomato; and so on). For this analysis, it was considered all combinations with carrot, since the region studied is a traditional carrot producer. The main target is to identify the combination with carrot that leads to good economic result for carrot producers. It was collected data on production cost, productivity, profitability for carrot; on what vegetables were cropped along with carrot; the region; home location; and the experience in agriculture for those 133 producers. Production cost, productivity, and profitability for carrot exploration were calculated as shown in Figure 3. Considering that CNM requires discrete data and that the data matrix present a high degree of sparseness, the numeric data were transformed into 2 intervals, using the statistical mean μ : POSITIVE ($Value \geq \mu$) and NEGATIVE ($Value < \mu$). The ranges for discretization are shown in Figure 4.

Variable	Formula	Target
<i>Production cost</i>	<i>Total cost / Harvested amount</i>	Smaller production cost
<i>Productivity</i>	<i>Amount produced / Area</i>	Bigger productivity
<i>Profitability</i>	<i>Total profit / Area</i>	Bigger profitability

Fig. 3. Calculation of production costs, productivity, profitability for carrot

Class	Production cost (R\$) ($\mu = 4.96$)	Productivity (boxes/ha) ($\mu = 937.17$)	Lucrativity (R\$/ha) ($\mu = 5516.60$)
POSITIVE	$Value \geq 4.96$	$Value \geq 937.17$	$Value \geq 5516.60$
NEGATIVE	$Value < 4.96$	$Value < 937.17$	$Value < 5516.60$

Fig. 4. Discretization of production costs, productivity, profitability for carrot

3 Preliminary Data Analysis

In order to gain some insights on the data set, it was built some frequency distributions. Generally, the bigger the productivity the smaller is the cost for unit involved and bigger the profitability. However, not always this relation holds. For example, a producer can have a good productivity but have not obtained good prices for the inputs or, even, sold his products by bad prices. This possibility is made clear in Table 1: the distribution of Negative and Positive (below and above the mean) is very different among the target variables. Notice that 56,8% of the producers have *Productivity* above the mean, but only 44,9% have *Production cost* below the mean, while 51,7% got *profitability* above the mean.

Table 1. Distribution of the target variables

Class	Production cost	Productivity	Profitability
NEGATIVE	44.9%	43.2%	51.7%
POSITIVE	40.7%	56.8%	33.1%
<i>Non identified</i>	14.4%	0.0%	15.3%
TOTAL (118 producers)	100.0%	100.0%	100.0%

Table 2. Frequency distribution of region and home location

Variables		Absolute frequency	Relative frequency
Region	Brazlândia	66	55.9%
	<i>Não identificado</i>	14	11.9%
	Alexandre Gusmão	11	9.3%
	Pipiripau	9	7.6%
	Sobradinho	7	5.9%
	Rio Preto	5	4.2%
	Taguatinga	4	3.4%
	Ceilândia	2	1.7%
	TOTAL	118	100.0%
Home location	Rural	115	97.5%
	Urban	3	2.5%
	TOTAL	118	100.0%

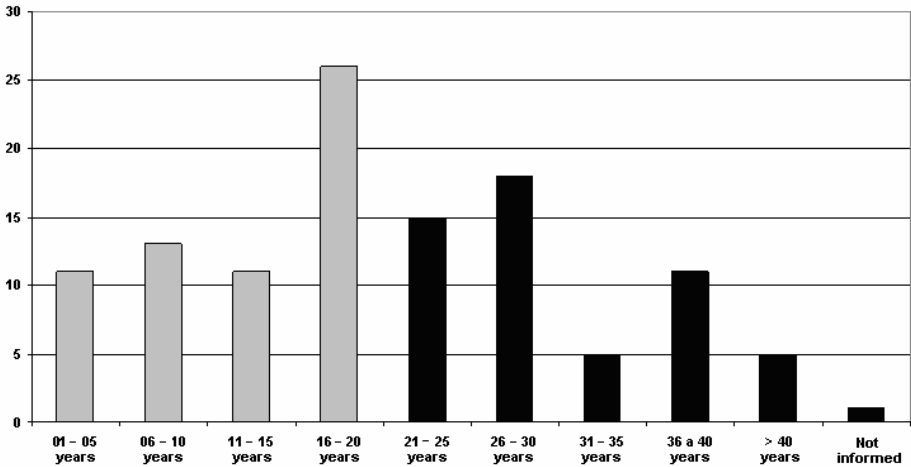


Fig. 5. Producer experience in agriculture

Table 2 and Figure 5 show the frequency distribution of *Region*, *Home location*, and *Experience in agriculture*. In Table 3 is shown the frequency distribution of the vegetables amount cultivated along with carrot. Notice that the 30 producers that do not cultivate other vegetable beyond carrot act as a comparative basis to understand how cropping other vegetables can influence the result of carrot. In other words, these producers allow one to evaluate the marginal gain (or loss) when other vegetables are introduced.

Table 3. Frequency distribution of amount of vegetables cultivated with carrot

Amount of vegetables cultivated with carrot	Absolute frequency	Relative frequency
0	30	25.4%
1	44	37.3%
2	17	14.4%
3	14	11.9%
4	5	4.2%
5	5	4.2%
6	0	0.0%
7	2	1.7%
8	0	0.0%
9	1	0.8%
Total of producers	118	100.0%

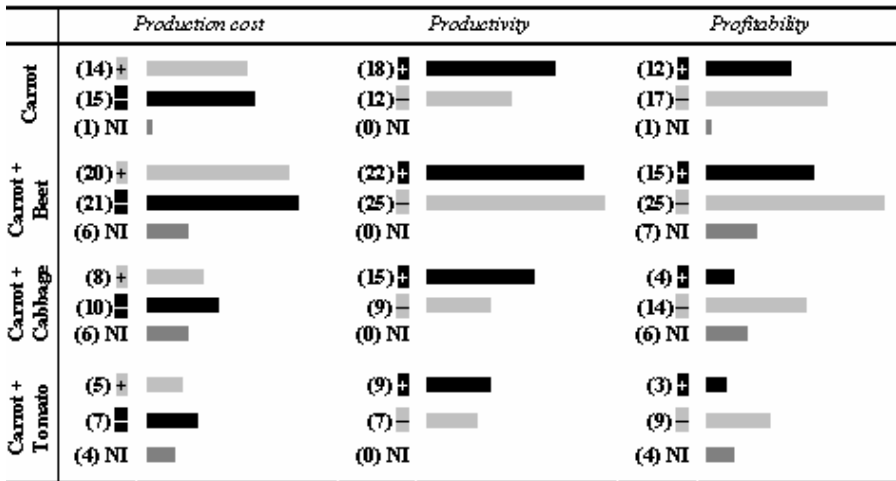


Fig. 6. Target variables performance for one or two combinations of vegetables (the black bars represents good results and the gray ones the bad)

Figure 6 allows one figure out some interactions of crops. Take, for example, the production of carrot alone. For *Productivity* we can see 18 producers above the mean (Positive), corresponding to 60% of the group. Considering the Carrot+Beet group, the chance to have good results decreases to 47%. The same happens when carrot is combined with tomato (56%). The best combination is Carrot+Cabbage, with the probability of 62.5% for obtaining good results in *Productivity*.

Regarding to *Production costs*, the chance of good results with carrot alone is present is 51.7%. It can be observed that the producers have not being able to convert good results of *Productivity* in smaller *Production costs*. Here, the better combination is Carrot+Tomato that improves the possibility of success to 58%. In terms of *Profitability*, this combination presents low probability of success (25%). It may represent difficulties of producers to commercialize their production.

The descriptive analysis presented considers only some combinations, it does not take into account all possibilities from the training set. This analysis does not allow one to express how significative is a relation, what can be easily obtained from CNM as we can see in the next section.

4 Model Results Analysis

To generate the model for this analysis, the parameters *Confidence* and *Support* were set to 50% and 3%, respectively. The maximum size of the rule left side was set to five conditions. From the rules generated, we focussed in those with the highest *Confidence*, as shown in Figure 7. Observe positive and negative results for low *Production cost*, with high confidence, for the combinations carrot + cassava + pepper (R_{C+01}), corn (R_{C+02}) and beet (R_{C+03}). With the same *Confidence*, but a slightly empirical support, carrot appears alone in the region of Rio Preto with a good *Production cost* (R_{C+04}). However, for the same region and carrot alone, the *Profitability* is negative (R_{L-05}). It may reveal difficulties to sell the production, reaching a low final price. Additionally, high *Production cost* for carrot in the region of Alexandre Gusmão for producer with high experience (R_{C-01}) is detected. This observation, despite the good *Confidence* and *Support* level of the rule, is contrary to the common sense that associates high experience with low *Production cost* and may be a result of noisy data. On the other hand, if this result is confirmed, it will point out the necessity of an intervention in the region to correct inadequate agricultural practices.

With regard to *Productivity*, it was not found relevant combinations neither positive nor negative. The only relation detected was the positive results for carrot cropping in the region of Pípiripau (R_{P+01}). However, the confidence and the support presented are not conclusive, only indicating that the relation requires a finer analysis, possibly, with more data.

REGRA	(Conf. /Sup.M.)
R_{C+01} : IF Cassava & Pepper THEN Production cost POSITIVE...	(100% / 3,8%)
R_{C+02} : IF Corn THEN Production cost POSITIVE...	(100% / 3,8%)
R_{C+03} : IF Eggplant..... THEN Production cost POSITIVE...	(100% / 3,8%)
R_{C+04} : IF Region = Rio Preto THEN Production cost POSITIVE...	(100% / 5,7%)
R_{C-01} : IF Region = Alexandre Gusmão & Experience > 20 THEN Production cost NEGATIVE	(100% / 6,3%)
R_{P+01} : IF Region = Pípiripau THEN Productivity POSITIVE.....	(77,8% / 10,5%)
R_{L+01} : IF Broccolis..... THEN Profitability POSITIVE	(100% / 5,1%)
R_{L-01} : IF Experience ≤ 20 & Beet & Tomato THEN Profitability NEGATIVE	(100% / 4,9%)
R_{L-02} : IF Region = Pípiripau & Experience ≤ 20 & Tomato THEN Profitability NEGATIVE	(100% / 4,9%)
R_{L-03} : IF Region = Alexandre Gusmão & Cabbage .. THEN Profitability NEGATIVE	(100% / 4,9%)
R_{L-04} : IF Yam THEN Profitability NEGATIVE	(100% / 4,9%)
R_{L-05} : IF Region = Rio Preto THEN Profitability NEGATIVE	(100% / 4,9%)
R_{L-06} : IF Experience ≤ 20 & Cabbage THEN Profitability NEGATIVE.....	(100% / 11,5%)

Fig. 7. Rules with the highest *Confidence*

For *Profitability*, broccolis is clearly a good candidate to combine with carrot (R_{L+01}), associated with a good confidence and support. On the other hand, the low *Profitability* appears strongly associated with low experience and cabbage (R_{L+06}). The further explanations for low *Profitability* are associated with low experience and combination of beet and tomato (R_{L-01}), region of Pípiripau, low experience and tomato (R_{L-02}), region of Alexandre Gusmão and cabbage (R_{L-03}), yam (R_{L-04}) and region of Rio Preto (R_{L-05}), the latter already discussed when we approached the *Production cost*.

5 Conclusion

It was presented a reasoning-based approach for a problem solution when many combinations of factors are presented. The knowledge generation task is based in data mining techniques. The approach was applied to a problem from the farm management context, of how to take advantage from the synergies existing when cultivating multiple vegetables. The study can provide useful information to the farm manager regarding to the target variables Production cost, Productivity, and Profitability, that can feed a decision support system for planning the farm exploration.

References

1. Beckenkamp, F.G., Pree, W., Feldens, M.A.: Optimizations of the Combinatorial Neural Model. In: Proceedings of the 5th Brazilian Symposium on Neural Networks, p. 49 (1998)
2. Feldens, M.A., Castilho, J.M.V.: Data Mining with the Combinatorial Rule Model: An Application in a Health-Care Relational Database. In: Proceedings of the XXIII Latin-American Conference on Informatics (CLEI), Valparaiso, Chile (1997)
3. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice-Hall, Englewood Cliffs (1999)
4. Machado, R.J., Rocha, A.F.: Handling knowledge in high order neural networks: the combinatorial neural network. IBM Rio Scientific Center, Brazil (1989) (Technical Report CCR076)
5. Machado, R.J., Rocha, A.F.: The combinatorial neural network: a connectionist model for knowledge based systems. In: Bouchon-Meunier, B., Zadeh, L.A., Yager, R.R. (eds.) IPMU 1990. LNCS, vol. 521, pp. 578–587. Springer, Heidelberg (1991)
6. Machado, R.J., Carneiro, W., Neves, P.A.: Learning in the combinatorial neural model. IEEE Transactions on Neural Networks 9, 831–847 (1998)
7. Noivo, R., Prado, H.A., Ladeira, M.: Yet Another Optimization of the Combinatorial Neural Model. In: Proceedings of the XXX Latin-American Conference on Informatics (CLEI), Arequipa, Chile, pp. 706–711 (2004)
8. Prado, H.A., Frigeri, S.R., Engel, P.M.: A Parsimonious Generation of Combinatorial Neural Model. In: IV Argentine Congress on Computer Science (CACIC 1998), Neuquén, Argentina (1998)
9. Prado, H.A., Machado, K.F., Frigeri, S.R., Engel, P.M.: Accuracy Tuning on Combinatorial Neural Model. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS (LNAI), vol. 1574, pp. 247–251. Springer, Heidelberg (1999)
10. Prado, H.A., Machado, K.F., Engel, P.M.: Alleviating the complexity of the Combinatorial Neural Model using a committee machine. In: Proceedings of the International Conference on Data Mining. WIT Press, Cambridge (2000)

Reasoning Elements for a Vehicle Routing System

Edilson Ferneda¹, Bernardo A. Mello², Janaína A.S. Diniz³, and Adelaide Figueiredo¹

¹ Graduate Program on Knowledge Management & IT,
Catholic University of Brasília, Brazil

² Physic Department, University of Brasília, Brazil

³ Undergraduate Course of Agribusiness Management, University of Brasilia, Brazil
eferneda@pos.ucb.br, bernardo@fis.unb.br,
janadiniz@unb.br, adelaid@pos.ucb.br

Abstract. Logistic processes are a challenge to collective organizations of rural producers. The lack of technological support in those processes is an obstacle for the maintenance of the small producers in the distribution channels. A research theme is the adaptation of some logistic models normally adequate for urban reality to the context of small rural organizations. This adaptation seems pertinent, since some particularities of this kind of organization can be associated to complex problems related to the supply chain. In the case presented in this work, the first stage in the development of the system consisted in defining the main problems related to the inefficiency of the collection and distribution processes. Costs concerning the logistics processes were also measured. In this paper we propose the reasoning elements regarding a routing system and methods for cost estimation adapted to small rural organizations.

Keywords: Intelligent routing, Logistics, Ant Colony Optimization.

1 Introduction

The inefficiency in product collection and delivery route planning has been a major problem for some associations, co-operatives and other groups of rural producers, especially those represented by small family farmers. This deficiency causes an increase in total costs and a decrease in competitive advantage due to the lack of infrastructure and planning for products transportation, with important impacts generally on the sparing budget of small producers.

To improve the levels of services and client attendance of these small organizations, they are almost obliged to modify their processes in the management of the flow of materials and information. Preferably they are aided by software to support supply chain management and the physical distribution operations. However, the majority of existing routing systems are not appropriate for the specifications of small rural organizations. Usually they were originally developed to deal with specific regions or even countries.

Taking into account this context, the software rotAgro was developed, in order to give access to a system offering the best collecting and delivery routes for the products of the engaged organizations. In this work the description of some essential

details in the development of the rotAgro system is presented: the architecture of the system, the best route calculation and the calculation of the best way between two points in the route.

2 The Architecture of rotAgro

The routing system presented here provides the best collection and distribution routes, based on rapidity, cost optimization and improvement of service. The development of this system is justified by the necessity of revising certain current routing systems, presenting solutions to the real problems mentioned by the chain members.

The developed production environment consists of a computer executing with a Linux platform, containing the following components (Figure 1), which were chosen for being open-source, counting on an ample community of support:

- A Web server for receiving HTTP requisitions over the Internet, transferring them, according to the necessity, to the application server. The chosen server was the Apache 2. This Server has a module (mod-jk2) for communication with the Java application servers (JEE).
- An application Server for giving support to the execution of applications. JBOSS 4.x was chosen.
- A repository where information will be recorded, brought up to date or consulted, including the georeferred information. The PostgreSQL 8.3 Data Base Management System was chosen, specifically the PostGIS module for georeferred data.

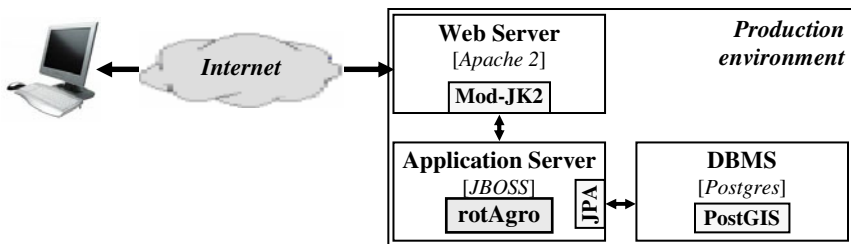


Fig. 1. Production environment of the rotAgro system

The rotAgro application was developed in Java, according to the best practices, following the model MVC (*Model-View-Controller*). The system's architecture is shown in Figure 2.

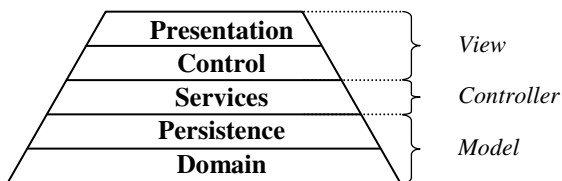


Fig. 2. The organization in layers of the rotAgro system

The presentation and control layers were implemented following the JSP's and Servlet's specifications, through the framework Spring-MVC. This open-source framework allows functionalities such as data transformation from HTML forms to Java objects, message control of the interface, configuration via XML, among others. In this layer, besides the Spring-MVC, there are also other components applied, like JSTL to implement customized components for form control and the Log4J for control of the audit system.

Among the services offered by the system rotAgro, there are the collection and data storage concerning users, clients, employees, commercial points, vehicles, products, nodes, edges, lanes and routes.

For rotAgro proposes a route, all the data infrastructure must be available, which means that the route must have nodes, edges, lanes and commercial points to cover; clients responsible for the routes; employees that will use the vehicle to distribute the products at the commercial points; and a user responsible to keep all these data.

The Services layer is characterized by being POJO (Plain Old Java Object). This means that only the standard API of Java was used, assuring that this layer can be reused in any other application, including a Desktop Version of rotAgro.

The Persistence layer uses a new technology standard known as JPA, which was already incorporated in the JEE 5 specification. This API gives us the flexibility of database independence during implementation, thus assuring one of the principles of Java language: portability. The Persistence Layer is an abstraction layer on the JDBC (Java Database Connectivity), an access standard to the database in Java language. The functions used to convert data into georeferred data are also implemented in this layer in order to persist in PostGIS.

Like the Persistence layer, the Domain layer is represented by POJOs that usually do not depend on other libraries. However, since this application depends on a georeferred base, the openGIS libraries and geotools were used.

Among the services offered by the system, those directly related to the vehicle routing problem can be highlighted. The order in which the points are visited is calculated from the relative data to the production collection points. The system calculates the best routes pairwise. To calculate the order of the points, an algorithm based on the Ant Colony Systems (Dorigo & Gambardella, 1997) is used. For the second phase, the Dijkstra algorithm (Zhan & Noon, 1998) is used.

3 The Best Route Calculation

The vehicle routing problem (VRP) has been considered from diverse angles in the Optimization area (Yang, 2008). Many works have shown the persistence of the VRP approach by bio-inspired systems (Pereira & Tavares, 2009). One of the most promising is situated in Swarm Intelligence area (Engelbrecht, 2005). Among its diverse branches, there is the Ant Colony Systems (ACS) (Dorigo, 1992).

ACS was inspired by the alimentary behavior of real an ant colony. The ants leave the colony searching for food, initially walking in a random way. When an ant finds food, it goes back to the colony and leaves a pheromone trace in the track. If another ant finds this trace, the tendency is that it follows it. In the case of finding food, the ant returns to the colony, reinforcing the pheromone trace in the track. But as time

passes the pheromone traces evaporate. So, the more ants pass over the track, more time will be necessary for the pheromone trace to evaporate. The pheromone evaporation prevents the convergence to a local optimal solution. If the evaporation does not occur, all the tracks marked by the first ants would be very attractive to other ants, reducing the attractiveness of the solution space.

ACS has been employed in many applications, including the vehicle routing problem (VRP). A classic VRP algorithm must find vehicle routes with minimal costs, such that (Bullheimer, Kotsis & Strauss, 1997): (i) every customer is visited once only, and by only one vehicle; (ii) for every vehicle, total demand should not exceed the capacity of the vehicle; (iii) every tour (done by a single vehicle) starts and ends at a unique depot, and (iv) the total tour length does not exceed a given limit. The cost of a solution is generally expressed as multiple objectives, mainly to minimize the total travel time and to maximize customer satisfaction. Additionally, the number of vehicles can be minimized.

A VRP can be represented by the following form, adapted from Engelbrecht (2005, p. 453-454) for only one vehicle. A complete weighted digraph, $G = (V, E)$, is used as representation of the search space, where V is the set of nodes $\{V_0, V_1, \dots, V_n\}$, one for each of the customers (nodes V_1, \dots, V_n), and V_0 represents the depot; $E = \{(r,s) \mid r, s \in V\}$ is the set of links between customers and between the depot and customers. With each link (r,s) is associated a cost $d_{r,s}$, which may represent the distance between customers r and s ($\delta(r,s)$). With each node $V_i, i = 1, \dots, n$, the following information is maintained for each customer i : the demand, $d_i(t)$, of that customer.

The algorithm was implemented in Java based on the framework described by Chirico (2004). According to the ACS model for VRP, virtual ants go along the streets (edges of the graph) until arriving at a stopping point (nodes of the graph). At each point, the ant must renew the quantity of pheromone in the street walked (local update) and define which will be the next point to be visited (state transition rule). When the ant arrives at the end of its trip, the quantity of pheromone in the streets of the produced route (global update) is updated. A fraction of pheromone evaporates in all the routes.

By the state transition rule, an ant situated in the node r chooses the next point s (next point to be visited) following the rule defined by Equation 1, where q is a random number uniformly distributed in $[0,1]$, q_0 is a parameter ($0 \leq q_0 \leq 1$). $\tau(r,s)$ is the desirability measure of the edge (r,s) , or pheromone. For this version of rotAgro, it was considered that $\tau(r,s)$ and $\tau(s,r)$ are symmetric ($\tau(r,s) = \tau(s,r)$), $\eta = 1/\delta(r,s)$ and β is the parameter that determines the relative importance of pheromone versus distance ($\beta > 0$). S is a random variable selected according to the probability distribution given by Equation 2, that gives the probability with which ant k in r chooses s as goal, where $J_k(r)$ is a set of nodes that were not yet visited by the ant k from r . The resulting state transition rule from these two equations is called pseudo-random proportional rule.

While a solution is built, the ants visit edges/streets and change their pheromone level, according to the local updating rule defined by Equation 3, where $0 < \rho < 1$ is the pheromone evaporation rate and L_{gb} is the size of the globally best tour from the beginning of the trial.

The global updating rule is defined by Equation 4, where $0 < \alpha < 1$ is the pheromone decay parameter.

$$s = \begin{cases} \arg \max_{u \in J_k(r)} \{ [\tau(r,u)] [\eta(r,u)^\beta] \} & \text{if } q \leq q_0 \\ S & \text{otherwise} \end{cases} \tag{1}$$

$$P_k(r,s) = \begin{cases} \frac{[\tau(r,s)] [\eta(r,s)^\beta]}{\sum_{u \in J_k(r)} [\tau(r,u)] [\eta(r,u)^\beta]} & \text{if } s \in J_k(r) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\tau(r,s) \leftarrow (1-\rho)\tau(r,s) + \rho\Delta\tau(r,s)$$

$$\text{where } \Delta\tau(r,s) = \begin{cases} (L_{gb})^{-1} & \text{if } (r,s) \in \text{global best tour} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$\tau(r,s) \leftarrow (1-\alpha)\tau(r,s) + \alpha\Delta\tau(r,s) \tag{4}$$

4 Calculation of the Best Way between Two Points in the Route

For the calculation of the best way between two adjacent points in the route, the classical algorithm of Dijkstra (Zhan & Noon, 1998) was used. This algorithm solves the problem of the shortest way in a graph driven or not driven with edges of non-negative weight. As this concerns a routing problem, some parameters must be considered for this calculation. The parameters applied to the system in rural areas are not very different from the ones used in urban sites. However, before and simultaneously to the implementation of the system, local features must be carefully considered.

Among many factors related to vehicle cost, those influencing the total cost of a route can be divided into three groups:

- *Personal Cost:* This mainly refers to the payment of the driver, who is also responsible for loading and unloading the products. The driver is paid a monthly salary (S_M) that corresponds to a T_D working hours per day. If the time necessary to finish a route (T_R) is greater than T_D , the additional cost per extra-hour is S_{HE} . We also consider the driver's preference in finishing the route in the minimum time. Even if this factor does not directly influence the route cost, its effect on the satisfaction of the employee cannot be neglected, mostly because he is a key element for the acceptance of the system. This will be evaluated by the cost S_H , much smaller than the extra-hour value. In Equation 5, $H(x)$ is the heavy side function: $H(x) = 0$ if $x < 0$ and $H(x) = 1$ if $x > 0$.

$$C_p = \frac{S_M}{22.5} + S_H T_R + S_{HE} (T_R - T_D) H(T_R - T_D) \tag{5}$$

- *Fuel cost*: As the fuel consumption per kilometer (L_{KM}) depends on the road condition and the total weight of the vehicle, its cost can be calculated by Equation 6, where P_L is the price per liter of fuel and the sum is calculated for each road and load condition, each one measuring d kilometers. Archondo-Callao (1994) presents a detailed model for fuel consumption. Due to the impossibility of obtaining all the parameters, a simplification is proposed by Equation 7, where L_0 is the fuel consumption in neutral, a is the energetic efficiency factor and F is the force required to move the vehicle. The main factors that affect that force are the rolling friction, ascents, descents and aerodynamics. The total force is the sum of these effects (Equation 8), where R_R is the rolling resistance, I_R is the sine of the road slope (positive for ramps and negative for declining surfaces), M_T is the total mass transported (vehicle mass plus load mass, in Kg), g is the gravity acceleration, ρ_{Ar} is the air density, C_{Ar} is the aerodynamic coefficient, A_{Ar} is the frontal area and V is the speed.

$$C_F = P_L \sum_i L_{KM_i} d_i \quad (6)$$

$$L_{KM} = L_0(1 + aF) \quad (7)$$

$$F = (R_R + I_R)M_T g + \rho_{Ar} C_{Ar} A_{Ar} V^2 \quad (8)$$

- *Maintenance cost*: Lubricant oils and tires represent a maintenance cost that is considered dependent only on the travel distance. The oil substitution occurs every d_{oil} kilometers and its cost is P_{oil} . In the same way, each tire costs P_{tire} , including the acquisition price and retreading during the life cycle of the tire. During this period each tire travels d_{tire} kilometers, the maintenance cost of a vehicle with N_{tire} tires is defined by Equation 9. There are other factors included in the vehicle maintenance cost, such as load deterioration, mechanic maintenance, insurance, depreciation, interest and taxes. However, these costs are either fixed, i.e., do not depend on the time or the distance of a given route, or they are much smaller than other costs.

$$C_M = \frac{P_{oil}d}{d_{oil}} + \frac{N_{tire}P_{tire}d}{d_{tire}} \quad (9)$$

An important parameter in route characterization is time. Although the Personal Cost (CP) defined above depends on the total route duration (TR), there are restrictions concerning the instant when the vehicle visits one of the loading and unloading stations. Consequently, it is important to estimate the vehicle speed at each point of the route. We will use the model presented by Archondo-Callao (1994) that defines that vehicle speed is approximately given by Equation 10, where the parameter β determines the Weibull distribution format. Each one of the speeds is related to at least one of these limiting factors: the maximum motor power, the maximum capacity of the brake system, the incidence of curves, the roughness of the route and the velocity desired by the driver in ideal conditions. When transporting liquid products, it is necessary to change the V_{curve} and V_{rough} calculation in order to include the effects of the liquid movement on the vehicle stability.

$$V = \left(\frac{1}{V_{drive}^{1/\beta}} + \frac{1}{V_{brake}^{1/\beta}} + \frac{1}{V_{curve}^{1/\beta}} + \frac{1}{V_{rough}^{1/\beta}} + \frac{1}{V_{desir}^{1/\beta}} \right)^\beta \quad (10)$$

When calculating the instant when each route point is reached, it is necessary to include the vehicle loading and unloading times at the collection and delivery points and at the beginning and end of the route. These times depend on the facilities available at each point and on the volume of cargo involved. Three variables must be computed while the route is being traveled: the fuel remaining, the load and the time of each step of the route. The first two are used for calculating the travel time and the fuel spent at each step. The load is constant between two nodes, and the fuel may be considered constant, allowing the cost of traveling between two nodes to be easily calculated. The third is important because restrictions may exist about the acceptable time for visiting each node. Routes that don't satisfy the restrictions are discarded.

5 Conclusion

Most of the problematic points searched in the rotAgro project were shown in the logistic chain and costs. The main problems of inefficiency in the planning were identified concerning collection and distribution of family farming products. This can be used as a base for feeding a model of routing software for cooperatives, associations and formal or informal groups of rural producers.

There were many difficulties to overcome to assure a first prototype of a routing system adapted to small organizations of rural producers. However, once the reality of each organization was known, it was possible to define the most adequate parameters to be applied in each case. The first version of the rotAgro system does not allow changes in the parameters. It is expected, however, to evolve in this sense, via future research and as new users get to know the system better, proposing changes and adaptations in the system.

Besides choosing an adequate routing model, it is necessary to overcome deficiencies in process management and in the integration of collection and distribution processes. The field research has proved that a well-planned route has not only a positive impact on costs, but can also improve the quality of the services delivered to clients. Consequently, an adequate model for rural organizations can help in the distribution of products, and can be a strategic differential to facilitate their insertion in the market.

Concerning the model, efforts are made to improve the geographical interface and to test other VRP algorithms, such as VRP with time windows (Gambardella, Tillard & Agazzi, 1999) and Dynamic VRP (Donati et al., 2003; Montemanni et al., 2002; Tian et al., 2003).

Acknowledgements

The authors are grateful to the Brazilian National Council for Scientific and Technological Development (CNPq), which partially financed this research.

References

1. Archondo-Callao, R.S., Faiz, A.: Estimating Vehicle Operating Costs. World Bank Technical Paper Number 234. Washington (1994)
2. Bullnheimer, B., Kotsis, G., Strauss, C.: Parallelization Strategies for the Ant System. In: De Leone, R., et al. (eds.) *High Performance Algorithms and Software in Nonlinear Optimization*. Kluwer Series on Applied Optimization, vol. 24, pp. 87–100 (1997)
3. Chirico, U.: A Java Framework for Ant Colony Systems (2004), <http://www.ugosweb.com/Download/JavaACSFframework.pdf>
4. Donati, A.V., Montemanni, R., Gambardella, L.M., Rizzoli, A.E.: Integration of a robust shortest path algorithm with a time dependent vehicle routing model and applications. In: *Proceedings of the IEEE International Symposium on Computational Intelligence for Measurement Systems and Applications*, pp. 26–31 (2003)
5. Dorigo, M.: Optimization, learning and natural algorithms. PhD Thesis, DEI, Politecnico di Milano, Italy, In Italian (1992)
6. Dorigo, M., Gambardella, L.M.: Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. *IEEE Trans. on Evolutionary Computation* 1(1) (1997)
7. Engelbrecht, A.P.: *Fundamentals of Computational Swarm Intelligence*. John Wiley & Sons Ltd, West Sussex (2005)
8. Gambardella, L.M., Taillard, E., Agazzi, G.: MACS-VRPTW: a multiple Ant Colony System for Vehicle Routing Problems with time windows. Technical report, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland (1999)
9. Montemanni, R., Gambardella, L.M., Rizzoli, A.E., Donati, A.V.: A new algorithm for a dynamic Vehicle Routing Problem based on Ant Colony System. Technical report, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland (2002)
10. Pereira, F.B., Tavares, J.: Bio-inspired algorithms for the Vehicle Routing Problem. In: *Studies in Computational Intelligence*, vol. 161, Springer, Berlin (2009)
11. Tian, Y., Song, J., Yao, D., Hu, J.: Dynamic vehicle routing problem using hybrid and system. *IEEE Intelligent Transportation Systems* 2, 970–974 (2003)
12. Zhan, F.B., Noon, C.E.: Shortest Path Algorithms: An Evaluation Using Real Road Networks. *Transportation Science* 32(1), 65–73 (1998)
13. Yang, X.-S.: *Introduction to Mathematical Optimization: From Linear Programming to Metaheuristics*. Cambridge Int. Science Publishing (2008)

A Mechanism for Converting Circuit Grammars to Definite Clauses

Takushi Tanaka

Department of Computer Science and Engineering
Fukuoka Institute of Technology
3-30-1 Wajiro-Higashi Higashi-ku, Fukuoka 811-0295, Japan
tanaka@fit.ac.jp
<http://www.fit.ac.jp/~tanaka>

Abstract. The circuit grammar is a logic grammar developed for knowledge representation of electronic circuits. Knowledge of circuit structures and their functions are coded as grammar rules. Those grammar rules, when converted into definite clauses, form a logic program that can parse given circuits and derive their electrical behavior. This paper shows a mechanism for converting circuit grammar rules into definite clauses.

1 Introduction

As a step toward automatic circuit understanding, we developed a new method for analyzing circuit structures [5], [6]. We viewed circuits as sentences and their elements as words. Electrical behavior and functions are meaning of the sentences. Knowledge of circuit structures and their electrical behavior are coded as grammar rules. A set of grammar rules, when converted into definite clauses, forms a logic program which performs top-down parsing.

The circuit grammar is a descendant of the logic grammar called DCSG (Definite Clause Set Grammar) [2] which was developed for analyzing word-order free languages. The circuit grammar consists of several DCSG extensions that are useful for analyzing electronic circuits.

We first introduce DCSG and show the mechanism for converting grammar rules into definite clauses. Next we introduce extensions for circuit analysis and show mechanisms for converting such circuit grammars to definite clauses.

2 Logic Grammar DCSG

2.1 Word-Order Free Language

Most implementations of the computer language Prolog provide a mechanism for parsing context-free languages called DCG (Definite Clause Grammar) [1]. A set of grammar rules, when converted into definite clauses, forms a logic program that performs top-down parsing. While, a logic grammar DCSG [2] was developed for word-order free language similar to the method of DCG.

A word-order free language $\mathbf{L}(\mathbf{G}')$ is defined by modifying the definition of a formal grammar. We define a context-free word-order free grammar \mathbf{G}' to be a quadruple $\langle V_N, V_T, P, S \rangle$ where: V_N is a finite set of non-terminal symbols, V_T is a finite set of terminal symbols, P is a finite set of grammar rules of the form:

$$\begin{aligned} A &\longrightarrow B_1, B_2, \dots, B_n. & (n \geq 1) \\ A \in V_N, \quad B_i &\in V_N \cup V_T & (i = 1, \dots, n) \end{aligned}$$

and $S(\in V_N)$ is the starting symbol. The above grammar rule means that the symbol A is rewritten not with the string of symbols “ B_1, B_2, \dots, B_n ”, but with the set of symbols $\{B_1, B_2, \dots, B_n\}$. A sentence in the language $\mathbf{L}(\mathbf{G}')$ is a set of terminal symbols derived from S by successive application of grammar rules. Here the sentence is a multi-set which admits multiple occurrences of elements taken from V_T . Each non-terminal symbol used to derive a sentence can be viewed as a name given to a subset of the multi-set.

2.2 DCSG Conversion

The general form of the conversion procedure from a grammar rule

$$A \longrightarrow B_1, B_2, \dots, B_n. \tag{1}$$

to a definite clause is:

$$\begin{aligned} subset(A, S_0, S_n) : - & subset(B_1, S_0, S_1), \\ & subset(B_2, S_1, S_2), \\ & \dots \\ & subset(B_n, S_{n-1}, S_n). \end{aligned} \tag{1}'$$

Here, all symbols in the grammar rule are assumed to be non-terminal symbols. If “[B_i]” ($1 \leq i \leq n$) is found in the right hand side of grammar rules, where “ B_i ” is assumed to be a terminal symbol, then “ $member(B_i, S_{i-1}, S_i)$ ” is used instead of “ $subset(B_i, S_{i-1}, S_i)$ ” in the conversion.

The arguments S_0, S_1, \dots, S_n in (1)' are multisets of V_T , represented as lists of elements. The predicate “ $subset$ ” is used to refer to a subset of an object set which is given as the second argument, while the first argument is the name of its subset. The third argument is a complementary set which is the remainder of the second argument less the first; e.g. “ $subset(A, S_0, S_n)$ ” states that “ A ” is a subset of S_0 and that S_n is the remainder.

The predicate “ $member$ ” is defined by the definite clauses (2) and (3) below. It has three arguments. The first is an element of a set. The second is the whole set. The third is the complementary set of the first argument.

$$member(M, [M|X], X). \tag{2}$$

$$member(M, [A|X], [A|Y]) : - member(M, X, Y). \tag{3}$$

When the clause (1)' is used in parsing, an object sentence (multiset of terminal symbols) is given as the argument S_0 . In order to find the subset A in S_0 , the first sub-goal finds the subset B_1 in S_0 then put the remainder into S_1 , the next sub-goal finds B_2 in S_1 then put the remainder into S_2 , ..., and the last sub-goal finds B_n in S_{n-1} then put the remainder into S_n . That is, when a grammar rule is used in parsing, each non-terminal symbol in the grammar rule makes a new set from the given set by removing itself as its subset. While, each terminal symbol used in the grammar rule also makes a new set from the given set by removing itself as its member.

DCSG uses the predicates *subset* and *member* to convert grammar rules into definite clauses, but the differences between DCG and DCSG are minimal. If we replace the predicate *subset* with *substring* and remove the clause (3) from the definition of *member*, the conversion will be equivalent to DCG conversion, although ordinary DCG does not use the predicate *substring* for simplification.

DCSG allows the symbol “;” as abbreviation of two grammar rules with the same left hand side. The following rule (4) which generates B or C_1, C_2 from A is converted to the definite clause (4)' as follows:

$$A \longrightarrow B; C_1, C_2. \quad (4)$$

$$\begin{aligned} \text{subset}(A, S_0, S_2) :- & \text{subset}(B, S_0, S_2); \\ & \text{subset}(C_1, S_0, S_1), \\ & \text{subset}(C_2, S_1, S_2). \end{aligned} \quad (4)'$$

3 Mechanism for DCSG-Conversion

The following List 1 shows a basic DCSG-converter written in Prolog. The line 01 defines the main predicate “dcsConv” which converts a grammar rule into a definite clause. It separates the grammar rule into a left-hand side and a right-hand side, and generates a definite clause from a head part and a body part. The first subgoal “conv(Lhs,Head,S0,S1)” generates the head part from the left-hand side using the definition of line 10, and the second subgoal “conv(Rhs,Body,S0,S1)” generates the body part from the right-hand side. If the right-hand side consists of more than two grammar symbols, the predicate “conv” defined by the line 02 and 03 separates these symbols into the first one and others. The line 02 generates a conjunctive subgoals from grammar symbols connected by “;”, while the line 03 generates a disjunctive subgoals from “;”. The line 09 defines the conversion of a single terminal symbol. The line 10 defines the conversion of a single non-terminal symbol, which is used both for converting the left-hand side and the right-hand side.

List 1: Basic DCSG-converter

```
-----
01 dcsConv((Lhs --> Rhs), (Head :- Body)) :-
    conv(Lhs,Head,S0,S1), conv(Rhs,Body,S0,S1).
02 conv((CompA,CompB), (CA,CB),S0,S1) :- !,
```

```

conv(CompA,CA,S0,S), conv(CompB,CB,S,S1).
03 conv((CompA;CompB),(CA;CB),S0,S1):-!,
conv(CompA,CA,S0,S1), conv(CompB,CB,S0,S1).
09 conv([Component],member(Component,S0,S1),S0,S1):-!.
10 conv(Component,subset(Component,S0,S1),S0,S1):-!.
-----

```

4 Extensions for Context-Dependent Features

4.1 Condition for Absence

When we define grammar rules for actual circuits, several extensions for context-dependent features are needed.

The circuit *ca39* in Figure 1 is represented by the fact (5). Here, the list "[*resistor*(*r1*, 1, 2), *resistor*(*r2*, 2, 3), ...]" is a word-order free sentence. The compound terms such as *resistor*(*r1*, 1, 2) and *terminal*(*t1*, 1) are words which represent the resistor *r1* connected to the nodes 1 and 2 and the external terminal *t1* connected to the node 1 respectively.

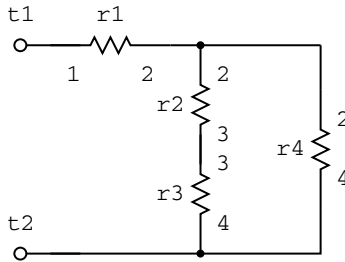


Fig. 1. Circuit ca39

$$ca39([resistor(r1, 1, 2), resistor(r2, 2, 3), resistor(r3, 3, 4), resistor(r4, 2, 4), terminal(t1, 1), terminal(t2, 4)]). \quad (5)$$

First we define the non terminal symbol *res*(*R*, *A*, *B*) by the rule (6) which enables us to refer to a resistor regardless its node order, because the resistor is a non-polar element. The rule is converted to the definite clause (6)'.

$$res(R, A, B) \longrightarrow [resistor(R, A, B)]; [resistor(R, B, A)]. \quad (6)$$

$$subset(res(R, A, B), S0, S1) :- member(resistor(R, A, B), S0, S1); member(resistor(R, B, A), S0, S1). \quad (6)'$$

The following conjunctive goal attempts to find a series connection of resistors (Figure 2) in the circuit *ca39*. The first subgoal *ca39(CT0)* binds the circuit to the variable *CT0*. The second subgoal *subset(res(X, A, B), CT0, CT1)* finds the non-terminal symbol *res(r1, 1, 2)* as a subset of *CT0*. The variable *CT1* is bound to the difference set which does not contain *resistor(r1, 1, 2)*. The third subgoal *subset(res(Y, B, C), CT1, -)* finds the non-terminal *res(r2, 2, 3)* in the circuit *CT1* as:

```
?- ca39(CT0),
    subset(res(X,A,B),CT0,CT1), subset(res(Y,B,C),CT1,-).
```

```
X = r1
Y = r2
A = 1
B = 2
C = 3
```

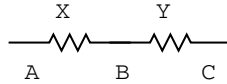


Fig. 2. Resistors connected in series

But this is not a correct answer because the central node *B*(= 2) of the series connection is also connected to another resistor *r4*. The central node of series connection must not be connected to any other element. This constraint is realized by introducing the condition of absence into grammar rules as follows:

$$anyElm(X, A) \longrightarrow [terminal(X, A)]; res(X, A, -). \tag{7}$$

$$rSeries(rs(X, Y), A, C) \longrightarrow \begin{array}{l} res(X, A, B), \\ res(Y, B, C), \\ not\ anyElm(-, B). \end{array} \tag{8}$$

The grammar rule (7) defines the non-terminal “*anyElm(X, A)*” which represents any element *X* connected to the node *A*. The grammar rule (8) defines two resistors connected to the same node *B* with the condition that no other elements may be connected to the node *B*. The grammar rule (8) is converted to the definite clause (8)′.

$$subset(rSeries(rs(X, Y), A, C), S0, S2) : - \begin{array}{l} subset(res(X, A, B), S0, S1), \\ subset(res(Y, B, C), S1, S2), \\ not\ subset(anyElm(-, B), S2, -). \end{array} \tag{8}'$$

The following goal successfully fined the series circuit of resistors “*rSeries(rs(r2, r3), 2, 4)*” in the circuit *ca39*:

```
?- ca39(CT0), subset(rSeries(X,A,C),CT0,-).
```

The conversion with this extension is implemented by adding the following lines after the line 03 of List 1. Here, “not” must be declared as a prefix-operator.

```
-----
04 conv(not [Component],not member(Component,S0,_),S0,S0) :- !.
05 conv(not Component,not subset(Component,S0,_),S0,S0) :- !.
-----
```

4.2 Conditions for Existence

Consider the circuit design process in contrast with sentence generation. Suppose a circuit goal generates two current sources as its sub-goals. Each current source needs a regulated voltage source, so two voltage regulators are generated. When one of the voltages is derived from the other, an engineer may combine two voltage regulators into one voltage regulator for simplicity. That is, he has the ability to use context dependent circuit generation rules.

In our system [5], the V_{be} -voltage regulator and the current source (sink-type) in Figure 3 are defined by the grammar rules (9) and (10).

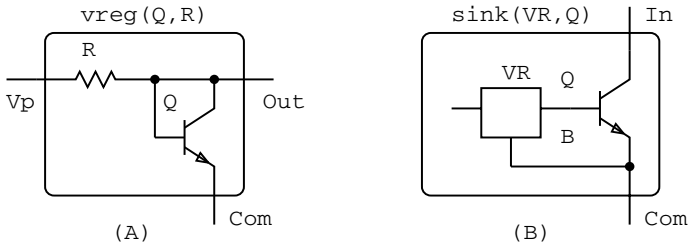


Fig. 3. V_{be} -voltage regulator (A) and Current source (B)

$$vbeReg(vreg(Q, R), Vp, Com, Out) \longrightarrow res(R, Vp, Out), [npnTr(Q, Out, Com, Out)]. \quad (9)$$

$$cSource(sink(VR, Q), In, Com) \longrightarrow vbeReg(VR, -, Com, B), [npnTr(Q, B, Com, In)]. \quad (10)$$

Figure 4 shows part of an analog IC circuit. Two transistors (q3, q5) form two current sources (sink-type) sharing one V_{be} -voltage regulator $vreg(q4, r1)$. When a goal needs to identify two current sources during parsing, the voltage regulator is used to identify one current source, and no voltage regulator remains to identify another current source. So the goal fails.

In order to solve this problem, we introduce a new mechanism which tests for the existence of a symbol but does not reduce it to an upper non-terminal symbol as follows:

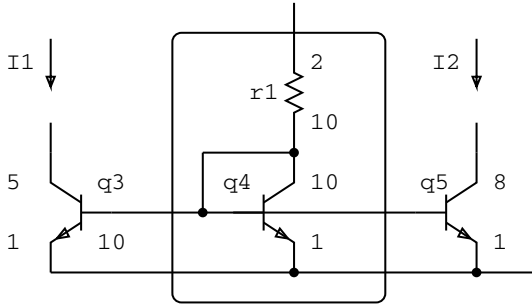


Fig. 4. Two current sources sharing one voltage regulator

$$\begin{aligned}
 cSource(sink(VR, Q), In, Com) \longrightarrow \\
 test\ vbeReg(VR, _, Com, B), \\
 [npnTr(Q, B, Com, In)].
 \end{aligned}
 \tag{11}$$

The grammar rule (11) is converted to (11)′.

$$\begin{aligned}
 subset(cSource(sink(VR, Q), In, Com), S0, S1) : - \\
 subset(vbeReg(VR, _, Com, B), S0, _), \\
 member(npnTr(Q, B, Com, In), S0, S1).
 \end{aligned}
 \tag{11}'$$

Although the current source has two different definitions (10) and (11), appropriate rules are selected during parsing by the non-deterministic mechanism of logic programming.

This extension is implemented by adding the following two lines into List 1. The “test” must be declared as a prefix-operator.

```

-----
06 conv(test [Component], member(Component, S0, _), S0, S0) :- !.
07 conv(test Component, subset(Component, S0, _), S0, S0) :- !.
-----

```

5 Extensions for Equivalent Circuits

In circuit analyses, we often rewrite object circuits into equivalent circuits, for example, with DC equivalent circuits and small signal equivalent circuits. This can be done by combining the parsing process which removes elements and a generating process which adds elements. Usually parsing programs can also work for generation by simply swapping input and output in logic programming, so we can consider an operator *invert* which exchanges input and output of the predicate *subset*. The following grammar rule defines a rewriting process *A* which first identifies the non-terminal *B* in the object circuit *S0* and removes *B* to make *S1*, then adds *C* to the circuit *S1* to make *S2*.

$$A \longrightarrow B, \text{invert } C. \tag{12}$$

$$\begin{aligned} \text{subset}(A, S0, S1) : - \quad & \text{subset}(B, S0, S1), \\ & \text{subset}(C, S2, S1). \end{aligned} \tag{12}'$$

But this method is problematic. Since we use Prolog lists to represent word-order free sentences, the same sentence has many different expressions of lists consisting permutation of words, and this causes useless backtracking.

Instead of this method, we introduce a simple mechanism for adding elements, as typified by (13).

$$A \longrightarrow B, \text{add } [C]. \tag{13}$$

$$\begin{aligned} \text{subset}(A, S0, S2) : - \quad & \text{subset}(B, S0, S1), \\ & S2 = [C|S1]. \end{aligned} \tag{13}'$$

This simple method is useful because equivalent circuits for devices usually consist of a small number of elements. Since this extension enables us to rewrite circuits during parsing, the predicate $\text{subset}(A, S0, S2)$ no longer means "A is a subset of S0". This extension is implemented by adding the following line to the List with the declaration of prefix operator "add".

08 conv(add [Component], S1=[Component|S0], S0, S1) :- !.

6 Extensions for Circuit Functions

6.1 Semantic Term in Left-Hand Side

Different from DCSG, the new circuit grammar defines not only syntactic structures but also relationships with their electrical behavior and functions using semantic terms in grammar rules [5], [6]. The semantic terms such as electrical states and voltage-current dependencies are placed in curly brackets in a grammar rule as follows.

$$A, \{F_1, F_2, \dots, F_m\} \longrightarrow B_1, B_2, \dots, B_n. \tag{14}$$

This grammar rule can be read as stating that the symbol A with meaning $\{F_1, F_2, \dots, F_m\}$ consists of the syntactic structure B_1, B_2, \dots, B_n . This rule is converted into a definite clause as follows:

$$\begin{aligned} \text{ss}(A, S0, S_n, E_0, [F_1, F_2, \dots, F_m|E_n]) : - \\ \text{ss}(B_1, S0, S_1, E_0, E_1), \\ \text{ss}(B_2, S_1, S_2, E_1, E_2), \\ \dots, \\ \text{ss}(B_n, S_{n-1}, S_n, E_{n-1}, E_n). \end{aligned} \tag{14}'$$

Since the conversion differs from that used in DCSG, we use the predicate “*ss*” instead of “*subset*”. When this rule is used in parsing, the goal $ss(A, S_0, S_n, E_0, E)$ is executed, where the variable S_0 is bound to an object set (object circuit) and the variable E_0 is replaced by the empty set. The subsets “ B_1, B_2, \dots, B_n ” are successively identified in the object set S_0 . After all of these subsets are identified, the remainder of these subsets (the complementary set) is stored in S_n . While, the semantic information of B_1 is added with E_0 and stored in E_1 , the semantic information of B_2 is added with E_1 and stored in E_2, \dots , and the semantic information of B_n is added to E_{n-1} and stored in E_n . Finally, the semantic information $\{F_1, F_2, \dots, F_m\}$, which is the meaning associated with symbol A , is added and all of the semantic information is stored in E .

6.2 Semantic Term in Right-Hand Side

Semantic terms in the right-hand side define the semantic conditions, such as the electrical states of transistors, which enable the circuit function. For example, the following rule (15) is converted into the definite clause (15)’ as follows.

$$A \longrightarrow B_1, \{C\}, B_2. \quad (15)$$

$$ss(A, S_0, S_2, E_0, E_2) : - \quad \begin{array}{l} ss(B_1, S_0, S_1, E_0, E_1), \\ member(C, E_1, -), \\ ss(B_2, S_1, S_2, E_1, E_2). \end{array} \quad (15)'$$

When the clause (15)’ is used in parsing, the semantic condition C is tested to see if the semantic information E_1 fills this condition after identifying the symbol B_1 . If it succeeds, the parsing process goes on to identify the symbol B_2 .

Appendix shows the converter for this new circuit grammar with these extensions. It was used to convert grammar rules which derive electrical dependencies from circuit topology [6].

7 Conclusions

We have developed a converter that transforms circuit grammar rules into definite clauses, which implement top-down parsing of circuits that can be described by those rules. These circuit grammar rules not only define syntactic structures for circuits, but also define relationships between structures and circuit functions as the meanings of the structures. Therefore, when a circuit is given, not only its structure but also its functions can be derived through parsing.

We see a circuit’s function to basically consist of the behaviors of the voltages and currents in that circuit. Therefore we attempted to define grammar rules to formalize dependencies of voltages and currents in circuits [6]. The electrical dependencies derived through parsing will be useful for understanding electrical behavior and troubleshooting circuits. These derived dependencies, however, only describe the surface behavior of circuits. We are currently developing grammar rules which define circuit behaviors and functions more precisely using this converter.

References

1. Pereira, F.C.N., Warren, D.H.D.: Definite Clause Grammars for Language Analysis. *Artificial Intell.* 13, 231–278 (1980)
2. Tanaka, T.: Definite Clause Set Grammars: A Formalism for Problem Solving. *J. Logic Programming* 10, 1–17 (1991)
3. Tanaka, T., Bartenstein, O.: DCSG-Converters in Yacc/Lex and Prolog. In: Proc. 12th Int. Conference on Applications of Prolog, pp. 44–49 (1999)
4. Tanaka, T.: A Logic Grammar for Circuit Analysis - Problems of Recursive Definition. In: Apolloni, et al. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 852–860. Springer, Heidelberg (2007)
5. Tanaka, T.: Circuit grammar: knowledge representation for structure and function of electronic circuits. *Int. J. Reasoning-based Intelligent Systems* 1, 56–67 (2009)
6. Tanaka, T.: Deriving Electrical Dependencies from Circuit Topologies Using Logic Grammar. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009. LNCS, vol. 5712, pp. 325–332. Springer, Heidelberg (2009)

A Converter for Circuit Grammars

```

-----
cgConv((Lhs --> Rhs), (Head :- Body)) :-
    lconv(Lhs,Head,C0,C1,E0,E1), rconv(Rhs,Body,C0,C1,E0,E1).
lconv((Component,{Es}),ss(Component,C0,C1,E0,E2),C0,C1,E0,E1):-!,
    makelist(Es,E1,E2).
lconv(Component,ss(Component,C0,C1,E0,E1),C0,C1,E0,E1):-!.
makelist((E,Es),E1,[E|E2]):-!,makelist(Es,E1,E2).
makelist(E,E1,[E|E1]).
rconv((CompA,CompB),(CA,CB),C0,C1,E0,E1):-!,
    rconv(CompA,CA,C0,C,E0,E),rconv(CompB,CB,C,C1,E,E1).
rconv((CompA;CompB),(CA;CB),C0,C1,E0,E1):-!,
    rconv(CompA,CA,C0,C1,E0,E1),rconv(CompB,CB,C0,C1,E0,E1).
rconv(not [Component],
    (not member(Component,C0,_),C1=C0,E1=E0),C0,C1,E0,E1):-!.
rconv(not Component,
    (not ss(Component,C0,_,E0,_),C1=C0,E1=E0),C0,C1,E0,E1):-!.
rconv(test [Component],
    (member(Component,C0,_),C1=C0,E1=E0),C0,C1,E0,E1):-!.
rconv(test Component,
    (ss(Component,C0,_,E0,_),C1=C0,E1=E0),C0,C1,E0,E1):-!.
rconv(quote Component,(Component,C1=C0,E1=E0),C0,C1,E0,E1):-!.
rconv(add [Component],
    (C1=[Component|C0],E1=E0),C0,C1,E0,E1):-!.
rconv({Component},
    (member(Component,E0,_),C1=C0,E1=E0),C0,C1,E0,E1):-!.
rconv([Component],
    (member(Component,C0,C1),E1=E0),C0,C1,E0,E1):-!.
rconv(Component,ss(Component,C0,C1,E0,E1),C0,C1,E0,E1):-!.
-----

```


Constructive Discursive Reasoning

Seiki Akama¹, Kazumi Nakamatsu², and Jair Minoro Abe³

¹ C-Republic, Tokyo, Japan

akama@jcom.home.ne.jp

² University of Hyogo, Himeji, Japan

nakamatus@shse.u-hyogo.ac.jp

³ Paulista University, Sao Paulo, Brazil

jairabe@uol.com.br

Abstract. Discursive reasoning is based on the nature of our ordinary discourse. Namely, several *participants* exist and have some information, beliefs, and others. Then, truth is formalized by means of the sum of opinions supplied by participants in discursive reasoning. Even if each participant has consistent information, some participant could be inconsistent with other participants. We propose a constructive discursive logic with strong negation *CDLSN* based on Nelson's constructive logic N^- as a refinement of Jaskowski's discursive logic. We give an axiomatic system and Kripke semantics with a completeness proof. We also discuss some applications in decision making.

Keywords: discursive reasoning, discursive logic, constructive logic, para-consistency.

1 Introduction

Discursive reasoning is based on the nature of our ordinary discourse. Namely, several *participants* exist and have some information, beliefs, and others. Then, truth is formalized by means of the sum of opinions supplied by participants in discursive reasoning. Even if each participant has consistent information, some participant could be inconsistent with other participants.

In discursive reasoning, we expect that $A \wedge \sim A$ does not hold while both A and $\sim A$ do. This means that the so-called *adjunction*, i.e. from $\vdash A, \vdash B$ to $\vdash A \wedge B$ is invalid. Jaskowski's *discursive logic* (or *discussive logic*) is the first formal *paraconsistent logic* for discursive reasoning, which is classified as a *non-adjunctive system*; see Jaskowski [4].

Jaskowski modeled the idea founded on modal logic S5 and reached the discursive logic in which adjunction and *modus ponens* cannot hold. In addition, Jaskowski introduced discursive implication $A \rightarrow_d B$ as $\diamond A \rightarrow B$ satisfying *modus ponens*.

The rest of this paper is as follows. Section 2 is devoted to an exposition Jaskowski's discursive logic. In section 3, we introduce constructive discursive logic with strong negation *CDLSN* with an axiomatic system. Section 4 outlines a Kripke semantics. We establish the completeness theorem. The final section gives some conclusions.

2 Jaskowski's Discursive Logic

Discursive Logic was proposed by a Polish logician S.Jaskowski [4] in 1948. It was a formal system J satisfying the conditions: (a) from two contradictory propositions, it should not be possible to deduce any proposition; (b) most of the classical theses compatible with (a) should be valid; (c) J should have an intuitive interpretation.

Such a calculus has, among others, the following intuitive properties remarked by Jaskowski himself: suppose that one desires to systematize in only one deductive system all theses defended in a discussion. In general, the participants do not confer the same meaning to some of the symbols.

One would have then as theses of a deductive system that formalize such a discussion, an assertion and its negation, so both are “true” since it has a variation in the sense given to the symbols. It is thus possible to regard discursive logic as one of the so-called *paraconsistent logics*.

Jaskowski's D_2 contains propositional formulas built from logical symbols of classical logic. In addition, possibility operator \diamond in S5 is added. Based on possibility operator, three discursive logical symbols can be defined as follows:

discursive implication: $A \rightarrow_d B =_{def} \diamond A \rightarrow B$

discursive conjunction: $A \wedge_d B =_{def} \diamond A \wedge B$

discursive equivalence: $A \leftrightarrow_d B =_{def} (A \rightarrow_d B) \wedge (B \rightarrow_d A)$

Additionally, we can define discursive negation $\neg_d A$ as $A \rightarrow_d false$. Jaskowski's original formulation of D_2 in [4] used the logical symbols: $\rightarrow_d, \leftrightarrow_d, \vee, \wedge, \neg$, and he later defined \wedge_d in [5].

The following axiomatization due to Kotas [6] has the following axioms and the rules of inference.

Axioms

- (A1) $\Box(A \rightarrow (\neg A \rightarrow B))$
- (A2) $\Box((A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C)))$
- (A3) $\Box((\neg A \rightarrow A) \rightarrow A)$
- (A4) $\Box(\Box A \rightarrow A)$
- (A5) $\Box(\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B))$
- (A6) $\Box(\neg \Box A \rightarrow \Box \neg \Box A)$

Rules of Inference

- (R1) substitution rule
- (R2) $\Box A, \Box(A \rightarrow B) / \Box B$
- (R3) $\Box A / \Box \Box A$
- (R4) $\Box A / A$
- (R5) $\Box \neg \Box \neg A / A$

There are other axiomatizations of D_2 , but we omit the details here.

3 Constructive Discursive Logic with Strong Negation

The gist of discursive logic is to use the modal logic S5 to define discursive logical connectives which can formalize a non-adjunctive system. It follows that discursive logic can be seen as a paraconsistent logic, which is not *explosive*, i.e. $\{A, \neg A\} \models B$ for any A and B , where \models is a consequence relation.

We say that a system is *trivial* iff all the formulas are provable. Therefore, paraconsistent logic is useful to formalize inconsistent but *non-trivial* systems.

Most works on discursive logic utilize classical logic and S5 as a basis, but we do not think that these are essential. For instance, an intuitionist hopes to have a discursive system in a constructive setting.

To make the idea formal, it is worth considering Nelson's constructive logic with strong negation N^- of Almkudad and Nelson [2]. In N^- , \sim denotes *strong negation* satisfying the following axioms:

- (N1) $\sim\sim A \leftrightarrow A$
- (N2) $\sim(A \wedge B) \leftrightarrow (\sim A \vee \sim B)$
- (N3) $\sim(A \vee B) \leftrightarrow (\sim A \wedge \sim B)$
- (N4) $\sim(A \rightarrow B) \leftrightarrow (A \wedge \sim B)$

and the axiomatization of the intuitionistic positive logic Int^+ with *modus ponens* (MP), i.e. $A, A \rightarrow B / B$ as the rule of inference.

If we add (N0) to N^- , we have N of Nelson [?].

$$(N0) (A \wedge \sim A) \rightarrow B$$

In N , *intuitionistic negation* \neg can be defined as follows:

$$\neg A =_{def} A \rightarrow \sim A$$

Indeed, N^- is itself a paraconsistent logic, but can also be accommodated as a version of discursive logic.

The constructive discursive logic with strong negation $CDLSN$ is an extension of N^- with discursive negation \neg_d . D_2 dispenses with discursive negation, but it plays an important role in $CDLSN$.

\neg_d is similar to \neg , but these are not equivalent. The motivation of introducing \neg_d is to interpret discursive negation as the negation used by an intuitionist in the discursive context. Unfortunately, intuitionistic negation is not a discursive negation. And we need to re-interpret it as \neg_d . Based on \neg_d , we can define \rightarrow_d and \wedge_d .

Discursive implication \rightarrow_d and discursive conjunction \wedge_d can be respectively introduced by definition as follows.

$$\begin{aligned} A \rightarrow_d B &=_{def} \neg_d A \vee B \\ A \wedge_d B &=_{def} \sim \neg_d A \wedge B \end{aligned}$$

Note that $A \rightarrow (\sim A \rightarrow B)$ is not a theorem in $CDLSN$ while $A \rightarrow (\neg_d A \rightarrow B)$ is a theorem in $CDLSN$. The axiomatization of $CDLSN$ is that of N^- with the following three axioms.

- (CDLSN1) $\neg_d A \rightarrow (A \rightarrow B)$
 (CDLSN2) $(A \rightarrow B) \rightarrow ((A \rightarrow \neg_d B) \rightarrow \neg_d A)$
 (CDLSN3) $A \rightarrow \sim \neg_d A$

Here, an explanation of these axioms may be in order. (CDLSN1) and (CDLSN2) describe basic properties of intuitionistic negation. By (CDLSN3), we show the connection of \sim and \neg_d . The intuitive interpretation of $\sim \neg_d$ is like possibility under our semantics developed below.

We use $\vdash A$ to denote that A is provable in $CDLSN$. Here, the notion of a proof is defined as usual. \neg_d is weaker than \neg . If we replace (CDLSN3) by the axiom of the form $\sim \neg_d A \leftrightarrow A$, \neg_d agrees with \neg in $CDLSN$ with (N0). This is because $\sim \neg A \leftrightarrow A$ is an axiom of N with \neg . Thus, it is not possible to identify them in our axiomatization.

Notice that \neg_d has some similarities with \neg , as the following lemma indicates.

Lemma 1. The following formulas are provable in $CDLSN$.

- (1) $\vdash A \rightarrow \neg_d \neg_d A$
 (2) $\vdash (A \rightarrow B) \rightarrow (\neg_d B \rightarrow \neg_d A)$
 (3) $\vdash (A \wedge \neg_d A) \rightarrow B$
 (4) $\vdash \neg_d (A \wedge \neg_d A)$
 (5) $\vdash (A \rightarrow \neg_d A) \rightarrow \neg_d A$

It should be, however, pointed out that the following formulas are not provable in $CDLSN$.

- $\not\vdash \sim (A \wedge \sim A)$
 $\not\vdash A \vee \sim A$
 $\not\vdash (A \rightarrow B) \rightarrow (\sim B \rightarrow \sim A)$
 $\not\vdash \neg_d \neg_d A \rightarrow A$
 $\not\vdash A \vee \neg_d A$
 $\not\vdash (\neg_d A \rightarrow A) \rightarrow A$
 $\not\vdash \sim \neg_d A \rightarrow A$
 $\not\vdash A \rightarrow_d A$

4 Kripke Semantics

It is possible to give a Kripke semantics for $CDLSN$ which is a discursive modification of that for N provided by Thomason [?] and Akama [II]. Let PV be a set of propositional variables and p be a propositional variable, and For be a set of formulas. A $CDLSN$ -model is a tuple $\langle W, w_0, R, V \rangle$, where $W \neq \emptyset$ is a set of worlds, $w_0 \in W$ satisfying $\forall w (w_0 R w)$, $R \subseteq W \times W$ is a reflexive and transitive relation, and $V : PV \times W \rightarrow \{0, 1\}$ is a partial valuation satisfying:

$$\begin{aligned} V(p, w) = 1 \text{ and } w R v &\Rightarrow V(p, v) = 1 \\ V(p, w) = 0 \text{ and } w R v &\Rightarrow V(p, v) = 0 \end{aligned}$$

for any formula $p \in PV$ and $w, v \in W$. Here, $V(p, w) = 1$ is read “ p is true at w ” and $V(p, w) = 0$ is read “ p is false at w ”, respectively. Both truth and falsity are independent statuses given by a constructive setting.

Observe that in a *CDLSN*-model, p and $\sim p$ may be true at some w since $V(p \wedge \sim p, w) \neq 0$.

We can now extend V for any formula A, B in a tandem way as follows.

$$\begin{aligned}
V(\sim A, w) &= 1 && \text{iff } V(A, w) = 0. \\
V(A \wedge B, w) &= 1 && \text{iff } V(A, w) = 1 \text{ and } V(B, w) = 1. \\
V(A \vee B, w) &= 1 && \text{iff } V(A, w) = 1 \text{ or } V(B, w) = 1 \\
V(A \rightarrow B, w) &= 1 && \text{iff } \forall v(wRv \text{ and } V(A, v) = 1 \Rightarrow V(B, v) = 1) \\
V(\neg_d A, w) &= 1 && \text{iff } \forall v(wRv \Rightarrow V(A, v) \neq 1) \\
V(\sim A, w) &= 0 && \text{iff } V(A, w) = 1 \\
V(A \wedge B, w) &= 0 && \text{iff } V(A, w) = 0 \text{ or } V(B, w) = 0 \\
V(A \vee B, w) &= 0 && \text{iff } V(A, w) = 0 \text{ and } V(B, w) = 0 \\
V(A \rightarrow B, w) &= 0 && \text{iff } V(A, w) = 1 \text{ and } V(B, w) = 0 \\
V(\neg_d A, w) &= 0 && \text{iff } \exists v(wRv \text{ and } V(A, v) = 1)
\end{aligned}$$

Additionally, we obtain the following

$$V(A \wedge \sim A, w) = 1 \text{ for some } A \text{ and some } w.$$

Here, observe that truth and falsity conditions for $\sim \neg_d A$ are implicit in the above clauses from the equivalences such that $V(\sim \neg_d A, w) = 1$ iff $V(\neg_d A, w) = 0$, and $V(\sim \neg_d A, w) = 0$ iff $V(\neg_d A, w) = 1$. We say that A is *valid*, written $\models A$, iff $V(A, w_0) = 1$ in all *CDLSN*-models.

Lemma 2. The following hold for any formula A which is not of the form $\neg_d B$, and worlds $w, v \in W$.

$$\begin{aligned}
V(A, w) = 1 \text{ and } wRv &\Rightarrow V(A, v) = 1, \\
V(A, w) = 0 \text{ and } wRv &\Rightarrow V(A, v) = 0.
\end{aligned}$$

Lemma 2 does not hold for $\neg_d A$. This is intuitive because $\neg_d A$ is paraconsistent negation. Next, we present a soundness theorem.

Theorem 3 (soundness). $\vdash A \Rightarrow \models A$.

Now, we give a completeness proof. We say that a set of formulas Γ^* is a *maximal non-trivial discursive theory* iff (1) Γ^* is a theory, (2) Γ^* is *non-trivial*, i.e. $\Gamma^* \not\vdash B$ for some B , (3) Γ^* is *maximal*, i.e. $A \in \Gamma^*$ or $A \notin \Gamma^*$, (4) Γ^* is *discursive*, i.e. $\neg_d A \notin \Gamma^*$ iff $\sim \neg_d A \in \Gamma^*$. Here, discursiveness is needed to capture the property of discursive negation.

Lemma 4. For any $\Gamma \in \Gamma^*$ and every formula A, B , the following hold:

- (1) $A \wedge B \in \Gamma$ iff $A \in \Gamma$ and $B \in \Gamma$
- (2) $A \vee B \in \Gamma$ iff $A \in \Gamma$ or $B \in \Gamma$
- (3) $A \rightarrow B \in \Gamma$ iff $\forall \Delta \in \Gamma^* (\Gamma \subseteq \Delta \text{ and } A \in \Delta \Rightarrow B \in \Delta)$
- (4) $\neg_d A \in \Gamma$ iff $\forall \Delta \in \Gamma^* (\Gamma \subseteq \Delta \Rightarrow A \notin \Delta)$
- (5) $\sim (A \wedge B) \in \Gamma$ iff $\sim A \in \Gamma$ or $\sim B \in \Gamma$
- (6) $\sim (A \vee B) \in \Gamma$ iff $\sim A \in \Gamma$ and $\sim B \in \Gamma$
- (7) $\sim (A \rightarrow B) \in \Gamma$ iff $A \in \Gamma$ and $\sim B \in \Gamma$
- (8) $\sim \sim A \in \Gamma$ iff $A \in \Gamma$
- (9) $\sim \neg_d A \in \Gamma$ iff $\exists \Delta \in \Gamma^* (\Gamma \subseteq \Delta \text{ and } A \in \Delta)$.

Based on the maximal non-trivial discursive theory, we can define a canonical model $(\Gamma^*, \Gamma, \subseteq, V)$ such that $\Gamma \in \Gamma^*$, where $\forall \Delta \in \Gamma^* (\Gamma \subseteq \Delta)$, and satisfying the conditions that $V(p, \Gamma) = 1$ iff $p \in \Gamma$ and that $V(p, \Gamma) = 0$ iff $\sim p \in \Gamma$.

Next lemma is a truth lemma.

Lemma 5 (truth lemma). For any $\Gamma \in \Gamma^*$ and any A , we have the following:

$$\begin{aligned} V(A, \Gamma) &= 1 \text{ iff } A \in \Gamma \\ V(A, \Gamma) &= 0 \text{ iff } \sim A \in \Gamma \end{aligned}$$

Finally, we can state the completeness of *CDLSN* as follows:

Theorem 6 (completeness). $\models A \Rightarrow \vdash A$.

5 Conclusions

We proposed a constructive version of discursive logic with an axiomatization and semantics. We set up it as a natural extension of Almukdad and Nelson's N^- [2] with \neg_d . Alternatively, it can be interpreted as a fragment of N^- with \neg with some restrictions. We believe that this system seems to be new in the literature.

The advantage of the proposed system is constructively intuitive and it dispenses with modal operators to define discursive connectives. However, it may be possible to introduce other types of discursive connectives.

There are several applications of *CDLSN* to be worked out. First, *CDLSN* can deal with inconsistency in an intelligent system since it is a paraconsistent logic. Despite the presence of inconsistency in a system, it is not trivial, that is, inconsistency can be localized.

Second, *CDLSN* can be modified to formalize *non-monotonic reasoning*, which is a form of common-sense reasoning. Although new information can invalidate old conclusions in a discourse, some old conclusions cannot be modified. Such a situation, paraconsistency plays an important role.

Third, *CDLSN* can serve as a basis for decision theory. It is interesting to study decision theory in a paraconsistent setting. In addition, discursive reasoning is a natural for multi-agent systems.

References

1. Akama, S.: Constructive predicate logic with strong negation and model theory. *Notre Dame Journal of Formal Logic* 29, 18–27 (1988)
2. Almukdad, A., Nelson, D.: Constructible falsity and inexact predicates. *Journal of Symbolic Logic* 49, 231–233 (1984)
3. da Costa, N.C.A., Dubikajtis, L.: On Jaskowski's discursive logic. In: Arruda, A.I., da Costa, N.C.A., Chuaqui, R. (eds.) *Non-Classical Logics, Model Theory and Computability*, pp. 37–56. North-Holland, Amsterdam (1977)

4. Jaskowski, S.: Propositional calculus for contradictory deductive systems. *Studia Societatis Scientiarum Torunensis, Sectio A* 1, 55–77 (1948) (in Polish)
5. Jaskowski, S.: On the discursive conjunction in the propositional calculus for inconsistent deductive systems. *Studia Societatis Scientiarum Torunensis, Sectio A* 8, 171–172 (1949) (in Polish)
6. Kotas, J.: The axiomatization of S. Jaskowski's discursive logic. *Studia Logica* 33, 195–200 (1974)

Formal Concept Analysis of Medical Incident Reports

Takahiro Baba¹, Lucing Liu¹, and Sachio Hirokawa²

¹ Graduate School of Information Science and Electrical Engineering,

² Research Institute for Information Technology

Kyushu University, Fukuoka 812-8581, Japan

hirokawa@cc.kyushu-u.c.jp

<http://matu.cc.kyushu-u.ac.jp>

Abstract. It is known that a lot of incidents has happened ahead of a serious accident. Such experiences have been collected in medical sites as incident reports. The text mining is expected as a method that discovers the factors of incidents and the improvement of the situation. This paper proposes a method to analyse the co-occurrence relation of the words that appear in the medical incident reports using concept lattice.

1 Introduction

Organizational efforts to collect incident reports are being made in hospitals, factories, traffic controls and network security where a small mistake would cause a terrible damage to the society or to individuals. Incident reports contain detailed situations which are very close to accidents. The incident reports contain not only the situation and the reason of the incident, but also how it was prevented to occur an accident. So, it is worthwhile to analyse incident reports to discover some hits to prevent accidents.

In the field of medicine, even a small mistake may cause a death of an client. Many hospitals are collecting these incident reports systematically. Indeed, the ministry of health and labor of Japanese government published the announcements #0330008 of medical policy, and #0330010 of medicine and food, to strengthen the activity to file the incident reports. However, analysing the reports is time consuming hard work for doctors and nurses who are working the actual situation. They do not have separate time to consider the report deeply. Nowadays, the reports are being kept as digital texts from the beginning. Hence, the number of reports are increasing [6]. There are strong needs in introducing ICT to support analysing reports.

In [8], keyword extraction methods are applied to incident reports to analyse particular reports that are specified by a metadata. In [7], the method of SOM(Self Organizing Map) are applied to discover similar reports. However, the method of [8] is restricted to metadata. The method of [7] does not explain the meaning of documents with characteristic words. Neither of the methods provide interaction with the user. There are many researches, e.g. [2,3], for clustering

documents. However, the most of these approach adapts vector space model to represent document. The results vary according to the definition of similarity of the vectors and to the threshold. There is no general rule to justify how we formulate.

In this paper, we demonstrate that the theory of FCA (formal concept analysis) is applicable to the analysis of incident reports.

2 Incident Reports

An incident report describes an experience of doctors and of nurses, where they faced a dangerous situation that was very close to an accident. The reports contain meta data such as date, time, place, type of an incident, the name of person who reports. They contain free texts that explain the situation. Table 1 is a typical example. Incident reports are similar to accident reports in nature. Both describe the situation of the accident/incident and the considerable factors of the accident/incident. The most important difference between incident reports and accident reports is that the former may contain some reason that prevented an accident. There are many lessons that we can learn from these incidents.

It is known, as the Heirig law, that there are much large number of incidents behind an accident. So, we can expect in collecting large number of incident reports than that of accident reports, if we do our efforts.

Table 1. An Example of Incident Report

date	yyyy/mm/dd
location	Medical Examination Room
contents	The syringe drivers for vein injection and for epidural injection were in the same tray. I almost gave the wrong injection by mistake.

This paper analyses the 47 incident reports that are described and analysed in the book [5]. The reports contain not only free texts but also metadata that specify the causes of the incident and possible improvement factors. The table [2] shows these metadata.

3 Related Keywords for Metadata and Their Weight

The analysis is based on the term*document matrix. Firstly, the set of the documents that contain a metadata are searched using the matrix. Then, the top 5 words of highest weight are retrieved using the matrix conversely as related words for the metadata. Table [3] shows the related words for each metadata. For example, the characteristic words for the metadata x:C, which indicates the

Table 2. Metadata for Cause and Improvement

x:A	Patient and their environment
x:B	Stuff and their experience
x:C	Organization
x:D	Other
y:1	Communication between Stuff
y:2	Design of Commodity
y:3	Design of Equipment and Operation
y:4	Maintenance
y:5	Inspection
y:6	Nursing Procedure
y:7	Paper Work
y:8	Information Sharing
y:9	Physical Environment
y:10	Workplace
y:11	Personnel
y:12	Management of Equipment and Medicine
y:13	Other Management Issue in Hospital
y:14	Educational Activity
y:15	Organizational Culture

Table 3. Top 5 Related Words for Metadata

freq	cause/improvement	word(weight)
32	x:C (organization)	drip(8.375) injection(8.171) bottle(8.033) direction(7.761) patient(7.553)
21	x:B (stuff)	injection(7.125) inject(6.876) confirm(6.578) ample(6.418) new stuff(6.217)
5	x:A (patient)	fit(3.834) spill(2.558) slip(2.558) return(2.558) reach(2.558)
12	y:6 (nurse)	nurse(5.101) prepare(4.558) tube(3.767) merge(3.546) mistake(3.522)
10	y:8 (information sharing)	told(6.057) bottle(4.643) new stuff(4.620) drip(4.376) intramuscular injection(4.117)
8	y:14 (education)	new stuff(6.217) ample(4.934) location(4.164) told(4.117) intramuscular injection(4.117)
5	y:12 (eq. management)	location(4.164) same(3.577) inject(3.268) connect(2.477) color sringe(2.477)
3	y:7 (paper work)	sharp(3.784) infection(2.249) without notice(2.249) break(2.249) duty(2.249)
2	y:1 (comm. stuff)	cards(2.477) "ha"(2.477) lubrine(2.477) se(2.477) check(2.477)

cause of incidents in the organization, are drip, injection, bottle, direction and patient. We might imagine a situation where a doctor gives a direction to a nurse to make an injection to a patient, or to connect a bottle to "drip". Without these

actual words, we cannot think of the real situation from the metadata "x:C". In this sense, these related word is useful. However, we do not know if the 5 words is enough or not. There is no criteria to determine the number of related words. There is no justification to determine the threshold of the weight.

4 Analysis System Using Formal Concept Lattice

4.1 Formal Concept Lattice

Given a set of documents and the set of keywords that appear in the documents, the formal concept lattice represents the relationship of documents and keywords and further, represents the hierarchical structure of keywords.

In the theory of concept lattice [14], a tuple (G, M, I) of a finite set G of objects, a finite set M of attributes and a relation $I \subseteq G \times M$ is called a context. When an object g has an attribute m , we denote $(g, m) \in I$ or gIm . Given an object g , the set of all attributes of g is represented as $nbr(g)$, i.e., $nbr(g) = \{m \in M \mid (g, m) \in I\}$. For a set of objects $X \subseteq G$, by $attr(X)$, we denote the set of attributes that are common to all objects g in X , i.e., $attr(X) = \bigcap_{g \in X} nbr(g)$. Given an attribute $m \in M$, the set $nbr(m)$ of objects that has the property is defined by $nbr(m) = \{g \in G \mid (g, m) \in I\}$. For a set of attribute $J \subseteq M$, by $obj(J)$, we denote the set of all objects that have all the attributes in J , i.e., $obj(J) = \bigcap_{m \in J} nbr(m)$. A pair (A, B) of a set $A \subseteq G$ of objects and a set $B \subseteq M$ of attributes are said to be a concept iff $A = obj(B)$ and $B = attr(A)$. An order relation \prec for two concepts $C_1 = (A_1, B_1)$ and $C_2 = (A_2, B_2)$ are defined by $C_1 \prec C_2 \iff (A_1 \subseteq A_2) \wedge (B_1 \supseteq B_2)$. When $C_1 \prec C_2$, C_2 is said to be a lower concept of C_1 and C_1 is said to be an upper concept of C_2 . When $C_1 \prec C_2$ and there is no concept E except C_1 or C_2 such that $C_1 \prec E \prec C_2$, C_1 is said to be a direct lower concept of C_2 and C_2 is said to be a direct upper concept of C_1 . When (A, B) is a concept, the set of objects A and the set of attributes B characterize each other. The direct lower concepts represent a clustering of the objects A according to attributes. The direct upper concepts represent a clustering of the attribute B according to objects.

Table 4.1 shows a context of reports D_1, D_2, D_3, D_4 , and attributes $x : a, x : b, x : c, y : d, y : e$, where "a", "b" etc represent words and "x", "y" represent the category of the words. Fig 2 is the concept lattice of this matrix. We can see that the set of reports $\{D_3, D_4\}$ is characterized by the attribute $\{y : e\}$ and is classified into the two direct lower concepts D_3 and D_4 according to the attributes $x : a$ and $x : b$.

4.2 Clustering with Words and Clustering with Attributes

If we consider the reports as objects and the words that appear in the reports as attributes, we can construct a concept lattice by which we can analyse the relation of reports and keywords and the relation between keywords. However, the analysis should depend on the purpose of the analyser. The viepoint of the analysis is crucial. In the case of the incident report of Table 3, the metadata

	x:a	x:b	y:c	y:d	y:e
D_1			√	√	
D_2			√		
D_3	√				√
D_4		√			√

Fig. 1. Object*Attribute Matrix

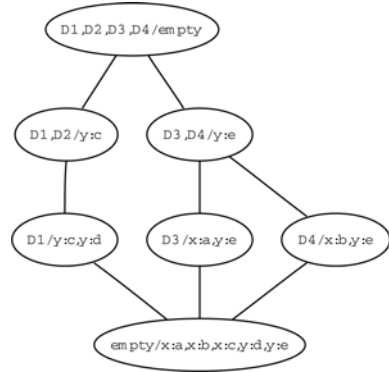


Fig. 2. Concept Lattice for Table 1

$x:A,x:B,x:C,y:1,y:2,\dots$ and $y:15$ represent the view points. The metadata $x:A, x:B$ and $x:C$ classifies the cause of the incidents. The metadata $y:1,\dots, y:15$ represent the improvement points to prevent the incidents.

A concept lattice is constructed from a context matrix that represents the relation of the objects and the attributes. A novelty of our approach is that we do not consider the documents as the objects. The attributes of the incident report are chosen as the objects. Thus, the row of the context matrix consists of attributes and the columns consists of terms. In other words, we use the term*attribute matrix as the context for the concept lattice.

Let X be the term*document matrix where $X[i, j] = 1$ iff the word or the attribute w_i occurs in a document d_j . To construct the term*attribute matrix, we have to determine the set of words that correspond to an attribute $x : P$. There are two way to formulate the set. The standard formulation (disjunctive construction) is to choose all the words in the documents that contains the attribute $x : P$. Another formulation (conjunctive construction) is to choose the words that appear in all documents that contain the attribute $x : P$.

Given a term*document matrix X , we construct attribute*term matrices Y (disjunctive construction) and Z (conjunctive construction) as follows. Here, the document d_k ranges under the condition $x : P_i \in d_k$ and $y : P_j \in d_k$. In this paper, we adapt the disjunctive construction.

$$Y[x : P_i, y : Q_j] = \Pi_{d_k} \{X[x : P_i, d_k] * X[y : Q_j, d_k]\}$$

$$Z[x : P_i, y : Q_j] = \Sigma_{d_k} \{X[x : P_i, d_k] * X[y : Q_j, d_k]\}$$

We constructed a system that accepts words and attributes as input, and that outputs the concept. The user can combine multiple words to form a "AND query" and "OR query". The system searches the documents that satisfy the query q , and then obtain the set $obj(q)$ of attributes that co-occur with the query q . Be Aware that the objects of the concept are metadata $x:A,x:B,x:C,y:1,y:2,\dots,$ and $y:14$ and that the attributes are keywords. The system retrieve the set $attr(obj(q))$ keywords that co-occur with all of the metadata in $obj(q)$. The pair $(obj(q), attr(obj(q)))$ is the concept for the query.

A user can use the system iteratively and interactively. Once a user send an input, the system displays related keywords and attributes. The user only has to click one of the words for further analysis. The user easily can reach an upper concept and a lower concept. An adjacent concept is shown with a keyword which is used as ankar text. So, the user can obtain the corresponding concept simply by clicking the keyword.

5 Case Studies

5.1 Concepts for Cause

Table 4 shows the concepts for each metadata. All words are listed in a line if they belong to the concept. This is the most crucial distinction compared to Table 3 of basic analysis where only top 5 words are chosen according to their

Table 4. The Concepts for Metadata

freq	metadata	Conjunctive Construction
32	x:C	A B C 1 11 12 14 2 5 6 7 8 one "5 minutes" IVH give after ...(378)
21	x:B	B C 1 11 12 14 2 6 7 8 one "5 minutes" IVH saying said ...(246)
5	x:A	A C 6 7 give after in convulsion spill ...(74)
12	y:6	A B C 12 6 8 one "5 minutes" IVH give always ...(178)
10	y:8	B C 12 14 5 6 8 "5 minutes" saying always say told ...(127)
8	y:14	B C 12 14 8 told understood up atoniun alarm ...(77)
5	y:12	B C 12 14 2 6 8 IVH saying told do connect ...(75)
3	y:7	A B C 7 after said in please moreover ...(69)
2	y:1	B C 1 not after-operation before-operation card Ha ...(29)
freq	metadata	Disjunctive Construction
32	x:C	C:organizational
21	x:B	B:stuff C:organizational
5	x:A	A:patient C:organizational
12	y:6	C:organizational 6:nurse
10	y:8	C:organizational 8:message
8	y:14	B:stuff C:organizational 14:education
5	y:12	B:stuff C:organizational 12:equipment management
3	y:7	C:organizational 7:official procedure
2	y:1	B:stuff C:organizational 1:"stuff communication" direction

weight. For example, the metadata m6 that corresponds to the nurse procedure is completely characterized by the metadata x:B,x:C,y:12,y:14,y:2,y:6,y:8 and other 75 words according to the conjunctive construction. On the other hand, no words appear in the concept except for y:1 in the disjunctive construction. In other words, the disjunctive construction is too restrictive that the metadata for cause and improvement cannot be characterized by words. However, the relation of metadata can be seen in the disjunctive construction. For example, $\{x : B, x : C\}$ and $\{x : A, x : C\}$ form concepts, but $\{x : A, x : B\}$ does not. This implies that no incident is observed with respect to the patient and the stuff. They are observed only when the relation of stuff is concerned.

Each case is worthwhile to analyse. The concept for metadata mB, that represents some cause of incident in stuff, contains mC that represents organizational cause. This can be interpreted that we should pay much attention to the relationship of stuff than the individual problems of stuff to prevent incidents. Organizational improvement is expected to solve the individual problems.

The concept for y:14, that represents improvement by education, consists of x:B, x:C and y:14. This implies that no incidents that could be prevented are observed unless organizational problems are concerned.

5.2 Analysis by Cause and Improvement

Given an attribute of cause or improvement, we can obtain the related words that characterize the attribute. Thus, we can analyse depending on particular target of analysis. However, these detailed analysis do not give a birds-eye view of the whole reports. We cannot grasp the whole picture of the reports. The visualization of the lattice is useful for this purpose. Fig 3 is the concept lattice

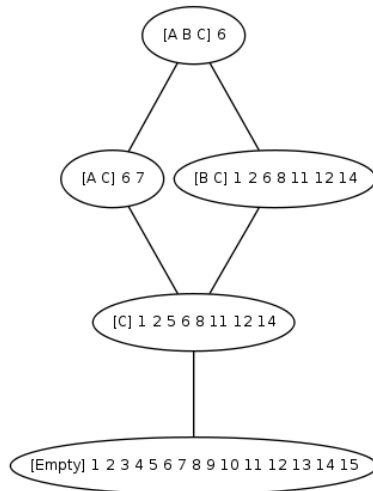


Fig. 3. Concept Lattice for Cause*Improvement

for cause*improvement. It is worthwhile to notice that the lattice is very small and that there are only 5 concepts. It is not because the number of the document is 47 and is very small. It is because there are only 3 attributes $x:A, x:B$ and $x:C$ as the attributes to describe the cause. The attributes are determined according to the purpose of the analysis. Therefore, the number of the attributes is considered to small enough compared to the number of reports. Since the number of concepts are bound by the exponential of the number of attributes, we can expect that the concept lattice is small enough to draw.

6 Conclusion and Further Work

This paper proposed a method to analyse medical incident reports using concept lattice. The metadata that specify the cause of incidents or the possible improvements are considered as objects and the words are considered as attributes. For each metadata, The set of words that characterizes the metadata are obtained and analysed.

The number of incident reports that analysed in this paper is 47. They may be too small for general evaluation. However, each sample was selected from expert point of view in the book [5]. They are valuable examples worthwhile to analyse in detail. Nonetheless, we need quantitative evaluation of the proposed method.

References

1. Carpineto, C., Romano, G.: *Concept Data Analysis Theory and Application*. John Wiley and Sons, Chichester (2004)
2. Cheng, Y., Church, G.M.: Biclustering of expression data. In: *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103 (2000)
3. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-Theoretic Co-clustering. In: *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 89–98 (2003)
4. Ganter, B., Wille, R., Franzke, C.: *Formal Concept BAnalysis:Mathematical Foundation*. Springer, Heidelberg (1999)
5. Kawamura, H.: Willing to write incident reports – lessons learned to prevent accidents, Igaku-Shoin (2000) (in Japanese)
6. Kaneko, K.: Reports on activities in Kyushu University Hospital for Medical Safety Management Database. In: *12th Web Intelligence Workshop (2008)* (in Japanese)
7. Kawanaka, H., Otani, Y., Yamamoto, K., Shinogi, T., Tsuruoka, S.: Tendency discovery from incident report map generated by self organizing map and its development. In: *IEEE International Conference on Systems, Man and Cybernetics*, pp. 2016–2021 (2007)
8. Okabe, T., Yoshikawa, T., Furuhashi, T.: A Proposal of Analysis System for Medical Incident Reports using Metadata and Co-occurrence Information. *Journal of Knowledge, Information and Fuzzy Society* 18(5), 689–700 (2006) (in Japanese)

Compression of Multispectral Images with Inverse Pyramid Decomposition

Roumen Kountchev¹ and Kazumi Nakamatsu²

¹ Technical University – Sofia, Department of Radio Communications and Video Technologies,
Boul. Kl. Ohridsky 8, Sofia 1000, Bulgaria
rkountch@tu-sofia.bg

² School of H.S.E. University of Hyogo, 1-1-12 Shinzaike Himeji, 670-0092 Japan
nakamatu@shse.u-hyogo.ac

Abstract. In the paper is presented a new method for compression of multispectral images, based on the Inverse Difference Pyramid decomposition. The method is applicable for any number of multispectral images of same object. The processing is performed as follows. First, the histograms of the multispectral images are calculated and compared. The image, whose histogram is most similar with these of the remaining ones, is chosen to be a reference one. The image decomposition starts with the reference image, which is processed with some kind of orthogonal transform, using a limited number of transform coefficients only. With the so obtained coefficients values is calculated the coarse approximation of the processed image. The IDP decomposition then branches out into several directions, corresponding to the number of multispectral images. The first approximation for all multispectral images is that of the reference image. Each branch is developed individually, using the same approximation. In result of this processing is obtained high compression and very good visual quality of the restored images. This approach gives better results than these, obtained with methods, based on the JPEG and JPEG 2000 standards.

Keywords: Multi-spectral images, Modified inverse pyramid decomposition.

1 Introduction

The contemporary research in different application areas sets the task of the efficient archiving of multispectral images. For this, in most cases is necessary to process several images of the same object(s). Multispectral images are characterized by very high spatial, spectral, and radiometric resolution and, hence, by ever-increasing demands of communication and storage resources. Such demands often exceed the system capacity like, for example, in the downlink from satellite to Earth stations, where the channel bandwidth is often much inferior to the intrinsic data rate of the images, some of which must be discarded altogether. In this situation the high-fidelity image compression is a very appealing alternative. As a matter of fact, there has been intense research activity on this topic [1–5], focusing, particularly, on transform-coding techniques, due to their good performance and limited computational complexity. Linear transform coding, however, does not take into account the nonlinear dependences

existing among different bands, due to the fact that multiple land covers, each with its own interband statistics, are present in a single image. Based on this observation, a class-based coder was proposed in [1] that addresses the problem of interband dependences by segmenting the image into several classes, corresponding as much as possible to the different land covers of the scene. As a consequence, within each class, pixels share the same statistics and exhibit only linear interband dependences, which can be efficiently exploited by conventional transform coding.

Satellite-borne sensors have ever higher spatial, spectral and radiometric resolution. With this huge amount of information comes the problem of dealing with large volumes of data. The most critical phase is on-board the satellite, where acquired data easily exceed the capacity of the downlink transmission channel, and often large parts of images must be simply discarded, but similar issues arise in the ground segment, where image archival and dissemination are seriously undermined by the sheer amount of data to be managed. The reasonable approach is to resort to data compression, which allows reducing the data volume by one and even two orders of magnitude without serious effects on the image quality and on their diagnostic value for subsequent automatic processing. To this end, however, is not possible to use the general purpose techniques as they do not exploit the peculiar features of multispectral remote-sensing images, which is why several ad hoc coding schemes have been proposed in recent years.

The transform coding is one of the most popular approaches for several reasons. First, transform coding techniques are well established and deeply understood; they provide excellent performances in the compression of images, video and other sources, have a reasonable complexity and besides, are at the core of the famous standards JPEG and JPEG2000, implemented in widely used and easily available coders. The common approach for coding multispectral images [8] is to use some decorrelating transforms along the spectral dimension followed by JPEG2000 on the transform bands with a suitable rate allocation among the bands.

Less attention has been devoted to techniques based on vector quantization (VQ) because, despite its theoretical optimality, VQ is too computationally demanding to be of any practical use. Nonetheless, when dealing with multiband images, VQ is a natural candidate, because the elementary semantic unit in such images is the spectral response vector (or spectrum, for short) which collects the image intensities for a given location at all spectral bands. The values of a spectrum at different bands are not simply correlated but strongly dependent, because they are completely determined (but for the noise) by the land covers of the imaged cell. This observation has motivated the search for constrained VQ techniques [2], which are suboptimal but simpler than full-search VQ, and show promising performances.

Multispectral images require large amounts of storage space, and therefore a lot of attention has recently been focused to compress these images. Multispectral images include both spatial and spectral redundancies. Usually we can use vector quantization, prediction and transform coding to reduce redundancies. For example, hybrids transform/vector quantization (VQ) coding scheme is proposed [6, 9]. Instead, Karhunen-Loeve transform (KLT) is used to reduce the spectral redundancies, followed by a two-dimensional (2D) discrete cosine transform (DCT) to reduce the spatial redundancies [2]. A quad-tree technique for determining the transform block size and the quantizer for encoding the transform coefficients was applied across KLT-DCT

method [7]. In [10] and [11] the researchers used a wavelet transform (WT) to reduce the spatial redundancies and KLT to reduce the spectral redundancies, and then encoded using the 3-dimensional (3D) SPIHT algorithm [11].

The state-of-the-art analysis shows that despite of the vast investigations and various techniques used for the efficient compression of multispectral images, a recognized general method able to solve the main problems is still not created.

In this paper is offered one new method for compression of multispectral images, based on the modified Inverse Difference Pyramid (IDP) Decomposition. The paper is arranged as follows: Section 2 provides brief information about the IDP decomposition and presents the modified version for processing of multispectral images; in Section 3 is given the way for the reference image selection; Section 4 presents some experimental results and comparison with the famous standards JPEG and JPEG2000 and Section 4 the Conclusions of this work are summarized.

2 Basic Principles of the Modified Inverse Difference Pyramid (IDP) Decomposition

The Inverse Difference Pyramid (IDP) decomposition [12] was selected for the compression of multispectral images because it offers significant flexibility of the image processing and together with this, in result is obtained very efficient image representation. The IDP decomposition is presented in brief, as follows. The initial approach is that the matrix $[X]$ of the original image is divided into sub-images of size $2^n \times 2^n$ and each is processed with some kind of two-dimensional (2D) orthogonal transform using a limited number of spectrum coefficients only. The values of the coefficients, calculated in result of the transform, build the first pyramid level. The sub-image is then restored (using the values of these coefficients only) with the same inverse orthogonal transform and the result is subtracted pixel by pixel from the original. The difference sub-image with elements $e_p(i, k)$ in IDP level p is defined as:

$$e_p(i, k) = \begin{cases} x(i, k) - \tilde{x}(i, k) & \text{for } p = 0; \\ e_{p-1}(i, k) - \tilde{e}_{p-1}(i, k) & \text{for } p = 1, 2, \dots, P, \end{cases} \quad (1)$$

where $x(i, k)$ is the pixel (i, k) in the sub-image of size $2^n \times 2^n$ of the input image $[X]$; $\tilde{x}(i, k)$ and $\tilde{e}_{p-1}(i, k)$ are correspondingly the pixels of the recovered input and difference sub-images in the IDP level $p=1, 2, \dots, P$, obtained with inverse orthogonal transform on the selected coefficients only, P - the total number of IDP levels. The so calculated difference sub-image is divided into 4 sub-images of size $2^{n-1} \times 2^{n-1}$. Each sub-image is processed with the 2D orthogonal transform again; the values of the transform coefficients build the second pyramid level. The image is then restored again and the corresponding difference image is calculated. The process continues in similar way with the next pyramid levels.

The approximation models of the input or difference image in the level p is represented by the relations:

$$\begin{aligned} \tilde{x}(i, k) / \tilde{e}_{p-1}(i, k) &= \text{IOT}[y_p(u, v)]; \\ y_p(u, v) &= \text{OT}[x(i, k) / e_{p-1}(i, k)], \end{aligned} \tag{2}$$

where $\text{OT}[\bullet]$ is the operator, representing the “truncated” direct two-dimensional orthogonal transform applied on the input block of size $2^n \times 2^n$, or on the difference sub-image of size $2^{n-p} \times 2^{n-p}$ from the pyramid level $p = 1, 2, \dots, P$; $\text{IOT}[\bullet]$ is the operator for the inverse orthogonal transform of the spectrum coefficients, $y_p(u, v)$ from the level p of the “truncated” transform $2^{n-p} \times 2^{n-p}$, obtained in result of the transformation of every $1/4$ part of the difference sub-image, $e_{p-1}(i, k)$.

The components of all IDP levels are used for the recursive calculation of the sub-image $x(i, k)$:

$$x(i, k) = \tilde{x}(i, k) + \sum_{s=1}^P \tilde{e}_{s-1}(i, k) \tag{3}$$

The IDP modification proposed for the adaptive processing of multispectral images is based on the assumption that the basic objects in all images are the same, i.e. – they have similar structure and correspondingly – their spatial spectrum is the same. This characteristic permits to use same set of transform coefficients for all layers and for all multispectral images. All images in the processed set are similar and on the grounds of this one of them is used as a reference in the lowest decomposition layer. The restored image obtained as first approximation for the reference one is then used to build the next decomposition layers of the remaining multispectral images.

The mathematical representation of the Modified IDP decomposition follows.

The matrix $[B_n]$ of the spectral image $n = 1, 2, \dots, N$ of the multispectral image is initially divided into sub-images of size $r \times r$ pixels (usually, 8×8 or 4×4 pixels each). The total number of spectral images is N , and one of them is used as a reference R , represented by the matrix $[B_R]$ (the reference image selection is given in Section 3 of the paper). For the lower decomposition layer $p = 0$ is calculated the transform $[S_0^R]$ of the reference image $[B_R]$ using direct orthogonal transform:

$$[S_0^R] = [T_0][B_R][T_0] \tag{4}$$

where $[T_0]$ is the matrix for direct orthogonal transform of size $r \times r$ (same as the sub-image size).

$$[\hat{S}_0^R] = [m_0(u, v) s_0^R(u, v)] \tag{5}$$

where $m_0(u, v)$ is an element of the matrix-mask $[M_0]$ for the layer $p = 0$, which defines the retained coefficients used for the approximation:

$$m_0(u, v) = \begin{cases} 1, & \text{if } s_0^R(u, v) \text{ - retained coefficient,} \\ 0 & \text{in all other cases,} \end{cases} \tag{6}$$

The approximated reference image $[\hat{B}_0^R]$ is calculated, using inverse orthogonal transform in correspondence with the relation:

$$[\hat{B}_0^R] = [T_0]^t [\hat{S}_0^R] [T_0]^t \quad (7)$$

where $[T_0]^t = [T_0]^{-1}$ is the matrix of the inverse orthogonal transform, of size $r \times r$.

The difference matrix is calculated in accordance with the relation:

$$[E_{0,R}] = [B_R] - [\hat{B}_0^R] \quad (8)$$

The difference matrix is then divided into 4 sub-matrices:

$$[E_{0,R}] = \begin{bmatrix} [E_{0,R}^1] & [E_{0,R}^2] \\ [E_{0,R}^3] & [E_{0,R}^4] \end{bmatrix} \quad (9)$$

where $[E_{0,R}^i]$ for $i = 1, 2, 3, 4$ are sub-matrices of size $(r/2) \times (r/2)$ each.

For the next decomposition layer $p = 1$ is calculated the transform $[S_{1,0}^R]$ of the sub-matrix i of the difference image $[E_{0,R}^i]$, using direct orthogonal transform:

$$[S_{1,i}^R] = [T_1][E_{0,R}^i][T_1] \text{ for } i = 1, 2, 3, 4, \quad (10)$$

where $[T_1]$ is the matrix of the direct orthogonal transform of size $(r/2) \times (r/2)$.

The approximating i^{th} transform is calculated:

$$[\hat{S}_{1,i}^R] = [m_1(u, v) s_{1,i}^R(u, v)], \quad (11)$$

where $m_1(u, v)$ is an element of the matrix-mask $[M_1]$ for the layer $p = 1$, which defines the retained spectrum coefficients (in correspondence with Eq. 6):

The retained coefficients in the second decomposition layer are usually different from these in the initial layer and their number depends on the restored image quality, needed for the application.

In decomposition layer $p = 1$ is calculated the difference for every sub-image $[B_n]$:

$$[E_{0,n}] = [B_n] - [\hat{B}_0^R] \text{ for } n \neq R. \quad (12)$$

The difference matrix is divided into 4 sub-matrices in accordance with Eq. 9.

The i^{th} transform $[S_{1,i}^n]$ of the difference $[E_{0,n}^i]$ sub-matrix is calculated, using direct orthogonal transform:

$$[S_{1,i}^n] = [T_1][E_{0,n}^i][T_1] \text{ for } i = 1, 2, 3, 4. \quad (13)$$

The approximated i^{th} transforms are:

$$[\hat{S}_{1,i}^n] = [m_1(u, v) s_{1,i}^n(u, v)] \quad (14)$$

The difference matrices of the approximated transforms are:

$$[\Delta \hat{S}_{0,i}^n] = [\hat{S}_{0,i}^R] - [\hat{S}_{0,i}^n] \quad (15)$$

The coefficients obtained in result of the orthogonal transform from all pyramid levels are sorted in accordance with their spatial frequency, and scanned sequentially. The so obtained one-dimensional sequence of coefficients for the s^{th} frequency band of the two-dimensional linear transform of the original or of the difference image for the IDP level p is represented by the relation:

$$y_p(s) = y_p[u=\varphi(s), v=\psi(s)] \quad (16)$$

where $u=\varphi(s)$ and $v=\psi(s)$ are functions, which define the transformation for the two-dimensional massif of spectrum coefficients in the s^{th} frequency band for the level p . In order to achieve higher compression ratio the data is processed, applying adaptive entropy and run-length coding, performed in two steps: adaptive coding of the lengths of the series of equal symbols (run-length encoding, RLE), and adaptive coding with modified Huffman code (HE). The compressed and decompressed data of the one-dimensional massif of spectrum coefficients for the s^{th} spectrum band in the level p is presented as follows:

$$\alpha_p(s) = \text{RLE/HE}[y_p(s)]; \quad y_p(s) = \text{RLD/HD}[\alpha_p(s)], \quad (17)$$

where $\text{RLE/HE}[\bullet]$ and $\text{RLD/HD}[\bullet]$ are operators for entropy coding/decoding with Run-Length and Huffman coding.

After the end of the processing the coded image is represented in a new format, specially developed for the IDP decomposition. A special header is added, which contains information about the number of decomposition layers, the retained coefficients for each decomposition layer, etc.

The block diagram of the coder based on the common pyramid decomposition for a set of multispectral images is given in Fig. 1 – an example for modified IDP decomposition of 2 layers only. The matrix $[B_R]$ which represents the reference image R is processed first and the approximation $[\hat{B}_R]$ is used by the remaining spectral images for their further decomposition. The 2D OT (Orthogonal Transform) and 2D IOT (Inverse Orthogonal Transform) for all images are calculated for same set of retained coefficients. The difference images are represented by matrices $[E]$. The processing of the next decomposition layers is performed in similar way.

The decoding is done performing the so described operations in reverse order.

3 Selection of the Reference Image

The image which will be chosen to be used as a reference one is selected on the basis of the histogram analysis: the image, whose histogram is closest to all the remaining images in the processed set, is selected to be the reference one. The analysis is made using the correlation coefficient. The correlation coefficient ρ_{xy} [13] between vectors $\vec{X} = [x_1, x_2, \dots, x_m]^t$ and $\vec{Y} = [y_1, y_2, \dots, y_m]^t$, which represent the histograms of

the two images is given in Eq. 18, where $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$ and $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$ are the mean

values of the two histograms and m is the number of brightness levels for the both spectral images:

$$\rho_{x,y} = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^m (y_i - \bar{y})^2}} \quad (18)$$

The decision for the reference image selection is taken after the histograms of the multispectral images had been calculated and then the correlation coefficients for all couples of histograms are calculated and evaluated. For a multispectral image of N components the number of these couples (p,q) is:

$$L = \sum_{p=1}^{N-1} \sum_{q=p+1}^N 1(p,q) \quad (19)$$

When all L coefficients ρ_{pq} are calculated, is defined the index p_0 in ρ_{p_0q} , for which is satisfied the requirement:

$$\sum_{q=1}^N \rho_{p_0q} \geq \sum_{q=1}^N \rho_{pq} \quad \text{for } p, q = 1, 2, \dots, N, \quad \text{when } p \neq q \quad \text{and } p \neq p_0. \quad (20)$$

The reference image then is $[B_R] = [B_{p_0}]$.

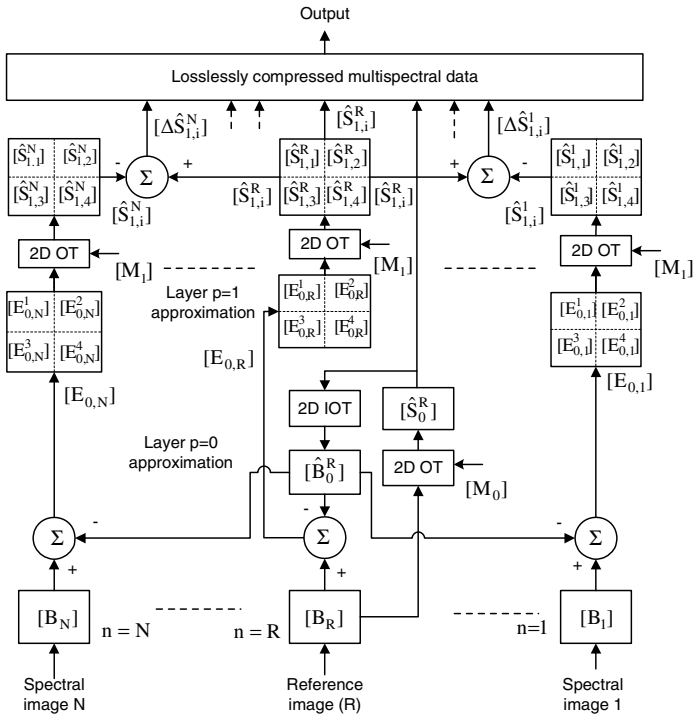
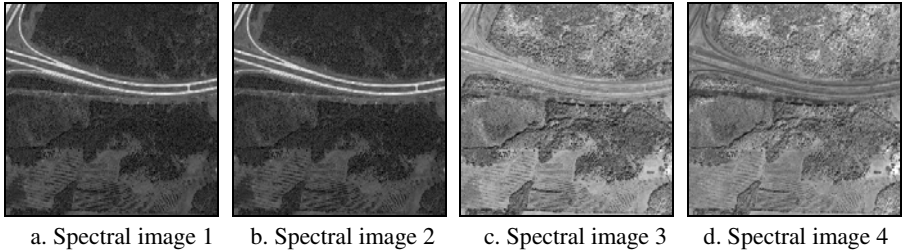


Fig. 1. Block diagram of the multispectral images coder

4 Experimental Results

For the experiments was used the software implementation of the method in Visual C++ (Windows environment). In accordance with the new method was developed a special format for multispectral images compression. For the experiments were used more than 100 sets of multispectral images (each set comprises 4 images, corresponding to the main colors – red, green, blue and one regular grayscale image).

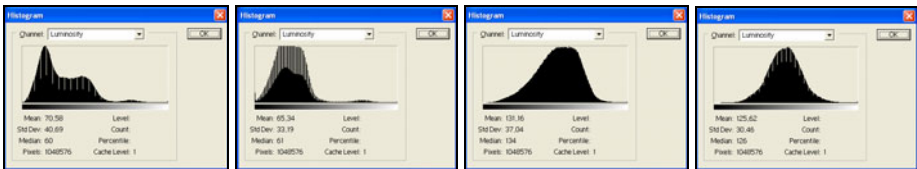
One set of 4 test images is shown in Fig. 2.



a. Spectral image 1 b. Spectral image 2 c. Spectral image 3 d. Spectral image 4

Fig. 2. A set of 4 spectral images of size 1000×1000 pixels, 8 bpp each

On Fig. 3 are shown the histograms of the test images from Fig. 2 (Adobe Photo-Point 10). As a reference image for the experiments was used the test Image 1.



a. Image 1 b. Image 2 c. Image 3 d. Image 4

Fig. 3. Histograms of the test images

For the experiments was used IDP decomposition of 2 layers. The size of the sub-image in the lower layer was 8×8, and in the next layer – 4×4 pixels. The number of coefficients for the lower layer was 4, and for the next layer – 7. For all experiments was used the 2D Walsh-Hadamard transform.

The experimental results for one of the test sets of spectral images are given in Table 1. The size of the compressed first approximation (Lv11) for the Reference Test image 1 is 91 KB. The size of the next-layer approximations (Lv12) depends on the similarity of the corresponding spectral image and the Reference one.

For some sets of spectral images as a reference image was used the grayscale one.

The results for all sets of test images are close to these, given below.

Table 1. Results obtained for a set of test spectral images with the new compression

Image	Lvl1 [KB]	Lvl2 [KB]	Total size	CR	PSNR [dB]
1	91	33,1			32,1
2		24,0	334,8 KB	11,95	75,0
3		36,7			36,0
4		150,0			34,0

The comparison for the evaluation of the method efficiency was performed with the widely used JPEG standard. The results obtained for the same set of spectral images with JPEG compression are given in Table 2. The quality of the restored images was selected to be close to that of the images obtained with the new compression (exact correspondence is not possible, but the values are similar). The results for JPEG 2000 (Table 3) are better than these for JPEG, but not as good as these obtained with the new method.

Table 2. Results obtained for a set of spectral images with JPEG-based compression

Image	JPEG size [KB]	Total size	CR	PSNR [dB]
1	114			32,2
2	136	543	7,36	31,3
3	157			31,3
4	136			32,2

Table 3. Results obtained for a set of spectral images with JPEG2000-based compression

Image	JPEG2K size [KB]	Total size	CR	PSNR [dB]
1	120			32,4
2	115	496	8,06	32,1
3	125			31,5
4	130			33,5

The compression ratio was calculated in accordance with:

The results thus obtained confirmed the expected high efficiency of the new method: for better quality of the restored image the compression ratio was much higher also.

5 Conclusions

The software implementation of the new method for compression of multispectral images based on the Inverse Difference Pyramid decomposition proved the method efficiency. The flexibility of the decomposition permits the creation of a branched structure, which suits very well the characteristics of multispectral images and makes the most of their similarity.

The main advantages of the new method are the high efficiency and the relatively low computational complexity. The high efficiency of the method was proved by the experimental results: it offers higher compression and better quality than that of the JPEG and JPEG2000 standards. The quality which is obtained with the new method is

high (visually lossless); its efficiency was not compared with that of the JPEG standards for very high compressions.

The method is suitable for a wide variety of applications: processing of video sequences, 3D object representation, efficient archiving of medical information and satellite images and many others, i.e. in all cases when objects are moving relatively slowly and the quality of the restored images should be very high. Some experiments had already been performed for 3D object representation [14] and for processing of video sequences, which confirmed the method flexibility and efficiency.

Acknowledgement

This work was supported by the National Fund for Scientific Research of the Bulgarian Ministry of Education and Science, Contract VU-I 305.

References

1. Gelli, G., Poggi, G.: Compression of multispectral images by spectral classification and transform coding. *IEEE Trans. Image Process.* 8(4), 476–489 (1999)
2. Dragotti, P., Poggi, G., Ragozini, A.: Compression of multispectral images by three-dimensional SPIHT algorithm. *IEEE Trans. Geosci. Remote Sens.* 38(1), 416–428 (2000)
3. Fowler, J., Fox, D.: Embedded wavelet-based coding of 3D oceanographic images with land masses. *IEEE Trans. Geosci. Remote Sens.* 39(2), 284–290 (2001)
4. Tang, X., Pearlman, W., Modestino, J.: Hyperspectral image compression using three-dimensional wavelet coding. In: *Proc. SPIE*, vol. 5022, pp. 1037–1047 (2003)
5. Cagnazzo, M., Poggi, G., Verdoliva, L., Zinicola, A.: Region-oriented compression of multispectral images by shape-adaptive wavelet transform and SPIHT. In: *Proc. IEEE Int. Conf. Image Process.*, pp. 2459–2462 (2004)
6. Gersho, A., Gray, R.: *Vector quantization and signal compression*. Kluwer AP, Dordrecht (1992)
7. Kaarna, A.: Integer PCA and wavelet transform for lossless compression of multispectral images. In: *Proc. of IGARSS 2001*, pp. 1853–1855 (2001)
8. Markas, T., Reif, J.: Multispectral image compression algorithms. In: Storer, J., Cohn, M. (eds.), pp. 391–400. *IEEE Computer Society Press*, Los Alamitos (1993)
9. Aiazzi, B., Baronti, S., Lastri, C.: Remote-Sensing Image Coding. In: Barni, M. (ed.) *Document and Image Compression*, ch. 15. *CRC Taylor&Francis* (2006)
10. Cagnazzo, M., Parrilli, S., Poggi, G., Verdoliva, L.: Improved Class-Based Coding of Multispectral Images With Shape-Adaptive Wavelet Transform. *IEEE Geoscience and Remote Sensing Letters* 4(4), 565–570 (2007)
11. Wu, J., Wu, C.: Multispectral image compression using 3-dimensional transform zeroblock coding. *Chinese Optic Letters* 2(6), 1–4 (2004)
12. Kountchev, R., Kountcheva, R.: Image Representation with Reduced Spectrum Pyramid. In: Tshirintzis, G., Virvou, M., Howlett, R., Jain, L. (eds.) *New Directions in Intelligent Interactive Multimedia*, pp. 275–284. *Springer*, Heidelberg (2008)
13. Bronshtein, I., Semendiyayev, K., Musiol, G., Muehlig, H.: *Handbook of Mathematics*, 5th edn. *Springer*, Heidelberg (2007)
14. Kountchev, R., Todorov, V., Kountcheva, R.: Multi-view Object Representation with Inverse Difference Pyramid Decomposition. *WSEAS Trans. on Signal Processing* 5(9), 315–325 (2009)

Econometric Approach for Broadband Market in Japan^{*}

Takeshi Kurosawa¹, Hiromichi Kawano³, Motoi Iwashita⁴,
Shinsuke Shimogawa², Shouji Kouno², and Akiya Inoue⁴

¹ Department of Mathematical Information Science, Tokyo University of Science

² NTT Service Integration Laboratories, NTT Corporation

³ Network Technology Center, NTT Advanced Technology Corporation

⁴ Department of Management Information Science,
Faculty of Social Systems Science, Chiba Institute of Technology

Abstract. We investigated the growth in the Japanese broadband market. We analyzed price elasticity and changes in the broadband market in 2009 by comparing those of the market in 2005, which we consider the growth stage of Fiber-To-The-Home (FTTH) service in terms of econometrics. Through analysis, we found that users feel FTTH service fees are reasonable and that the service is used among high- and low-income households. We also believe that FTTH service is no longer for users who use the Internet for heavy downloading activities, such as P2P file sharing. Furthermore, FTTH fees have become less elastic than in 2005.

1 Introduction

Broadband services have been increasing in Japan (See Fig. 1), especially with regards to Fiber-To-The-Home (FTTH), which has been rapidly increasing compared with other countries. We investigated the Japanese broadband market by dividing it into three stages: 1. Initial, 2. Growth, and 3. Stable. We define the initial stage as a period before the middle of 2005. In this stage, ADSL service lead the broadband market in Japan. During the middle of 2005 (growth stage), the market situation changed. That is, the number of ADSL users reached a peak and started to decrease. Meanwhile the number of FTTH users increased dramatically. The migration from ADSL to FTTH occurred. Recently, however, the number of subscribers who are switching from ADSL to FTTH seems to be decreasing. Therefore, we define the growth stage from the middle of 2005 to the middle of 2008, after which the stable stage begins. FTTH service has been used by not only “Innovators” or “Early adopters” but also by “Early majority” or “Late majority” users (See 2). The fact is that the net increase in FTTH subscribers has been decreasing since the middle of 2008 (See Fig. 2). Figure 2 shows the change in the market structure from the growth to stable stage. It is important to clarify this change in the market structure.

^{*} This work was conducted when the first author was belonging to NTT Service Integration Laboratories.

Network carriers or business managers estimate service demand for the sake of capital investment and management decisions. As shown in Fig. 1, the increase in demand of broadband services, including FTTH or ADSL, cannot be forecasted with only traditional methods, such as time-series analysis, because the increase is affected by many factors such as customer requirement diversification (switching from ADSL to FTTH, for example), new service introduction, and deployment. For an accurate demand estimation which is one of the objectives, we need to perform market analysis by considering such diversified factors. We used a micro econometric analysis. The micro econometric approach requires service-choice behavior modeling of customers.

Our main purpose is to analyze the broadband market in the stable stage by applying the micro econometric method. With this method we can clarify the change from the growth to the stable stage with regards to price elasticity and characteristics of FTTH users.

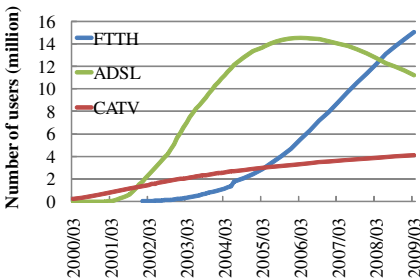


Fig. 1. Number of broadband users in Japan (Source: Ministry of International Affairs and Communication <http://www.soumu.go.jp/>)

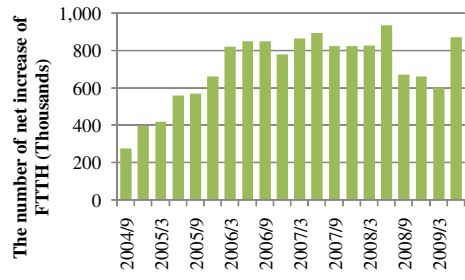


Fig. 2. Net increase in FTTH (Source: Ministry of International Affairs and Communication <http://www.soumu.go.jp/>)

2 Related Work

2.1 Econometric Approach for Telecom Services

There have been several empirical studies on the analysis of the Internet market. It is, however, difficult to predict service demand using conventional techniques such as time-series analysis based on revealed data. There has been research related to *Framework for Scenario Simulation* (FSS) [2,3] as one method of solving this problem. Our objective is not to obtain a demand-forecasting result but to simulate scenarios under assumed situations. FSS needs to construct a service-choice behavior model to express the actual market structure. The service-choice behavior model is the most important factor for improving the accuracy of the evaluation results in FSS. Usually, we use *discrete choice analysis* (DCA) [4,5] to construct a service-choice behavior model. DCA (see Section 2.2) is a technique derived from the civil engineering field and is applied in various fields. To use this modeling, consideration of customer preference variation is required.

Several types of empirical models related to the service-choice behavior model have been studied for telecommunication services. Kridel et al. described a binary choice model of Internet access [6]. Rappoport et al. extended this model to a nested logit specification (See Section 2.2) that includes other Internet access options [7]. These models are based on revealed preferences (RP) data and do not include the attributes of Internet access alternatives such as price and speed. Chaudhuri et al. [8] proposed an Internet access model based on RP data including price as well as socioeconomic factors and regional characteristics. Savage and Waldman proposed an Internet access model based on stated preferences (SP) [9] and [10]. Their results indicate the sensitivities and willingness to pay for the service attributes of Internet access alternatives. Varian analyzed the trade-off between price and speed in the choice of Internet access services [11]. There has been research specifically on the analysis of the broadband market in Japan [12,13,14,15,16,17,18]. As mentioned in Section 1, our aim is to understand the change in the Japanese broadband market from the growth to stable stage. Ida [12] analyzed the Japanese broadband market during the growth stage. His focus was mainly on analysis of migration from ADSL to FTTH. In 2005, the growth stage, high income users and the users who watched video via the Internet preferred FTTH service. This paper focuses on the change of broadband market from the growth stage to the stable stage by comparing Ida's results [12] from the viewpoint of econometrics.

2.2 Discrete Choice Analysis

We give an overview of DCA. In which each alternative has a utility function. It is assumed with the random utility maximization (RUM) [19] model that a customer chooses the alternative that has the highest utility from all the selectable alternatives. Let U_{in} be a utility function of alternative i of customer n . These explanatory variables are individual, service, and environmental attributes. The utility function U_{in} consists of a systematic term V_{in} and error term ε_{in} . The service-choice set C_n , which is a subset of the universal choice set C , differs from person to person. A universal set means a set of all alternatives. Generally, customer decision-making processes are complex. That is, the decision-making process has a multidimensional structure, which is expressed like a decision tree.

The most popular and simplest DCA model is the multinomial logit (MNL) model. It is assumed with this model that alternatives are on the same level in a hierarchy. That is, customer n compares all the alternatives contained in C_n and chooses one alternative. For a simple multinomial choice problem, if we take *extreme value type I* (EV1) as a random error term ε_n , the probability $P_n(i|C_n; \beta)$ of alternative i being chosen is expressed by

$$P_n(i|C_n; \beta) = \frac{\exp(\mu V_{in})}{\sum_{k \in C_n} \exp(\mu V_{kn})}, \quad (1)$$

where μ is a scale parameter in EV1 that is usually normalized by 1, and β is an unknown parameter vector that appears in V_{kn} ($k \in C_n$). Furthermore,

Ben-Akiva proposed the nested logit model, which is used to analyze a hierarchy structure of the decision-making process (See [4]).

We determine the value of β by using maximum likelihood estimation. This function is defined by

$$L(\beta; \mathbf{y}) = \prod_{n=1}^N \prod_{i \in C_n} P_n(i|C_n; \beta)^{y_{in}},$$

where N is the number of samples and y_{in} is a dummy variable that is 1 when a customer n chooses alternative i ; otherwise, 0. Let vector $\hat{\beta}$ be the estimated vector value of β . In the DCA model, $\bar{\rho}^2$, which shows the goodness, suitability, of an index, is often used. This is defined by $\bar{\rho}^2 = 1 - (\mathcal{L}(\hat{\beta}) - K)/\mathcal{L}(\mathbf{0})$, where $\mathcal{L}(\beta) = \log L(\beta)$ and K is the number of variables. The value lies between 0 and 1. The model improves if the value is close to 1. Generally, the model is good if $\bar{\rho}^2$ takes the value around 0.3, although this value depends on the data.

As a generalization of the DCA model, a mixed logit (MXL) model (logit kernel or logit mixture) [5,20] is attracting attention. The model includes mixtures of Logits for randomly distributed parameters and error components. The model corresponds to our a-priori expectations about the distribution of the random utilities of alternatives. To understand the randomness of the parameter, the mixed model is expressed by an integral form. For the simulation of this integral form, the MXL model must generate draws followed by a certain distribution.

3 Modeling of Service-Choice Behavior

3.1 Market Survey

We performed a market survey in August 2009 using a Web questionnaire system. We asked 2,043 Internet users what kind of broadband services the respondents are using now, what related services they are using, and their preference for the services. We denote the broadband services as FTTH, ADSL, or CATV. To analyze the migration from ADSL to FTTH, The percentage of FTTH users we questioned actually use this service more than the ratio of FTTH to Internet users estimated in the current market. Moreover, we restricted the participation of heavy users to no more than 10% of this study group to avoid Web questionnaire bias.

3.2 Modeling

We analyzed how the market changed (from the growth to stable stage). As mentioned in Section [1], the growth in FTTH has been gradually slowing down. To understand the difference between the growth stage (second stage) and the stable stage (third stage), we compared the recent broadband market with that 2005 (second stage). The first study of the Japanese broadband market was studied by Ida [12]. We discuss the difference between these stages by applying

Table 1. Summary of results

	BB 2005	BB 2009	
Model	MXL	MNL	MXL
stage	Nov. 2005	Aug. 2009	
Observations	1890	2043	2043
Num. of draws	250	–	500
$\mathcal{L}(\mathbf{0})$	-2076.4	-2224.5	-2244.5
$\mathcal{L}(\hat{\beta})$	-1647.6	-1665.5	-1639.0
ρ^2	0.206	0.258	0.270

the same modeling structure as the one he developed. We call his model BB 2005 and call our model BB 2009. Table 1 compares the modeling results of this market survey with Ida’s [12].

In addition to the MXL model, we used an estimation of the MNL model. Although we tried to reconstruct the BB 2005 model as much as possible, there may be minute differences between them, for example, the sampling method, and how to generate draws. Therefore, we cannot make a sweeping judgment about whether the model is better than others. We believe, however, that the results have the same level of goodness as modeling with respondents. The following equations are utility functions of FTTH, ADSL, and CATV, respectively.

$$\begin{cases} U_{FTTH} = \beta_{FT} + \beta_{\text{monCh}} \cdot X_{\text{monChFO}} + \beta_{FO}^T \mathbf{X}_{FO} + \varepsilon_{FO} \\ U_{ADSL} = \beta_{AD} + \beta_{\text{monCh}} \cdot X_{\text{monChAD}} + \beta_{AD}^T \mathbf{X}_{AD} + \varepsilon_{AD} \\ U_{CATV} = \beta_{\text{monCh}} \cdot X_{\text{monChCT}} + \varepsilon_{CT} \end{cases},$$

where β_{AD}, β_{FT} are specific alternative constants and β_{monCh} is a parameter for a common variable, β_{AD}, β_{FO} are vectors of the individual and household characteristics, and T is a transpose operator of the vector.

3.3 Results of Estimated Parameters

We compared results of the estimated parameters to understand the change in the market. We use t -value for the validity of the parameters as one of the statistics. Excluding the parameter related with the monthly charge (β_{monCh}), a positive sign of the estimated parameters means that the variables related to the positive parameter increase the utility functions for choosing (ADSL or FTTH) compared with the utility function for choosing CATV. In contrast, the minus sign indicates that the variables related to the minus parameter decrease. If the t -value of the parameters, excluding the monthly charge, is estimated near 0, the variable with the low absolute value of the t -value has the same utility for the alternative CATV since the utility functions of ADSL and FTTH are expressed by the difference between the utility function of CATV.

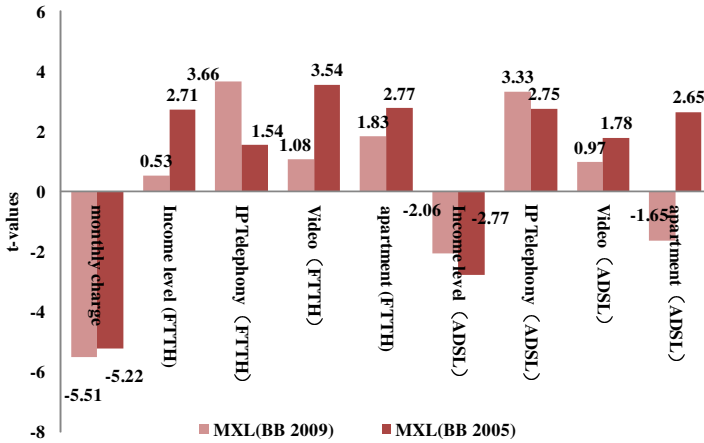


Fig. 3. t -values for parameters

From Fig. 3 we see that

1. high-income users and users who frequently view videos on the Internet do not affect the choosing of FTTH as much as they do in BB 2005 and
2. the sign of the estimated parameter (apartment), which is for users living in a apartment, is inverted.

As a result of the spread of FTTH service as a popular broadband service in 2009, the first fact mentioned above occurred. Note that although FTTH service has been the most expensive Internet access service, the monthly charge has been decreasing. Because of this decrease, users at the same income level of CATV users chose FTTH service. This confirms the results of the questionnaire (See Fig. 4). Moreover, viewing videos on the Internet has increased among not only FTTH users but also ADSL and CATV users (See Fig. 5). The second fact mentioned above may come about as a result that a cheaper FTTH service (VDSL) for apartments has been more popular in Japan. Therefore, the parameter of ADSL users living in apartments may be small compared with that of BB 2005.

Next we observe the randomness of the parameter, which comes about because of the difference among customers. The MXL model has random in addition to fixed parameters. Namely, the parameters have variance. Only the fixed parameters are shown in Fig. 3. We constructed a model with random parameters in the same manner as BB 2005. Some of the estimated results are listed in Table 2.

We see that the standard derivation (std) of the IP Telephony (ADSL) parameter is bigger than the one for FTTH. This comes from the fact that ADSL service has increased with developments in IP telephony. In fact, IP telephony was one of the killer services of ADSL. Although IP telephony based on FTTH access line has dramatically increased recently, there are still many ADSL users with IP telephony. For constructing the MXT model, we generated 500 draws

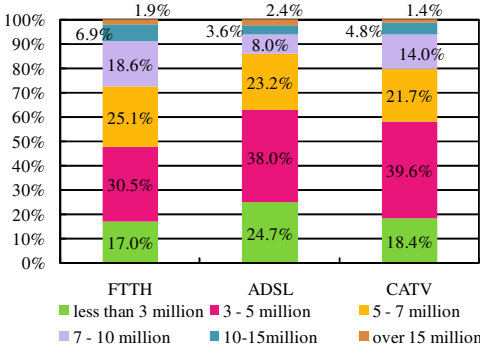


Fig. 4. Income level

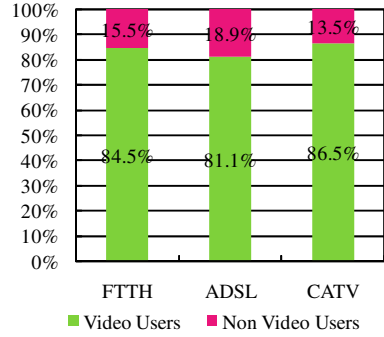


Fig. 5. Users viewing videos on Internet

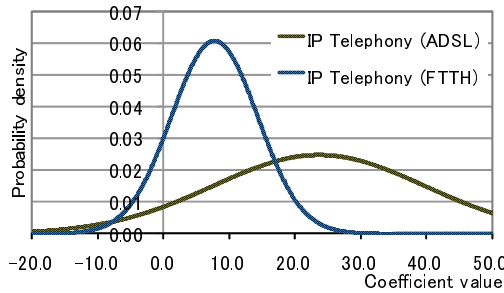


Fig. 6. Distribution of coefficient

Table 2. Average value and standard derivation

		Value
IP Telephony (FTTH)	Average	23.6
	Std	16.1
IP Telephony (ADSL)	Average	7.71
	Std	6.56

followed by a normal distribution. Figure 6 shows the distribution using the results from Table 2. The conditional distribution of the IP Telephony parameter for each individual using his/her choice of service can be calculated using Bayes Theorem in the MXL model (See 21).

3.4 Price Elasticity

Finally, we analyzed price elasticity. Table 3 is a comparison of self price elasticity between 2005 and 2009.

Table 3. Comparison of self price elasticity

	BB 2005 (MXL)	BB 2009 (MNL)
FTTH	-1.308	-0.149
ADSL	-0.439	-0.162
CATV	-1.675	-0.169

The values in Table 3 means, for example, that the choice probability of FTTH goes below 1.308% when the fee of FTTH is reduced by 1%. Usually, the market is called inelastic when the absolute value of price elasticity is less than 1. The price elasticity of BB 2009 was simply evaluated using the MNL model, not the MXL model, although the MNL model has the independence of irrelevant alternatives (IIA) property (See 4). Since the models are different (MXL vs MNL) and the dataset is different, a direct comparison is not sufficient. However, we can assume that the price elasticity of FTTH decreased about 1/10, from -1.308 to -0.149, and the estimated results of the two models (BB 2009 (MXL) and BB 2009 (MNL)) are not extremely different (See \bar{p}^2 in Table 4). This result relatively shows the decreasing of the price elasticity compared with 2005. Moreover the market is called as *inelastic* if the price elasticity is estimated as less than 1%. Therefore we conclude the market is inelastic with regard to the price. Through the evaluation of elasticity, service providers or network careers can take a suitable action for the market.

We can also see this fact from the questionnaire (See Figs. 7, 8). From Fig. 7, we can see that there are many people who thought that the speed and fee of Internet access line is reasonable when they choose the Internet access line. When they switch the Internet access line, they placed importance on the speed of the

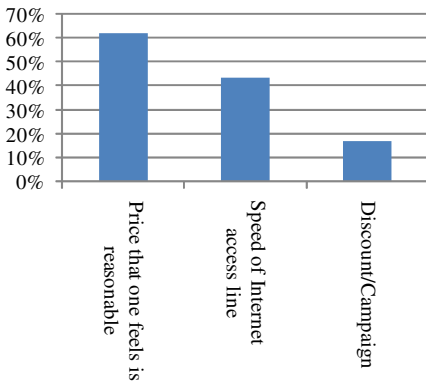


Fig. 7. Top three reasons of having selected existing Internet access line

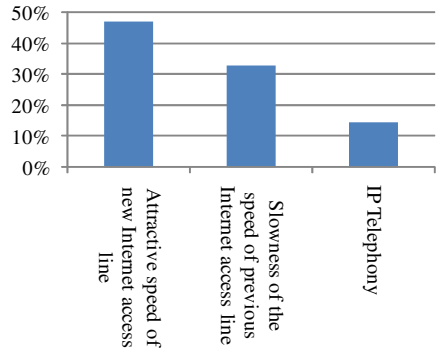


Fig. 8. Top three reasons of having switched Internet access line

Internet access line. Therefore, we can assume that price elasticity decreases. As a result of this decrease, customers feel that

- the fee and
- speed of the Internet access line are important.

4 Conclusion

We focused on analyzing the difference between the growth and stable stages of FTTH service in Japan. In general we think that the fee and the speed are important for Internet access. Through our analysis, the degree of importance was shown quantitatively. Furthermore we found changes in the broadband market from the analysis results. In comparison with the broadband market in 2005, FTTH service is in widespread use for not only high income users but also lower income users. As a result, video service via Internet is not always for FTTH users. It is inferred that early majority or late majority users are the main FTTH users at the beginning. Such a growth in FTTH occurred because FTTH provides stable and high-speed Internet connection, which is the most attractive characteristic of FTTH. Although IP telephony is one of the killer applications of FTTH service, the sensitivity, or importance, of IP telephony for the utility function of FTTH is less than that of ADSL, and it widely varies among users. The other reason of the increase in FTTH service is that the fee has been decreasing, which users believe is reasonable. The strategy of competitive pricing may be no longer useful because the price elasticity of FTTH is less than that in 2005. However, this strategy may be useful for competition among service providers. The results do not reflect brand competition. This is for future work.

References

1. Rogers, E.M.: *Diffusion of Innovations*, 5th edn. Free Press, New York (2003)
2. Inoue, A., Nishimatsu, K., Takahashi, S.: Multi-attribute learning mechanism for customer satisfaction assessment. *Intelligent Engineering Systems Through Artificial Neural Networks* 13, 793–800 (2003)
3. Inoue, A., Takahashi, S., Nishimatsu, K., Kawano, H.: Service demand analysis using multi-attribute learning mechanisms. In: *IEEE International Conference on Integration of knowledge intensive multi-agent systems*, pp. 634–639. IEEE, Cambridge (2003)
4. Ben-Akiva, M., Lerman, S.R.: *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, Cambridge (1985)
5. Train, K.E.: *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge (2003)
6. Kridel, D.J., Rappoport, P.N., Taylor, L.D.: The demand for high-speed access to the internet: The case of cable modems. In: *International Telecommunications Society*, Buenos Aires (July 2000)
7. Rappoport, P., Kridel, D.J., Taylor, L.D., Alleman, J.: A residential demand for access to the internet. In: Madden, G. (ed.) *The International Handbook of Telecommunications Economics*, vol. II. Edward Elgar Publishers, Cheltenham (2002)

8. Chaudhuri, A., Flamm, K.S., Horrigan, J.: An analysis of the determinants of internet access. *Telecommunications Policy* 29, 731–755 (2005)
9. Savage, S.J., Waldman, D.: Broadband internet access, awareness, and use : Analysis of United States household data. *Telecommunications Policy* 29, 615–633 (2005)
10. Savage, S.J., Waldman, D.: United States demand for internet access. *Review of Network Economics* 3(3), 228–247 (2004)
11. Varian, H.R.: Broadband: Should we regulate high-speed internet access? In: Allaman, J., Crandall, R. (eds.) *The Demand for Bandwidth: Evidence from the INDEX Project*. Brookings Institution, Washington (2002)
12. Ida, T.: Fiber to the Home. In: *Broadband Economics: Lessons from Japan*, pp. 199–217. Routledge, New York (2009)
13. Kurosawa, T., Inoue, A., Nishimatsu, K., Ben-Akiva, M., Bolduc, D.: Customer-choice behavior modeling with latent perceptual variables. *Intelligent Engineering Systems Through Artificial Neural Networks* 15, 419–426 (2005)
14. Kurosawa, T., Inoue, A., Nishimatsu, K.: Telephone service choice-behavior modeling using menu choice data under competitive conditions. *Intelligent Engineering Systems Through Artificial Neural Networks* 16, 711–720 (2006)
15. Ben-Akiva, M., Gershfeld, S.: Multi-featured products and services: analysing pricing and bundling strategies. *Journal of Forecasting* 17, 175–196 (1998)
16. Ida, T., Kuroda, T.: Discrete choice analysis of demand for broadband in Japan. *Journal of Regulatory Economics* 29(1), 5–22 (2006)
17. Ida, T.: The broadband market in Japan. In: Taplin, R., Wakui, M. (eds.) *Japanese Telecommunications Market and Policy in Transition*, pp. 37–64. Routledge, New York (2006)
18. Ida, T., Kinoshita, S., Sato, M.: Conjoint analysis of demand for IP telephony: The case of Japan. *Applied Economics* 18(40), 1279–1287 (2006)
19. Manski, C.F.: The structure of random utility models. *Theory and Decision* 8, 229–254 (1977)
20. Louviere, J.J., Hensher, D.A., Swait, J.D., Adamowicz, W.: *Stated Choice Methods: Analysis and Applications*. Cambridge University Press, Cambridge (2000)
21. Revelt, D., Train, K.: Specific taste parameters and mixed logit. Working Paper E00-274. Department of Economics. University of California, Berkeley (2000)

Opinion Exchange Support System by Visualizing Input History

Yukihiro Tamura, Yuuki Tomiyama, and Wataru Sunayama

Graduate School of Information Sciences, Hiroshima City University,
3-4-1 Ozuka-Higashi, Asa-Minami-Ku, Hiroshima, 731-3194, Japan

Abstract. The development of network communications has been fueled by the utilization of network services. Opinion exchanges by the general public are frequently realized in such network services as electronic bulletin board systems (BBS), social network services (SNS), blogs, and so on. However, grasping the discussion flow is time-consuming, because there are too many logs of previously written comments. Therefore, a system is required that shows the discussion state and its history.

In this paper, an opinion exchange support system is proposed that visualizes input history. Each opinion is input by moving images on the interface, and moving histories are replayed as animation. In this study, a target discussion is imagined as a role-sharing problem. The experimental results showed that the system effectively realized flourishing discussion and discussion convergence.

Keywords: opinion exchange support, input history visualization, online communication.

1 Introduction

Communication using such network services as bulletin board systems (BBS) and social networking services (SNS) continue to increase. They encourage idle chatter, opinion exchanges, and discussion. Users continuously provided new opinions and informations with shared opinions.

However, BBS services, which submit comments to show the series of comments entered on all of the opinions so far and to grasp their overall changes, are time-consuming. Problems include the difficulty of intuitively understanding the overall picture of the opinions.

Therefore, in this research, we propose a system that users can intuitively understand the opinions and input histories of other users and supports idea exchange between users. In particular, we determine such roles as division of labor and assignment of duties within small groups of 4-8 people.

Moreover, this system is not the standard type of character discussion that handles an image by exchanging opinions on a two-dimensional interface. The opinion input in our study is performed by moving a picture of person A on an interface to a domain showing role B; the movement corresponds to opinion input

that considers person A suitable for role B. We propose an opinion exchange support system equipped with the following features so that such opinion exchanges can be performed smoothly.

1. Asynchronous: Not all members need to simultaneously discuss.
2. Anonymity: Opinions are not specified by topic or to whom they were made.
3. Visualization: Situation of entire current opinion and its history can be checked.

2 Opinion Exchange Support System That Visualizes Input History

In this chapter, we give the composition and details of our proposed opinion exchange support system.

2.1 Problem Setting

Our system's target opinion exchange solves problems in which a group of people must mutually divide roles or responsibilities, including such examples as work in an organization, the positions on a softball team, casting a theater production, cleaning assignments, and responsibilities for a barbecue.

2.2 System Outline

The composition of our opinion exchange support system is shown in Fig. 1. With the opinion exchange support interface, the following functions can be performed: processing the picture at the time of opinion input, preserving an opinion input history, and replay.

Next we describe the details.

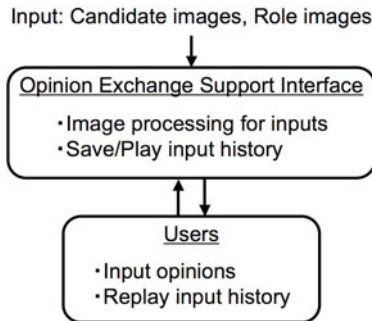


Fig. 1. Opinion exchange support system

2.3 Input

Pictures that express each role [\[1\]](#) and the candidates are given as input in this system. In addition, we assume that all group members who perform opinion exchanges are candidates for each role.

2.4 Interface

In this section, we state the composition of the interface in which users perform opinion exchange. The interface is shown in Fig. [2](#). Candidate pictures are first arranged in the center in the four areas that correspond to each role and display pictures that express each role in each area.

Below, image processing at the time of an opinion input, the preservation of the opinion input history, and replay are described.

Image processing at time of opinion input. The following three image processing steps are performed during an opinion input:

1. picture movement
2. picture division
3. adjustment of picture brightness

Picture movement is performed when a user drags a candidate picture. However, a user ID, which grants others the permission to move the picture to each candidate's picture, can only be used to move the picture (one candidate picture per sheet) to which a user's own ID was assigned. Pictures to which other user IDs were assigned cannot be moved. As an exception, a gforced moveh can also move the picture to which other user IDs have been assigned under certain situations (see [2.5](#)).

Picture division involves user A who performs an opinion input for the first time, for example. If user A drags another user picture when the candidate picture to which user A's ID was assigned does not exist on the interface, the picture will divide and generate a candidate picture with user A's ID. The pictures of other users who became the dividing agency remain in the original state without moving.

Adjustment of picture brightness refers to the quantity of pictures arranged in the domain that show each role of each candidate picture. Rate $Light_{ik}(\%)$ of the brightness in domain k of each candidate picture i is given by Formula [\(1\)](#). However, the total of candidate pictures i by which num_i was divided and num_{ik} express the number of pictures i arranged at domain k :

$$Light_{ik}(\%) = \frac{num_{ik}}{num_i} \times 100. \quad (1)$$

When a certain candidate picture on the interface is arranged so that any one domain shows the four roles, it is the blightest, distributed to other roles, and arranged to express the distribution of intuitive opinions to support their comprehension.

¹ Although presently a maximum of four roles can be set up, we believe increasing the number of roles is possible by changing the screen composition.

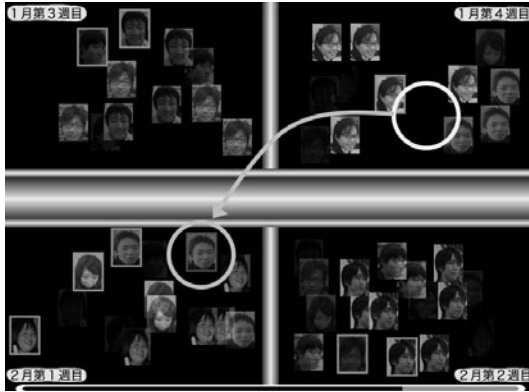


Fig. 2. Opinion input by moving an image. Member picture with a picture frame (green) has an operational user under login, and arrow expresses candidate picture on which forced movement was carried out.

Preservation of opinion input history and replay. While each user is performing opinion input, the following picture information moved by the user is saved as opinion input history: starting time of an opinion input, picture ID being moved, its xy coordinates, and the user ID used for the movement.

At the time of the replay of the opinion input history, animation replays each user’s opinion input operation based on the saved opinion input history. Since animation is connected and created in the order into which all user input histories were input, the reproducing time increases with an increase in an input history; but it is shortened by fast forward or/and skip functions.

2.5 Users

Users are the group members who exchange opinions to determine roles by opinion input with our proposed system. Below, opinion input and animation replay are described.

Opinion input. Users move a candidate picture to the domain that shows each role and input their opinions about “I think that this candidate is suitable for this role”. (Fig. 2). The picture that user can operate is surrounded and expressed as a thick green frame on the interface. In Fig. 2, opinion exchanges determined weekly cleaning duties for four weeks, using the following roles: 3rd week of the month, 4th week of the month, and 1st and 2nd weeks of February.

Moreover, the time when one user can input one opinion is displayed in the bar that shows the residual time on the interface’s lower part as one minute to restrict the saved input history and to condense the replayed animation time.

The period during which opinion input can be performed is considered a separate set among members. If each user remains within a particular period, they can input their own opinion any number of times (from the 2nd time to correction).

Table 1. Themes for role sharing

Theme	Four Roles	Users
Cleaning duty 1	Four weeks	8
Cleaning duty 2	Four weeks	8
Cleaning and Seminar presentation	Cleaning for two weeks and two presentations	8
Seating chart	Seats in lab	4
Voice actor 1	Dragon Ball	7
Voice actor 2	Slam Dunk	4
Voice actor 3	Gegege-no-Kitaro	7

Opinion input by forced moves. To perform opinion exchanges, interaction between users is required that considers operations in which a counterargument and a proposal can react to other opinions. So the gforced moveh button, which moves a picture with the IDs of other users who cannot move it, was prepared in the interface. Users can only move one candidate picture that only other users moved to the domain. Opinions that express opposition and suggest a different role for this candidate can be input. However, if the fixed time lapses² before the user logs out, this function cannot be used.

Animation replay. Users can replay and check the history of the inputted opinions by animation. The following verification can be attained with this function: which candidates went to the domain showing each role, the time it became the present state, what process was involved, and which users performed what kind of opinion input.

Moreover, the picture's movement speed during opinion input expresses the degree of the confidence of each user's opinion. By acquiring such information from animation, a more effective opinion input can be supported to assign the intended roles.

3 Evaluation Experiment: Results of Opinion Exchange Support System

In this section, people exchange opinions about the division of roles about problems to decide roles with our proposed interface. We experimentally verified the existence of an effect that supports opinion exchanges between users.

3.1 Experiment Description

We conducted an experiment on mutual roles determined by opinion exchanges with our proposed interface on the seven themes shown in Table 1 with 16 information science and graduate students. The experiment performed opinion

² Its default was set up as one hour.

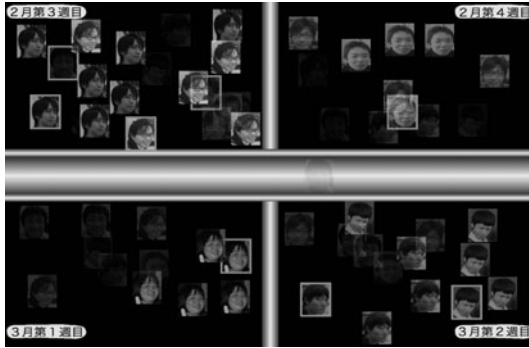


Fig. 3. Interface after opinion exchanges for Cleaning Duty 2

exchanges that equaled the number shown on the right of Table II, and roles were assigned to the four participants.

The experiment procedure was carried out as follows.

1. All users input their opinions of candidates when there is no opinion input from others.
2. All users input their opinions until the term³ ended during which all user opinions were input in Procedure 1 for every theme using the initial state and the proposed interface. All were then combined.
3. All users were informed of their roles that were determined by the majority.

Evaluation verified the effect of our proposed interface by comparing the majority results of Procedures 1 and 2.

3.2 Experimental Result and Consideration

The interface image plane at the end of the opinion exchanges about “Cleaning Duty 2” is shown in Fig. 3. After the opinion exchanges, the persons, who participated and whose pictures can be seen in each domain by agreement of the opinion of eight persons, can be intuitively distinguished as the remaining candidates. One independent candidate was also determined by the majority for all four roles.

We show the candidate picture ranked first by the majority before/after opinion exchanges in Table 2. For the majority when there are no other user opinions, more candidates are introduced for one role, or a candidate is nominated for two or more roles. This is because each user only input opinions based on individual preferences without considering other opinions.

In almost all roles, a candidate was not assigned to two or more roles while a role was being narrowed down to one person. After this system, users repeated opinion input to reach a fixed agreement based on shared ideas.

³ It took one or two days.

Table 2. Results of majority votes between before/after opinion exchanges: underlined members are included in both before and after

Theme	Before	After
Cleaning Duty 1: Role 1 (A-H) Role 2 Role 3 Role 4	B, <u>D</u>	<u>D</u> ,F
	E	B,H
	A,C, <u>E</u> ,F,G,H	<u>E</u>
	D,H	A
Cleaning Duty 2: Role 1 (A-H) Role 2 Role 3 Role 4	<u>A</u> ,B,E,G	<u>A</u>
	A,C,F,H	H
	H	D
	<u>E</u> ,G	<u>E</u>
Cleaning and Seminar presentation: Role 1 (A-H) Role 2 Role 3 Role 4	G	F
	F,G	A
	A	E
	B,D	H
Place of seats: Role 1 (A-D) Role 2 Role 3 Role 4	<u>D</u>	<u>D</u>
	<u>A</u>	<u>A</u>
	<u>C</u>	<u>C</u>
	<u>B</u>	<u>B</u>
Voice actor 1: Role 1 (A-G) Role 2 Role 3 Role 4	<u>D</u> ,G	<u>D</u>
	B, <u>E</u>	C, <u>E</u>
	<u>A</u> ,C,E	<u>A</u>
	C,F	B
Voice actor 2: Role 1 (A-D) Role 2 Role 3 Role 4	<u>B</u>	<u>B</u>
	<u>D</u>	<u>D</u>
	<u>C</u>	<u>C</u>
	<u>A</u>	<u>A</u>
Voice actor 3: Role 1 (A-G) Role 2 Role 3 Role 4	B	D,G
	<u>E</u>	<u>E</u>
	B	A
	F	C

When attention was paid to the common candidates before and after opinion exchanges, as shown by the underlining in Table 2, except for the “Seating chart” and “Voice actor 2” tasks, almost all roles were changed by the candidates before and after opinion exchanges. Even when just a single candidate was presented to the opinion exchange, other candidates eventually surfaced. During the opinion exchange process, in many cases they would be a different candidate. After an active opinion exchange with this interface that examined various possibilities, a final conclusion was reached.

For the “Seating chart” and “Voice actor 2” tasks, the number of users was four, and when four persons were needed to perform one of the roles, all user opinions were considered. No difference was found in the results before and after opinion exchanges because agreement was high.

All the user logins and the times other user pictures were forced to move are shown in Table 3 by theme. Each user performed an average of two or three

Table 3. Number of logins and forced moves

Theme	Logins	Forced moves	Users
Cleaning duty 1	18	16	8
Cleaning duty 2	24	18	8
Cleaning and seminar presentation	19	17	8
Seating chart	8	7	4
Voice actors 1	17	14	7
Voice actors 2	6	5	4
Voice actors 3	21	15	7

logins, and forced moves were performed at a rate of about 80%. When we questioned users about why they used forced movements, one user answered that he wanted to do a particular role. Another user said that he wanted to avoid a particular role and another said that a more suitable candidate for a certain role was acquired.

Each user positively interfered in other user opinions, precisely grasping the situation of all opinions with the output interface. Our function that inputs opinions promotes mutual communication by displaying all the user opinions and forced movements on the proposed interface.

4 Related Research

In this section, we review the related research of opinion exchange support systems from a viewpoint of communication support and information visualization.

4.1 Communication Support

FreeWalk^[4], which communicates with other users with sound in three-dimensional virtual space, is expressed as a three-dimensional object in which the user himself is placed in a camera shot. Photo Chat^[6] communicates by writing shared comments on each user's view image. Although these support real-time communication in shared distant places, they are aimed at cases where time can't be shared.

In our research, even though we didn't use text information as a medium for opinion input but used pictures instead, all opinions can be intuitively grasped and visually understood. Moreover, in asynchronous opinion exchanges, they are also tied to time shortening when checking an opinion's input history.

Although some research^[5] aims for communication by sending a picture using a virtual drawer and other research^[3] supports communication using a face mark between people whose native languages are different, we seek opinion exchanges that give meaning to movement and picture arrangement rather than reveal the picture contents as a message.

4.2 Reappearance History

One work [2] visualizes user look history to replay user action history and transmits view skills to it. By replaying action history as it is, unconveyed skills are transmitted in language, suggesting that the feeling and the situation at that time can be grasped by seeing and experiencing the actions of others.

In our research, we built an opinion input situation in which opinion exchange participants replayed the movement history of pictures used as opinion input. The environment can be visually checked where the history changes of all participant opinions are set to support opinion exchanges.

Moreover, although another research [1] visualizes the history of who talked to whom within a certain group, our research supports opinion exchanges based on the opinions of all the members in a group without anonymously concentrating on a specific individual.

5 Conclusion

We proposed a system that supports opinion exchanges by opinion input that arranges candidate pictures for suitable roles to problems determined by mutual division of roles. With our proposed interface, users can replay opinion input histories by animation and intuitively understand the entire current opinion situation as well as the process underlying the results. Our experiment showed opinion exchanges with this active system. An effect supported the convergence of opinions about the division of roles.

References

- [1] Beale, M., Einstein, M., McCrickard, S., North, C., Saraiya: Visualizing Communication Timelines Containing Sparsely Distributed Clusters. In: Proc. of the IEEE Symposium on Information Visualization (InfoVis 2001), pp. 16–19 (2001)
- [2] Yamaguchi, T., Nakamura, T., Sunayama, W., Yachida, M.: Mirror Agent: An Interface Agent that Mirrors and Supports User's Behaviors by Visualizing Gazing Lines. In: Proc. of International Conference on Human-Computer Interaction(HCI 2001), vol. 4, pp. 394–398 (2001)
- [3] Itou, J., Hoshio, K., Munemori, J.: A Prototype of a Chat System Using Message Driven and Interactive Actions Character. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 212–218. Springer, Heidelberg (2006)
- [4] Nakanishi, H., Yoshida, C., Nishimura, T., Ishida, T.: FreeWalk: A 3D Virtual Space for Casual Meetings. IEEE MultiMedia 6(2), 20–28 (1999)
- [5] Siio, I., Rowan, J., Mynatt, E.: Peek-A-Drawer: Communication by Furniture. In: Proc. of the ACM Conference on Human Factors in Computing System (CHI 2002), vol. 2, pp. 582–583 (2002)
- [6] Sumi, Y., Ito, J., Nishida, T.: PhotoChat: Communication Support System based on Sharing Photos and Notes. In: Proc. of the ACM Conference on Human Factors in Computing Systems (CHI 2008), pp. 3237–3242 (2008)

Extracting Promising Sequential Patterns from RFID Data Using the LCM Sequence

Takanobu Nakahara¹, Takeaki Uno², and Katsutoshi Yada³

¹ Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka, Japan
nakapara@kansai-u.ac.jp

² National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan
uno@nii.jp

³ Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka, Japan
yada@kansai-u.ac.jp

Abstract. Recently, supermarkets have been using RFID tags attached to shopping carts to track customers' in-store movements and to collect data on their paths. Path data obtained from customers' movements recorded in a spatial configuration contain valuable information for marketing. Customers' purchase behavior and their in-store movements can be analyzed not only by using path data but also by combining it with POS data. However, the volume of path data is very large, since the position of a cart is updated every second. Therefore, an efficient algorithm must be used to handle these data. In this paper, we apply LCMseq to shopping path data to extract promising sequential patterns with the purpose of comparing prime customers' in-store movements with those of general customers. LCMseq is an efficient algorithm for enumerating all frequent sequence patterns. Finally, we construct a decision tree model using the extracted patterns to determine prime customers' in-store movements.

Keywords: Sequential pattern mining, Path data, LCM sequence, Decision tree, Data mining.

1 Introduction

In an attempt to understand customer purchasing behavior in the retail industry, much research has been carried out on customer purchase history data, such as POS data with IDs. There have also been attempts in recent years to attach RFID tags to shopping carts to determine customer in-store movement behavior by collecting customer tracking data called RFID data [1], [2]. Such research uses POS data with IDs combined with RFID data, making it possible to identify which customers bought what products at what prices and when via what movement paths. However, because RFID data is updated every seconds as the coordinates of the cart position change, there is a huge volume of data. Therefore, it is necessary to analyze the data by using an efficient algorithm or method. In this study we use an algorithm called the linear time closed itemset

miner sequence (LCMseq) [3], [4], which can enumerate frequent sequential patterns with high efficiency. We use LCMseq to extract patterns that can identify the movement paths and purchasing behavior of prime customers and other customers. Analyzing the sequences makes it possible to clarify shopping behavior that was previously unobtainable: in what order customers buy things, under what circumstances they exhibit indecisive behavior, etc. Finally, we use the extracted patterns to build a decision tree model and identify the characteristic shopping behavior of prime customers. We then show that one can use the patterns extracted from as explanatory variables to improve the accuracy of the model.

2 Analysis Data and Basic Analysis

The RFID data used in this study were obtained from one store of a supermarket chain in Japan. The data acquisition period was from 30th September to 31st October 2008. The data obtained from RFID tags attached to shopping carts were used to determine the movement behavior of approximately 1,000 customers. Note that the RFID tags identify the customers using each shopping cart; thus, only data on customers who used a cart were saved. On the other hand, customer purchase history data for 4 months from July to October 2008 in that store were available, enabling the use of purchase data of approximately 25,000 people. Among the customers in the RFID data, there were some customers who could not be merged with purchase history data; therefore, we finally analyzed 625 customers who could be identified from both data sets. Thus, we were able to use both RFID data and purchase history data for these customers.

Figure 1 shows the layout of the store, which is classified into 16 sections. As can be seen from the layout, this store has carts placed in the upper left corner, and a product layout with the vegetable section (V) next to the section of daily foods such as tofu and fermented soybeans (G), next to the meat section (M), followed by the seafood section (F), and the precooked food section (Z), with customers finally reaching the cash registers (R) in the top center. This layout is the same as that found in most supermarkets in Japan.

Our basic analysis shows a large gender bias in the customers, with approximately 90% females. Male and female customers are also primarily age 30 to 69. For the customers in the RFID data, slightly more are age 60-69 than age 30-39. This data was obtained from shopping cart users, which may be why it tends to have more older users than young users (in Japan, some shoppers use hand-carried baskets). As a similar tendency, customers in the RFID data tended to have a slightly higher than average purchase value per shopping trip. However, no significant difference was seen between customers in the RFID data and customers in the customer purchase history data; therefore, customers in the customer in-store RFID data are usable as a sample for this analysis.

Next, regarding the product sections in the store, each section has several characteristics. The vegetable section had the highest number of visitors and the highest purchase ratio of approximate 78% (purchasers/visitors) among the

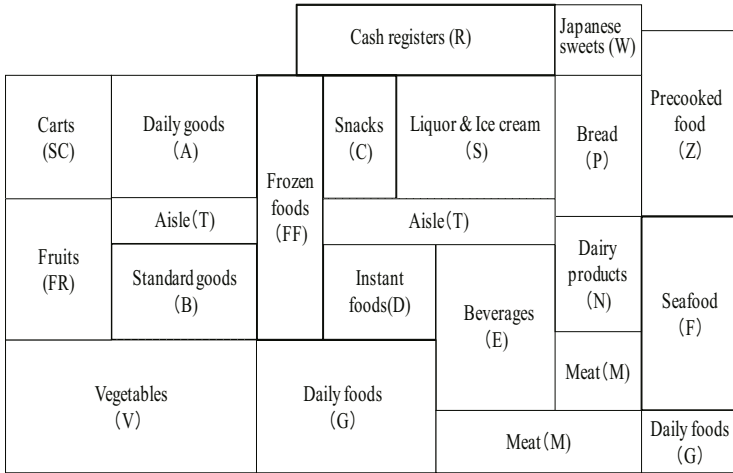


Fig. 1. Layout of store

16 sections. Following the vegetable section, the daily foods section and meat section had the next highest numbers of visitors and purchase ratios. Although the precooked food section had the fifth highest number of visitors, its purchase ratio was tenth at approximately 38%, showing that for this section there is a low correlation between the number of visitors and purchases. Sections such as this, where many customers visit but the purchase ratio is low, are likely to be used as a route to the cash registers, and improvements to the product line and other aspects have the potential to result in higher sales. On the other hand, the bread section is 13th in terms of the number of visitors, but its purchase ratio is fourth at approximately 56%, making it a section with relatively few visitors but most of whom purchase something. It appears that customers visiting this type of section already intend to make a purchase when visiting.

Figure 2 is a directed graph showing movement routes in the store. The nodes represent each section, and the paths linking each section show the numbers of movements between sections. The node size expresses the relative frequency of visits to the section, and the thickness of the path linking sections expresses the frequency of movements between sections. Here, the frequency represents the number of times the customers stops at the section, and only routes with at least 5% of the maximum movement frequency are drawn. A relatively common route is from the cart section, to the vegetables, daily foods, meat, seafood, precooked food, Japanese sweets, and ending at the cash registers. This confirms that many customers start from the cart section and move in a counterclockwise circle around the outer edge of the store. Moreover, a relatively common inner circular route is from the vegetable section, to the standard goods section, then to frozen foods. However, the inner circle routes have lower movement frequencies than the outer circle route; therefore, increasing the number of movements along such routes is important for boosting sales.

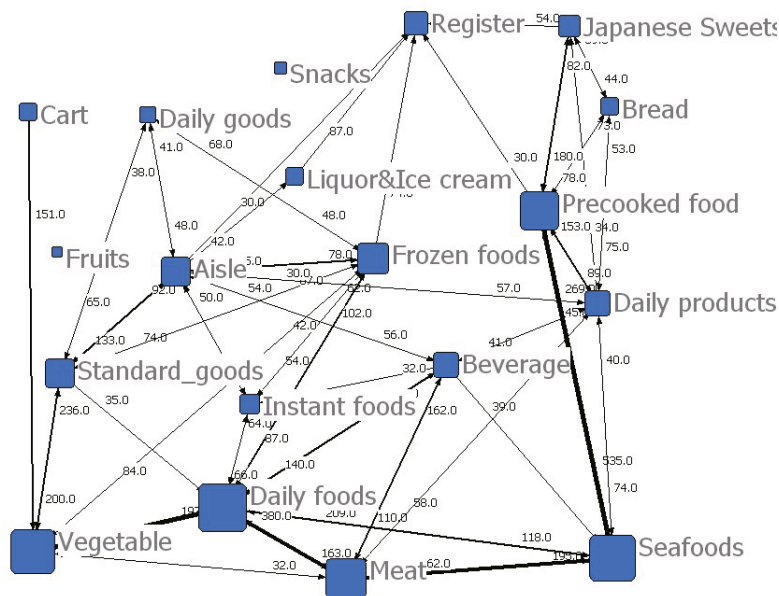


Fig. 2. In-store movements

From these basic analyses, we conclude that there are sections where customers already intend to make a purchase when visiting, sections where customers consider purchasing after seeing the products, and sections visited on the way to the cash registers. Sections visited on the way to the cash registers have many visitors, and linking this to purchases at these sections is important for boosting sales.

In this study we classify customers into high-value purchasing customers (prime customers) and other customers (general customers). We identify differences in the in-store purchasing behavior of these two groups from the perspective of movement routes. We then use LCMseq to determine the characteristic shopping behavior of prime customers.

3 Using LCMseq to Find Characteristic Movement Routes

3.1 Determining Customer Groups to Analyze

In classifying the 625 customers into prime customers and general customers, we used the purchase history data for the 3 months before obtaining RFID data, and performed a decile analysis on the purchase values of all customers. Table 1 shows the range of purchase values in each rank obtained when evenly dividing the customers into deciles. There were approximately 2,220 customers in each rank. The numbers of customers in the table show the number of customers

Table 1. Purchase Value Deciles

Decile Rank@	Total Purchase Value in 3 Months	# of Customers
1	>46,974	285
2	26,345-46,973	116
3	16,245-26,344	73
4	10,245-16,244	52
5	6,552-10,244	32
6	4,155-6,551	20
7	2,612-4,154	12
8	1,573-2,611	6
9	807-1,572	6
10	<806	5

in the RFID data in each rank, with 607 of the 625 customers classified into ranks. The remaining 18 customers only came to the store during the tracking data collection period, so they were removed from the customers to be analyzed. From the relation between customer numbers and purchase value, we performed an analysis on the 285 high-value customers belonging to rank 1, referred to as prime customers, and the remaining 322 customers, referred to as general customers. Prime customers spend approximately an average of 15,000 yen or more per month. Hereafter, we only use data from the RFID data collection period to compare prime customers and general customers and find characteristic movement behaviors of each customer group.

3.2 Application of LCMseq

LCMseq is an efficient algorithm for enumerating frequent sequence patterns from a sequential database. In addition to its high speed, LCMseq can be applied in a variety of ways, as it can assign a positive or negative weight to each sequence and only extract frequent sequence patterns that appear in a specified window width.

For an alphabet Σ , the set of finite sequences on Σ is expressed by Σ^* . A sequence pattern is an arbitrary sequence $s = a_1 \cdots a_n \in \Sigma^*$, and $P = \Sigma^*$ expresses the set of all sequence patterns on Σ . The sequence database on Σ is the sequence set $S = \{s_1, \dots, s_m\}$. We denote the size of S by $|S|$. For sequence pattern $p \in P$, a sequence database including p is called an *occurrence* of p . The *denotation* of p , denoted by $\tau(p)$ is the set of the occurrences of p . $|\tau(p)|$ is called the *frequency* of p , and denoted by $frq(p)$. For given constant $\sigma \in \mathbb{N}$, called a *minimum support*, sequence pattern p is *frequent* if $frq(p) \geq \sigma$.

When using frequency patterns in the classification of data, a pattern that often appears in one set but not often in another set can be used as a pattern to characterize two sets. In this case, a pattern p in the two sets with a difference in the appearance frequency at or above a certain threshold is called a contrast pattern [5], and a pattern p with an appearance frequency ratio at or above a certain threshold is called an emerging pattern [6].

In this study, when extracting frequency sequence patterns from RFID data, the sections visited starting from the cart section until reaching the cash register are expressed as a time series; thus, customer movement behavior regarding the sections visited is handled as a sequence database S . Because 607 customers are analyzed, $|S| = 607$. Next, the sequence data of the prime customer set are given weight w_h , and the sequence data of the general customer set are given weight w_g .

Usually, when frequency patterns are enumerated, all sequences are regarded as equivalent, and weights are calculated using unit costs. However, in LCMseq, it is possible to give each sequence a different weight to extract patterns. In this case, it is possible to extract sequence patterns with a difference between $\sum_{s \in Hc} w_h$ and $\sum_{s \in Gc} w_g$ that is at least $minDiff$. Here, Hc and Gc signify sub-sequence sets including pattern p for prime customer sequence data and general customer sequence data, respectively. Therefore, this is equivalent to extracting a contrast pattern.

When performing pattern extraction by this method, if the numbers of elements in the sets differ, then problems arise. Consider, for example, if we use unit costs in weights, and a pattern appears in all sequences. This pattern is completely contained in both sets, thus it is not a characteristic pattern set in one of the sets. However it is treated as a characteristic pattern in the set with the larger number of elements. To resolve this problem, we introduce the weights $w_h = 1/|Hc|$ and $w_g = 1/|Gc|$; thus, $\sum_{s \in Hc} w_h$ and $\sum_{s \in Gc} w_g$ each range from 0 to 1, and even if a pattern appears in all sequences, their differences between $\sum_{s \in Hc} w_h$ and $\sum_{s \in Gc} w_g$ will be 0. Therefore, it becomes possible to extract a characteristic pattern that does not depend on the number of elements. In this study, we also use a window width ($1 \leq win \leq n$), which is another function of LCMseq, so that after a customer visits a certain section, we limit the movement behavior to sections within the window width, in an attempt to extract characteristic sequence patterns of the two customer sets.

4 Calculation Results

4.1 Extraction of Sequence Patterns

Figure 3 shows the sequence patterns extracted by LCMseq with $minDiff \geq 0.05$. The patterns for which the frequency does not change by attaching an item at their ends are removed since they can be considered as redundant patterns. Each point corresponds to one sequence pattern, and the point's color expresses the value of each window width win . "Other" consists of the patterns extracted when the window width is n , equivalent to the condition where the window width restriction is removed. A total of 9,622 sequence patterns were extracted.

The diagram's horizontal axis expresses the total weight of general customers, and the vertical axis expresses the total weight of prime customers. If we draw a 45 degree line passing through the origin, points above the line represent characteristic sequence patterns in the prime customer group, whereas those below the line represent characteristic sequence patterns in the general customer group.

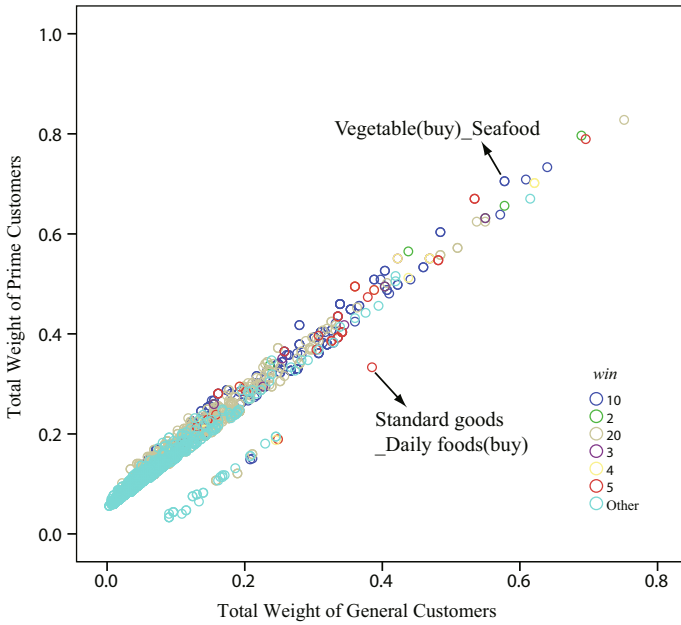


Fig. 3. Extracted sequence patterns

The figure shows there are more sequence patterns corresponding to prime customers; thus, it appears that prime customers exhibit more diverse purchasing behavior than general customers. To use sequence patterns to distinguish the characteristics of customer groups, we require sequence patterns for which the total weight of one customer group is high, and that of the other group is low. For example, “Standard goods_daily foods (buy)” is a pattern when the window width is 5; this pattern signifies that after visiting the standard goods section the customer visited and bought product(s) at the daily goods section among five possible sections. For this pattern, general customers have a large weight, making it one of the characteristic patterns of general customers. On the other hand, “Vegetable (buy)_Seafood” is a pattern with a window width of 10, in which after visiting and buying at the vegetable section, the customer visited the fish section among ten possible sections. For this pattern, prime customers have a total weight of 0.7 and the general customer weight is 0.57; the difference of approximately 0.13 makes this a pattern with a relatively strong ability to distinguish between the two types of customers. Note that the sample patterns include sections only visited and sections where a purchase was made, i.e., patterns that cannot be obtained solely from purchase history such as POS data. However, these patterns are obtainable by using RFID data.

4.2 Decision Tree Analysis Using Sequence Patterns

We used the sequence patterns extracted by LCMseq to perform decision tree analysis and clarify the characteristics of prime customers and general customers.

Then, among the approximately 10,000 extracted patterns, we set a large *minDiff* and extracted 137 patterns with *minDiff* ≥ 0.1 as characteristic patterns associated with prime customers. However, among the patterns characteristic of general customers, the maximum value of *minDiff* was 0.07, and there were only 36 patterns characteristic of general customers; therefore, we used all the sequence patterns associated with general customers. In the decision tree analysis, 173 sequence patterns are used as 0,1 dummy variables. In addition to these, we also use the 12 explanatory variables showing in table 2.

Table 2. Used explanatory variables (exclude patterns)

	Explanatory variables@
1	The number of visits per section
2	The visit time ratio by section
3	The purchase value by section
4	The number of sections where a purchase was made
5	The number of section visited
6	The purchase/visit ratio
7	The total purchase time (minutes)
8	The total number of visits
9	The total purchase value
10	The total purchase quantity
11	The average purchase value
12	The average purchase quantity

Figure 4 shows the model generated by the decision tree. As a result of performing 10-fold cross-validation, the model’s accuracy was 62.43%. When we generated the model without containing sequence patterns, a similar cross-validation gave 57.83% accuracy; thus, including the extracted sequence patterns improved its accuracy.

Pattern (1) shows a pattern in which the sections visited are beverages, then frozen foods, then the cash register; this is a characteristic pattern that identifies prime customers. The label (N/A) in each leaf indicates the number of customers who are distinguished correctly (N) and the total numbers of customers included in the leaf (A). This pattern has a relatively high distinction rate of 68.7%. Interpreting this pattern, these appears to be a general purchasing trend in which customers visit the beverage section followed by the frozen foods section when buying temperature-sensitive products, before going to the cash register; it appears that customers who exhibit such behavior are often prime customers. Next, pattern (2) expresses visits to the meat section, to the seafood section twice, and then to the dairy section. The total shopping value per visit to a section is rather high (above 822 yen), and cases with this pattern tends to be characteristic of prime customers. The fresh food sections contribute the most to supermarket revenue, and visiting them multiple times is important to distinguish prime customers. It may be that customers with this pattern are

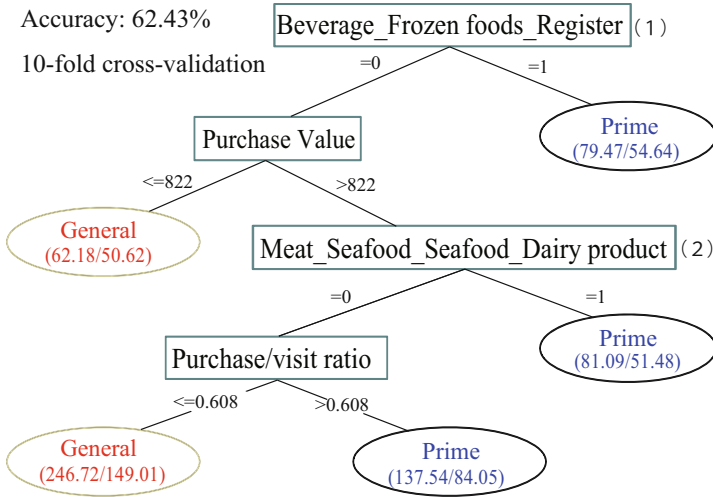


Fig. 4. Decision tree model for prime customers & general customers

loyal to this store. Finally, the lowest branch is the purchase/visit ratio, which is the number of sections where purchases were made divided by the number of sections visited. If this value is high, the customer has a strong tendency to buy at the sections stopped at. In contrast, a low value indicates that even if the customer stops, he/she does not make a purchase, suggesting a tendency of confusion and indecision when shopping. By interpreting the decision tree based on these patterns, we can summarize the characteristics of prime customers as follows.

- Customers who exhibit the general purchase behavior of visiting the beverages section, directly followed by the frozen foods section, and then the cash registers.
- Customers with a reasonably high purchase value who visit the fresh food sections several times and may be loyal to the store.
- Customers whose purchase value is rather high and who often buy products when stopping.

In contrast, the characteristics of general customers are as follows.

- Customers with a low purchase value.
- Customers whose purchase value is rather high but who exhibit indecision purchasing behavior at the sections stopped at.

Clarifying the characteristics of customers in this way makes it possible to determine the movement behavior and purchasing tendencies of prime customers and general customers in stores. In this study, such characteristics were clarified for the first time using RFID data, which enable the clarification of behavior other than purchasing, which is difficult using only purchase history data.

5 Conclusion

In this study we used RFID data on in-store movement behavior and purchase history data with LCMseq to extract characteristic sequence patterns among prime customers and general customers. These patterns were also used as explanatory variables in decision tree analysis; using the patterns improved the distinguishing accuracy by approximately 5% compared with not using these patterns. The example of applying RFID data shown in this study demonstrated the usability of sequence patterns obtained from RFID data, and showed that it is effective to use sequence patterns to extract characteristics and to use the extracted characteristics in classification problems. The results obtained in this study were clearly understandable, and enabled us to understand the characteristics of prime customers and general customers. However, issues remain concerning the potential business applications of the results obtained. Although the characteristics of prime customers were clarified, we were unable to clarify how they should be utilized in business, nor whether this will ultimately lead to higher sales. We aim to address these points in the future.

References

1. Larson, J.S., Bradlow, E.T., Fader, P.S.: An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing* 22(4), 395–414 (2005)
2. Yada, K.: String analysis technique for shopping path in a supermarket. *Journal of Intelligent Information Systems* (2009)
3. Ohtani, H., Kida, T., Uno, T., Arimura, H.: Efficient serial episode mining with minimal occurrences. In: *Proceedings of the third ICUIMC*, pp. 457–464. ACM Press, New York (2009)
4. <http://research.nii.ac.jp/~uno/code/LCMseq.htm>
5. Bay, S.D., Pazzani, M.J.: Detecting change in categorical data: Mining contrast sets. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 302–306. ACM Press, New York (1999)
6. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 43–52. ACM Press, New York (1999)

Relation between Stay-Time and Purchase Probability Based on RFID Data in a Japanese Supermarket

Keiji Takai¹ and Katsutoshi Yada²

¹ Data Mining Laboratory, Kansai University, Suita, 564-8680, Japan
r098032@ipcku.kansai-u.ac.jp

² Data Mining Laboratory, Kansai University, Suita, 564-8680, Japan
yada@kansai-u.ac.jp

<http://www2.ipcku.kansai-u.ac.jp/~yada/>

Abstract. Radio Frequency Identification (RFID) technology uses radio waves to track an object to which a small tag is attached. In a Japanese supermarket, we attach the RFID device to the cart and collect data on purchase behavior. In this article, we clarify the relation between purchase probability and the time customers spend in the store section by analyzing the RFID data with main use of descriptive methods. We clarify the way how the stay-time explains the purchase probability and characteristics of each store section. Based on the result, some implications for business are made as well.

1 Introduction

Radio Frequency Identification (RFID for short) is the technology to use a small tag for tracking an object with radio waves. The trace of an object to which the small tag is attached is remotely recorded. This technology has been widely used, for example, in supply chain management and libraries, and has recently attracted much attention. RFID tags give us much information which was unavailable before.

In a supermarket, this RFID technology has also brought us much information on customer purchase behavior such as the location of a customer, the place where a customer stops and the time a customer spends there. Combined with POS data, the RFID data provide us with more minute information on purchase behavior. We use this technology in a supermarket in Japan and collect the data of customers. The tags of RFID are attached to carts. Various information is recorded about the customer who use the cart. Using POS data and RFID data, we have information on the location, the time of a customer at each location, what the customer buys and does not buy and so on.

In this article, we focus our interest on the relation of the time a customer spends and the purchase probability. Especially, our interest is in the relations for each section and the comparison among them. The article [2] and the past research [3] have proposed a model to explain customers' behavior with a hierarchical Bayesian method. This article shall be a basis for a springboard to

construct a frequentist modeling of customers' purchase behavior instead of a Bayesian method.

This article is organized as follows. In the next section, we explain the details on the data to be analyzed. The names of sections and their characteristics are given. In Section 3, we classify the distribution of the people over the time spent in the section. The relation of time to stay in the section to the purchase probability is explained using a linear regression model. In Section 4, we form a conclusion on the analysis and provide some implications for business.

2 The RFID Data

The data are collected in a supermarket in Japan. Most of the carts in the supermarket have RFID devices attached. The path of a customer using the cart is recorded. Also, the record of what the customer buys is registered. Although the RFID device is not necessarily accurate enough to report the precise location, it is capable of reporting a rough location.

We divide the supermarket into 28 sections as shown in Figure 1. The sections are Household 1 to 3, Food 1 to 6, Snacks & Sweets 1 to 3, Liquor 1 and 2, Entrance, Seafood 1 and 2, Prepared Food, Central Aisle, Western Deli, Japanese Deli, Frozen Food, Drinks, Meat, Register, Event Space, and Fresh Produce 1 and 2. Examples of products sold in Household 1 are cat food and household cleaning products. In Household 2 are diapers, napkins and laundry products. In Household 3 are shampoos, toiletry, hair dyes. In Food 1 is cup noodles. In Food 2 are miso (Japanese traditional soybeans paste) and soup stocks. In

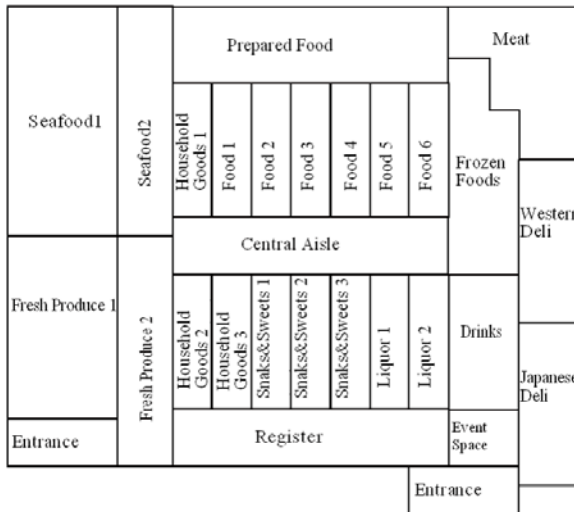


Fig. 1. Store sections: the store is divided into 28 sections. Entrance, Central Aisle and Register and Event Space are excluded from our analysis.

Food 3 are mayonnaise and soy sauce. In Food 4 are retort curry and spaghetti. In Food 5 are coffee, cocoa and vegetable juice. In Snacks & Sweets 1 are chocolate, gums, low-end sweets. In Snacks & Sweets 2 are snacks such as crisps. In Snacks & Sweets 3 are rice crackers, jelly, appetizers. Liquor 1 deals with high-end alcohol, while Liquor 2 deals with beer and Shochu (distilled spirit)-based beverages. Seafood 1 sells, for example, sashimi, sushi and seaweeds. In Western Deli, western delicatessen products such as cheese and margarine are sold while in Japanese deli Japanese delicatessen products such as Tofu, Natto (fermented beans) and Takuan (Japanese pickles). Drinks stocks soft drinks. Fresh Produce 1 sells vegetables, while Fresh Produce 2 sells fruits.

After processing the data, the number of customers is 6977 records. The variables we use for analysis are the time spent in each section measured by the second (which we call “stay-time” below) and a binary purchase indicator indicating whether the customer buys at least one product in a given section. Some of the sections are excluded from our analysis. Entrance, Central Aisle and Register are removed because they have no relation with the purchase probability. Event Space is also removed since only three customers bought products in the section. This may be because most of customers visit the section without using carts. Thus, we have data consisting of 48 variables (the stay-times and purchase indicators for 24 sections).

3 Analysis of RFID Data

In this section, we analyze the data using the statistical software R [6]. We are interested in describing how the stay-time is related to the purchase probability in each section. By clarifying this relation, we can use it not only to promote the purchase but also to obtain a deeper insight into the purchase behavior related to the stay-time in a section.

Instead of considering the joint distribution of stay-times and purchase indicators in all the sections, the joint distribution of the stay-time and the purchase indicator in each section is employed. We compare the distributions of the sections and probe the relation of the stay-time to the purchase probability. In what follows, we use the notation X to indicate the purchase indicator, Y to indicate the stay-time, and $P(\cdot)$ to indicate the probability of the event in the parentheses. We define $X = 1$ if a purchase occurs and $X = 0$ otherwise. Fundamentally, there should be variables (X, Y) for each section and, thus, we should use, for example, $(X_1, Y_1), \dots, (X_{24}, Y_{24})$. For notational simplicity, we suppress the suffix and use (X, Y) .

3.1 Basic Statistics

Table 1 shows the means and standard deviations of the groups of customers buying something and not buying anything and the results of the statistical

Table 1. This table shows the means, standard deviations (std.dev) and the sizes of the groups of the customers who do not buy anything and who buy something in each section.

	not buying anything			buying something			Reject
	mean	std.dev	size	mean	std.dev	size	
Food 1	22.61	25.63	485	.93	9.65	6492	Yes
Food 2	38.24	43.87	1553	3.19	13.3	5424	Yes
Food 3	43.75	56.02	1673	3.51	13.13	5304	Yes
Food 4	37.09	42.74	1163	3.64	13.63	5814	Yes
Food 5	9.72	11.58	175	.23	2.24	6802	Yes
Food 6	59.29	63.06	3235	16.53	37.08	3742	Yes
Drinks	23.31	26.00	1444	7.71	18.22	5533	Yes
Snacks & Sweets 1	61.27	81.81	515	4.68	33.02	6462	Yes
Snacks & Sweets 2	34.41	35.38	912	3.07	13.57	6065	Yes
Snacks & Sweets 3	53.21	57.84	879	2.57	13.49	6098	Yes
Liquor 1	30.78	47.58	134	.98	6.13	6843	Yes
Liquor 2	26.16	33.91	678	3.08	12.41	6299	Yes
Household Goods 1	42.27	50.68	445	1.59	12.39	6532	Yes
Household Goods 2	40.29	84.83	319	1.75	12.78	6658	Yes
Household Goods 3	43.38	52.38	260	1.28	10.66	6717	Yes
Meat	28.72	33.52	3205	3.59	15.42	3772	Yes
Fresh Produce 1	74.99	68.48	5499	23.41	44.79	1478	Yes
Fresh Produce 2	55.04	55.02	3771	19.21	33.66	3206	Yes
Seafood 1	44.25	48.00	3320	8.24	22.18	3657	Yes
Seafood 2	55.86	57.74	3269	22.61	41.45	3708	Yes
Prepared Food	44.28	53.27	2982	16.9	29.22	3995	Yes
Western Deli	20.74	23.35	750	6.07	16.10	6227	Yes
Frozen Food	33.47	37.19	832	7.42	16.84	6145	Yes
Japanese Deli	31.93	33.67	4058	4.77	15.51	2919	Yes

test. In order to advance our discussion, our first priority is to examine whether stay-time significantly changes between the group of customers buying something and the group leaving the section without buying anything. Thus, our question is whether there is a difference in the mean of stay-time between those groups. To examine this, a statistical test is conducted using the central limit theorem¹. The null hypothesis here is $E[Y|X = 1] = E[Y|X = 0]$, where $E[Y|X = 1]$ means the expectation of Y under the condition $X = 1$ and $E[Y|X = 0]$ is similarly defined. In the rightmost column, the result is given. All the null hypotheses are rejected with a significant level 5%. This means that the mean of the customers buying something significantly differs from that of those not buying anything in each section. The result partially supports our supposition that longer stay-time affects purchase probability.

¹ Though it is theoretically necessary for the distribution of stay-time to meet some conditions for application of the central limit theorem, we assume that the conditions hold to avoid discussing the technical details.

3.2 Classification of Stay-Time Distribution

To investigate the characteristics of the distributions of stay-time, $P(Y|Y > 0)$, in each section, we classify them using cluster analysis (See the book [4] for details). To conduct cluster analysis, there are two issues to discuss. The first issue is the distance between the distributions. The second is the interval to be used to construct the distributions.

The distance must be defined between two distributions. One of the most popular distances is “Euclidian distance.” In this example, such a distance cannot be applied. This is because the data have a special characteristic in that their sum is one. This type of data is called “compositional data.” See the book [1] on the theoretical details and the defects by use of the usual distances. We use the distance introduced in the article [5], which defines the distance between the two distributions, $P = [p_1, \dots, p_K]$ and $Q = [q_1, \dots, q_K]$ as

$$D(P, Q) = \sum_{k=1}^K (\sqrt{p_k} - \sqrt{q_k})^2,$$

where $\sum_{j=1}^K p_j = \sum_{j=1}^K q_j = 1$. This distance features symmetry between the two distribution, P and Q . With this distance, we have a symmetric distance matrix between any two of the purchase probability distributions.

The remaining issue is the decision about the interval for creating the distribution of stay-time for each section. The distribution data used for the cluster method consist of the stay-time intervals and the probability of purchasing belonging to those intervals. The issue is how to decide the interval. In this case, we take the way to determine the interval to make the classification of the distribution as clear as possible. From this perspective, it is found that the interval of 2 seconds is best to classify the distributions.

The resulting classification is given in Table 2. The first group consists of short stay-time customers in the section. A typical distribution to Group 1 is given in Figure 2. Those who belong to this group have tendency to leave the section within a relatively short period. The second group is composed of the sections where the customers spend a moderate stay-time compared to Groups 1 and 3. A typical distribution of the group is that of “Drinks” as given in Figure 3. The third group consist of the sections where the customers tend to stay longer than the sections of the other groups. A typical distribution is that of Food 6 as shown in Figure 4.

Next, we explore the relation of purchase probability to the stay-time. The plot of purchase probability along stay-time in Food 1 is shown in Figure 5. This plot indicates a special feature. It is that the purchase probability monotonously goes up to one as stay-time becomes longer. All the other sections, which we do not show here due to limitation of the space, have the same feature: the purchase probability monotonously goes up to one along stay-time. A difference among the distributions is the purchase probability at the beginning second and the

Table 2. Classification of Distributions: Group 1 consists of the sections where customers tend to stay shortly. Group 2 consists of the sections where customers stay neither longer nor shortly. Group 3 consists of the sections where customers tend to stay longer.

Sections in Group 1	Section in Group 2	Sections in Group 3
Food 1	Drinks	Food 6
Food 5	Food 2	Fresh Produce 1
Liquor 2	Food 3	Fresh Produce 2
Liquor 1	Food 4	Seafood 1
Western Deli	Frozen Foods	Seafood 2
	Household Goods 1	Snacks&Sweets 1
	Household Goods 2	
	Household Goods 3	
	Japanese Deli	
	Meat	
	Prepared Food	
	Snacks & Sweets 2	
	Snacks & Sweets 3	

change in the probability rates along stay-time. To analyze this, we introduce a following model for each section,

$$P(X = 1|Y = y) = \beta_0 + \beta_1 y. \tag{1}$$

This model aims to explain the relation between the purchase probability $P(X = 1|Y = y)$ and stay-time y . Although the linearity may not be enough to capture the exact relation and it causes some difficulties in estimating the parameters, still it is sufficient to approximately explain the relation which we want to know when the estimation can be made. The parameters appearing in this model are the intercept β_0 and the slope β_1 . When standardizing the data, this model explains the correlation between $P(X = 1|Y = y)$ and y . More importantly, when not standardizing the data, the parameters in this model have special interpretations. The first parameter explains the basic purchase probability more than zero seconds in the section. This probability means that a customer who enters the section buys something with more than that probability if we assume that the purchase probability continues to increase in accordance with the stay-time. The second parameter indicates the probability increase per one second. If we make a customer stay one second longer in the section by some sales promotion, it is expected that the purchase probability increases by β_1 . Therefore, it is important not to standardize the data.

In applying this model, we make a selection of the data in each section. When using all the data of the customers whose stay-time is more than zero, the regression parameters do not give the information we need to describe the behavior of purchase probability in line with stay-time. In Food 1, for example, the purchase probability distribution as shown in Figure 5, becomes one after a certain point

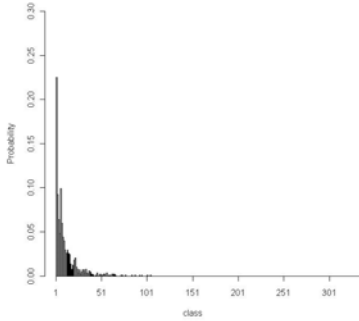


Fig. 2. Stay-time Distribution of Food 1: customers tend to leave the section during earlier seconds

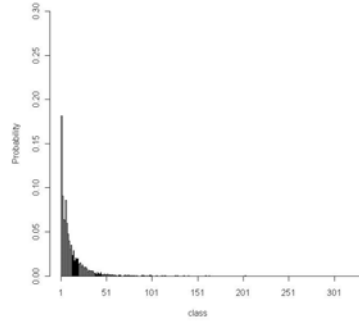


Fig. 3. Stay-time Distribution of Drinks: customers tend to stay for a moderate time compared to the other distributions

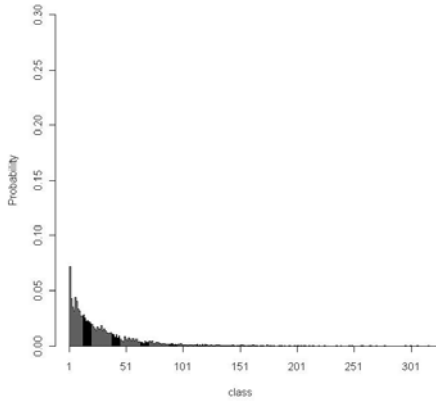


Fig. 4. Stay-time Distribution of Food 6: customers tend to stay in the section longer

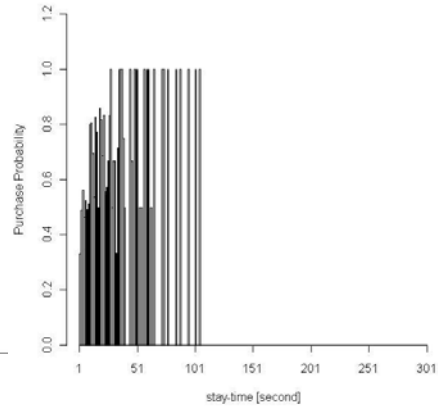


Fig. 5. Purchase probability Distribution of Food 1: After a point passes (here 18 seconds), the purchase probability connectively becomes one

in stay-time. In any section, the purchase probability becomes consecutive ones after a specific stay-time. We call this point of stay-time Time. This shows that a few customers stay in the section after that point of stay-time, and still buy something with probability one, and therefore the purchase probability becomes one in most of the following stay-times. We call the maximum stay-time one second before achieving a purchase probability of one “Time.” For instance, in the Food 1 section, Time is 18 seconds. Over 18 seconds, the purchase probability is one. Therefore, we remove the data that exceeds Time in order to explain the relation between stay-time and purchase probability.

Table 3. The results of each section: Time is the time one second before the purchase probability becomes one. Cor is the Pearson correlation between the stay-time and the purchase probability. (β_0, β_1) are the parameters of the linear regression model given in equation 1. The column named 5 sec gives the predicted increase in the purchase probability by lengthening the stay-time by 5 seconds. The column named 10% next to 5 sec gives the purchase probability for lengthening the stay-time by 10% of Time.

	Time	Cor	β_0	β_1	5 sec	10%
Food 1	18	.70	.25	.029	14.49	5.22
Food 2	48	.74	.40	.009	4.65	4.46
Food 3	62	.70	.48	.006	3.04	3.77
Food 4	51	.67	.32	.007	3.68	3.76
Food 5	18	-.01	.40	.000	-.19	-.07
Food 6	92	.69	.47	.004	1.99	3.67
Drinks	61	.55	.29	.005	2.36	2.87
Snacks & Sweets 1	34	.69	.23	.011	5.57	3.79
Snacks & Sweets 2	53	.57	.35	.006	3.22	3.41
Snacks & Sweets 3	50	.72	.30	.009	4.57	4.57
Liquor 1	31	.42	.06	.007	3.73	2.31
Liquor 2	46	.45	.30	.005	2.51	2.31
Household Goods 1	32	.57	.25	.010	5.24	3.36
Household Goods 2	46	.29	.24	.005	2.27	2.09
Household Goods 3	27	.29	.21	.007	3.47	1.88
Meat	40	.55	.65	.007	3.68	2.95
Fresh Produce 1	66	.76	.59	.006	3.21	4.23
Fresh Produce 2	95	.69	.52	.003	1.72	3.28
Seafood 1	66	.62	.60	.004	2.15	2.84
Seafood 2	111	.59	.47	.003	1.35	2.99
Prepared Food	91	.73	.39	.000	2.36	4.29
Western Deli	49	.19	.20	.002	.78	.77
Japanese Deli	34	.61	.64	.010	4.98	3.38
Frozen Food	61	.65	.15	.005	2.73	3.33

Table 3 gives Time, the correlation of the stay-time and the purchase probability, the estimated regression parameters (β_0, β_1) , 5 sec and 10%, the latter two of which will be explained below. Time exceeds the median, which is given in this article, in each section. This indicates that more than half of the customers visiting the section leave there before Time.

In the next column, the correlation between the stay-time and the purchase probability is given for reference. The value is also the regression coefficient when regression analysis is conducted for standardized data. The correlation does not necessarily provide information helpful to understand the relation of purchase probability along stay-time. What is noteworthy about this correlation is that in Food 5 there is almost no correlation. This means that longer stay-time does not lead to increase in the purchase probability in Food 5.

The regression parameter β_0 is, as described earlier, the basic purchase probability. Once the customers enter the section, they buy something there with a

higher probability than the probability β_0 . The sections with high values of β_0 are, for example, Meat and Japanese Deli. Whenever customers go into those sections, they purchase something with high probability. The regression slope parameter β_1 indicates the predicted increase in the purchase probability by spending another second in the section. Food 1 has the highest slope .029 of all the sections. This means that in Food 1 making a customer stay one second longer is expected to increase the purchase probability by about three percent. The next column shows “5 sec.” This indicates that an increase in stay-time by 5 seconds in the section is expected to result in an increase in the purchase probability by the amount listed in the column. For instance, in Food 1, making stay-time 5 seconds longer results in a 14.49% increase of purchase probability. Lengthening the stay-time in each section by 5 seconds results in different stay-times in each section. Taking into account that Time differs from section to section, it turns out that the value of 5 sec differs from section to section. In other word, if we take a strategy to make a customer spend a longer time in a section, it would cost differently from section to section. For this reason, we compute the predicted increase when lengthening the stay-time by 10% of Time. These values are given in the rightmost column. These values explain how much the purchase probability is increased by lengthening the stay-time by 10% of Time.

4 Summary and Implication

In summary, we have classified the distributions $P(Y|Y > 0)$ into three groups. Each group differs in the stay-times of the sections belonging to the group. Next, we have described the data whose purchase probability is less than one, based on the model given in (II). It is found that some of the sections have relatively high purchase probabilities and a low expected purchase probability increase when making a customer spend 10% more Time.

Table 4 summarizes the results analyzed above. The rows of the table are the result of the cluster analysis. The columns are the result of roughly halving the basic probabilities into high probability groups and low probabilities groups.

From the Table 4, we draw two implications for business based on the confirmed covariate relations. The first one is to increase the purchase probability by increasing stay-time, while the second one is to decrease waste time of buying. In Group 1 and Group 2, making a customer stay longer in the section would lead to a higher purchase probability. Though the same thing is true in Group 3, it is better to try to make stay-time shorter while keeping the same basic probability. In the sections with a high basic probabilities, a customer buys something with the high probability once he or she goes into the section. For this reason, it will not pay off to make efforts to make customers stay longer in the section. It is better to take a strategy to make a customer stay longer in those sections with low basic probabilities. In Group 3, customers tend to stay longer than in the sections of any other group. It is possible that customers waste time in these sections. In these sections, we should find a way to decrease stay-time while

Table 4. Summary of the analysis: In each group decided by the cluster analysis, the sections are classified into the two groups. This classification is just obtained by roughly halving the sections belonging to the group. The sections are placed with high basic probability in the left column in each row, while the sections with low basic probability are placed in the right one. The probabilities in parentheses next to each section name are for $(\beta_0, 10\%)$.

		Basic Probability β_0	
		High	Low
cluster	Group 1	Food 5 (40,-.07)	Food 1 (25,5.22)
		Liquor 2 (30,2.31)	Western Deli (20,.77) Liquor 1 (.06,2.31)
	Group 2	Meat (65,2.95)	Food 4 (32,3.76)
		Japanese Deli (64,3.38)	Snacks & Sweets 3 (30,4.57)
		Food3 (48,3.77)	Drinks (29,2.87)
		Food 2 (40,4.46)	Household Goods 1 (25,3.36)
		Prepared Food (39,4.29)	Household Goods 2 (24,2.09)
		Snacks & Sweets 2 (35,3.41)	Household Goods 3 (21,1.88) Frozen Food (15,3.33)
	Group 3	Seafood 1 (60,2.84)	Food 6 (47,3.67)
		Fresh Produce 1 (59,4.23)	Seafood 2 (47,2.99)
Fresh Produce 2 (52,3.28)		Snacks & Sweets 1 (23,3.79)	

keeping the purchase probability as nearly to the current one as possible and then we should lead the customers there to one of the other sections belonging to Group 1 and Group 2.

References

1. Aitchison, J.: The statistical Analysis of Compositional Data. Chapman & Hall, London (1986)
2. Hui, S.K., Bradlow, E.T., Fader, P.S.: Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior. *Jour. Cons. Res.* 36, 478–493 (2009)
3. Hui, S.K., Fader, P., Bradlow, E.: Path data in marketing: an integrative framework and prospectus for model-building (August 7, 2007), SSRN, <http://ssrn.com/abstract=930141>
4. Izenman, A.J.: Modern Multivariate Statistical Techniques. Springer, New York (2008)
5. Matusita, K.: Decision rule, based on the distance, for problem of fit, two samples, and estimation. *Ann. Math. Stat.* 26, 631–640 (1956)
6. R Development Core Team.: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2009), ISBN 3-900051-07-0, <http://www.R-project.org>

Implementing an Image Search System with Integrating Social Tags and DBpedia

Chie Iijima, Makito Kimura, and Takahira Yamaguchi

Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522 Japan
{c_ijima,m_kimura,yamaguti}@ae.keio.ac.jp

Abstract. Although the number of recommending system has increased, many of the existing recommending systems often only offer general-purpose information. In the case of multimedia searches, novelty and unexpectedness are seen as particularly important. In this paper, we propose an image search method with a high degree of unexpectedness by integrating the social tag of Flickr and DBpedia, and using preference data from search logs. We also propose an image search system named Linked Flickr Search, which implemented the proposed method. By evaluation with an unexpectedness index, and by comparing the basic Flickr search functions and flickr wrappr, which is related research, we confirmed that particularly in the initial stages of the search, our proposed system was possible to recommend highly unexpected images.

1 Introduction

The volume of data from content posted on the Internet has increased with explosive momentum. As a result, the information a user seeks is typically buried the much unnecessary data, in another word, searching for information that matches the user's preferences becomes difficult to obtain, in addition, the data volume creates overload of information. Content sharing systems such as SNS, photograph sharing systems, and social bookmarking systems have been developed in an attempt to share useful contents in the case, where overloaded information is created. However, the content will increase as the number of users increases, and there are many cases in which overloaded data exists even within the content sharing systems.

On the other hand, the number of web services that have mechanisms for recommending useful data to users from among this huge volume of data, together with normal searches, will also increase. However, many of the existing recommendation methods, although well known, often offer general-purpose information, and for users there is little benefit to lists of information already known [1]. In the case of multimedia searches, it is said that novelty and unexpectedness are seen as particularly important. However, it is difficult to accomplish success multimedia searches with unexpectedness that is accomplished enough to gain user satisfaction [2].

In this paper, we propose an image search method with a high degree of user unexpectedness by integrating folksonomy social tags and DBpedia, and using the preferences of the user.

2 Proposed Method

2.1 Overview of the Proposed Method

As a web service to search images, this study uses Flickr[3], which is a folksonomy service that shares image data uploaded by a user. By linking image tags on Flickr and DBpedia[4] data and by offering the user class-instance relationships that use tags as an instance, the user can be guided in searching for the photographs of interest. In addition, the preference information obtained from the user from the search log is used to bring about a search that is highly satisfying to the user. DBpedia data has a large volume of RDF data based on Wikipedia categories and YAGO[5] categories. The dataset is interlinked on RDF level with various other Linked Open Data datasets on the Web.

2.2 Photograph Tags and DBpedia Linkage

For photographs that a user uploads to Flickr, their characteristics and tags, representing additional data, are assigned to Flickr. Multiple tags can be assigned to one photograph. Moreover, the data expressed in photographs can be handled as useful data as metadata, but the data cannot be handled from the approach of image content analysis. The present study attempts to make semantic processing possible by linking Flickr's tag data to DBpedia, which contains a variety of data, as classes and instances. Linkages are made using the following procedures.

(1) Acquire the DBpedia data

Acquire the label (title), class, and instance data to be used in this experiment from DBpedia. In this instance, the dump data that DBpedia offers is stored in a database.

(2) Match the tag and DBpedia labels

Use the Flickr API and acquire photographs from Flickr and the tags attached to each photograph. Then match the acquired photograph tag to the DBpedia label data. For DBpedia, since the label designation is described by the user according to a number of implicit criteria, there may be some difficulties to follow special rules to extract terms. Thus, for this case, we conducted the matching by taking into account notational variants; for example, an under bar or space between terms, and upper- and lowercase characters designations. They were not easy to identify the difference and the terms didn't match completely.

(3) Acquiring class-instance Data

The class data belonging to a concept is acquired from the matched DBpedia article page in response to the tag assigned to a photograph. That is, if the assumed tag is assigned to the photograph to be an instance, the class data to which that instance belongs can be acquired.

2.3 User Preferences

The user's preferences for searching are acquired from the user search log. The preferences are divided into two categories: those between a class and another class which

is selected next, and those between a class and an instance which is selected among the instances belonging to the target class. Figure 1 shows the relation between the two types of preferences.

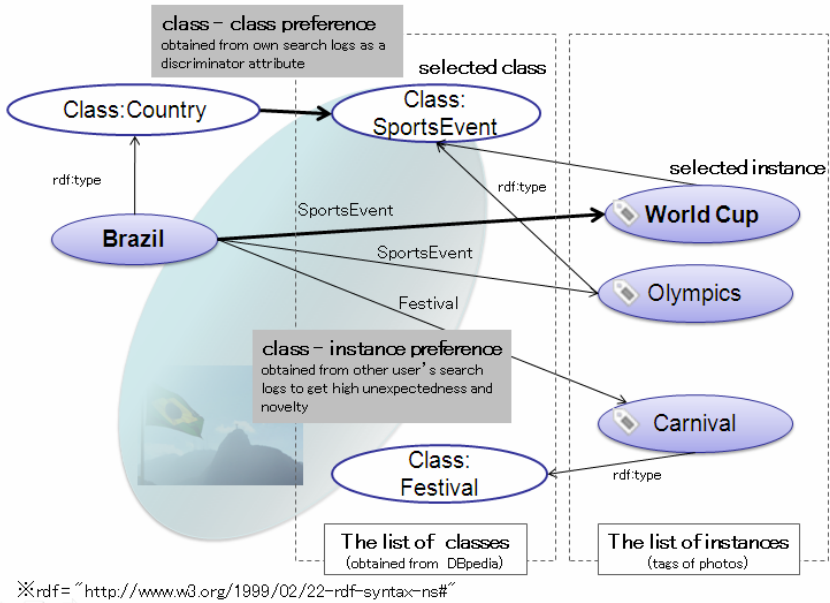


Fig. 1. Class-class preference and class-instance preference

2.3.1 Class-Class Preferences

Class-class preferences represent the classes that were most selected until retrieving the photograph sought during the search process. These represent the kinds of decision criteria used to classify the photograph from the time the user initiated the search until making the final decision. They also approach the user's discriminatory attributes. Using this class-class preference makes it possible to present user-specific results. The class data acquired from DBpedia has been subdivided into minute details. The data is sufficient enough to reflect the detailed preferences of the user.

When presenting a class to the user, we recommend using the class-class preferences the user employed to execute the search to allow smooth searching in line with the user's own discriminatory attributes.

2.3.2 Class-Instance Preferences

This preference shows which instance is preferred when selecting among instances of a selected class. The more frequently the user actually selects an instance, the higher this instance is valued.

These are the preferences used when presenting the user with instances. To enhance the unexpectedness and novelty, instances not in the user's own search log but are very popular with other users (many searches) are presented. In addition, to offer

unknown data, the display of data in one's own search log and data not in that log (unknown data) are distinguished and presented.

3 Implementation

3.1 System Overview

We implemented the Linked Flickr Search, which uses the method proposed in Section 2 to search images. Figure 2 shows the system overview of Linked Flickr Search.

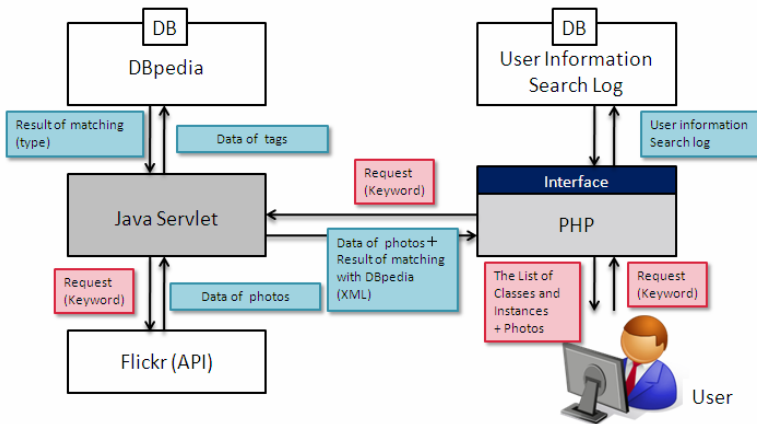


Fig. 2. System Overview of Linked Flickr Search

3.2 Procedures of the Proposed System

The following procedures are used to search images.

(1) Enter query

The user enters a keyword query concerning the desired photographs to be viewed. At the same time the user enters the keyword, he can choose in the following step whether to have the tag and DBpedia to be a complete match or whether to allow the absorption of notational variants. Moreover, he can also specify the number of photographs to acquire from Flickr. Under current specifications, the choice is between 40 and 100 photographs. The photographs related to the keyword entered are acquired at the limit specified.

(2) Display class

In number (1), taking all of the tags assigned to the acquired photographs and matching them through DBpedia will display a list of classes associated with each tag in the list. A list of the acquired photographs is also shown below the list of classes. If the photographs the user desires to see exist at this point, then selecting more class-instances to further refine the search is not required.

When a class is presented, the number of class-class search logs is referenced from the user's own log and sorting takes place in sequence from the class with the greatest number of searches. If the number of logs is the same, sorting takes place from the greatest number of instances (number of tags) of the class. The case of the number of logs being insufficient is considered if the number of uses is low. The class recommendation takes place by referencing only the user's own class-class log. The class-class logs of other users are not considered. This is what makes it possible to a search along the lines of the user's discriminator attributes.

(3) Display instance

In number (2), if one class among those displayed is selected, the instances belonging to that class will be displayed in a list. An instance displayed in this case indicates the tag assigned to the photograph acquired in number (1) among the instance lists of the selected class possessed by DBpedia.

When instances are presented, both those of the user's search log and those of other user's search log are used for recommendations. Particularly in the case of multimedia searches, such as photographs, presenting unknown content that the user himself can not find is highly beneficial; this type of searching can present instances when the user is not yet able to select but are highly acclaimed by other users. The number enclosed in parentheses and displayed beside the instance name shows the number of other users' search logs and represents an instance that highly preferred.

If the total number of other users' logs for a certain instance is L_{others} and the total number of all other users of the instances shown in the list is $L_{othersAll}$, then we can assume the higher the value of $L_{others} / L_{othersAll}$, the higher the support rate among other users. In this system, by displaying the size of this value and the star symbol of the corresponding quantity next to the instance name, we were able to show the popularity among other users. In this manner, we were able to carry out searches that were unknown to the user but popular among other users, that is, searches that were novel and quite unexpected.

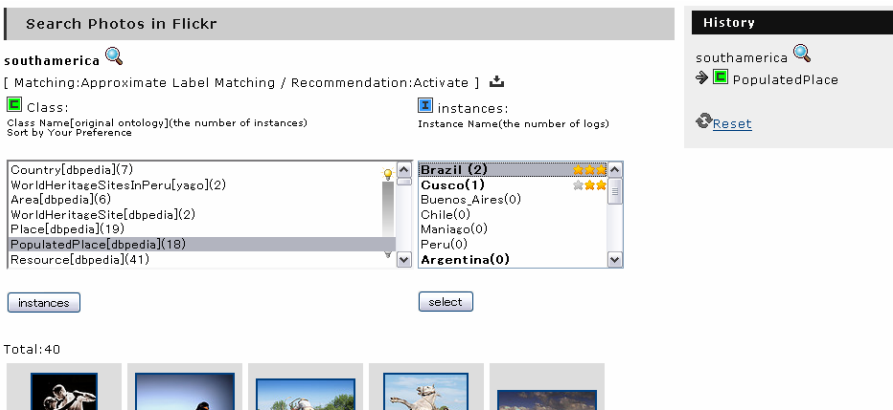


Fig. 3. Screenshot of Linked Flickr Search System

Finally, making a selection from the instance list shown in number (3) will allow the acquisition of photographs related to two keywords, the initial keyword entered and the selected instance, by repeating the same process. Subsequently, the class selection and instance selection will alternately repeat, and increasing the keywords when the photograph is acquired refines the search for the photograph. For one entry query, the path the user takes when searching and the selected instance are compiled as an individual user log. This log is then used in numbers (2) and (3) as class and instance displays. Figure 3 shows a screenshot of Linked Flickr Search System.

4 Experimental Evaluation

We have conducted some experiments to evaluate the unexpectedness, the preference matching, and precision and recall of search results of the proposed system.

The experiment took place during the two-week period from January 10 through January 24, 2010. To compile search logs for the preferences of 12 users, they were asked to execute searches that matched their preferences.

The evaluation took place by comparing the flickr wrappr[6], which is a Flickr image search service that uses DBpedia, as carried out in the present study, and the related tag display function of Flickr itself. The flickr wrappr is the focus of the Flickr image searching using DBpedia data; it improves the search accuracy by comparing geotags that indicate the region data within the data and the position data written in DBpedia. In addition, it handles multiple languages when searching using multiple language labels.

4.1 Unexpectedness

The highly unexpected data was evaluated and presented to the user relative to the classes and instances displayed at the initial stage of the search. The twelve users were selected with different preference entering four queries. The unexpectedness of the classes and instances gained from the results were evaluated.

For the evaluation, the unexpectedness and unexpectedness_r, which are unexpectedness indices [7]. These indices measure the unexpectedness of the recommended system advocated by Murakami et al. The indices indicate that data that can be predicted using a primitive prediction method has a low level of unexpectedness, whereas data that is impossible to predict using a primitive prediction method has a high level of unexpectedness.

For L number of recommendation lists, let the recommended item in position i of the list be s_i and the prediction probability value of s_i due to the target prediction method be $P(s_i)$. Also let the predictive probability value of s_i due to the primitive prediction method be $prim(s_i)$. Let $isrel(s_i)$ represent the degree of conformity of s_i to the interests of the user. As follows, it calculated by (1).

$$\text{unexpectedness} = \frac{1}{L} \sum_{i=1}^L \max(P(s_i) - prim(s_i), 0) \cdot isrel(s_i) \quad (1)$$

In addition, *unexpectedness_r* can be considered as the sequence level of the search item in terms of unexpectedness. If we assume a conformity item in position *i* or higher to be *count(i)*, then as follows, it calculated by (2).

$$\text{unexpectedness}_r = \frac{1}{L} \sum_{i=1}^L \max(P(s_i) - \text{prim}(s_i), 0) \cdot \text{isrel}(s_i) \cdot \frac{\text{count}(i)}{i} \quad (2)$$

In this experiment, we assumed the primitive prediction method would be the display of the Flickr related tags. As for $P(s_i)$, if the instance which was recommended was in the list of the upper-level 10 sets of the class data presented, it was 1, if not 0. As for $\text{prim}(s_i)$, if it was displayed by a Flickr related tag, it was 1, if not, 0. In the case of $\text{isrel}(s_i)$, if the instance existed in the user logs compiled previously by the user (the instance has selected previously), it was 1, if not 0.

For the flickr wrapper, which was used in making the comparisons, the related tags were obtained from the top ten photographs of the input query and were compared.

The results of the comparisons with flickr wrapper are shown in Table 1. Where unexpectedness is concerned, the results of the method proposed in the present study exceed those of flickr wrapper.

Table 1. Comparison of Unexpectedness Averages

Unexpectedness Evaluation	unexpectedness	unexpectedness_r
Proposed Method	0.2886	0.1151
flickr wrapper	0.1824	0.0422

In the experiment, as for the result of the user who had the preference of the travel, the value of unexpectedness was high. It seems that there are a lot of photographs in various places and photographs of user's going on a trip in the photograph of Flickr. For instance, the user who wants to go to South America, it is possible to start from the retrieval of the key word initial "South America", and to draw out the user's potential demand through an appropriate interaction, that is, South America → country(class) → Peru(instance) → World Heritage(class) → Machu Picchu(instance) with the user. Therefore, it thinks the decision making of the user who tries to go on a trip to only have to be able to support it by combining the trip plan generation service and this system.

4.2 Preference Matching

As with the unexpectedness evaluation, it was evaluated whether the top ten classes gained from querying and whether the instances associated with them matched the user preferences. In the case of a related tag, for which Flickr independently calculates the degree of relevance and displays it, and in the case of the flickr wrapper, ten photographs obtained using flickr wrapper and their tag data were compared. The upper limit of the tag data was assumed to be the total number of classes and instances obtained for each search query. We asked each user to determine if there was a match with their preferences.

The results of comparing matches are shown in Table 2. The average of the total number of classes and instances for each search word was 42.7. In terms of the compatibility of the proposed method, the results exceed those of the Flickr related tag or flickr wrappr, although the number of logs is not so great.

Table 2. Comparing Preference Matches

	Proposed Method		Flickr related tag	flickr wrappr
	instance	class		
Average value	0.7348	0.7076	0.3385	0.2314

4.3 Precision and Recall of Search Results

In the case of the search results the user obtains by narrowing down the search, it was confirmed whether appropriately relevant data is presented and the final decision making has taken place. By confirming that precision and recall were obtained, this resulted in confirming whether the class and instance were overloaded or insufficient as the search was refined over a number of tries by the user.

In general, precision indicates the ratio of the number of items that were matches in the instances of each class presented and recall indicates the ratio of matched items among all items included in the presented list. Listing all of the instances of classes that are relevant in order to calculate the total number of related items is difficult. For that reason, in this case, the user were asked to select items that were of interest from among those obtained from DBpedia as instances of the target class. Then considered those instance sets as matches. For a refined class, flickr wrappr, the target of comparison, carries out a search using the class name and calculates the level of matching with all matching items for a maximum of five photographs from the acquired list of photographs. Further, the experiment allowed the users to select one query and use interaction until the target instance was found. The results of the average values of precision and recall are shown in Table 3.

Table 3. Comparison of Precision and Recall Averages

	Precision	Recall
Proposed Method	0.5327	0.2433
flickr wrappr	0.6483	0.6833

For both precision and recall, the proposed method showed results below that of related study, flickr wrappr. This is thought to be due to cases in which the tags assigned to the photographs had meanings that were different from those of DBpedia and to be due to flickr wrappr being researched with a high level of precision, which means that flickr wrappr is better in a situation of fewer matching items.

5 Conclusions

In this paper, we proposed an image search method with a high degree of unexpectedness by integrating the social tag of Flickr and DBpedia, and using preference data from search logs. We also proposed a photograph search system named Linked Flickr Search, which implemented the proposed method.

By evaluation with an unexpectedness index, and by comparing the basic Flickr search functions and flickr wrapp, which is related research, we confirmed that particularly in the initial stages of the search, it was possible to show highly unexpected selections matched the user preferences.

Although we handled only the number of logs obtained as users' preferences, in future studies, we think it is possible to build a better user preference model by applying user clustering methods at the preference data building stage and expanding the presentation of data by handling other users with interests similar to those of the user. Moreover, although the precision and recall showed results below those of related studies, we believe the necessary improvements can be made using a better method: one that absorbs notational variants when matching Flickr photograph tags with DBpedia, and by pre-processing keywords peculiar to photographs, which become problems in the photograph domain.

References

1. Resnick, P., Varian, R.H.: Recommender systems. *Communications of the ACM* 40(3), 56–58 (1997)
2. McNee, S.M., Riedl, J., Konstan, J.A.: Being accurate is not always good: How Accuracy Metrics have hurt Recommender Systems. In: *Proc. of ACM Special Interest Group on Computer-Human Interaction (ACM SIGCHI)*, pp. 997–1001 (2006)
3. Flickr: <http://www.flickr.com/>
4. DBpedia: <http://dbpedia.org/>
5. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO - A Large Ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics* (2007)
6. Becker, C., <http://www4.wiwiwiss.fu-berlin.de/flickrwrapp/>
7. Murakami, T., Mori, K., Orihara, R.: Metrics for Evaluating the Serendipity of Recommendation Lists. In: Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T. (eds.) *JSAI 2007. LNCS (LNAI)*, vol. 4914, pp. 40–46. Springer, Heidelberg (2008)
8. Linked Data, <http://linkeddata.org/>

The Influence of Shopping Path Length on Purchase Behavior in Grocery Store*

Marina Kholod¹, Takanobu Nakahara², Haruka Azuma², and Katsutoshi Yada²

¹ Data Mining Laboratory, Research Institute for Socionetwork Strategies, Kansai University

² Faculty of Commerce, Kansai University

3-3-35 Yamate, Suita, Osaka, 564-8680 Japan

{r098057, nakapara, da60016, yada}@kansai-u.ac.jp

Abstract. In this paper we analyze the new type of information, namely RFID (Radio Frequency Identification) data, collected from the experiment in one of the supermarkets in Japan in 2009. This new type of data allows us to capture different aspects of actual in-store behavior of a customer, e. g. the length of her shopping path. The purpose of this paper is to examine more closely the effect of shopping path length on sales volume, which is one of the established ideas in RFID research as well as in retailing industry. In this paper we developed a simple framework, based on criteria of Wandering Degree and Purchase Sensitivity, in order to see how the relationship between distance* walked within the store and sales volume interacts with walking behavior of customers. As a result, in this paper we came up with some useful suggestions for more efficient in-store area management.

Keywords: RFID (Radio Frequency Identification) data, Wandering Degree, Purchase Sensitivity, In-Store Area Management.

1 Introduction

Analysis of RFID data has become an attention-getting topic in the research field of in-store customer behavior. The experiments on tracking actual purchasing behavior by using RFID tags have been conducted in Europe, the US and Japan (see details in Larson et al. 2005, Hui et al. 2009, Yada 2009). In our paper we use the unique dataset, obtained from the experiment carried out in one of the supermarkets in Japan.

One of the main advantages of RFID data is that it allows us to capture the exact shopping path of each particular customer, which is different from one another, as customers are different e.g. those, who are in hurry or elderly customers, etc. Thus, Larson et al. (2005) discovered 14 path types and found out that shoppers make short trips into the aisles instead of crossing areas in their full size. Their research treats exclusively the path itself, without taking into consideration the purchasing activity of the customer. However, it is natural that there is a difference in actual purchasing activity depending on the path walked. While Hui et al. (2009) observe the deviations from the optimal shopping path and find the positive relationship with purchased

* In this paper we use the terms “shopping path”, “distance” and “shopping route” interchangeably.

quantity, in our paper we investigate further into the relationship between the length of the shopping route and sales volume with the purpose to make concrete suggestions for the efficient in-store area management, resulting in sales growth.

This paper develops a simple framework which allows the use of the relationship between the length of shopping path and sales with the purpose to improve the latter. We base our analysis on two criteria – Wandering Degree and Purchase Sensitivity. Wandering Degree demonstrates the extent of wandering in the store, therefore, identifying two types of customers—those who wander around while shopping, as they do not have clear shopping plans and need to think what to buy, and those who don't wander, as they come to the store with clear shopping goals. We call such behavioral patterns wandering and decisive, respectively. Purchase Sensitivity is the correlation of Wandering Degree with purchased quantity, so that it indicates areas with different strength of positive relationship between two variables. Finally, we classify areas in the WD-PS Matrix, which helps us spot the areas for the improvement by changing the behavioral pattern of their customers. From the matrix it becomes clear that in the case of high Purchase Sensitivity, it is preferable to have wandering customers, and in the case of low Purchase Sensitivity – decisive customers.

The remainder of this paper is organized as follows. In the second section, we describe the data obtained from the experiment in more detail, validate the relationship between shopping path length and sales, and postulate the necessity of taking size of supermarket areas into account, when analyzing the length of customers shopping route. In the third section, we introduce the criterion of Wandering Degree with the purpose to overcome the area size difference problem, identifying different types of shopping behavior in different areas, and then we compute Purchase Sensitivity of each area. In the fourth section, we present the WD-PS Matrix and give some area improvement suggestions. The fifth section summarizes main results and brings in the future tasks.

2 Does Longer Shopping Path Result in Sales Growth?

2.1 Basic Analysis of RFID and POS Data

RFID experiment, which data we analyze in this paper, was conducted in one of the supermarkets in Japan during almost 6 weeks on May 11-June 15, 2009. RFID tags were attached to shopping carts in order to track the trajectory of a customer within the store, to record her departure from the entrance, her paths from one area to another, stationary visits to different areas until she reaches the checkout register, where her shopping trip completes and POS (Point-of-Sale) transaction data on products, prices and quantity purchased is generated. As a result, we have two main types of information – on what was bought and how those purchases were made. Our dataset has 6997 customers (110492 purchases) and the store, where the experiment was conducted, has sales of 45-60 million yen per day. Majority of customers in this store are females of 40-60 years old. An average customer in our dataset spends 3525 yen and buys 20 items per one shopping trip.

The supermarket has 2 entrances, 25 areas, Central Aisle and Checkout Register as seen in Figure 1. This layout, which is typical for a supermarket in Japan, is reproduced from x and y coordinates, registered by RFID sensors placed around the perimeter of the store.

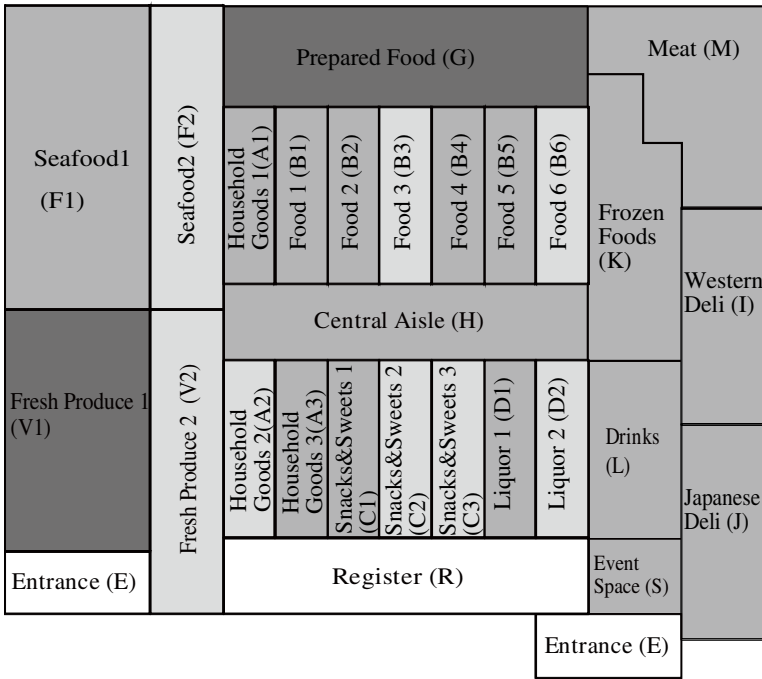


Fig. 1. Grocery Store Layout

2.2 Relationship between Length of Shopping Path and Sales

As mentioned in the introduction, Hui et al. (2009) showed positive relationship between the distance customers walk during shopping trips and the quantity they buy. This gives us a strong motivation to investigate further into the relationship between length of shopping paths and sales volume. This relationship holds for our data with Pearson Correlation coefficient equal to 0.8457, which can be confirmed from Figure 2. The chart shows average sales (in yen) per one shopping trip for each interval of distance, computed from x and y coordinates contained in RFID data by using Euclid's algorithm. As seen from the graph, there is stable growth in sales as shopping path length increases. Thus, a supermarket manager might think to “make” a customer walk longer distances in order to improve sales. However, he should be careful and check the validity of the same relationship at the area level.

In our data the correlation between average shopping path and average sales on an area level was found out to be poor, which can be seen from Figure 3. To demonstrate why the relationship does not hold, let's have a look at the area V1 (vegetables) located at the extreme right end of the shopping path axis. As we can see customers do walk long distances in this area but average sales are relatively low. We relate this poor correlation to the fact that the plot in Figure 3 does not take into consideration the difference in area sizes, e.g. V1 is the second biggest area, which is obvious from Figure 1. In this case it would be useful to have a criterion based on the size of the area and distance walked in it.

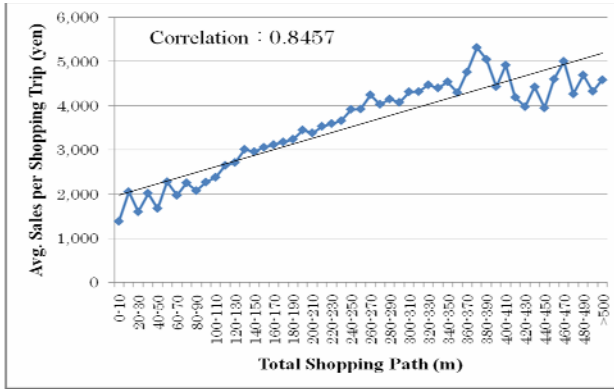


Fig. 2. Average Sales per Shopping Trip vs. Total Shopping Path

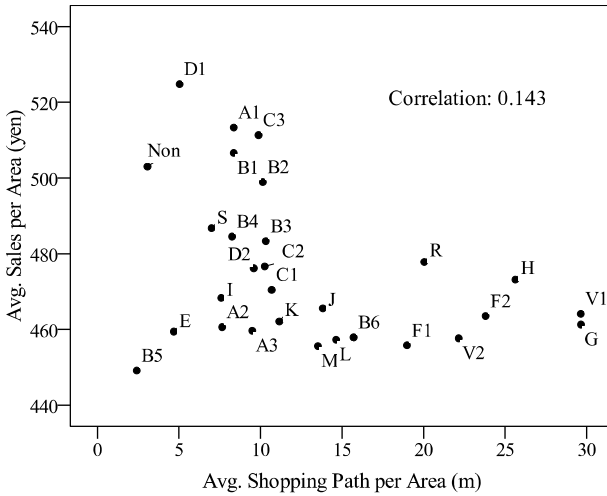


Fig. 3. Average Shopping Path vs. Average Sales per Area

3 Wandering Degree and Its Influence on Sales

3.1 The Definition of Wandering Degree and Its Interpretation

In order to take area size into consideration, as mentioned above, we present Wandering Degree (*WD*), the ratio of distance and area size, which shows the extent of wandering for customer c ($c=1, \dots, n$) in area a ($a=1, \dots, m$), and is computed by the following equation.

$$WD = \frac{D_{ca}}{\sqrt{A_{size}}}, \text{ where } WD - \text{Wandering Degree of customer } c \text{ in area } a ,$$

D_{ca} - distance walked by customer c in area a ,

A_{size} - size of area a

If ratio WD is equal to 1, this means that the length of the side of square-approximated area is equal to the distance walked within that area. If the ratio is less than 1, then it means that distance walked by a customer within that area is shorter than the side of the squared area, thus a customer probably simply made a short trip into the area, instead of crossing it in full size. If the ratio is greater than 1 than it means that distance walked within that area is longer than the length of one side, thus demonstrating the fact that customer was wandering around the area.

3.2 Wandering Degree and Three Types of Purchasing Behavior

Furthermore, when looking at each area, we found out that the distributions of WD among purchasers can be grouped into three different shapes as in Figure 5, reflecting three types of purchasing behavior: wandering, decisive and mixed.

- a) The peak of the distribution curve of WD corresponds to the value greater than 1, which shows that customers do wander around, while thinking what to purchase. The areas with wandering customers are colored dark grey in Figure 1.
- b) The peak of the distribution curve corresponds to the value less than 1, meaning that customers do not walk a lot within the area, as they come to the store with clear shopping goals. The areas with these customers are colored grey in Figure 1.
- c) The distribution has two peaks, lying on both sides from value of 1, corresponding to the presence of both above-mentioned types of customers. These areas are light grey in Figure 1.

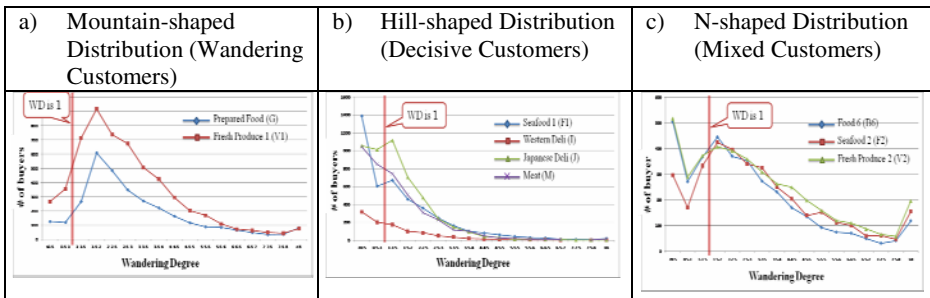


Fig. 4. Three Types of WD Distributions

3.3 Wandering Degree and Purchase Sensitivity

In order to shed light on the strength of the relationship between Wandering Degree and quantity purchased in each area, we check for their correlation, computed by Pearson Correlation coefficient, and call this correlation Purchase Sensitivity (PS). In

Figure 5 areas are divided into four groups and colored according their PS. High correlation implies that the longer distance a customer walks in a corresponding area the more she buys, and low PS corresponds to the situation when even if a customer walks longer distance, she won't buy more. Thus using such criterion as PS can help us judge whether it is appropriate or not to extend the shopping route of a customer for each particular area.

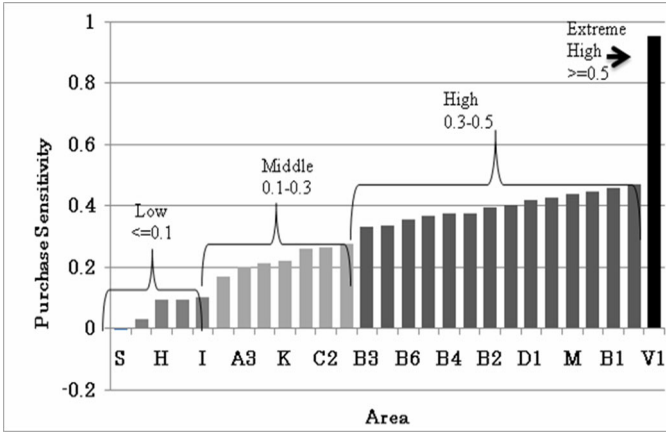


Fig. 5. Purchase Sensitivity of Supermarket Areas

4 Implications for In-Store Area Management

In this section, we present a matrix in Table 1, basing on criteria of *WD* and *PS*, from which we can draw conclusions regarding the effect of walking distance on purchased quantity in each area and make some implications for more efficient in-store area management. This matrix allocates each supermarket area in a corresponding cell according to its *WD* and *PS*. The rows of the matrix correspond to behavioral patterns of customers (wandering, decisive and mixed), implied by the distribution of *WD* among purchasers in the area. The columns correspond to *PS* of the area, which shows the strength of relationship between the degree of wandering of customers and quantity purchased. Each cell contains areas similar to each other according their behavioral pattern and *PS*, thus helping us to understand how *PS* interacts with the behavioral pattern of the area. According to the meaning of *PS*, it is profitable to have wandering shoppers as they buy more while thinking and wandering around. If *PS* of the area is high, it means that customers who come to such areas buy more as they wander around. Thus it is profitable to have wandering customers in these areas. As we can see from the matrix, areas V1 and G reflect this idea. Furthermore, if *PS* of the area is low, it means that customers who come to such areas already have clear shopping goals and just pick up the items they need, without wandering around. In this case it is profitable to have decisive customers. From the matrix it is clear that this is exactly the case for areas I, B5, H, R and S. The cells which contain two above-mentioned types of areas present the ideal situations which do not require any improvement.

Table 1. The WD-PS Matrix

	PS High	PS Middle	PS Low
Wandering Customers	V1, G		
Decisive Customers	M,F1,J,C1	L,B2,B4,K,A2, A3	I, B5,H,R,S
Mixed Customers	C3,D1,B3,B6	F2,V2, D2,A2	

However, as we can see from the matrix, for such areas as M, F1, J and C1, it is important to make some improvements, because in spite of high PS value they have decisive customers. Hence it is necessary to change their behavioral pattern into wandering e.g. by capturing their attention with in-store media.

5 Conclusion

In this paper we performed the analysis of new type of data, namely RFID data, recently available as a result of RFID technology development. This type of data allows to study the actual in-store behavior of customers. Combined with the purchase data, it gave us a powerful source of information on how a customer made her purchasing decisions. One of the main aspects of in-store behavior is a path that the customer takes during her shopping trip. In this paper we explore the relationship between the length of the shopping route and the purchased quantity. Instead of path length, we use the standardized measure, Wandering Degree, which demonstrates the extent of area wandering, bringing out three types of areas with wandering, decisive and both types of customers. Then we compute Purchase Sensitivity of each area by correlation between *WD* and purchased quantity and summarize in the WD-PS Matrix. This Matrix helps us understand the necessity of improvements for each particular area. Guided by the value of PS we identify which type of customer behavior makes the area profitable. E.g. if PS of the area is high then it is desirable to have wandering customers in this area, and on the contrary, if PS is low, then – decisive customers. Thus we made some concrete suggestions about the areas which needed to be improved by changing customer behavior from decisive into wandering.

The further hypothesis about the effects of proposed measures on purchasing behavior, as well as verification of the proposed criteria in other settings except for supermarkets, is left as a future task.

Acknowledgements. This work was partially supported by MEXT.KAKENHI 22243033 and “Strategic Project to Support the Formation of Research Bases at Private Universities”: Matching Fund Subsidy from MEXT (Ministry of Education, Culture, Sports, Science and Technology), 2009-2013.

References

1. Larson, J.S., Bradlow, E.T., Fader, P.S.: An Exploratory Look at Supermarket Shopping Paths. *International Journal of Research in Marketing* 22(4), 395–414 (2005)
2. Hui, S.K., Fader, P.S., Bradlow, E.T.: Path Data in Marketing: An Integrative Framework and Prospectus for Model Building. *Marketing Science* 28(2), 320–335 (2009)
3. Hui, S.K., Fader, P.S., Bradlow, E.T.: The Travelling Salesman Goes Shopping: The Systematic Deviations of Grocery Paths from TSP Optimality. *Marketing Science* 28(3), 566–572 (2009)
4. Hui, S.K., Bradlow, E.T., Fader, P.S.: Testing Behavioral Hypotheses using An Integrated Model of Grocery Store Shopping Path and Purchase Behavior. *Journal of Consumer Research* 36(3), 478–493 (2009)
5. Yada, K.: String Analysis Technique for Shopping Path in a Supermarket. *Journal of Intelligent Information Systems* (2009) (Online publication)

Existence of Single Input Rule Modules for Optimal Fuzzy Logic Control

Takashi Mitsuishi¹, Hidefumi Kawakatsu¹, and Yasunari Shidama²

¹ University of Marketing and Distribution Sciences, Kobe 651-2188, Japan

² Shinshu University, Nagano 380-8553, Japan

`takashi_mitsuishi@red.ums.ac.jp`

Abstract. The nonlinear feedback control whose feedback law is constructed by the SIRM's method is structured and presented mathematically. The set of SIRM's formed from the membership functions, important degrees and some parameters is compact by considering it to be a family of them. And by considering the fuzzy inference calculations as a composite functional on the family of the membership functions, its continuity is proved in functional analysis. Then, the existence of optimal solution of fuzzy feedback control using SIRM's method is derived from these facts.

Keywords: Fuzzy logic control, SIRM's model, optimal control, functional space.

1 Introduction

In 1965 Zadeh introduced the notion of fuzziness [1] [2] and then Mamdani has applied fuzzy logic to the field of control theory [3]. After that, fuzzy control has been increased and studied widely, since it able to realize numerically the control represented by human language and sensitivity [4].

In practical use, fuzzy membership functions, which represent input and output states in optimal control system, are decided on the basis of the experience of experts in each peculiar plant. Since it is very difficult that the design of them which demonstrates satisfactory performance, they should be adjusted to optimum. It is necessary to make an effort so that the tuning by the person. Moreover, its evaluation has not yet been discussed sufficiently. Recently many studies about automatic tuning of fuzzy rules and membership functions has become considerable interest [7]-[9].

The optimization of fuzzy control discussed in this paper is different from conventional method such as classical control and modern control. We consider fuzzy optimal control problems as problems of finding the minimum (maximum) value of the performance function with feedback law constructed by IF-THEN rules through a fuzzy inference [10]-[13].

In this study, we analyzed a single input rule modules connected fuzzy inference method (SIRM's method) proposed by Yubazaki [5], that can decrease the number of fuzzy rules and membership functions radically in comparison with

the usual fuzzy inference methods. The SIRMs method is applied to clustering field. As an example, Watanabe suggests that the SIRMs method is useful for discriminant analysis of iris data set [14].

To guarantee the convergence of optimal solution, the compactness of the set of membership functions is proved. And assuming fuzzy inference to be a functional on the set of membership functions and some constants, its continuity is obtained. Then, it is shown that the system has an optimal feedback control by essential use of compactness of sets of fuzzy membership functions. The tuple of membership functions, in other words SIRMs, which minimize the integral cost function of fuzzy logic control exists.

2 SIRMs Fuzzy Model

In this section we briefly explain the fuzzy control method for the convenience of the reader.

2.1 Single Input Rule Modules (SIRMs)

In this study, following SIRMs model is considered.

$$\begin{aligned}
 \text{SIRM-1} &: \{R_j^1 : \text{if } x_1 = A_j^1 \text{ then } y = c_j^1\}_{j=1}^{m_1} \\
 &\dots \\
 \text{SIRM-}i &: \{R_j^i : \text{if } x_i = A_j^i \text{ then } y = c_j^i\}_{j=1}^{m_i} \\
 &\dots \\
 \text{SIRM-}n &: \{R_j^n : \text{if } x_n = A_j^n \text{ then } y = c_j^n\}_{j=1}^{m_n}
 \end{aligned} \tag{1}$$

Here, n is the number of SIRMs and premise variables x_1, x_2, \dots, x_n , and $m_i (i = 1, 2, \dots, n)$ are the numbers of single input rules in each SIRM- i . y is the output of SIRMs connected fuzzy inference model and consequent variable. Let $A_j^i(x_i) (i = 1, 2, \dots, n; j = 1, 2, \dots, m_i)$ be fuzzy grade of each fuzzy set A_j^i for i -th input x_i . c_j^i is consequent output of j -th rule in i -th module, which is crisp value in this model. The membership function of fuzzy set A_j^i is written as same character A_j^i in this paper.

For simplicity, we write “if” and “then” parts in the rules by the following notation:

$$\begin{aligned}
 \mathcal{A}^i &= (A_1^i, A_2^i, \dots, A_{m_i}^i), \quad c^i = (c_1^i, c_2^i, \dots, c_{m_i}^i) (i = 1, 2, \dots, n), \\
 \mathcal{A} &= (\mathcal{A}^1, \mathcal{A}^2, \dots, \mathcal{A}^n), \quad c = (c^1, c^2, \dots, c^n).
 \end{aligned}$$

Then, SIRMs (II) is called a fuzzy controller in this paper, and is denoted by (\mathcal{A}, c) which is the pair of the membership functions and the consequent outputs.

2.2 The Set of Membership Functions and Its Properties

In this section, we introduce the set of fuzzy membership functions and study their topological properties. Then we can show that a set of admissible fuzzy controllers is compact and metrizable with respect to an appropriate topology on fuzzy membership functions.

Denote by $C[a, b]$ be the Banach space of all continuous real functions on $[a, b]$ with the norm $\|\mu\| = \max_{x \in [a, b]} |\mu(x)|$.

Let $\Delta_{ij} > 0$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m_i$). We consider the following sets of fuzzy membership functions.

$$G_{\Delta_{ij}} = \{ \mu \in C[a, b] : 0 \leq \mu(x) \leq 1 \text{ for } \forall x \in [a, b], \\ |\mu(x) - \mu(x')| \leq \Delta_{ij}|x - x'| \text{ for } \forall x, x' \in [a, b] \}$$

The set $G_{\Delta_{ij}}$ above contains triangular, trapezoidal and bell-shaped fuzzy membership functions with gradients less than positive value Δ_{ij} . Consequently, if $\Delta_{ij} > 0$ is taken large enough, $G_{\Delta_{ij}}$ contains almost all fuzzy membership functions which are used in practical applications. In next section, we shall assume that the fuzzy membership functions A_j^i in premise parts of the SIRM (1) belong to the set $G_{\Delta_{ij}}$ for all $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m_i$.

In the following, we endow the space $G_{\Delta_{ij}}$ with the norm topology. Then, by Ascoli Arzelà's theorem [16], for each $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m_i$, $G_{\Delta_{ij}}$ is a compact subset of the Banach space $C[a, b]$ of real continuous functions on $[a, b]$ with the supremum norm $\|\cdot\|_\infty$.

Put

$$\mathcal{G} = \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} G_{\Delta_{ij}} \right\}.$$

Then, by the Tychonoff theorem, we can have following proposition.

Proposition 1. *\mathcal{G} is compact and metrizable with respect to the product topology.*

3 Nonlinear Fuzzy Feedback Control

3.1 Nonlinear Feedback System

In this section, we propose a nonlinear fuzzy feedback control system for the purpose of showing the optimization. In this study we assume that the feedback part in this system is calculated by approximate reasoning presented in the next subsection. Using an idea and framework mentioned in the following section 4, the existence of optimal control based on fuzzy rules will be designed.

\mathbb{R}^n denotes the n -dimensional Euclidean space with the usual norm $\|\cdot\|$. Let $f(v_1, v_2) : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ be a nonlinear vector valued function which is Lipschitz continuous. In addition, assume that there exists a constant $M_f > 0$ such that

$\|f(v_1, v_2)\| \leq M_f (\|v_1\| + |v_2| + 1)$ for all $(v_1, v_2) \in \mathbb{R}^n \times \mathbb{R}$. Consider a system given by the following state equation:

$$\dot{x}(t) = f(x(t), u(t)),$$

where $x(t)$ is the state and the control input $u(t)$ of the system is given by the state feedback $u(t) = \rho(x(t))$. For a sufficiently large $r > 0$,

$$B_r = \{x \in \mathbb{R}^n : \|x\| \leq r\}$$

denotes a bounded set containing all possible initial states x_0 of the system. Let T be a sufficiently large final time. Then, we have

Proposition 2. [12] *Let $\rho : \mathbb{R}^n \rightarrow \mathbb{R}$ be a Lipschitz continuous function and $x_0 \in B_r$. Then, the state equation*

$$\dot{x}(t) = f(x(t), \rho(x(t))) \tag{2}$$

has a unique solution $x(t, x_0, \rho)$ on $[0, T]$ with the initial condition $x(0) = x_0$ such that the mapping

$$(t, x_0) \in [0, T] \times B_r \mapsto x(t, x_0, \rho)$$

is continuous.

For any $r_2 > 0$, denote by Φ the set of Lipschitz continuous functions $\rho : \mathbb{R}^n \rightarrow \mathbb{R}$ satisfying

$$\sup_{u \in \mathbb{R}^n} |\rho(u)| \leq r_2.$$

Then, the following a) and b) hold.

a) For any $t \in [0, T]$, $x_0 \in B_r$ and $\rho \in \Phi$, $\|x(t, x_0, \rho)\| \leq r_1$, where

$$r_1 = e^{M_f T} r + (e^{M_f T} - 1)(r_2 + 1). \tag{3}$$

b) Let $\rho_1, \rho_2 \in \Phi$. Then, for any $t \in [0, T]$ and $x_0 \in B_r$,

$$\|x(t, x_0, \rho_1) - x(t, x_0, \rho_2)\| \leq \frac{e^{L_f(1+L_{\rho_1})t} - 1}{1 + L_{\rho_1}} \sup_{u \in [-r_1, r_1]^n} |\rho_1(u) - \rho_2(u)|, \tag{4}$$

where L_f and L_{ρ_1} are the Lipschitz constants of f and ρ_1 .

3.2 Fuzzy Inference (Approximate Reasoning)

Assume the feedback law ρ of the nonlinear system (2) is constructed based on the SIRMs (1). In this study, when an input information

$$x = (x_1(t), x_2(t), \dots, x_n(t)) \in [-r_1, r_1]^n$$

is given to the fuzzy controller (\mathcal{A}, c) (SIRMs (1)), then one can obtain the amount of operation $\rho_{\mathcal{A}c}(x)$ from the controller through the following calculation:

Procedure 1: The degree of premise part of j -th rule in SIRM- i is following:

$$h_j^i = A_j^i(x_i) \quad (j = 1, 2, \dots, m_i; i = 1, 2, \dots, n).$$

Procedure 2: The inference result of the rule group SIRM- i as a weighted average of the consequent output using the agreement degrees h_j^i is calculated by

$$\beta_{\mathcal{A}^i c^i}(x_i) = \frac{\sum_{j=1}^{m_i} h_j^i \cdot c_j^i}{\sum_{j=1}^{m_i} h_j^i} \quad (i = 1, 2, \dots, n).$$

In SIRMs inference method, the importance degrees $d_i (i = 1, 2, \dots, n)$ are introduced to give each SIRMs weight of contribution. In the same way as \mathcal{A} and c , put

$$d = (d_1, d_2, \dots, d_n).$$

Procedure 3: Using importance degrees, the inference result of all rules is calculated by

$$\rho_{\mathcal{A}cd}(x) = \sum_{i=1}^n d_i \cdot \beta_{\mathcal{A}^i c^i}(x_i).$$

Let the n -tuple of importance degrees d join with fuzzy controller (\mathcal{A}, c) . Then it is denoted by (\mathcal{A}, c, d) . We can consider (\mathcal{A}, c, d) as the pair of SIRMs and importance degree and newly call it SIRMs fuzzy controller. In this paper, it is assumed that each $d_i (i = 1, 2, \dots, n)$ is belonging to closed interval $[0, 1]$, and satisfies $\sum_{i=1}^n d_i \leq 1$. Yubazaki does not need this condition [5], but for the existence of solution of the state equation (2) it is needed in this study. Moreover each membership function $A_j^i (i = 1, 2, \dots, n)$ is a element of following set:

$$F_{\Delta_{ij}} = \{ \mu \in C[-r_1, r_1] : 0 \leq \mu(x) \leq 1 \text{ for } \forall x \in [-r_1, r_1], \\ |\mu(x) - \mu(x')| \leq \Delta_{ij}|x - x'| \text{ for } \forall x, x' \in [-r_1, r_1] \}$$

Here, the constant r_1 is given by (3). Put

$$\mathcal{F} = \prod_{i=1}^n \left\{ \prod_{j=1}^{m_i} F_{\Delta_{ij}} \right\} \times [-r_2, r_2]^{(\sum_{i=1}^n m_i)} \times [0, 1]^n.$$

Here, closed interval $[-r_2, r_2]$ is the domain of the control variable y which is also output of previous inference method. Then \mathcal{F} is Cartesian product and consists of SIRMs and importance degrees of inference calculations. And it is obvious that $(\mathcal{A}, c, d) \in \mathcal{F}$.

3.3 Admissible Fuzzy Controller

To avoid making the denominator of the expression $\beta_{\mathcal{A}^i c^i}$ in procedure 2 equal to 0, for any $\delta > 0$, we consider the set

$$\mathcal{F}_\delta = \left\{ (\mathcal{A}, c, d) \in \mathcal{F} : \forall i = 1, 2, \dots, n, \forall x \in [-r_1, r_1]^n, \sum_{j=1}^{m_i} A_j^i(x_i) \geq \delta \right\}, \quad (5)$$

which is a slight modification of \mathcal{F} . If δ is taken small enough, it is possible to consider $\mathcal{F} = \mathcal{F}_\delta$ for practical applications. We say that an element (\mathcal{A}, c, d) of \mathcal{F}_δ is an admissible fuzzy controller. Then, we have the following

Proposition 3. *The set \mathcal{F}_δ of all admissible fuzzy controllers is compact and metrizable with respect to the product topology.*

Proof. Assume that a sequence $\{(\mathcal{A}^k, c^k, d^k)\}$ in \mathcal{F}_δ converges to $(\mathcal{A}, c, d) \in \mathcal{F}$. Fix $x \in [-r_1, r_1]^n$. Then, it is easy to show that for any $i = 1, 2, \dots, n$

$$\sum_{j=1}^{m_i} A_j^i(x_i) = \lim_{k \rightarrow \infty} \sum_{j=1}^{m_i} (A_j^i)^k(x_i) \geq \delta,$$

and this implies $(\mathcal{A}, c, d) \in \mathcal{F}_\delta$. Therefore, \mathcal{F}_δ is a closed subset of \mathcal{F} , and hence it is compact metrizable.

3.4 Lipschitz Continuity and Unique Solution of State Equation

In this paper, for any $(\mathcal{A}, c, d) \in \mathcal{F}_\delta$, we define the feedback function

$$\rho_{\mathcal{A}cd}(x) = \rho_{\mathcal{A}cd}(x_1, x_2, \dots, x_n) : [-r_1, r_1]^n \rightarrow \mathbb{R}$$

on the basis of the SIRMs by the approximate reasoning in the subsection 3.2. To apply the proposition 2, the following proposition is needed for the existence of unique solution of the state equation (2).

Proposition 4. *Let $(\mathcal{A}, c, d) \in \mathcal{F}_\delta$. Then, the following 1) and 2) hold.*

- 1) $\rho_{\mathcal{A}cd}$ is Lipschitz continuous on $[-r_1, r_1]^n$.
- 2) $|\rho_{\mathcal{A}cd}(x)| \leq r_2$ for all $x \in [-r_1, r_1]^n$.

Proof. 1) For any $x = (x_1, x_2, \dots, x_n), x' = (x'_1, x'_2, \dots, x'_n) \in [-r_1, r_1]^n$ and any $i = 1, 2, \dots, n$, we have

$$|\beta_{\mathcal{A}^i c^i}(x_i) - \beta_{\mathcal{A}^i c^i}(x'_i)| \leq \frac{2m_i^2 r_2 \Delta_i}{\delta^2} |x_i - x'_i| \tag{6}$$

Here, $\Delta_i = \max_{j=1,2,\dots,m_i} \{\Delta_{ij}\}$. Hence the mapping $\beta_{\mathcal{A}^i c^i}$ is Lipschitz continuous on $[-r_1, r_1]$. Noting that $d_i \leq 1$ for all $x \in [-r_1, r_1]^n$, it follows from (6) that

$$\begin{aligned} |\rho_{\mathcal{A}cd}(x) - \rho_{\mathcal{A}cd}(x')| &\leq \sum_{i=1}^n |\beta_{\mathcal{A}^i c^i}(x_i) - \beta_{\mathcal{A}^i c^i}(x'_i)| \\ &\leq \frac{2r_2}{\delta^2} \sum_{i=1}^n m_i^2 \delta_i |x_i - x'_i| \leq \frac{2r_2}{\delta^2} \sum_{i=1}^n m_i^2 \delta_i \|x - x'\| \end{aligned}$$

and the Lipschitz continuity of $\rho_{\mathcal{A}cd}$ is proved.

2) For any $x \in [-r_1, r_1]^n$, noting that $\sum_{i=1}^n d_i \leq 1$ and $\forall i, |\beta_{\mathcal{A}^i c^i}(x_i)| \leq r_2$, we can have

$$|\rho_{\mathcal{A}cd}(x)| = \left| \sum_{i=1}^n d_i \beta_{\mathcal{A}^i c^i}(x_i) \right| \leq r_2 \left| \sum_{i=1}^n d_i \right| \leq r_2.$$

Then 2) is obtained.

Let (\mathcal{A}, c, d) be a fuzzy controller given by the SIRMs (III). We say that the system (2) is a fuzzy feedback system if the control function $u(t)$ is given by the state feedback $u(t) = \rho_{\mathcal{A}cd}(x(t))$, where $\rho_{\mathcal{A}cd}(x(t))$ is the amount of operation from the fuzzy controller (\mathcal{A}, c, d) for an input information $x(t)$.

It is easily seen that every bounded Lipschitz function $\rho : [-r_1, r_1]^n \rightarrow \mathbb{R}$ can be extended to a bounded Lipschitz function $\tilde{\rho}$ on \mathbb{R}^n without increasing its Lipschitz constant and bound. In fact, define $\tilde{\rho} : \mathbb{R}^n \rightarrow \mathbb{R}$ by

$$\tilde{\rho}(x) = \tilde{\rho}(x_1, \dots, x_n) = \begin{cases} \rho(x_1, \dots, x_n), & \text{if } x \in [-r_1, r_1]^n \\ \rho(\varepsilon(x_1)r_1, \dots, \varepsilon(x_n)r_1), & \text{if } x \notin [-r_1, r_1]^n, \end{cases}$$

where

$$\varepsilon(u) = \begin{cases} 1, & \text{if } u > r_1 \\ -1, & \text{if } u < -r_1. \end{cases}$$

Let $(\mathcal{A}, c, d) \in \mathcal{F}_\delta$. Then it follows from proposition 3 and the fact above that the extension $\tilde{\rho}_{\mathcal{A}cd}$ of $\rho_{\mathcal{A}cd}$ is Lipschitz continuous on \mathbb{R}^n with the same Lipschitz constant of $\rho_{\mathcal{A}cd}$ and satisfies $\sup_{u \in \mathbb{R}^n} |\tilde{\rho}_{\mathcal{A}cd}(u)| \leq r_2$. Therefore, by proposition 2 the state equation (2) for the feedback law $\tilde{\rho}_{\mathcal{A}cd}$ has a unique solution $x(t, x_0, \tilde{\rho}_{\mathcal{A}cd})$ with the initial condition $x(0) = x_0$ [15]. Though the extension $\tilde{\rho}_{\mathcal{A}cd}$ of $\rho_{\mathcal{A}cd}$ is not unique in general, the solution $x(t, x_0, \tilde{\rho}_{\mathcal{A}cd})$ is uniquely determined by $\rho_{\mathcal{A}cd}$ using the inequality (4) of b) of proposition 2. Consequently, in the following the extension $\tilde{\rho}_{\mathcal{A}cd}$ is written as $\rho_{\mathcal{A}cd}$ without confusion.

4 Application to Optimal Control Problem

The performance index of this fuzzy feedback control system is evaluated with the following integral performance function:

$$J = \int_{B_r} \int_0^T w(x(t, \zeta, \rho_{\mathcal{A}cd}), \rho_{\mathcal{A}cd}(x(t, \zeta, \rho_{\mathcal{A}cd}))) dt d\zeta, \tag{7}$$

where $w : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$ is a positive continuous function. The following theorem guarantees the existence of a SIRMs fuzzy controller (\mathcal{A}, c, d) which minimizes and maximize the previous performance function (7).

Theorem 1. *The mapping*

$$(\mathcal{A}, c, d) \mapsto \int_{B_r} \int_0^T w(x(t, \zeta, \rho_{\mathcal{A}cd}), \rho_{\mathcal{A}cd}(x(t, \zeta, \rho_{\mathcal{A}cd}))) dt d\zeta$$

has a minimum (maximum) value on the compact space \mathcal{F}_δ defined by (5).

Proof. Since compactness of \mathcal{F}_δ is already derived by proposition 3, it is sufficient to prove that the performance function is continuous on \mathcal{F}_δ . Routine calculation gives the estimate

$$\begin{aligned} \sup_{x \in [-r_1, r_1]^n} |\rho_{\mathcal{A}^k c^k d^k}(x) - \rho_{\mathcal{A}cd}(x)| &\leq \frac{2}{\delta^2} \sum_{i=1}^n \left\{ m_i \sum_{j=1}^{m_i} |A_j^{i,k}(x) - A_j^i(x)| \right\} \\ &+ \frac{1}{\delta^2} \sum_{i=1}^n \left\{ m_i \sum_{j=1}^{m_i} |c_j^{i,k} - c_j^i| \right\} + r_2 \sum_{i=1}^n |d_i^k - d_i|. \end{aligned}$$

Assume that $(\mathcal{A}^k, c^k, d^k) \rightarrow (\mathcal{A}, c, d)$ in \mathcal{F}_δ and fix $(t, \zeta) \in [0, T] \times B_r$. Then it follows from the estimate above that

$$\lim_{k \rightarrow \infty} \sup_{x \in [-r_1, r_1]^n} |\rho_{\mathcal{A}^k c^k d^k}(x) - \rho_{\mathcal{A}cd}(x)| = 0. \tag{8}$$

Hence, by b) of proposition 2, we have

$$\lim_{k \rightarrow \infty} \|x(t, \zeta, \rho_{\mathcal{A}^k c^k d^k}) - x(t, \zeta, \rho_{\mathcal{A}cd})\| = 0. \tag{9}$$

Further, it follows from (8), (9) and a) of proposition 2 that

$$\lim_{k \rightarrow \infty} \rho_{\mathcal{A}^k c^k d^k}(x(t, \zeta, \rho_{\mathcal{A}^k c^k d^k})) = \rho_{\mathcal{A}cd}(x(t, \zeta, \rho_{\mathcal{A}cd})). \tag{10}$$

Noting that $w : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$ is positive and continuous, it follows from (9), (10) and the Lebesgue’s dominated convergence theorem [17] that the mapping is continuous on the compact metric space \mathcal{F}_δ . Thus it has a minimum (maximum) value on \mathcal{F}_δ , and the proof is complete.

5 Conclusion

Fuzzy logic control which the feedback law is constructed by fuzzy inference method in the nonlinear feedback control was studied. Two kinds of continuity of the inference method (approximate reasoning) on the single input rule modules (SIRMs) are proved. One is Lipschitz continuity on the premise variables, the other is the continuity as a functional on the compact set of membership functions. Then it is concluded that the set of membership functions, in other words SIRMs, which gives optimal control to the nonlinear feedback control exists.

Seki proposed the conversion of the SIRMs connected type fuzzy model into the simplified reasoning method [6]. And furthermore, the same result as this study concerning the simplified reasoning method has obtained [11]. So in the future work, the relation between both should be analyzed.

On the other, these properties will build the framework of optimization for the decision making system from the observation of real phenomena. In the data mining field, the experts judged and extracted the outliers previously. Its automation with a human linguistic knowledge is not so easy, and includes the problems of ambiguousness (fuzziness). Therefore the inference system in this study is useful as the subsystem to find the outliers from large-scale data.

References

1. Zadeh, L.A.: Fuzzy Sets. *Information and Control* 8, 338–353 (1965)
2. Zadeh, L.A.: Fuzzy algorithms. *Information and Control* 12, 94–102 (1968)
3. Mamdani, E.H.: Application of fuzzy algorithms for control of simple dynamic plant. *Proc. IEE* 121(12), 1585–1588 (1974)
4. Akiyama, T., Takaba, T., Mizutani, K.: Fuzzy Travel Behaviour Analysis for Navigation. *Journal of Japan Society for Fuzzy Theory and Systems* 11(2), 21–30 (1999)
5. Yubazaki, N., Yi, J., Hirota, K.: A Proposal of SIRMs (Single Input Rule Modules) Connected Fuzzy Inference Model for Plural Input Fuzzy Control. *Journal of Japan Society for Fuzzy Theory and Systems* 9(5), 699–709 (1997)
6. Seki, H., Mizumoto, M., Yubazaki, N.: On the Property of Single Input Rule Modules Connected Type Fuzzy Reasoning Method. *The Transactions of the Institute of Electronics, Information and Communication Engineers (A)* J89-A(6), 557–565 (2006)
7. Ishibuchi, H., Nii, M.: Generating Fuzzy Classification Rules from Trained Neural Networks. *Journal of Japan Society for Fuzzy Theory and Systems* 9(4), 512–524 (1997)
8. Nomura, H., Wakami, N.: A Method to Determine Fuzzy Inference Rules by a Genetic Algorithm. *The Transactions of the Institute of Electronics, Information and Communication Engineers (A)* J77-A(9), 1241–1249 (1994)
9. Gonda, E., Miyata, H., Ohkita, M.: Self-Tuning of Fuzzy Rules with Different Types of MSFs. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics* 16(6), 540–550 (2004)
10. Shidama, Y., Yang, Y., Eguchi, M., Yamaura, H.: The compactness of a set of membership functions and its application to fuzzy optimal control. *The Japan Society for Industrial and Applied Mathematics* 6(1), 1–13 (1996)
11. Mitsuishi, T., Wasaki, K., Ohkubo, K., Kawabe, J., Shidama, Y.: Fuzzy Optimal Control Using Simple Inference Method and Function Type Inference Method. In: *Proc. American Control Conference 2000*, pp. 1944–1948 (2000)
12. Mitsuishi, T., Kawabe, J., Wasaki, K., Shidama, Y.: Optimization of Fuzzy Feedback Control Determined by Product-Sum-Gravity Method. *Journal of Nonlinear and Convex Analysis* 1(2), 201–211 (2000)
13. Mitsuishi, T., Endou, N., Shidama, Y.: Continuity of Nakamori Fuzzy Model and Its Application to Optimal Feedback Control. In: *Proc. IEEE International Conference on Systems, Man and Cybernetics*, pp. 577–581 (2005)
14. Watanabe, S., Seki, H., Ishii, H.: Application for Discriminant Analysis by Kernel Type SIRMs Connected Fuzzy Reasoning Method. In: *Proc. The 2008 Fall National Conference of Operations Research Society of, Japan*, pp. 332–333 (2008)
15. Miller, R.K., Michel, A.N.: *Ordinary Differential Equations*. Academic Press, New York (1982)
16. Riesz, F., Sz.-Nagy, B.: *Functional Analysis*. Dover Publications, New York (1990)
17. Dunford, N., Schwartz, J.T.: *Linear Operators Part I: General Theory*. John Wiley & Sons, New York (1988)

Temporality and Reference Place: Discovering Chances for Conflict Avoidance in Teamwork

Ruediger Oehlmann

Kingston University London,
Faculty of Computing, Information Systems and Mathematics
Penrhyn Road, Kingston upon Thames, KT1 2EE, UK
R.Oehlmann@Springer.com

Abstract. This paper considers teams of experts who evaluate the output of a chance discovery system. The objective of the reported investigation is to identify strategies of conflict avoidance in the Japanese tea ceremony and to transfer these to such teamwork settings. Two *Usucha* and two *Koicha* ceremonies with between three and ten participants have been video recorded. Interpretive phenomenological analyses (IPA) of all video recordings have been conducted. The analyses have shown that temporal reflections affect conflict avoidance if they are related to a reference place, a conceptual space of agreed references. The remainder of the paper considers possible consequences of the identified reflective strategies for the management of chance discovery teams as well as wider team work settings and the design of intelligent systems that support team work.

Keywords: Teamwork, Conflict Avoidance, Temporality, Reference Place.

1 Introduction

Chance Discovery is concerned with the identification and the management of rare, but significant events, such as potential risks or opportunities, in some domain(1). The process of Chance Discovery combines automatic data analysis that provided candidate chances and a sequence of group discussions to identify the most promising opportunity(2). It is this teamwork context that bears the risk of conflicts. Conflicts typically arise from contradicting needs and goals, which are often related to an individual's identity and resources to maintain that identity. As in any teamwork scenarios, this raises the question of how to avoid such conflicts between team members. Whereas there is a considerable literature on conflict resolution, inquiries on conflict avoidance are rare. If conflict avoidance has been investigated, the focus was just on walking away from a conflictual situation. This is different from creating a harmonious situation where the conflict is less likely to occur in the first place.

However, as Katai et al. (3) have noted, it is difficult to achieve harmony and balance even between children, because nature becomes more and more removed from our environment, which leads to "an erosion of native culture, tradition and folkways." But still some cultural practices exist and a practice which is closely related to

the concept of harmony is the tea ceremony in Japan, subsequently referred to as *Chanoyu*. This practice has been formed since the 16th century.

Previously, the concept of the reference place has been introduced. It denotes a conceptual place outside of the current perceptual focus. Oehlmann (4) has identified examples of the reference place in Chanoyu, such as seasons, temples, paintings, and areas outside of the tea house. Such references form a reference place if the participants agree about the reference and share a common emotion related to the reference. Under these conditions, the shared reference place may increase harmony.

Various authors have suggested that in addition to the place also the temporality plays an important role in Chanoyu. An example is the concept of impermanence. Authors inspired by Buddhism as well as by Shinto equally emphasized this concept. Fujiwara (5) saw impermanence as the notion that all things are bound to change. He pointed to the juxtaposition between the fragility of human life and the permanence of nature. This sense is also encapsulated in *mono no aware*, which is often translated as the pathos of things and has been ascribed to Motoori Norinaga (1730 – 1801), a Japanese philosopher during the Tokugawa period (6).

Kenko (7), a Buddhist priest and essayist in the late Kamakura and early Muromachi period in Japan, saw impermanence in terms of the uncertainty that is associated with the beauty of life. For instance he reflected on the sudden change from the blossoming flowers in spring, which are suddenly scattered due to the effects of wind and rain. Such a situation caused him grief.

Takeuchi (8) emphasized acceptance and sadness which he both associated with the farewell of *Sayounara*. He contrasted this meaning with western expressions that negate the acceptance of the finality and loss that is associated with parting and instead emphasize the forward looking aspect. For instance, the French *au revoir* or the German *Aufwiederssehen* focus on the hope for a better time, when the separation ends. Whereas Takeuchi focuses on the linguistic differences, it should be noted that these different attitudes are also reflected in the various literatures. For instance, Sei Shounagon (9), a court lady in 10th century Japan, likens cherry blossoms to the faces of tearful lovers who are forced to say farewell. There is no positive outlook to the future. The blossoms fall and are gone. By contrast, authors from western cultures often convey hope or at least the refusal of acceptance, as Dylan Thomas (10) did in his poem *Do not go gentle into that good night*. Also the German poet Joseph von Eichendorff (11) in his poem *winter night* describes a tree that dreams of spring time when it receives a new dress of blossoms and green leaves. As Takeuchi (8) pointed out, this constitutes two fundamentally different positions, where one position emphasizes the acceptance of the current situation and the resulting sadness, whereas there is no acceptance in the western view, sometimes even denial, which is managed by hoping for a better future. Whereas Takeuchi fears that the use of *Sayounara* and its meaning vanishes, Ochi (12) found in her study that 80% of the Japanese participants used the term in formal situations. However, these participants were all university students from the Yokohama area. Therefore the result cannot be generalized to all Japanese people.

Although it has been established that the concept of impermanence plays a role in the wider Japanese culture and is related to Chanoyu, it is less clear how it is used and whether this concept contributes to harmony in Chanoyu.

Another aspect of temporality is continuity. Chanoyu as it is practiced today basically has not changed since the 17th century, although the concept itself is much older. Also during Chanoyu various references to folkloric tales and legends are made. Sometimes items are used which are several hundred years old. All these activities also emphasize a temporal dimension, which if shared may contribute to a harmonious situation.

The remainder of the paper describes an interpretive phenomenological study of various settings of Chanoyu that has investigated these questions. The next section will highlight a few aspects of the analysis method. Section 3 will describe the study and Section 4 will discuss the conclusions from that study and will in particular argue that there is a strong interrelationship between temporality and spaciality in the form of the reference place. This interrelationship will then give rise for a design that will support conflict avoidance in Chance Discovery teams and in teamwork in general.

2 Interpretive Phenomenological Analysis

The method of interpretive phenomenological analysis (IPA) focuses on a person's experience and how the world appears to that person (13). This subjective experience in the world is labeled *lifeworld*, which is characterized in terms of

- Temporality (the experience of time)
- Spaciality (the experience of space)
- Embodiment (the experience of one's own body)
- Intersubjectivity (the experience of relationships with other people)

Often a person's experience is accessed in the form of semi-structured interviews (13). Dangers of this approach are that the experience is no longer accessible in memory or that it has been filtered because the context in which the experience was made differs from the context of the interview situation. Therefore the study below relies on the transcripts of video recordings, which include experience in the form of direct actions and speech that describes and qualifies that experience.

3 Study

The study aims at identifying elements of the *lifeworld* that are related to Chanoyu and are used to support harmony. The particular focus of the analysis will be on the different roles the concept of temporality plays in Chanoyu.

3.1 Method

Design. The study has investigated four settings of Chanoyu, denoted as NS, NK, IS and IK. Ceremony NS was conducted in a refectory room of the Nara Institute of Science and Technology (NAIST) in Nara, Japan. Ceremony NK was conducted in a dedicated tea room at NAIST. Ceremony IS and IK were conducted in a purpose built tea house which was situated in a Japanese style garden in Ikoma, Japan. Tea ceremonies NS and IS were based on light tea (Usucha), whereas tea ceremonies NK and IK

were based on strong tea (Koicha). The difference in the tea was also reflected in the degree of formality.

Participants. Tea ceremony NS consisted of 8 guests with an age range between 20 and 60, a host in her forties and an assistant in her twenties. Two of the guests were male, 6 guests were female. The guests were novices or had little experience with the tea ceremony. The host and the assistant, both female, were experienced members of the NAIST tea club, which is associated with the Omotesenke tradition in Kyoto. The format of host and assistant was used because the guests required more explanations than usual. The host gave the explanations and the assistant prepared the tea. In this setting additional helpers were used to distribute sweets (Okashi).

Tea ceremony NK consisted of 6 guests with an age range between 20 and 50, 2 male and 4 female, a host and an assistant, both in their twenties and female. Again host and assistant were experienced members of the NAIST tea club, whereas the guests were novices or had little experience with Chanoyu. The tea ceremonies IS and IK consisted of the same two guests and the same host. All three participants were in their twenties, female, and experienced members of the NAIST tea club.

Materials and Apparatus. The materials included the instruction sheets and consent forms for the participants, which informed about the purpose of the study and eight digital video tapes. Two SONY video cameras with mounted microphones were used, one to record the guests and one for the host/assistant.

Procedure. Prior to the tea ceremony, the participants were informed about the purpose of the study and asked for their consent. Two cameras were set up. The tea ceremonies were recorded as they happened without any interference from the researcher. The recordings were synchronized, transcribed, and analyzed. All speech elements were transcribed in Japanese and translated into English in collaboration with native Japanese speakers. Both versions were used in the analysis. In addition, the transcripts describe every action of the participants.

3.2 Results and Analysis

Time has been compared to a river. Temporality can then be seen in terms of a continuous flow. But the flow of time can also be seen in terms of discontinuities. For instance an elderly person may claim that in her youth life was better, which means different from today. In the video recordings, discontinuities could be identified with reference to impermanence. Furthermore temporality was expressed with reference to objects of a reference place, because such references made the abstract concept of time more concrete.

Temporal Continuity and Reference Place. The following citation refers to *Aiko*, Princess *Toshi*, daughter to HIH Crown Prince *Naruhito* and Crown Princess *Masako*. It also refers to Princess *Chujo*, who was so kind that her poem could calm down the *Tatsuta* River, which kept her sick father from sleeping. The tale continues with her becoming a Buddhist nun and being asked by two nuns to spin the stems of lotus flowers into threads. The nuns wove these threads into a beautiful sparkling cloth. As

a reward, her greatest wish of seeing her mother who died when she was only 5 years old was fulfilled.

IS-404 GUEST: Could you tell us the name of the
 IS-405 tea scoop?
 IS-406 HOST: It is "TATSUTA-HIME", because there are
 IS-407 two princesses today.

The Japanese people were very happy when Princess *Toshi* was born as first child to HIH the Crown Prince and HIH the Crown Princess. Like with other important events, many people still remember where they were when they received the good news. Therefore this is already a harmony enhancing reference.

However, additional feelings of harmony can be derived from the combination of two events. Princess *Chujo* is portrayed as a good person and an example for moral behavior in the Buddhist sense. This means that there is a mutual identification between two positive events in the sense of temporal continuity. But this abstract continuity is made concrete by projecting it on two people.

Temporal Discontinuity and Reference Place. Also the abstract concept of impermanence as a form of temporal discontinuity is made concrete by relating it to elements of the reference place.

IS-172 FIRST GUEST: Well, I think it is late
 IS-173 cherry blossom in this area.
 IS-174 HOST: Yes, though cherry trees in a river
 IS-175 area almost finish blooming, I think this
 IS-176 area is different from over there.
 IS-177 HOST: Have you been to HANAMI somewhere?
 IS-178 FIRST GUEST: Not yet, I went to KYOTO
 IS-179 last week, but it was earlier than main
 IS-180 season. It was starting to bloom at that
 IS-181 time.
 IS-182 HOST: Did it start?

The falling rain signals the end of the cherry blossoms. The reference that links the concept of impermanence with the cherry blossom event enables acceptance as opposed to the hope for better times.

IS-183 GUEST Yes, if it wouldn't be raining
 IS-184 today, maybe there were many people.
 IS-185 HOST: It would be a pity because of this
 IS-186 rain. The area holds many cherry trees.
 IS-187 First GUEST: This rain spoils them.
 IS-188 HOST: However, a carpet of cherry blossoms is
 IS-189 also beautiful isn't it?

Also another Chanoyu setting provides evidence for acceptance without expressing hope for the future.

NS-014 Host: It's a brief moment, but I hope you
 NS-015 will have a good time here.

The host accepts the brevity of the event and hopes for the enjoyment just as long as it lasts. There is no positive outlook towards the future or an attempt to prolong the moment.

In both examples, the temporal discontinuity has the potential to lead to disharmony among the participants. But the shared cultural agreement about impermanence and acceptance of the discontinuity leads to harmony.

Continuity and Discontinuity. Sometimes a reference to a complex event combines both, continuity and discontinuity. The following example refers to such a combination. The *Omizutori* (Water-Drawing Festival) ceremony is conducted every year in *Todaiji* temple in *Nara*, Japan (14). It includes monks drawing fresh water and offer it to the Buddhist deities. The water is thought to have healing power. It is mixed with water from previous ceremonies during the last 1200 years. This can be viewed as the continuity component.

NS-110 Host: The tea scoop is called "Mizu (water)"
 NS-111 because it is made of bamboo which is used
 NS-112 as a torch of Omizutori at Nigatsudo in
 NS-113 Todaiji Temple. Omizutori is finished just
 NS-114 the other day.

During the festival *Otaimatsu* takes place: Monks carry large burning torches made from bamboo. The rain of fire sparks is believed to protect from evil things. Of course the torches burn only for a short time and can be seen to represent the temporal discontinuity. The festival is finished after 15 days. It is this impermanence the host referred to (Line NS-113).

4 Discussion

The above study has investigated temporality as a conflict avoidance strategy. The results have confirmed the division of temporality concepts into temporal discontinuity and continuity. Discontinuity was exemplified by the concept of impermanence, whereas continuity was represented by historical references.

However, it was not the use of concepts such as impermanence or historical references alone that established harmonious relationships and avoided conflicts. It was rather the combination of these concepts with agreed reference places that led to harmony. For instance, combining the idea of impermanence with the commonly appreciated cherry blossoms led to acceptance that avoids conflicts. Not even the negative concept of the rainy day that spoiled the cherry blossoms could interfere with that acceptance.

It has already been pointed out that in western thinking acceptance is often replaced by hope for a better future. This may be a difficulty in transferring temporal strategies of conflict avoidance to teams with western members. In part, this can be addressed by a suitable training regime. It can also be supported by the design of computing tools.

Large projects sometimes draw time lines on the wall of a meeting room. These are mainly limited to deadlines and deliverables. A step further is *The Wall* (15), a large physical board that supports interaction in that it allows team members to pin artifacts at the wall whereas other team can add comments. In principle, such a scenario would allow temporal reflections as well. In addition, the wall could be envisaged as a large screen with interactive windows where team members could add reflections, either in

interaction with other team members or within a private window. Creating temporal awareness in such a way, may then lead to an understanding of impermanence and continuity, because the windows could establish agreed reference places. This would support a situation where the generation of conflicts becomes less likely.

The current study was restricted to the actions that had been performed and the utterances of the participants. Dialogues are situational and allow only a limited insight into the attitudes of the participants. Therefore the current study will be complemented by a psycho-social analysis of interviews with people who have experience with Chanoyu.

References

1. Ohsawa, Y.: Modeling the Process of Chance Discovery. In: Ohsawa, Y., McBurney, P. (eds.) *Chance Discovery*. Springer, Berlin (2003)
2. Ohsawa, Y., Nara, Y.: Decision Process Modeling across Internet and Real World by Double Helical Model of Chance Discovery. *New Generation Computing* 21, 109–121 (2003)
3. Katai, O., Minamizono, K., Shiose, T., Kawakami, H.: System design of “Ba”-like stages for improvisational acts via Leibnizian space-time and Peirce’s existential graph concepts. *Ai & Society* 22, 101–112 (2007)
4. Oehlmann, R.: Place and self: From harmony strategies in the Japanese tea ceremony to conflict avoidance in team work. In: *Proceedings of the 8th International Workshop on Social Intelligence Design*, pp. 214–240 (2009)
5. Fujiwara, M.: *Kokka no Hinkaku*. Bilingual edn. Shinchosha, Tokyo (2005) (*The Dignity of a Nation*, translated by G. Murray)
6. Mara, M.: Japanese Aesthetics: The construction of meaning. *Philosophy East and West* 45(3), 367–386 (1995)
7. Kenko, Y.: *Essays in Idleness*. Cosimo Classics, New York (2005 [1911])
8. Takeuchi, S.: *Nihonjin wa naze “Sayounara” to wakareru no ga*. Chikuma Shobo, Tokyo (2009) (in Japanese, *Why Japanese People depart from Sayounara*)
9. Shounagon, S.: *The Pillow Book of Sei Shounagon*. Penguin, London (1967) (translated and edited by I. Morris)
10. Thomas, D.: *The Poems*. Dent, London (1974) (edited with an introduction and notes by Daniel Jones)
11. von Eichendorff, J.: *Saemtliche Gedichte and Versepen*. Insel Verlag, Frankfurt (2007) . (in German, edited by H. Schultz)
12. Ochi, N.: A cognitive study of TIME conceptualizations in Spanish and Japanese. *Intercultural Communication Studies XII-1*, 113–121 (2003)
13. Langdridge, D.: *Phenomenological Psychology: Theory, Research and Method*. Pearson Education, Harlow (2007)
14. Sato, M.: Over 1200 years continuing wishful prayers for Omizutori. In: *Todaiji. Shukan Bukkyo Shin Hakken*, vol. 4, Asahi Shinbun-sha, Tokyo (2007) (in Japanese)
15. Fruchter, R., Bosch-Sijtsema, P.: The wall: participatory design workspace in support of creativity, collaboration, and socialization. In: *Proceedings of the 8th International Workshop on Social Intelligence Design*, pp. 174–186 (2009)

Discovering Research Key Terms as Temporal Patterns of Importance Indices for Text Mining

Hidenao Abe and Shusaku Tsumoto

Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org

Abstract. For researchers, it is important to continue discovering and understanding key topics on their own fields. However, the analysis is almost depended on their experiences. In order to support for discovering emergent key topics as key terms in given textual datasets, we propose a method based on temporal patterns in several data-driven indices for text mining. The method consists of an automatic term extraction method in given documents, three importance indices, and temporal patterns based on results of clustering and linear trends of their centroids. Empirical studies show that the three importance indices are applied to the titles of two academic conferences about artificial intelligence field as sets of documents. After extracting the temporal patterns of automatically extracted terms, we compared the trends of the technical terms among the titles of the conferences.

Keywords: Text Mining, Trend Detection, TF-IDF, Jaccard's Matching Coefficient, Temporal Clustering, Linear Regression.

1 Introduction

In recent years, the development of information systems in every field such as business, academics, and medicine, and the amount of stored data have increased year by year. Accumulation is advanced to document data by not the exception but various fields. Document data provides valuable findings to not only domain experts but also novice user. For such electrical documents, it becomes more important to sense emergent targets about researchers, decision makers, marketers, and so on. In order to realize such detection, emergent term detection (ETD) methods have been developed [1,2].

However, because the frequency of words was used in earlier methods, their detections were difficult as long as the word that became an object did not appear. Besides, emergent or new concepts are usually appeared as new combination of multiple words, coinages created by an author, and words with different spellings of current words. Most conventional methods did not consider abovementioned natures of terms and importance indices separately. This causes difficulties in text mining applications, such as limitations on the extensionality of time direction, time consuming post-processing, and generality expansions.

After considering these problems, we focus on temporal behaviors of importance indices of terms and their temporal patterns. Temporal behaviors of the importance indices of extracted phrases are paid attention so that a specialist may recognize emergent terms and/or such fields. In order to detect various temporal patterns of behaviors of terms in the sets of documents, we have proposed a framework to identify the remarkable terms as continuous changes of multiple metrics of the terms [3].

In this paper, we describe an integrated framework for detecting trends of technical terms by combining automatic term extraction methods, importance indices of the terms, and temporal clustering in Section 2. After implementing this framework with the three importance indices, we performed a case study to extract remarkable temporal patterns of technical terms on the titles of AAAI and IJCAI in Section 3. Finally, in Section 4, we summarize this paper.

2 An Integrated Framework for Identifying Temporal Patterns of Technical Terms Based on Importance Indices

In this section, we describe the difference between conventional ETD methods and our proposal; detecting continuous temporal patterns of terms in temporal sets of documents.

As illustrated in Fig. 1, in order to find remarkable temporal trends of terms, we developed a framework for detecting various temporal trends of technical

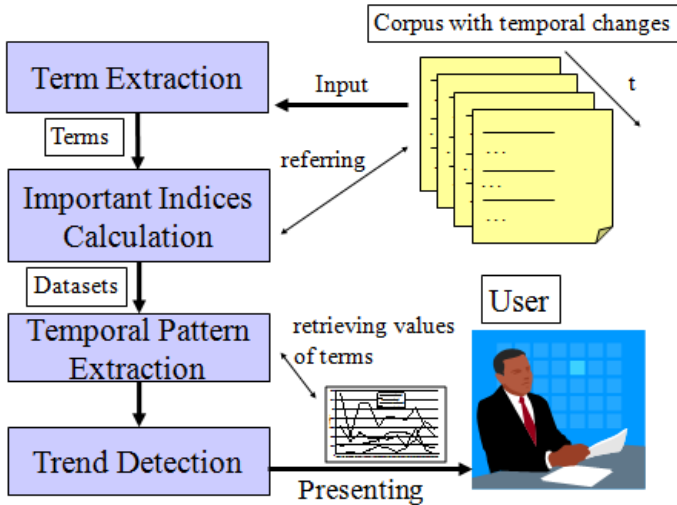


Fig. 1. Overview of the integrated temporal pattern extraction framework for technical terms

terms by using multiple importance indices consisting of the following four components:

1. Technical term extraction in a corpus
2. Importance indices calculation
3. Temporal pattern extraction
4. Trend detection

There are some conventional methods of extracting technical terms in a corpus on the basis of each particular importance index [2]. Although these methods calculate each index in order to extract technical terms, information about the importance of each term is lost by cutting off the information with a threshold value. We suggest separating term determination and temporal trend detection based on importance indices. By separating these phases, we can calculate different types of importance indices in order to obtain a dataset consisting of the values of these indices for each term. Subsequently, we can apply many types of temporal analysis methods to the dataset based on statistical analysis, clustering, and machine learning algorithms.

2.1 Extracting Technical Terms in the Given Corpus

First, the system determines terms in the given corpus. There are two reasons why we introduce term extraction methods before calculating importance indices. One is that the cost of building a dictionary for each particular domain is very expensive task. The other is that new concepts need to be detected in a given temporal corpus. Especially, a new concept is often described in the document for which the character is needed at the right time in using the combination of existing words.

Considering the difficulties of the term extraction without any dictionary, we apply a term extraction method that is based on the adjacent frequency of compound nouns. This method involves the detection of technical terms by using the following values for each candidate continued noun CN :

$$FLR(CN) = f(CN) \times \left(\prod_{l=1}^L (FL(N_l) + 1)(FR(N_l) + 1) \right)^{\frac{1}{2L}}$$

where $f(CN)$ means frequency of the candidates CN , and $FL(N_l)$ and $FR(N_l)$ indicate the frequencies of different words on the right and the left of combinations of nouns N_l in bi-grams including each CN . In this scoring method, N_l s are the parts of the continued nouns with $l(1 \leq l \leq L)$ nouns. In the experiments, we selected technical terms with this FLR score as $FLR(CN) > 1.0$. This threshold is important to select adequate terms at the first iteration. However, since our framework assumes the whole process as iterative search process for finding required trends of terms by a user, the user can input manually selected terms in the other iterations. In order to determine terms in this part of the process, we can also use other term extraction methods and terms/keywords from users.

2.2 Calculating Importance Indices of Text Mining in Each Period

As for importance indices of words and phrases in a corpus, there are some well-known indices. Term frequency divided by inversed document frequency (tf-idf) is one of the popular indices used for measuring the importance of the terms. tf-idf for each term $term_i$ can be defined as follows:

$$TFIDF(term_i, D_{t_j}) = TF(term, D_{t_j}) \times \log \frac{|D_{t_j}|}{DF(term_i, D_{t_j})}$$

where $TF(term, D_{t_j})$ is the frequency of each term $term_i$ in the set of t_j th documents D_{t_j} , and $|D_{t_j}|$ means the number of documents included in each period from t_{j-1} to t_j . $DF(term, D_{t_j})$ is the frequency of documents containing $term_i$.

As another importance index, we use Jaccard’s matching coefficient [4][1]. Jaccard coefficient can be defined as follows:

$$Jaccard(term_i, D_{t_j}) = \frac{DF(w_1 \cap w_2 \cap \dots \cap w_L, D_{t_j})}{DF(w_1 \cup w_2 \cup \dots \cup w_L, D_{t_j})}$$

where $DF(w_1 \cap w_2 \cap \dots \cap w_L, D_{t_j})$ is equal to $DF(term_i, D_{t_j})$, because each $term_i$ consists of word or one more words $\{w_l | 1 \leq l \leq L\}$. $DF(w_1 \cup w_2 \cup \dots \cup w_L, D_{t_j})$ means the frequency of documents that contains at least one word $w \in \{w_l | 1 \leq l \leq L\}$ that are included in the term $term_i$. Jaccard coefficient is originally defined as the ratio of the probability of an intersection divided by the probability of a union in a set of documents. In this framework, we applied this index by defining as the ratio of the frequency of an intersection divided by the frequency of a union in each set of documents D_{t_j} . Each value of Jaccard coefficient shows strength of co-occurrence of multiple words as an importance of the terms in the set of documents.

In addition to the above two importance indices, we used simple appearance ratio of terms in a set of documents.

$$Odds(term_i, D_{t_j}) = \frac{DF(term_i, D_{t_j})}{|D_{t_j}| - DF(term_i, D_{t_j})}$$

where, $DF(term_i, D_{t_j})$ means the frequency of the appearance of each term $term_i$ in each set of documents D_{t_j} .

Fig.2 shows an example of the dataset consisting of an importance index for each year.

2.3 Analyzing Temporal Patterns of Technical Terms Based on Importance Indices

Then, the framework provides the choice of some adequate trend extraction method to the dataset. In order to extract useful time-series patterns, there are

¹ Hereafter, we refer to this coefficient as “Jaccard coefficient”.

	t_1	t_2	...	t_n
$term_1$	v_{11}	v_{12}	... v_{1j} ...	v_{1n}
$term_2$	v_{21}	v_{22}	... v_{2j} ...	v_{2n}
	v_{i1}	v_{i2}	... v_{ij} ...	v_{in}
	.	.		.
	.	.		.
	.	.		.
$term_m$	v_{m1}	v_{m2}	... v_{mj} ...	v_{mn}

Fig. 2. Example of a dataset consisting of an importance index

so many conventional methods as surveyed in the literatures [5][6]. By applying an adequate time-series analysis method, users can find out valuable patterns by processing the values in rows in Fig. 2

After identifying temporal patterns in the dataset, we apply the linear regression analysis technique in order to detect the degree of existing trends for each importance index. The degree of each term $term_i$ is calculated as the following:

$$Deg(term_i) = \frac{\sum_{j=1}^m (y_j - \bar{y})(x_j - \bar{x})}{\sum_{j=1}^m (x_j - \bar{x})^2}$$

where the value in x-axis x_j means each timepoint as calculated $x_j = t_j - t_1$, and the value in y-axis y_j means the value v_{ij} of one importance index $Index_x$. In this definition, \bar{x} is the average of overall period with the m time points, and \bar{y} is the average of the values of each importance index for the period. Simultaneously, we calculate the intercept $Int(term)$ of each term $term_i$ as follows:

$$Int(term_i) = \bar{y} - Deg(term_i)\bar{x}$$

We observed the linear trends of each temporal pattern that are identified as temporal clusters. For each cluster c , the averaged degrees $Avg.Deg(c) = \sum_{term_i \in c} Deg(term_i) / num(term_i \in c)$ and intercepts $Avg.Int(c) = \sum_{term_i \in c} Int(term_i) / num(term_i \in c)$ of each term $term_i$ are used in the following experimentation.

3 Experiment: Extracting Temporal Patterns of Technical Terms by Using Temporal Clustering

In this experiment, we show the results temporal patterns by using the implementation of the method described in Section 2. As the input of temporal documents, we used the sets of the titles of the following two academic conferences [3], AAAI and IJCAI.

² These titles are the part of the collection by DBLP [7].

We determine technical terms by using the term extraction method [8]³ for each entire set of documents.

Subsequently, the values of tf-idf, Jaccard coefficient, and Odds are calculated for each term in the annual documents. To the datasets consisting of temporal values of the importance indices, we extract temporal patterns by using k-means clustering. Then we apply the meanings of the clusters based on their linear trends calculated by the linear regression technique for the timeline.

3.1 Extracting Technical Terms

We use the titles of the two artificial intelligence (AI) related conferences as temporal sets of documents. The numbers of the titles as the documents is shown in Fig. 3.

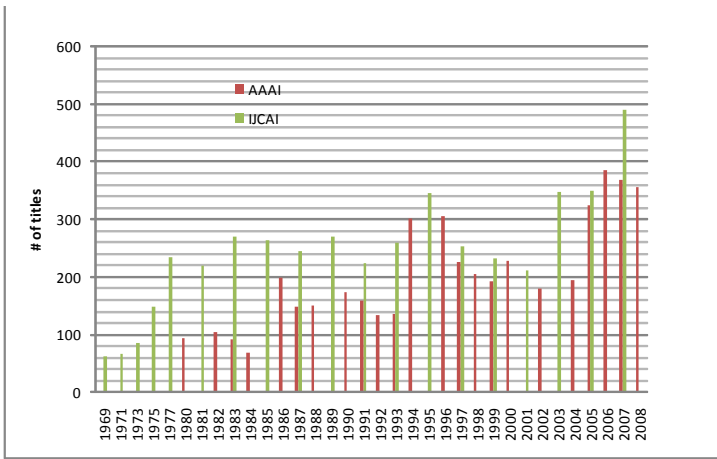


Fig. 3. The numbers of the titles of IJCAI and AAI

As for the sets of documents, we assume each title of the articles to be one document. Note that we do not use any stemming technique because we want to consider the detailed differences in the terms.

By using the term extraction method with simple stop word detection for English, we extract technical terms. From AAI titles, the method extracted 9035 terms, and 8659 terms from IJCAI titles.

3.2 Extracting Temporal Patterns by Using K-Means Clustering

In order to extract temporal patterns of each importance index, we used k-means clustering. We set up the numbers of one percent of the terms as the maximum

³ The implementation of this term extraction method is distributed in <http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html> (in Japanese).

number of clusters $k = 0.01 \times n$ for each dataset. Then, the system obtained the clusters with minimizing the sum of squared error within clusters. By iterating less than 500 times, the system obtains the clusters by using Euclidian distance between instances consisting of the values⁴ of the same index.

Table 1 shows the result of the SSE of k-means clustering. As shown in this table, the SSE values of Jaccard coefficient are higher than the other two indices: tf-idf and odds. Since we were not selected the terms with two or more words, the values of Jaccard coefficient of the terms with just one word, which are 0 or 1, are not suitable to make clusters.

Table 1. The sum of squared errors of the clustering for the technical terms in the titles of the two conferences

Dataset	tf-idf	Jaccard	Odds
AAAI	53.91	2538.29	4.10
IJCAI	29.78	1961.44	2.49

3.3 Details of a Temporal Pattern of the Key Technical Terms

By focusing on the result of the titles of IJCAI with tf-idf, in this section, we show the details of the temporal pattern. We obtained up to 86 clusters with this setting for extracting temporal patterns, the SSE is shown in Table 1.

As shown in Table 2, the k-means clustering found 33 temporal clusters based on the temporal sequence of tf-idf values of the technical terms. The centroid terms mean the terms that are the nearest location to the centroids. Then, by using the averaged degree and the averaged intercept of the terms within each cluster c , we attempt to determine the following three trends:

- Popular
 - the averaged degree is positive, and the intercept is also positive.
- Emergent
 - the averaged degree is positive, and the intercept is negative.
- Subsiding
 - the averaged degree is negative, and the intercept is positive.

The averaged values of the tf-idf as temporal patterns are visualized in Fig. 4.

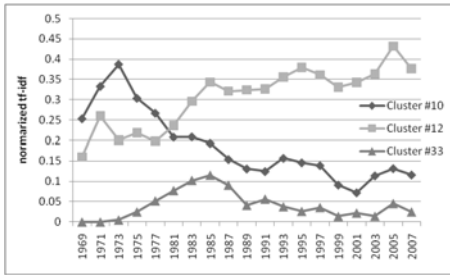
The pattern #10, #12 and #33 shows the high averaged tf-idf values in Fig. 4(a). The reason why these patterns have high scores is that they have been frequently used in the titles for the overall or the past period. Although the numbers of terms included in the patterns are not so uniform, the centroid terms are reasonable to understand their usages in the overall period.

According to the meanings based on the linear trend, the patterns #5, #6, and #8 have the emergent patterns in Fig. 4(b). The terms included in these patterns are appeared in recent years in the conference. At the same time, the

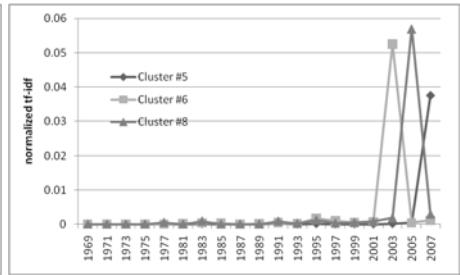
⁴ The system also normalized the values from 1 to 0 for each year.

Table 2. The centroid information about the temporal patterns of tf-idf of the terms on the titles of IJCAI from 1969 to 2007

Cluster#	# of members	Term	Avg.Deg(c)	Avg.Int(c)	Meaning
1	3457	probabilistic inference	0	0.0002	Popular
2	267	expert systems	-0.0001	0.0046	Subsiding
3	298	semantic analysis	0.0004	0.0009	Popular
4	246	network learning	0	0.0041	Popular
5	480	data	0.0003	-0.001	Emergent
6	382	using information integration	0.0004	-0.0004	Emergent
7	29	description logics	0.002	0.0076	Popular
8	363	using ai	0.0005	-0.001	Emergent
9	287	multi-agent systems	0.0005	0.0006	Popular
10	19	system	-0.0075	0.2574	Subsiding
11	308	learning system	0.0003	0.0017	Popular
12	11	learning	0.0066	0.2445	Popular
13	29	reinforcement learning	0.0022	0.0328	Popular
14	24	structure	-0.0031	0.1098	Subsiding
15	259	causal reasoning	-0.0001	0.0043	Subsiding
16	261	model of planning	0	0.0032	Popular
17	87	robot control	-0.0014	0.029	Subsiding
18	47	representations	0.001	0.0205	Popular
19	295	planning using	0.0002	0.0026	Popular
20	266	concept learning	-0.0002	0.0065	Subsiding
21	25	logic programming	0.0011	0.0206	Popular
22	100	fuzzy logic	-0.0013	0.0284	Subsiding
23	226	vision system	-0.0003	0.0067	Subsiding
24	24	information extraction	0.0029	0.0231	Popular
25	15	models	0.0055	0.084	Popular
26	34	machine learning	0.0019	0.0025	Popular
27	330	using knowledge	0.0002	0.0012	Popular
28	30	heuristic search	-0.0013	0.0563	Subsiding
29	113	automatic programming	-0.001	0.0203	Subsiding
30	20	solving	0.0001	0.1156	Popular
31	28	action	0.0033	0.0479	Popular
32	193	solving using	-0.0005	0.0116	Subsiding
33	27	expert system	0	0.0413	Popular
34	1	a	0	1	Popular



(a)



(b)

Fig. 4. The temporal patterns of tf-idf of terms from the titles of IJCAI: (a) the patterns have high averaged scores, (b) the patterns have emergent trend defined by using the averaged degree and intercept

appearances reflect the researchers interests, because these articles through the competitive peer-review process.

Focusing on the detail of the emergent pattern, Table 3 shows top ten terms included in the pattern #6, which is the biggest cluster with the emergent linear trend. Although they were not used in the titles from '80s or more recently, researchers become to use these terms to express their key topics in these years.

Table 3. Top ten emergent terms included in the pattern #6

RANK	term	tf-idf		Jaccard		Odds	
		Deg(term)	Int(term)	Deg(term)	Int(term)	Deg(term)	Int(term)
1	comparing	0.001	0.002	0.017	0.053	0.0003	-0.0001
2	web search	0.001	-0.001	0.001	0.000	0.0004	0.0003
3	local consistency	0.001	-0.001	0.003	-0.001	0.0003	-0.0003
4	unifying	0.001	-0.002	0.020	-0.032	0.0003	-0.0003
5	ensembles	0.001	-0.001	0.019	-0.023	0.0003	-0.0002
6	experimenting	0.001	0.000	0.015	0.012	0.0003	0.0000
7	querying	0.001	0.000	0.016	0.003	0.0003	0.0001
8	information integration	0.001	0.001	0.001	0.002	0.0003	-0.0001
9	data integration	0.001	-0.002	0.002	-0.003	0.0003	-0.0005
10	services	0.001	-0.002	0.014	-0.030	0.0002	-0.0005

As shown in Table 3, the linear trends of the three indices are almost the same, excepting ‘web search’ and ‘information integration’. By gathering similar temporal sequences of the terms, we can find more promising linear trends of the technical terms by using the three indices, compared to the result in our previous study [9].

4 Conclusion

In this paper, we proposed a framework to detect remarkable temporal patterns of technical terms appeared as the temporal behaviors of the importance indices in the sets of temporally published documents. We implemented the framework with the automatic term extraction, the three importance indices, and temporal pattern detection by using k-means clustering.

The empirical results show that the temporal patterns of the importance indices can detect the trends of each term, according to their values for each annual set of the titles of the two academic conferences that are famous on AI field. Regarding to the results, the temporal patterns indicate that we can find out not only the current key terms, but also historical keys of the research field by using the linear trends of the periodical values of the importance indices.

In the future, we will apply other term extraction methods, importance indices, and trend detection method. As for importance indices, we are planning to apply evaluation metrics of information retrieval studies, probability of occurrence of the terms, and statistics values of the terms. Then, we will apply this framework to other documents from various domains.

References

1. Lent, B., Agrawal, R., Srikant, R.: Discovering trends in text databases, pp. 227–230. AAAI Press, Menlo Park (1997)
2. Kontostathis, A., Galitsky, L., Pottenger, W.M., Roy, S., Phelps, D.J.: A survey of emerging trend detection in textual data mining. *A Comprehensive Survey of Text Mining* (2003)
3. Abe, H., Tsumoto, S.: Detecting temporal trends of technical phrases by using importance indices and linear regression. In: Rauch, J., Raś, Z.W., Berka, P., Elomaa, T. (eds.) *ISMIS 2009*. LNCS, vol. 5722, pp. 251–259. Springer, Heidelberg (2009)
4. Anderberg, M.R.: *Cluster Analysis for Applications*. Monographs and Textbooks on Probability and Mathematical Statistics. Academic Press, Inc., New York (1973)
5. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In: *In an Edited Volume, Data mining in Time Series Databases*, pp. 1–22. World Scientific, Singapore (2003)
6. Liao, T.W.: Clustering of time series data: a survey. *Pattern Recognition* 38, 1857–1874 (2005)
7. The dblp computer science bibliography, <http://www.informatik.uni-trier.de/~ley/db/>
8. Nakagawa, H.: Automatic term recognition based on statistics of compound nouns. *Terminology* 6(2), 195–210 (2000)
9. Abe, H., Tsumoto, S.: Detecting temporal patterns of importance indices about technical phrases. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) *KES 2009, Part II*. LNCS, vol. 5712, pp. 252–258. Springer, Heidelberg (2009)

Categorized and Integrated Data Mining of Medical Data from the Viewpoint of Chance Discovery

Akinori Abe¹, Norihiro Hagita², Michiko Furutani³,
Yoshiyuki Furutani³, and Rumiko Matsuoka⁴

¹ NTT Communication Science Laboratories

2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan

² ATR Intelligent Robotics and Communication Laboratories, Japan

³ Department of Pediatric Cardiology, Tokyo Women's Medical University, Japan

⁴ Wakamatsu Kawada Clinic, Japan

ave@cslab.kecl.ntt.co.jp, ave@ultimaVI.arc.net.my, hagita@atr.jp,
{michi,yoshi,rumiko}@imcir.twmu.ac.jp

Abstract. In this paper, we analyze the procedure of computational medical diagnosis based on collected medical data. Especially, we focus on features or factors which interfere with sufficient medical diagnoses. In order to reduce the data complexity, we introduce medical data categorization. Data are categorized into six categories to be analyzed and to generate rule sets for medical diagnosis. We analyze the relationships among categorized data sets within the context of chance discovery, where hidden or potential relationships lead to improved medical diagnosis. We then suggest the possibility of integrating rule sets derived from categorized data for improving the accuracy of medical diagnosis.

Keywords: medical diagnosis, categorized data mining, hidden relationships.

1 Introduction

We are able to live longer due to the advanced and innovative medical treatment. Living longer is very desirable; however long lives also cause problems. One of the best known problems is the increasing number of patients who suffer from cancer. Advancements in medical treatment will be able to overcome these illnesses. However, it is even more important to prevent them in the earlier place. Therefore it is necessary to detect the symptoms of cancer early and to predict their occurrence. In a project at the International Research and Educational Institute for Integrated Medical Science (IREIIMS), we are collecting medical data sets to analyze and generate relationships between medical data and the health status of patients [Abe et al., 2008a, Abe et al., 2008b]. Various investigations have been conducted that focussed on medical data mining. Most notable are the contributions of Tsumoto [Tsumoto, 2004] and Ohsawa [Ohsawa et al., 2004] which pointed out the importance of medical data mining for discovering rare or unknown (interesting) rules for a medical diagnosis. Of course the main aim of the

current project is the generation of plausible relationships between medical data and health status. However, during the analyses of complex medical data, which included more than 130 different items, it was found that it is rather difficult to simply analyze these data. Moreover, the data sets include various types of data. For instance, a data set may include data of patients with lung cancer and patients with stomach cancer. It is sometimes hazardous to use such mixed and complex data for data mining. Accordingly, it has been suggested that in the case of multiple categorized (mixed) data, it is difficult to discover hidden or potential knowledge [Abe et al., 2007a]. Therefore an approach of integrated data mining was proposed. In addition, Abe et al. found that parts of health level models, especially at health level 3, could not be built correctly, because the data distribution was unbalanced and the data size was insufficient [Abe et al., 2008a]. Additional findings indicated that the health level determination system cannot determine the health levels correctly by using such rule-based models, because combinations of certain factors are rare. In some cases, (over-)detailed or over-fitting learning situations were identified. If such type of model is used, it is difficult to determine a patient's health status properly. A solution for this type of problem involves the reduction of the data complexity. This paper describes the procedure of computational medical diagnosis based on collected medical data. It focuses particularly on features and factors that interfere with sufficient medical diagnoses. In order to reduce the data complexity, the method of medical data categorization is introduced. Data are categorized into six categories, which are then analyzed and used to generate the rule sets for medical diagnosis. Next the relationships between the categorized data sets are investigated in the context of chance discovery, where hidden or potential relationships lead to improved medical diagnosis. Subsequently the possibility of integrating the rule sets obtained from mining categorized data is suggested to improve the accuracy of medical diagnosis. To date, ROC (Receiver Operating Characteristic) curve analysis had mainly been used for diagnostic studies in clinical chemistry, pharmacology and physiology. In this analysis, a graphical technique is used to choose a cut-off point for the 'normal' and 'abnormal' categories [Akobeng, 2007]. The most effective parameter can be selected by ROC curve analysis. However this approach cannot discover hidden relationships between several parameters.

2 The Characteristics of the Medical Data

2.1 Clinical Data

Currently we are collecting various types of clinical data, for instance from blood and urine samples. Some individuals were requested to provide samples over a period of time. The data set of each individual contains more than 130 items. Examples of items are, for instance, β 2-microglobulin (serum), SCC antigen, γ -seminoprotein, TPA, TK activity, NSE, Ferritin, BFP (serum), Total protein, albumin, γ -GTP, ZTT, serum protein fraction- α 1-globulin, ALP Triglyceride, LDL Cholesterol, HDL Cholesterol, urine sugar, HbA1c, White blood cell count, Platelet, Cl, Cell immunity T Cell number, and ACP. In addition, health levels

are assigned to each data item by physicians, who consider the medical data and clinical interviews. Health levels that express the health status of patients are defined based on *Tumor stage* [Kobayashi and Kawakubo, 1994] and modified by Matsuoka. Originally, health levels consist of 5 levels (I–V). Persons at levels I and II can be regarded as being healthy, but those at levels III, IV, and V can possibly develop cancer. Kobayashi and Kawakubo have defined level III as the stage before the shift to preclinical cancer [Kobayashi and Kawakubo, 1994]. Level IV is defined as conventional stage 0 cancer (G0), and level V is defined as conventional stages 1–4 cancer (G1–G4). For further analysis, Matsuoka has described the more detailed categorization of I, II, III, IVa, IVb, IVc, Va and Vb. Since health levels IV and V include more data, they are categorized in a more fine-grained way. In the following sections, descriptions of 1,2,3,4, and 5 are used for health levels instead of I, II, III, IV, and V.

Table 1. Health levels

health level	I	II	III	IVa	IVb	IVc	Va	Vb
ratio (%)	0.0	0.0	2.5	13.4	49.4	24.5	8.5	1.6

Currently, more than 3500 data sets with health level assignments have been collected. Medical data have mainly been collected from the same patient group as described in [Abe et al., 2007a]; i.e. the data were collected from office workers with an age range from 40 to 50, but not from students. The health level distribution is therefore still unbalanced, in spite of the introduction of a new categorization (Table 1). So an imbalance in the data may still influence the analyses.

2.2 Data Categorization

For a detailed analysis, we currently categorize the above mentioned items into the six categories listed below. The categorization has been determined according to effects or functions related to the health status, such as cancer and metabolism.

- 1) peculiar tumor markers (21 items):
- 2) inflammatory tumor markers (11 items):
- 3) liver, pancreas, and kidney test data (28 items):
- 4) metabolic function test data (16 items):
- 5) blood and immunity test data (12 items):
- 6) others (47 items):

These categories include items such as those shown below

- 1) β 2-microglobulin (serum), SCC antigen, γ -seminoprotein, TPA
- 2) TK activity, NSE, Ferritin, BFP (serum)
- 3) Total protein, albumin, γ -GTP, ZTT, Serum protein fraction- α 1-globulin, ALP
- 4) Triglyceride, LDL Cholesterol, HDL cholesterol
- 5) Urine sugar, HbA1c, White blood cell count, Platelet
- 6) Cl, Cell immunity T Cell number, ACP

3 Medical Data Analysis and Medical Diagnosis

In this section, the analysis of the collected data is described from the viewpoint of the patient's health status.

3.1 Medical Data Analysis and Health Level Determination

First, entire medical data set (with 3590 records) was analyzed using the C4.5 program [Quinlan, 1993], which resulted in the following decision tree (Ht denotes Hematocrit):

```

Ht (%) <= 34.4
|   B cell (CD20) (%) <= 21
|   |   Na (mEq/l) <= 133: 5b
|   |   Na (mEq/l) > 133
|   |   |   SLX (U/ml) <= 32.4
|   |   |   |   Inirect bilirubin (mg/dl) <= 0.572546: 5a
|   |   |   |   Inirect bilirubin (mg/dl) > 0.572546: 4b
|   |   |   |   SLX (U/ml) > 32.4: 5b (3.0/1.0)
|   |   B cell (CD20) (%) > 21
|   |   |   K (mEq/l) <= 3.9: 4b
|   |   |   K (mEq/l) > 3.9: 4a
Ht (%) > 34.4
|   TPA (U/l) <= 62
(to be continued)

```

The decision tree was viewed as rule set for medical diagnosis where health levels can be determined based on medical data. The rule set was applied to the same data set¹. This resulted in a accuracy ratio of 94.9%. However, when the data set is divided into two subsets of 3000 and 590 records to generate a rule set from 3000 records and apply it to the remaining 590 records, the accuracy ratio is reduced to 45.5%. Actually both decision trees appear similar. A proper model might not be generated, if the samples are heterogeneous, scarce or not existing. For instance, sometimes useless or harmful rules, such as “TTT <= 0.7” are included in the rule set. For healthy individuals, the TTT value can be less than 4.0; i.e. the rule is correct but establishes a very strict condition. Individuals who will not be classified at health level 4x(x=a,b,c) will be classified at level 4x instead of 3. Therefore as shown in the Introduction section, (over-)detailed or over-fitting learning appears to occur, if the samples are heterogeneous or the sample size is too small. For instance, Table 2 shows a part of the results obtained by determining the health levels. There are differences between the results obtained from the rule set based on 3500 records and those obtained from the rule set based on 3000 records. However it is difficult to identify any trend in these differences.

¹ The rule set was applied to the subset of 590 records. Later it will be shown that this is equivalent to using the entire data set.

Table 2. Differences in determining health levels (based on rule sets derived from 3590 and 3000 records.)

from 3590	4a	4b	4b	4c	4b	4b	4c	4a	4a	4a	4c	4b	4b	4c	4c	4c	4a	4a	4a	3	4b	4a
from 3000	4b	4a	4b	4b	4c	4b	4c	3	3	3	4c	4b	4b	5a	4b	5a	3	3	3	3	4b	4b
actual HL	4a	4b	4b	4c	4b	4b	4c	4a	4a	4a	4b	4b	4b	4c	4c	4c	4a	4a	4a	3	4b	4a

3.2 Categorized Data Analysis

As shown in the previous section, currently (over-)detailed or over-fitting rule sets will be generated, if the entire data set is used for the rule set generation. Therefore, it would be better to generate rule sets from relatively simple data sets. In order to reduce the data complexity, the data categorization strategy that was mentioned in the previous section is introduced. This is considered to be a medically meaningful categorization. It is therefore expected, that the data can be divided into relatively small subsets according the to patient’s health situation characterised by illnesses such as lung cancer and stomach cancer.

After generating rule sets from the categorized medical data, the accuracies for determining health levels given in Table 3 can be obtained. Similarly, the health levels are determined by generating rule sets from the categorized data and applying them to the same data set of 590 records that was used in the previous section. For instance, a rule set generated from the data category of “peculiar tumor markers” was applied to the data set of 590 records to determine the health levels for these data.

Table 3. Accuracy ratios for medical diagnoses

category	accuracy ratio (%)
peculiar tumor markers	42.7
inflammatory tumor markers	38.5
liver, pancreas, and kidney test data	30.7
metabolic function test data	13.7
blood and immunity test data	47.5
others	2.03

The table shows that the accuracy ratio is better for the rule set that was generated from the “blood and immunity test data” than for that generated from the entire medical data, although both decision trees are similar. The result suggests that in these medical data the “blood and immunity test data” play a central role for the medical diagnosis. Consequently the “blood and immunity test data” are a very influential factor for determining the health level. Therefore, subsets of the medical data that are not part of the “blood and immunity test data” may interfere with a satisfactory medical diagnosis. Missing or unnecessary links may cause parts of the necessary medical data to be ignored. In order to perform a better medical diagnosis, it is necessary to add additional rule sets and to ignore unnecessary rule modules.

3.3 Data That Might Disturb Sufficient Medical Diagnosis

It was noted before that health levels were correctly determined by the rule set based on the entire medical data set, but were incorrectly determined by the rule set based on the data subset of 3000 records. In addition, cases were considered where neither of the rule sets was able to determine the health levels correctly. It is therefore necessary to discover relationships, which can suggest the removal of unnecessary links and can aid the recovery or supplement of missing rules or links that lead to the determination of the correct health levels. Several relationships that could improve the accuracy in determining health levels were discovered.

- 1) Cases where the health level (HL) 4b was assigned. This group includes cases where rule sets that were derived from the entire data set can correctly determine HL, but rule sets derived from the data set with 300 records often failed to determine the HL correctly. So rule sets derived from “blood and immunity data” could determine HL correctly, whereas rule sets derived from “metabolic function test data” could not correctly determine HL. The differences suggest that for health level 4b, that the first type of rule set functions effectively, but the second type may interfere with the determination of health levels.
- 2) Rule sets derived from “blood and immunity test data” can determine mainly health level 4b correctly. However for the other health levels, it is rather difficult to determine HL.
- 3) Rule set derived from “metabolic function test data” can only determine health levels correctly that are worse than 4c.
- 4) There are not so many cases, but rule sets from “peculiar tumor markers” can determine all health levels correctly. This applies even if rule sets derived from the data set with 3000 records cannot determine the health levels correctly.
- 5) Rule sets derived from “inflammatory tumor markers” and “liver, pancreas, and kidney test data” can correctly determine health levels 4b and 4c, even if the rule sets that were derived from the data set with 3000 records were not able to determine the health levels correctly.
- 6) Rule sets derived from the “others” category are usually able to determine the restricted health levels. This indicates that the category does not include necessary data to generate links to the other health levels.

The above relationships could not be discovered by analysing the entire medical data set. Moreover, those relationships will be important for generating suitable models for medical diagnosis. The discoveries could be achieved by dividing the data and changing the viewpoint. These relationships are not always aiding an improved determination of the health levels. For an improved medical diagnosis, it will be necessary to introduce suitable combinations of relationships with rule sets that were generated from the entire medical data set. These results demonstrate the possibility of integrated data mining that was proposed by Abe et al. [Abe et al., 2007a](#), [Abe et al., 2008b](#).

4 Chance Discovery in Medical Data Analysis

The results of the above analyses will make it possible to improve the quality of medical diagnosis by adding additional relationships to the rule set that was generated from the entire data set. For instance, the following relationships could be added to the rule set ((1') is stricter condition than (1)):

- 1) IF the health level determined by a rule set that was derived from the "blood and immunity test data" is 4b AND more than two of the health levels determined by the rule sets that were derived from "peculiar tumor markers", "inflammatory tumor markers", and "liver, pancreas, and kidney test data" are 4b, THEN the patient's health level will be 4b.
- 1') IF the health level determined by a rule set that was derived from the "blood and immunity test data" is 4b AND all of the health levels determined by the rule sets that were derived from "peculiar tumor markers", "inflammatory tumor markers", and "liver, pancreas, and kidney test data" are 4b, THEN the patient's health level will be 4b.

In the case of (1), the accuracy ratio for determining the health level increases by 7.5%, and in the case of (1')), the accuracy ratio for determining the health level increases by 3.0%. Compared with the above case, the health level is less frequently incorrectly determined. These incorrect health levels were correctly identified by the rule set that was derived from the entire data set. The addition of discovered relationships can therefore increase the accuracy ratio of determining the health levels. It depends on the user's choice which advantage is preferred: fewer incorrectly determined health levels or a higher accuracy ratio in determining health levels. In relation to the categorization of medical data it has been shown that hidden but necessary relationships could be discovered. This paper only examines one relationship, but the other relationships described in the previous section may also be effective. Also these relationships should be able to increase the accuracy ratio of determining health levels or generate better models for the medical diagnosis. Furthermore, there might be other categorizations, which can lead to better relationships, which can then be used to improve the accuracy ration of determining health levels further or the model generation for the medical diagnosis. It is therefore necessary to discover hidden or potential factors for improving the medical diagnosis.

5 Conclusions

The approach of categorized data mining has been introduced and it has been argued for an integration of results that are obtained by categorized data mining to generate better models. It is difficult to discover useful models that can be applied to many phenomena, if the mined data include many and various types of data. Therefore, the medical data were partitioned according to medically meaningful categorizations. This enabled the discovery of hidden relationships among the data, which can be used to improve the accuracy of the medical diagnosis.

Only one example of improving the accuracy of a medical diagnosis was shown and the improvement was not very large. However, if more hidden relationships and rule sets could be discovered, it would be possible to improve the medical diagnosis further. This paper also demonstrated the possibility of discovering hidden relationships during categorized data mining and suggested that the discovered relationships can be used for integrating the results of categorized data mining. As shown in the previous studies, categorized and integrated data mining is one of the promising strategies for complex data mining. There should be other useful relationships and rule sets that cannot be discovered easily. In the next stage of the research, it is necessary to propose a new type of strategy to discover such hidden and useful relationships and rule sets.

Acknowledgments

This research was conducted when the authors were members of IREIIMS at the Tokyo Women's Medical University. It was supported in part by the Program for Promoting the Establishment of Strategic Research Centers, Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sports, Science and Technology (Japan). We are particularly indebted to Dr. Yoshiaki Umeda (NTT-AT) for supporting the development of the analysis tools.

References

- [Abe et al., 2007a] Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Possibility of Integrated Data Mining of Clinical Data. *Data Science Journal* 6(supplement), S104–S115 (2007)
- [Abe et al., 2008a] Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Data mining of Multi-categorized Data. In: Raś, Z.W., Tsumoto, S., Zighed, D. (eds.) *MCD 2007. LNCS (LNAI)*, vol. 4944, pp. 182–195. Springer, Heidelberg (2008)
- [Abe et al., 2008b] Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Categorized and Integrated Data Mining of Clinical Data. In: Iwata, S., Ohsawa, Y., Tsumoto, S., Zhong, N., Shi, Y., Magnani, L. (eds.) *Communications and Discoveries from Multidisciplinary Data. Studies in Computational Intelligence*, vol. 123, pp. 315–330. Springer, Heidelberg (2008)
- [Akobeng, 2007] Akobeng, A.K.: Understanding diagnostic tests 3: Receiver Operating Characteristic Curves. *Acta Paediatr* 96(5), 644–647 (2007)
- [Kobayashi and Kawakubo, 1994] Kobayashi, T., Kawakubo, T.: Prospective Investigation of Tumor Markers and Risk Assessment in Early Cancer Screening. *Cancer* 73(7), 1946–1953 (1994)
- [Ohsawa et al., 2004] Ohsawa, Y., Fujie, H., Saiura, A., Okazaki, N., Matsumura, N.: Process to Discovering Iron Decrease as Chance to Use Interferon to Hepatitis B. In: *Proc. of AM 2004*, pp. 21–24 (2004)
- [Quinlan, 1993] Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco (1993)
- [Tsumoto, 2004] Tsumoto, S.: Mining Diagnostic Rules from Clinical Databases Using Rough Sets and Medical Diagnostic Model. *Information Sciences* 162(2), 65–80 (2004)

Support System for Thinking New Criteria of Unclassified Diseases

Yoko Nishihara¹, Yoshimune Hiratsuka², Akira Murakami³,
Yukio Ohsawa¹, and Toshiro Kumakawa²

¹ Department of Systems Innovation, School of Engineering, The University of Tokyo,
7-3-1, Hongo, Bunkyo, Tokyo 113-8656, Japan
nishihara@sys.t.u-tokyo.ac.jp

² Department of Management Science, National Institute of Public Health,
2-3-6, Minami, Wako, Saitama 351-0197, Japan

³ Department of Ophthalmology, Juntendo University, School of Medicine,
2-1-1, Hongo, Bunkyo, Tokyo 113-8421, Japan

Abstract. It is considered that many people are struggling with diseases that are difficult to diagnose. In general, it takes a long time to diagnose and to cure such diseases. There may be also people who are struggling with diseases that we have not classified yet. In Japan, methods for curing 56 incurable diseases have been studied. However, methods for discovering unclassified diseases have not been studied. This paper proposes a new system that supports doctors in thinking of new criteria for classifying diseases.

Keywords: unclassified disease, new criteria for unclassified diseases, analogy game, concept revision.

1 Introduction

There are many people who are struggling with diseases that are difficult to diagnose. According to the United Nations, at least four million people worldwide are suffering from Parkinson's disease [2]. In general, it takes a long time to diagnose and to cure such diseases. There may be also people who are struggling with unclassified diseases. In Japan, methods for curing 56 incurable diseases have been studied [1]. However, methods for discovering unclassified diseases have not been studied.

Previous investigations in information science have proposed many data mining methods. The methods can extract rare but important data from huge data sets. For example, Yamanishi et al. have proposed a method for detecting invasions in large database system [6]. Ohsawa et al. have proposed KeyGraph, an algorithm for discovering chances. These are rare but important events for decision making in business communications [4]. However, even medical doctors may not discover significant data in large data sets by these methods, because many noisy data are included in the sets. Therefore, new methods for discovering such data are required.

This paper proposes a new system that supports medical doctors in thinking about new criteria for unclassified diseases. A user of the system enters names of classified

diseases into the system. The system chooses 20 names from the input and shows them to the user on its visualizing interface. The maximum number of disease names is 20 because of the limitation of a display size. The user chooses some disease names whose diagnostics are similar to each other, and forms a group of the names. After forming the disease groups, if the user revises the decision of which group is appropriate for a disease name, the user writes the name and the reason for the revision. The system shows the reason as a candidate of a new criterion for classifying unclassified diseases.

In this research, we think of two types of new criteria for unclassified diseases. One of them is a criterion that is not usually considered by medical doctors, and the other involves symptoms of patients that have not been verbalized by the patients. In this paper, we focus on the former type and propose a new system to obtain the criteria.

The proposed system was created by improving the analogy game, which was invented by Nakamura et al. [3]. The analogy game evaluates human creativity from the playing result. The game gives 20 cards to a user from a card database. Nouns (strawberry, train, Mt. Fuji, university, baseball and so on) are written on each card. The user forms groups of cards whose characteristics are similar to each other, and describes conceptions for each group. A user who has formed original groups receives a high score in the game. In the analogy game, a user can move and color the cards by using a mouse. Groups are recognized based on their color. The maximum number of groups is five. The user writes conceptions in boxes in the bottom of its visualizing interface. Basically, a user of the analogy game has to follow two rules in grouping. One of requires forming less than five groups, and the other using all cards in the categorization. We improve the game by adding some rules on group formation and then describe our system.

2 Support System for Thinking of New Criteria for Unclassified Diseases

A user inputs more than 20 names of disease into the system. It chooses 20 names from the inputs randomly and shows the names to the user on its visualizing interface. The user makes groups of disease names whose characteristics are similar to each other and colors cards on which disease names are written with the same color (as shown in Figure 1). If the user has revised the group membership of a disease, the user writes the disease name and the reason why he/she has conducted the revision. The system outputs the reason as a criterion candidate for unclassified diseases.

The target users of the proposed system are limited to medical doctors because the general public does not know the details of diseases. Medical specialists usually diagnose patients considering symptom patterns of the patients. The patterns prevent to discover unclassified diseases because the symptom patterns may be the same between classified diseases and unclassified disease. We thought that seeing patient symptoms from many criteria was useful for finding unclassified diseases. A doctor as a user inputs disease names that are familiar to the doctor. The doctor makes groups four to six times. We intend to obtain rare criteria by forcing users thinking up to many criteria by grouping disease names many times.

The proposed system also measures line of an user sight in grouping diseases using an eye-tracking system in order to obtain information which disease relations the user recognizes / does not recognize. In previous research, it has been verified that line of sight runs between two parts before people verbalize the relations of the two parts [5]. Combinations of diseases whose relations are not usually recognized are obtained by using the eye-tracking system.

2.1 Input: Disease Names

A user inputs disease names into the system. The user should input the names that are familiar to the user. The number of the names is more than 20. The disease names are not limited to formal disease names because the names are just signals for the users to imagine the diseases.

2.2 Rules in Grouping Names of Disease

In making groups using the proposed system, users have to follow the four rules.

- 1) Form less than five groups.
- 2) Use all disease names in the group formation.
- 3) Do not use the same criteria of groups used in the previous session as long as possible.
- 4) In the last grouping, users have to distribute five disease names chosen by the proposed system over different groups.

The rule 1) and rule 2) are the original rules prepared in the analogy game [3]. The rule 3) and the rule 4) are rules added by the authors. We set the rules considering our experiences. The purpose of the added rules is to obtain criteria that are not usually considered by medical doctors.

3 Experiment

We experimented with the system and verified whether the system can support users in obtaining criteria that are not usually considered.

3.1 Procedures

The procedures of the experiment were as follows.

- (1) Participants of the experiment made groups of disease names using the system.
- (2) Participants wrote names and reasons why they had revised the group membership of items.
- (3) Participants were asked the details of the written reasons.

Participants used the system six times at a maximum until they could not think up new criteria. The lines of sight were also measured while grouping diseases. Some of participants measured their lines of sight were asked to group diseases in which they did not recognized the relations.

3.2 Participants

We asked six doctors as participants of the experiment. Two of them were eye doctors, three of them were hematology doctors, and one of them was a psychiatry doctor. The eye doctors and the hematology doctors have examined patients more than 10 years. The psychiatry doctor has examined patients less than two years.

3.3 Evaluation

In the experiment, participants were asked for the details of the reasons why they had revised the group membership of a disease. If a participant answered that the written reason was a criterion that was not usually considered, we considered that the written criterion was a candidate of criteria for classifying unclassified diseases, and the system worked efficiently.

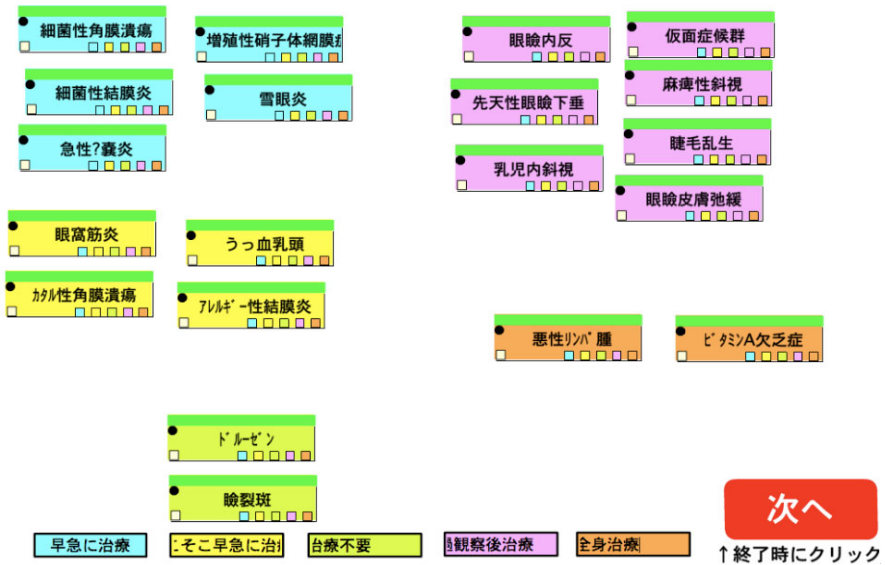


Fig. 1. Visualizing interface of proposed system. Disease names are written in each colored card. Disease groups are recognized by the same color. Boxes in the bottom of an interface are prepared for writing criterion for classifying diseases.

3.4 Consideration of Ethics

The participants have given their informed consent and have understood that they can withdraw from the experiment at any time.

3.5 Results

Table 1 shows an example of written criteria in grouping by an eye doctor. Five of the participants used the system six times, and one of them used the system four times.

Table 1. Examples of criterion written by an eye doctor in grouping

#	Criterion written by an eye doctor
1	eyeball disease, nerve of sight disease, cornea disease, cockeye, retina disease
2	eyeball disease, cornea disease, nerve disease, retina disease, palpebra disease
3	eyeground disease, elevation of intraocular pressure disease, nerve of sight disease, cornea disease, misalignment of the eyes and eye movement
4	curing is unnecessary, relapse disease, curing is difficult, generalized disease, operation of external eye is applied
5	retina disease, ischemic disease, inflammation of the uvea disease, cornea disease, palpebra disease
6	acute disease, chronic disease, recurrent disease between relapse and remission, sudden change in the condition is not seen

Table 2. Names and reasons written by participants

Participant	Names
Number of grouping	Reasons why participants have revised in grouping
Eye doctor 1 In 4 th grouping	lymphoma malignum This disease develops in many parts of eye: conjunctiva, eyeground, and eyeball.
Eye doctor 1 In 6 th grouping	keratoconus This disease occasionally brings on rupture of Descemet's membrane. If the rupture is brought on, this disease is considered acute disease. Otherwise, sudden change in the condition is not seen.
Eye doctor 2 In 3 rd grouping	drusen If this disease develops in macular area, patients lose their sight. Otherwise, patients do not lose their sight.
Hematology doctor 2 In 6 th grouping	thrombocytopenia Some of doctors consider that this disease is half-cured. The others consider that this disease is not cured.
Psychiatry doctor In 4 th grouping	periodic hypersomnia, Abnormal behavior in sexuality The participant was not sure the details of these diseases.
Psychiatry doctor In 6 th grouping	periodic hypersomnia, Prison reaction The participant was not sure the details of these diseases.

Table 2 shows names and reasons written by the participants. Table 3 shows the number of reasons that were not usually considered by the participants. In this experiment, we obtained six reasons. Two of them were criteria that were not usually considered.

Table 4 shows the results of grouping by an eye doctor following instructions obtained by an eye-tracking system. In the former grouping of Table 4, the doctor measured his lines of sight by an eye-tracking system, and in the latter grouping, the doctor asked to make disease groups in which the relations were not recognized.

Table 3. Number of written reasons and number of reasons that are not usually considered by participants

Participant	Number of reasons	Number of reasons not usually considered
Eye doctor	3	2
Hematology doctor	1	0
Psychiatry doctor	2	0

Table 4. Results of grouping by an eye doctor following the instruction obtained from an eye-tracking system

Grouping	Criteria	Names of disease
Former grouping	Aggressive treatment is needed.	hordeolum, corneal herpes, acute retinal necrosis
	Aggressive treatment is not needed.	palpebra papilloma, Drusen, cone dystrophin
Latter grouping	Troubling disease	corneal herpes, cone dystrophin, acute retinal necrosis
	Simple disease	hordeolum, palpebra papilloma, Drusen

4 Discussion

This section discusses the experimental results.

4.1 Obtained Criteria in Grouping

The eye doctors wrote criteria about eye parts in which diseases occur, treatment courses, disease advancements, and so on. Criteria written by both of the doctors were identical. The doctors said that they think firstly which parts have got diseases when they examined patients. Secondly, they considered disease causes. Thirdly, they thought how to cure them. The doctors usually think in such order to communicate with their patients to cure them. Therefore, both of them wrote identical criteria.

The hematology doctors wrote criteria about diseased cells, treatment courses, and so on. Criteria written by the three doctors were also identical because the doctors firstly think which cells are diseased, and secondly how to cure them. Therefore, all the hematology doctors also wrote the identical criteria.

The psychiatry doctor wrote criteria about disease causes, ages of disease occurring, treatment courses, disease presentations. The treatment courses were also written by the eye doctors and the hematology doctors. Though the doctors whose majors were different from each other, they wrote identical criteria (for example, diseased part, treatment course of patients, and so on). These results indicate that doctors think the same criteria even if their majors are different from each other.

4.2 Efficiency of Added Rules for Obtaining Criteria of Disease

We added the rule (3) in Section 2.3. The rule (3) was not to use the same criteria that were used in the previous session as long as possible. The participants revised items in the 4th grouping to the 6th grouping in Table 2. In the 1st grouping and the 2nd grouping, the participants had many candidates of criteria that were usually considered by the participants. However, after grouping many times, they could use few criteria that were usually considered, and they had to squeeze out other criteria that were not usually considered. Therefore, some of the participants revised items in 4th to 6th grouping. This result indicates that the added rule (3) makes grouping difficult, and helps users to think up many criteria that are / are not usually considered.

We also added the rule (4) in Section 2.2. The rule (4) was that in the last grouping, the user had to divide five disease names chosen by the proposed system into different groups. When the rule was applied, three of the participants revised the group membership of items and three of them did not. This result indicates that the added rule (4) does not make grouping difficult. It is also considered that the added rule (4) is not efficient for obtaining many criteria.

4.3 Candidates of New Criteria for Classifying Unclassified Diseases

In this experiment, six participants made groups 34 times in total. From the results, only six reasons why they revised the group formation were obtained. This indicates that participants rarely revised the group formation of disease names because they have many criteria for classifying diseases.

In Table 3, the number of the reasons written by eye doctors was the largest. The number of reasons that were not usually considered by the eye doctors was also the largest. The participants had to reconstruct groups by analogical reasoning if a disease name could not put into any groups. Criteria that were not usually considered were obtained in the reconstructions [3]. Though we could not judge whether the criteria obtained in this experiment was useful for classifying unclassified diseases or not, the criteria were the best ones that we could obtain because the criteria were obtained from analogical reasoning of doctor after revision and deliberation of groups.

4.4 Scenario Obtained from Thinking the Unrecognized Relations between Diseases

Table 4 shows the result of grouping by an eye doctor following the instruction obtained by an eye-tracking system. The obtained criteria were also about treatment methods.

After grouping, the participant said that he revised the group formation. The instruction was to make one group using corneal herpes and cone dystrophin that were not related to each other at all. The participant thought some suppositional scenarios connecting the two diseases because he could not remember appropriate criteria. One of the scenarios was that patients who have got cone dystrophins in childhood will get corneal herpes in adulthood. Though this scenario was a fiction, the patient said that he could consider the two diseases from different view points. It is considered that instructions obtained by eye-tracking system help doctors to imagine new scenarios that may be useful for discovering unclassified diseases.

5 Conclusion

This paper proposes a new system that supports doctors in thinking up new criteria for unclassified diseases. We experimented the system as preliminary and obtained two criteria that are not usually considered by eye doctors.

As a future work, we plan to collect information from doctors about the criteria obtained in this experiment. We also plan to experiment the system to other doctors belonging to other departments.

References

1. Japan Incurable Diseases Information Center, <http://www.nanbyou.or.jp/>
2. Lozano, A.M., Kalia, S.K.: New Movement in Parkinson's. *Scientific American* 293, 68–75 (2005)
3. Nakamura, J., Ohsawa, Y.: Insight or Trial and Error: Ambiguous Items as Clue for Discovering New Concepts in Constrained Environments. In: *Proceedings of the 12th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, pp. 742–749 (2008)
4. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. In: *Proceedings of the Advanced Digital Library Conference*, pp. 12–18 (1998)
5. Ohsawa, Y., Maeda, Y., Yoshida, T.: Eyes Draw Auxiliary Lines before Insight Moment. In: *Proceedings of IEEE Conference on Systems, Man, and Cybernetics*, pp. 3470–3475 (2007)
6. Yamanishi, K., Takeuchi, J., Williams, G., Milne, P.: On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. *Data Mining and Knowledge Discovery Journal* 8(3), 275–300 (2004)

Interpretation of Chance Discovery in Temporal Logic, Admissible Inference Rules

Vladimir Rybakov

Department of Computing and Mathematics,
Manchester Metropolitan University,
Chester Street, Manchester M1 5GD, UK
V.Rybakov@mmu.ac.uk

Abstract. Our paper [\[1\]](#) suggests a possible treatment of Chance Discovery (CD) via interpretation in temporal logic. We propose a semantic definition for computation CD-operations and consider the resulting logic. The main result of the paper is a necessary condition for rules to be admissible in considered logics.

Keywords: Chance Discovery, temporal logics, Kripke/Hintikka models, inference rules, admissible rules.

1 Introduction

This paper studies behavior of inference procedure in temporal logics describing Chance Discovery via temporal accessibility in future and (for symmetry) in past. Chance Discovery (CD in the sequel, cf. Ohsawa and McBurney [\[18\]](#), Abe and Ohsawa [\[1\]](#)) is a modern direction in Artificial Intelligence (AI) which analyzes important events with uncertain information, incomplete past data, so to say, *chance* events, where a *chance* is defined as some event which is significant for decision-making in a specified domain. The prime aim of CD is determination of methods for discovering various chance events. Research dedicated to CD for practical applications formed a solid branch in information sciences (cf. e.g. [\[2,3,4,19,20,21,13,17\]](#)).

This our paper attempts to interpret CD by technique of temporal logic. This logic, in various forms, has been successfully applied to problems arising in CS and AI (cf. Barringer, Fisher, Gabbay and Gough [\[5\]](#)), Gabbay and Hodkinson [\[8\]](#), Gabbay, Hodkinson and Reynolds [\[7\]](#)). One of logical problems is description of logical admissible rules, inference rules. Admissible rules were (perhaps, first time explicitly) introduced into consideration by Lorenzen [\[14\]](#). Initially there were only observations made on existence of interesting examples of admissible but not derivable rules (cf. Harrop [\[12\]](#), Mints [\[16\]](#)). Then Friedman (1975) [\[6\]](#) set the problem whether the intuitionistic logic IPC is decidable w.r.t. admissible inference rules. This problem (together with its counterpart for modal logic $S4$)

¹ Supported by Engineering and Physical Sciences Research Council (EPSRC), U.K., grant EP/F014406/1.

was solved affirmatively in Rybakov [23,24]. Algorithms deciding admissibility for some transitive modal logics and IPC, which are based on projective formulas and unification, were later discovered by Ghilardi [9,10,11]. Roziere [22] found a solution of the Friedman problem for IPC by methods of proof theory. Later many new techniques for study and computing admissible inference rules were suggested in Rybakov [25]. Decidability procedures for recognizing correct inference rules and theorems in temporal logics close to linear logics were found in Rybakov [26,27,28], the decidability of admissibility problem for modal logic S_4 with universal modality was proved in Rybakov [29]. It is interesting and difficult problem to identify how to treat CD in temporal logic and which inference rules are correct for such logics.

We aim of our present paper is (i) to suggest an interpretation of CD within framework of temporal logics and to built corresponding logics; (ii) to study inference rules for these logics, to find necessary conditions for inference rules to be admissible. The main results are necessary conditions for rules to be admissible in temporal transitive and reflexive logic $T_{S_4,U}$ with CD and uncertainty operations and the variant of this logic T_{S_4} with omitted uncertainty operation.

2 Notation, Definitions, Preliminary Information

For convenience of reading, we start from a recall of necessary definitions and notation. The language of standard temporal logics consists of the language of Boolean logic extended by two unary temporal operations \diamond^+ (will be in the future) and \diamond^- (was in the past). $\diamond^+\varphi$ to be read *there is a future state where φ is true*, $\diamond^-\varphi$ means *there was a state in past where φ was true*.

We extend the language by unary temporal operation \diamond_l for local possibility: $\diamond_l\varphi$ has meaning: φ is possible locally. The operation \diamond_l is needed to define our approach to model *uncertainty* in temporal logic.

A Kripke/Hintikka frame is a pair $\mathcal{F} := \langle F, R \rangle$, where F is the base of \mathcal{F} – a non-empty set, and R is a binary (accessibility by time) relation on F . $|\mathcal{F}| := F$, $a \in \mathcal{F}$ is a denotation for $a \in |\mathcal{F}|$. In this paper we consider only reflexive and transitive temporal logics, so R in the sequel is always reflexive and transitive. In what follows, R^{-1} is the relation converse to R . If, for a set of propositional letters P , a valuation V of P in $|\mathcal{F}|$ is defined, i.e. $V : P \rightarrow 2^F$, in other words, $\forall p \in P (V(p) \subseteq F)$, then the tuple $\mathcal{M} := \langle \mathcal{F}, V \rangle$ is called a Kripke/Hintikka model (structure). The truth values of formulas are defined at elements of \mathcal{F} by the following rules:

$$\begin{aligned} \forall p \in Prop, \forall a \in \mathcal{F}, (\mathcal{F}, a) \Vdash_V p &\Leftrightarrow a \in V(p); \\ (\mathcal{F}, a) \Vdash_V \varphi \wedge \psi &\Leftrightarrow (\mathcal{F}, a) \Vdash_V \varphi \text{ and } (\mathcal{F}, a) \Vdash_V \psi; \\ (\mathcal{F}, a) \Vdash_V \varphi \vee \psi &\Leftrightarrow (\mathcal{F}, a) \Vdash_V \varphi \text{ or } (\mathcal{F}, a) \Vdash_V \psi; \\ (\mathcal{F}, a) \Vdash_V \varphi \rightarrow \psi &\Leftrightarrow \neg[(\mathcal{F}, a) \Vdash_V \varphi] \text{ or } (\mathcal{F}, a) \Vdash_V \psi; \end{aligned}$$

$$(\mathcal{F}, a) \Vdash_V \neg\varphi \Leftrightarrow \neg[(\mathcal{F}, a) \Vdash_V \varphi];$$

$$(\mathcal{F}, a) \Vdash_V \diamond^+\varphi \Leftrightarrow \exists b \in \mathcal{F}((aRb) \wedge (\mathcal{F}, b) \Vdash_V \varphi);$$

$$(\mathcal{F}, a) \Vdash_V \diamond^-\varphi \Leftrightarrow \exists b \in \mathcal{F}((bRa) \wedge (\mathcal{F}, b) \Vdash_V \varphi).$$

$$(\mathcal{F}, a) \Vdash_V \diamond_l\varphi \Leftrightarrow \exists b \in \mathcal{F}((aRb) \wedge (bRa) \wedge (\mathcal{F}, b) \Vdash_V \varphi).$$

For any $a \in \mathcal{F}$, $Val_V(a) := \{p_i \mid p_i \in P, (\mathcal{F}, a) \Vdash_V p_i\}$. For any formula φ , $V(\varphi) := \{a \mid a \in \mathcal{F}, (\mathcal{F}, a) \Vdash_V \varphi\}$. Some more abbreviations will be used in the sequel: for any formula φ , $\square^+\varphi := \neg\diamond^+\neg\varphi$, $\square^-\varphi := \neg\diamond^-\neg\varphi$. For $a \in \mathcal{F}$, $C(a) := \{b \mid (aRb) \wedge (bRa)\}$, i.e. $C(a)$ is the cluster containing a .

Definition 1. For a Kripke-Hintikka structure $\mathcal{M} := \langle \mathcal{F}, V \rangle$ and a formula φ , φ is true in \mathcal{M} (denotation - $\mathcal{M} \Vdash \varphi$) if $\forall a \in \mathcal{F} (\mathcal{F}, a) \Vdash_V \varphi$. $\mathcal{F} \Vdash_V \varphi \Leftrightarrow \forall w \in \mathcal{F} ((\mathcal{F}, w) \Vdash_V \varphi)$.

Our suggestion to interpret uncertainty is as follows: we introduce the following new operation U definable in offered language:

$$U\varphi := \diamond_l^+\varphi \wedge \diamond_l^+\neg\varphi.$$

The meaning of $U\varphi$ is: *the statement φ is uncertain*. It seems this approach model very well the intuitive understanding of uncertainty: truth $U\varphi$ means that *today* - in current time cluster - both φ and $\neg\varphi$ may be true, what concludes that φ is *uncertain* (at least from temporal viewpoint).

Our interpretation for *Chance Discovery (CD)* is based on possibility to discover an event in future (local future) or past. So, $\diamond^+\varphi$ says that the truth of φ is discoverable (in future computation, via web linking etc.), $\diamond_l^+\varphi$ says about possibility to discover that φ is true in local future, and $\diamond^-\varphi$ says about past.

Definition 2. For a class \mathcal{K} of frames, the logic $\mathcal{L}(\mathcal{K})$ generated by \mathcal{K} is the set of all formulas which are true in all models based on frames from \mathcal{K} .

The temporal logic $T_{S4,U}$ is the logic $\mathcal{L}(\mathcal{K}_{r+tr})$, where \mathcal{K}_{r+tr} is the class of all reflexive and transitive frames, in described above language; T_{S4} is the logic $\mathcal{L}(\mathcal{K}_{r+tr})$ in the language with omitted \diamond_l .

Definition 3. A logic \mathcal{L} has the finite model property (fmp in the sequel) if $\mathcal{L} = \mathcal{L}(\mathcal{K})$, where \mathcal{K} is a class of finite frames.

As well known, $T_{S4} = \mathcal{L}(\mathcal{K}_{r+tr,f})$, where $\mathcal{K}_{r+tr,f}$ is the class of all finite frames from \mathcal{K}_{r+tr} , so, T_{S4} has fmp. By a bit refined technique close to filtration tools it is simple to show that $T_{S4,U}$ also has fmp.

For any logic $\mathcal{L}(\mathcal{K})$, a formula φ is a *theorem* of $\mathcal{L}(\mathcal{K})$ if $\varphi \in \mathcal{L}(\mathcal{K})$, φ is satisfiable in \mathcal{K} if, for some valuation V in some frame $\mathcal{F} \in \mathcal{K}$, φ is true w.r.t. V at some world of \mathcal{F} .

A *consecution* (or, synonymously, – a *rule*, *inference rule*) \mathbf{c} is an expression

$$\mathbf{c} := \frac{\varphi_1(x_1, \dots, x_n), \dots, \varphi_m(x_1, \dots, x_n)}{\psi(x_1, \dots, x_n)},$$

where $\varphi_1(x_1, \dots, x_n), \dots, \varphi_m(x_1, \dots, x_n)$ and $\psi(x_1, \dots, x_n)$ are some formulas constructed out of letters x_1, \dots, x_n . Letters x_1, \dots, x_n are called variables of \mathbf{c} . For any consecution \mathbf{c} , $Var(\mathbf{c}) := \{x_1, \dots, x_n\}$. The formula $\psi(x_1, \dots, x_n)$ is the *conclusion* of \mathbf{c} , formulas $\varphi_j(x_1, \dots, x_n)$ are the *premises* of \mathbf{c} .

Definition 4. A consecution \mathbf{c} is said to be valid in a Kripke structure $\langle \mathcal{F}, V \rangle$ (we will use notation $\langle \mathcal{F}, V \rangle \Vdash \mathbf{c}$, or $\mathcal{F} \Vdash_V \mathbf{c}$) if $(\mathcal{F} \Vdash_V \bigwedge_{1 \leq i \leq m} \varphi_i) \Rightarrow (\mathcal{F} \Vdash_V \psi)$. Otherwise we say \mathbf{c} is refuted in \mathcal{F} , or refuted in \mathcal{F} by V , and write $\mathcal{F} \not\Vdash_V \mathbf{c}$. A consecution \mathbf{c} is valid in a frame \mathcal{F} (notation $\mathcal{F} \Vdash \mathbf{c}$) if, for any valuation V of $Var(\mathbf{c})$, $\mathcal{F} \Vdash_V \mathbf{c}$.

Historically, introduction into consideration of admissible inference rules may be referred to Lorenzen, 1955, [14]. The definition of admissible rules is as follows. Let \mathcal{L} be a logic, $Form_{\mathcal{L}}$ be the set of all formulas in the language of \mathcal{L} and $\mathbf{c} := \varphi_1(x_1, \dots, x_n), \dots, \varphi_m(x_1, \dots, x_n) / \psi(x_1, \dots, x_n)$ be an inference rule.

Definition 5. Rule \mathbf{c} is admissible for (in) \mathcal{L} if, $\forall \alpha_1 \in Form_{\mathcal{L}}, \dots, \forall \alpha_n \in Form_{\mathcal{L}}$,

$$[\bigwedge_{1 \leq i \leq m} (\varphi_i(\alpha_1, \dots, \alpha_n) \in \mathcal{L})] \implies [\psi(\alpha_1, \dots, \alpha_n) \in \mathcal{L}].$$

Thus, \mathbf{c} is admissible if, for every substitution s , $s(\varphi_1) \in \mathcal{L}, \dots, s(\varphi_n) \in \mathcal{L}$ implies $s(\psi) \in \mathcal{L}$. We list below several examples of admissible rules: (i) The rule $\neg x \rightarrow y \vee z / (\neg x \rightarrow y) \vee (\neg x \rightarrow z)$ (Harrop [12]) is admissible in the intuitionistic logic *IPC* but not derivable in the Heyting axiomatic system for *IPC*; (ii) The rule $(x \rightarrow y) \rightarrow x \vee z / ((x \rightarrow y) \rightarrow x) \vee ((x \rightarrow y) \rightarrow z)$ (G.Mints [16]) has the same as above properties; (iii) The Lemmon-Scott rule

$$\Box(\Box(\Box \Diamond \Box p \rightarrow \Box p) \rightarrow (\Box p \vee \Box \neg \Box p)) / \Box \Diamond \Box p \vee \Box \neg \Box p$$

is admissible (but non-derivable in standard Hilbert-style axiomatic systems) for modal logics *S4*, *S4.1* and *Grz* (cf. Rybakov [25]).

The *admissibility problem* for inference rules in \mathcal{L} is to determine, given by arbitrary inference rule \mathbf{r} , whether \mathbf{r} is admissible in \mathcal{L} . If there is an algorithm solving this problem, \mathcal{L} is said to be decidable by admissibility.

3 Technique, Main Results

The aim of this section is to provide a necessary condition for rules to be admissible in temporal logics similar to *T_{S4}*. For any tense frame $\mathcal{F} := \langle F, R \rangle$ and any $w \in F$, the frame $\mathcal{F}(w)^g$ generated by w in \mathcal{F} is the set of all $a \in \mathcal{F}$, were $a = w$ or $wQ_{i1}w_1, w_1Q_{i2}w_2, \dots, w_{k-1}Q_k w_k, w_k = a$ for some $Q_{ij} \in \{R, R^-\}, w_i \in \mathcal{F}$.

Definition 6. A frame \mathcal{F} is said to be generated (or, synonymously, rooted) if $\mathcal{F} = \mathcal{F}(w)^g$ for a world $w \in \mathcal{F}$.

To proceed to admissibility, we need a reduction of consecutions (inference rules) to some equivalent reduced normal forms. A consecution (inference rule) \mathbf{c} is said to have the *reduced normal form* if $\mathbf{c} = \varepsilon_c/x_1$ where

$$\varepsilon_c := \bigvee_{1 \leq j \leq m} \left(\bigwedge_{1 \leq i \leq k} [x_i^{t(j,i,0)} \wedge (\diamond^+ x_i)^{t(j,i,1)} \wedge (\diamond^- x_i)^{t(j,i,2)} \wedge (\diamond_l x_i)^{t(j,i,3)}] \right),$$

all x_s are certain variables (letters), $t(j, i, z) \in \{0, 1\}$ and, for any formula α above, $\alpha^0 := \alpha$, $\alpha^1 := \neg\alpha$. For a consecution \mathbf{c} , for the rest of the paper, $Var(\mathbf{c})$ is the set of all variables from \mathbf{c} .

Definition 7. Let \mathbf{c} be a consecution \mathbf{c} and $\mathbf{c}_{\mathbf{nf}}$ be a consecution in the reduced normal form containing all variables of \mathbf{c} . A consecution $\mathbf{c}_{\mathbf{nf}}$ is said to be a normal reduced form for \mathbf{c} iff the following hold: (i) \mathbf{c} and $\mathbf{c}_{\mathbf{nf}}$ are equivalent w.r.t. admissibility in any logic, and (ii) \mathbf{c} and $\mathbf{c}_{\mathbf{nf}}$ are equivalent w.r.t. validity in any frame.

Notice that, in the sequel, we can consider only consecutions (rules) with a single premise. Following closely proofs for Lemma 3.1.3 and Theorem 3.1.11 from [25], by similar reasoning, we can derive

Theorem 1. There exists an algorithm running in (single) exponential time, which, for any given consecution \mathbf{c} , constructs a consecution $\mathbf{c}_{\mathbf{nf}}$, which is a normal reduced form of \mathbf{c} .

Definition 8. Let g be a computable function. A logic \mathcal{L} has the (g)-computable-compression property if the following holds. For any finite model $\mathcal{M} := \langle \mathcal{F}, V \rangle$ based on \mathcal{L} -frame \mathcal{F} , where

$$\mathcal{M} \Vdash_V \delta, \quad \delta = \bigvee_{1 \leq j \leq m} \left(\bigwedge_{1 \leq i \leq k} [x_i^{t(j,i,0)} \wedge (\diamond^+ x_i)^{t(j,i,1)} \wedge (\diamond^- x_i)^{t(j,i,2)} \wedge (\diamond_l x_i)^{t(j,i,3)}] \right),$$

there is a mapping f of $|\mathcal{F}|$ onto the base set of a finite model $\mathcal{M}_1 = \langle f(|\mathcal{F}|), R_1, V_1 \rangle$ with the following properties: for any disjunct $\varphi_j = \bigwedge_{1 \leq i \leq k} [x_i^{t(j,i,0)} \wedge (\diamond^+ x_i)^{t(j,i,1)} \wedge (\diamond^- x_i)^{t(j,i,2)} \wedge (\diamond_l x_i)^{t(j,i,3)}]$ of δ

$$\forall a \in \mathcal{F}(\mathcal{F}, a) \Vdash_V \varphi_j \Leftrightarrow (f(\mathcal{F}), f(a)) \Vdash_{V_1} \varphi_j,$$

and the size of $|\mathcal{M}_1|$ is at most $g(s)$ where s is the length of δ .

In the sequel, for any frame $\mathcal{F} := \langle F, R \rangle$ and any $a \in \mathcal{F}$, $C(a)$ denotes the cluster of \mathcal{F} containing a , i.e., $C(a) := \{b \mid b \in F, aRb, bRa\}$. For a logic \mathcal{L} , a frame \mathcal{F} is \mathcal{L} -frame, if $\mathcal{L} \subseteq \mathcal{L}(\{\mathcal{F}\})$.

Definition 9. We say a logic \mathcal{L} has the zigzag-ray property if the following holds. For any finite \mathcal{L} -frame $\mathcal{F} = \langle F, R \rangle$, any $a \in F$ and any finite set $\{d_1, \dots, d_{2m+3}\}$ disjoint with F , the frame $\langle F \cup \{d_1, \dots, d_{2m+3}\}, R_1 \rangle$, where R_1 is the transitive close of the extension of R by xR_1d_1 , for all $x \in C(a)$, $d_{2i}R_1d_{2i-1}$, $d_{2i}R_1d_{2i+1}$ and $d_iR_1d_i$ is also an \mathcal{L} -frame

For any finite frame $\mathcal{F} := \langle F, R \rangle$ and, for any number $k \in N$, we define frames k -stretched from \mathcal{F} as follows. First, $Sl_{min}(\mathcal{F})$ is the set of all R -minimal R -clusters from \mathcal{F} , and $Sl_{max}(\mathcal{F})$ is the set of all R -maximal R -clusters. Let $St_1(\mathcal{F}) := \langle F_1, R_1 \rangle$, where $F_1 := \{w_{1,1} \mid w \in F\}$, $R_1 := \{(w_{1,1}, v_{1,1}) \mid (w, v) \in R\}$.

$St_2(\mathcal{F}) := \langle |St_1(\mathcal{F})| \cup St_{2,1}(\mathcal{F}) \cup St_{2,2}(\mathcal{F}), R_2 \rangle$, where
 $St_{2,1}(\mathcal{F}) := \{w_{2,1} \mid w \in \mathcal{F}\}$, $St_{2,2}(\mathcal{F}) := \{w_{2,2} \mid w \in \mathcal{F}\}$, and

$$R_2 := R_1 \cup R_{2,2}, \text{ where } (w_{2,i}, v_{2,i}) \in R_{2,2} \Leftrightarrow (w, v) \in R,$$

$$(w_{1,1}, v_{2,1}) \in R_{2,2} \Leftrightarrow [w \in Sl_{min}(\mathcal{F}) \wedge (w, v) \in R],$$

$$(w_{2,2}, v_{2,1}) \in R_{2,2} \Leftrightarrow [v \in Sl_{max}(\mathcal{F}) \wedge (w, v) \in R].$$

If $St_k(\mathcal{F})$ is defined, $St_{k+1}(\mathcal{F}) := \langle |St_k(\mathcal{F})| \cup St_{k+1,1}(\mathcal{F}) \cup St_{k+1,2}(\mathcal{F}), R_{k+1} \rangle$, where $St_{k+1,1}(\mathcal{F}) := \{w_{k+1,1} \mid w \in \mathcal{F}\}$, $St_{k+1,2}(\mathcal{F}) := \{w_{k+1,2} \mid w \in \mathcal{F}\}$, $R_{k+1} := R_k \cup R_{k+1,k+1}$, where $(w_{k+1,i}, v_{k+1,i}) \in R_{k+1,k+1} \Leftrightarrow (w, v) \in R$,

$$(w_{k,2}, v_{k+1,1}) \Leftrightarrow [w \in Sl_{min}(\mathcal{F}) \wedge (w, v) \in R],$$

$$(w_{k+1,2}, v_{k+1,1}) \Leftrightarrow [v \in Sl_{max}(\mathcal{F}) \wedge (w, v) \in R].$$

Definition 10. A logic \mathcal{L} is said to have the zigzag stretching property if, for any generated finite \mathcal{L} -frame and any number $m \in N$, the frame $St_m(\mathcal{F})$ is again \mathcal{L} -frame.

Theorem 2. Let \mathcal{L} be a any tense logic in the language of $T_{S4,U}$ possessing: (1) fmp, (2) the (g)-computable-compression property, (3) the zigzag-ray property and (4) the zigzag stretching property. If a rule \mathbf{r} in the reduced normal form is not admissible in \mathcal{L} then there are finite frames $\mathcal{F}(b)^g$ and \mathcal{F}_1 such that \mathcal{F}_1 is \mathcal{L} -frame, \mathcal{F}_1 refutes \mathbf{r} by a valuation V and

- (i) $\mathcal{F}_1 = \mathcal{F}(b)^g \oplus_R [d_1, \dots, d_{2m}, d_{2m+1}, d_{2m+2}, d_{2m+3}]$, where $d_i \notin \mathcal{F}(b)^g$, $|\mathcal{F}_1| = |\mathcal{F}(b)^g| \cup \{d_1, \dots, d_{2m+3}\}$, and the relation R on \mathcal{F}_1 is the transitive close of the extension of the relation from $\mathcal{F}(b)^g$ by xRd_1 , for all x from the R -cluster $C(w)$, for some single fixed $w \in \mathcal{F}(b)^g$, $d_{2i}Rd_{2i-1}$, $d_{2i}Rd_{2i+1}$ and d_iRd_i ;
- (ii) $m = ||\mathcal{F}(b)^g|| + 4$;
- (iii) $||\mathcal{F}(b)^g|| \leq g(s_r)$, where s_r is the length of \mathbf{r} ;
- (iv) the conclusion of \mathbf{r} is false by V at b ;
- (v) $\forall i \geq 2m$, $Val_V(d_i) = Val_V(d_{i+1})$, in particular, the following holds: $\forall x_j \in Var(r) ([(\mathcal{F}_1, d_i) \Vdash_V \diamond^+ x_j \Leftrightarrow (\mathcal{F}_1, d_i) \Vdash_V x_j] \ \& \ [(\mathcal{F}_1, d_i) \Vdash_V \diamond^- x_j \Leftrightarrow (\mathcal{F}_1, d_i) \Vdash_V x_j])$.

To apply this theorem to $T_{S_4,U}$ note, first, that $T_{S_4,U}$ evidently has the zigzag-ray property and the zigzag stretching property.

Lemma 1. (i) *The logic $T_{S_4,U}$ has the (g) -computable-compression property, where g is exponential function. (ii) *The logic T_{S_4} has the (g) -computable-compression property, where g is linear function.**

Note that if we omit \diamond_l from the language, similar to above techniques work for the logic T_{S_4} . Therefore using this lemma we obtain that Theorem 2 gives a necessary condition for admissibility of inference rules in $T_{S_4,U}$, similar statement holds for T_{S_4} .

4 Conclusion, Future Work

Our paper suggests only one of possible interpretation of CD within temporal logic. Evidently it is only one of possible ways and there is a good avenue for future research. One of promising extensions is to engage first order temporal logic which possesses more expressible language; the area of propositional temporal logics, which are non-transitive, would be very popular base for future research as well. The problem how to compute admissible rules in logics already considered in this paper is open as well, because we suggested only a necessary condition.

Results of our paper may be useful for researchers from the field of information systems for verification correctness of reasonings about properties concerning CD in particular application areas.

References

1. Abe, A., Ohsawa, Y. (eds.): Readings in Chance Discovery. International Series on Advanced Intelligence (2005)
2. Abe, A., Kogure, K.: E-Nightingale: Crisis Detection in Nursing Activities. In: Chance Discoveries in Real World Decision Making, pp. 357–371 (2006)
3. Abe, A., Ohsawa, Y.: Special issue on chance discovery. KES Journal 11(5), 255–257 (2007)
4. Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Exceptions as Chance for Computational Chance Discovery. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 750–757. Springer, Heidelberg (2008)
5. Barringer, H., Fisher, M., Gabbay, D., Gough, G.: Advances in Temporal Logic. Applied logic series, vol. 16. Kluwer Academic Publishers, Dordrecht (1999)
6. Friedman, H.: One Hundred and Two Problems in Mathematical Logic. Journal of Symbolic Logic 40(3), 113–130 (1975)
7. Gabbay, D.M., Hodkinson, I.M., Reynolds, M.A.: Temporal Logic: - Mathematical Foundations and Computational Aspects, vol. 1. Clarendon Press, Oxford (1994)
8. Gabbay, D.M., Hodkinson, I.M.: An axiomatisation of the temporal logic with Until and Since over the real numbers. Journal of Logic and Computation 1, 229–260 (1990)

9. Ghilardi, S.: Unification in Intuitionistic logic. *Journal of Symbolic Logic* 64(2), 859–880 (1999)
10. Ghilardi, S.: Best solving modal equations. *Annals of Pure and Applied Logic* 102, 183–198 (2000)
11. Ghilardi, S., Sacchetti, L.: Filtering Unification and Most General Unifiers in Modal Logic. *Journal of Symbolic Logic* 69(3), 879–906 (2004)
12. Harrop, R.: Concerning Formulas of the Types $A \rightarrow B \vee C$, $A \rightarrow \exists xB(x)$ in Intuitionistic Formal System. *J. of Symbolic Logic* 25, 27–32 (1960)
13. Hahum, K.S.: The Window of Opportunity: Logic and Chance in Becquerel's Discovery of Radioactivity. In: *Physics in Perspective (PIP)*, vol. 2(1), pp. 63–99. Birkhäuser, Basel (2000)
14. Lorenzen, P.: *Einführung in die operative Logik und Mathematik*. Springer, Berlin (1955)
15. Mundici, D.: Foreword: Logics of Uncertainty. *Journal of Logic, Language, and Information* 9, 13 (2000)
16. Mints, G.E.: Derivability of Admissible Rules. *J. of Soviet Mathematics* 6(4), 417–421 (1976)
17. Magnani, L.: Abduction and chance discovery in science. *International J. of Knowledge-Based and Intelligent Engineering Systems* 12 (2008)
18. Ohsawa, Y., McBurney, P. (eds.): *Chance Discovery (Advanced Information Processing)*. Springer, Heidelberg (2003)
19. Ohsawa, Y.: Chance Discovery with Emergence of Future Scenarios. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) *KES 2004. LNCS (LNAI)*, vol. 3213, pp. 11–12. Springer, Heidelberg (2004)
20. Ohsawa, Y.: Chance Discovery, Data-based Decision for Systems Design. In: *ISDA (2006)*
21. Ohsawa, Y., Ishii, M.: Gap between advertisers and designers: Results of visualizing messages. *International J. of Knowledge-Based and Intelligent Engineering Systems* 12 (2008)
22. Roziere, P.: Admissible and Derivable Rules. *Math. Structures in Computer Science*, vol. 3, pp. 129–136. Cambridge University Press, Cambridge (1993)
23. Rybakov, V.V.: A Criterion for Admissibility of Rules in the Modal System S_4 and the Intuitionistic Logic. *Algebra and Logic* 23(5), 369–384 (1984) (Engl. Translation)
24. Rybakov, V.V.: Rules of Inference with Parameters for Intuitionistic logic. *J. of Symbolic Logic* 57(3), 912–923 (1992)
25. Rybakov, V.V.: Admissible Logical Inference Rules. *Studies in Logic and the Foundations of Mathematics*, vol. 136. Elsevier Sci. Publ., North-Holland (1997)
26. Rybakov, V.V.: Logical Consecutions in Discrete Linear Temporal Logic. *J. of Symbolic Logic* 70(4), 1137–1149 (2005)
27. Rybakov, V.V.: Linear Temporal Logic with Until and Before on Integer Numbers, Deciding Algorithms. In: Grigoriev, D., Harrison, J., Hirsch, E.A. (eds.) *CSR 2006. LNCS*, vol. 3967, pp. 322–334. Springer, Heidelberg (2006)
28. Rybakov, V.: Until-Since Temporal Logic Baed on Parallel Time with Common Past. Deciding Algorithms. In: Artemov, S., Nerode, A. (eds.) *LFCS 2007. LNCS*, vol. 4514, pp. 486–497. Springer, Heidelberg (2007)
29. Rybakov, V.: Logics with Universal Modality and Admissible Consecutions. *Journal of Applied Non-Classical Logics* 17(3), 381–394 (2007)

Faking Chance

Cognitive Niche Impoverishment

Lorenzo Magnani^{1,2} and Emanuele Bardone¹

¹ Department of Philosophy and Computational Philosophy Laboratory, University of Pavia

² Department of Philosophy, Sun Yat-sen University, Guangzhou, P.R. China

Abstract. The present paper will aim at bringing into light a particular issue related to those situations in which a chance is faked. That is, when an agent acts as if a chance were present in his cognitive niche, when it is not. In illustrating the idea of chance-faking, we will take advantage of the notion of bullshit introduced by Frankfurt. The notion of bullshit has quite recently acquired a theoretical and philosophical dignity. Described as the careless attitude that an agent has towards the truth-value of what he believes in, the notion of bullshit will be taken as a case in point for shedding light on the phenomenon of chance-faking. In the last part of the paper, we will investigate the confabulating dimension of chance-faking introducing the idea of *chance-confabulator*.

1 The Eco-Cognitive Dimension of Chance Discovery and the Role of Abduction

As defined by Oshawa and McBurney [1], a chance is a new event or situation conveying both an opportunity and a risk in the future. Recently, a number of contributions have acknowledged the abductive dimension of seeking chances with relation to science [2,3,4,5]. As maintained by Magnani and Bardone [3] and Abe [5], the process of chance detection (and creation) is resulting from an inferential process – mainly abductive – in which the agent exploits latent clues and signs signaling or informing the presence of an action opportunity [3]. In this case, as argued by Magnani [4] the abductive inferential dimension has to be considered beyond its sentential/computational one.

Accordingly, an inference is a form of sign activity in which the word sign encompasses several types of sign, for instance, symbol, feeling, image, conception, and other representation [6, 5.283]. The process of inferring – and so the activity of chance seeking and extracting – is carried out in a distributed and hybrid way [4]. This approach considers cognitive systems in terms of their environmental situatedness: instead of being used to build a comprehensive inner model of its surroundings, the agent’s perceptual capacities are seen as simply used to obtain “what-ever” specific pieces of information are necessary for its behavior in the world. The agent constantly “adjusts” its vantage point, updating and refining its procedures, in order to uncover a piece of information. This resorts to the need of specifying how to efficiently examine and explore and to the need of “interpreting” an object of a certain type. It is a process of attentive and controlled perceptual exploration through which the agent is able to collect the necessary information: a purposefully moving through what is being examined, actively

picking up information rather than passively transducing [7]. In this sense, humans like other creatures are ecological engineers, because they do not simply live their environment, but they actively shape and change it looking for suitable chances, epistemic for example, like in the case of scientific abductive thinking.

Generally speaking, the activity of chance-seeking as a plastic behavior is administered at the eco-cognitive level through the construction and maintenance of the so-called cognitive niches [4]. The various cognitive niches humans live in are responsible for delivering those clues and signs informing about an (environmental) chance. So, the mediating activity of inferring as sign activity takes place (and is enhanced) for the presence of the so-called eco-cognitive inheritance system [4,8]. That is, humans can benefit from the various eco-cognitive innovations as forms of environmental modifications brought about and preserved by the previous generations. Indeed, many chance-seeking capacities are not wired by evolution, but enter one's behavioral repertoire because they are secured not at the genetic level, but at the eco-cognitive one – in the cognitive niches. The second important point to mention is that humans as chance extractors act like eco-cognitive engineers [3,9]. Accordingly, they take part in the process of extracting chances by performing smart manipulation in order to turn an external constraint into a part of their extended cognitive system.

Humans as eco-cognitive engineers are also chances-maintainers. Humans act so as to preserve those chances that have been proved as successful and worth pursuing. That is what allows us to have an eco-cognitive inheritance that, in turn, enriches or even creates the behavioral chances a community or a group of people has for surviving and prospering.

Chances are provided by the continuous eco-cognitive activity of humans as chance extractors. But what happens when an agent fails as chance-seeker? Why does he fail? What are the main elements leading to impoverished decisions? In the present paper we will be dealing with the problem of *chance faking* as one of the most serious threats for the preservation and “curation” of our cognitive niches. First of all, we will present the notion of bullshit introduced by Frankfurt. Secondly, we will contend how bullshitting can actively promote what we call chance-faking. In the last part, we will illustrate the confabulatory dimension of chance-faking. More precisely, we will present the idea of bullshitter as chance-confabulator.

2 The Notion of Bullshit

The notion of bullshit introduced by Frankfurt [10] may help describe a fundamental feature of people who are not docile: they exhibit a carelessness about truth and can easily perpetrate violence – favoring deception and fraud – just thanks to this systematic undervaluing of truth [11], at least when we consider truth as bounded up with the concepts of reasons, evidence, experience, inquiry, reliability, and standard or rational belief, for example in agreement with a kind of scientific mentality. According to Frankfurt, there is an important distinction to make between a bullshitter and a liar. The difference between the two is that the liar has a general concern about truth. And this is just because, in order to tell a lie, he has to know what the truth is. Although the liar fails to be cooperative with respect to a certain state of things in the world, he is indeed

cooperative with respect to his attitude towards truth. The fabrication of a lie may require a great deal of knowledge and it is mindful: it requires the guidance of truth. More generally, a certain state of mind – namely, an intention to deceive – is required to the liar while making a statement. This attitude is what makes his statement potentially informative. For instance, consider the case of a person telling us that he has money in his pocket, when he has not. His lie is informative as one can guess whether he lied or not. What is interesting about lying is that there always is a reason why a person may not tell the truth: lies and deceit can be detected.

People, for instance, have at disposal both verbal and non verbal cues enabling them to detect potentially deceiving situations [12]. A minor detail about the way a person dresses may be a chance suggesting a man that his wife is cheating on him, and vice-versa. Sometimes people fear the consequences of knowing the truth, therefore they prefer not to investigate. But this does not mean that they would not succeed. Quite the opposite. So, given the fact that a liar is committed to the truth-value of one's belief, lying may contribute to hiding a chance, but at the same time it might be revealing it, as one detects the deceiving intention of the liar.

According to Frankfurt, the case of bullshit is different, as the bullshitter is supposed to lack any concern or commitment to the truth-value of what he says. What turns out to be extremely puzzling is not the content, but his attitude. For instance, as we have already pointed out above, a liar voluntarily gets something wrong. But in doing so he conveys a certain commitment to the truth-value of what he claims, and so the chance to be debunked. A bullshitter does not care about it. As just mentioned, a liar has a deceptive intention that can be detected. Whereas the case of bullshitter is different. When a person believes P , he intends to believe P . And that intention becomes meaningful to other people. In the case of bullshitter, he believes without a real intention to believe what he believes. So, what really defines a bullshitter is his attitude towards truth: he fails to be committed to truth. He simply does not care whether what he says is true or false, accurate or inaccurate.

3 Bullshitting as Chance-Faking

The illustration of bullshit we have presented so far allows us to argue that bullshitters are basically chance-fakers. Why are they so? The deceiving character of chance-faking is related to the fact that they act as if a chance were present, when it is not. Roughly speaking, what comes out of the bullshitter's mouth is *hot air* or *vapor*, meaning that the informative content transmitted is *nil*.

The deceiving character of chance-faking is particularly evident in the case of those cognitive processes involving a collaborative dimension, for instance, the case of second-hand knowledge, that is, the possibility of passing a chance on to another person.

As argued by Simon, we do lean on what other people say. That disposition to rely on social channels in problem-solving activities is what he called *docility* [13]. People tremendously benefit from aids and resources provided by their fellows. This has a major cognitive advantage: that almost any body can trust other people and so have at disposal chances that, first of all, he has never personally experienced, and, secondly,

that are already available to pick up. That is one of the most important assets describing cognitive economy, that is, the need to reach a sort of trade-off between accuracy of a decision and the limited time one is bounded to. Indeed, trust is not informatively empty. One decides to trust another person, because he has reasons to do so. There are a number of clues we make use of in order to consider a particular source of information (a person, for instance) as trustworthy or not. For instance, people usually tend to trust people that exhibit some authority [14]. What happens then to a bullshitter?

A bullshitter does not really intend what he says to believe in. He does not have any concern about the source of what he chooses to believe in. It just happens to him to believe. If so, then information transmission becomes highly *noisy*. Here we come up with another fundamental difference with lying. As already illustrated, a lie is not informatively empty, because people have various mechanisms for detecting lies. Our lie detector is based on our ability of reading others' minds. Basically, we can guess that a person might lie, because we know that we can lie. We read people's intentions. Would we say the same about bullshitters? Do we analogously have a sort of bullshit detector enabling us to see when one is faking a chance? However trivial this question might be, our answer is that we have nothing like that.

Following Frankfurt, we claim that a bullshitter is defined by the kind of attitude he has about truth: he exhibits no commitment regarding what he came to believe in. Our take is that we can infer that he is bullshitting only because we are already familiar (or expert) about what the bullshitter is talking. The cues that are meaningful to us are only related to what he is talking about. But, as one can easily note, in the case of second-hand knowledge it is precisely that thing that is missing. This would be a kind of vicious circle, as we would need what we lack (knowledge) in order to detect bullshitting.

We contend that bullshitting as chance-faking is a particular kind of semantic attack or, more generally, an example of cognitive hacking. By the term semantic attack, we refer to all those situations characterized by a more or less fraudulent/violent move aiming at hacking a chance potentially available to a person or a group of person. A chance might be an idea, a word, a new, a statement, or an explanation. In our terminology, a semantic attack is concerning with the manipulation of the meaning a person assign to something that he is going to use in his decision-making process.

Interestingly, Thompson [15] has recently argued that a semantic attack, and therefore – we would add – the notion of bullshitting as chance-faking, may be a threat for all those situations in which our cognitive performances are mediated by various technological cognitive artifacts like, for instance, computers or artificial agents. According to Thompson there are three categories of threats that can put a computer or a network of computer in danger. The first category groups those physical attacks aimed at physically destroying, for instance, hardware. Then, we have syntactic attacks that regard, for instance, virus and worms. They are thought to destroy software or alter its normal functioning. Semantic attacks belong to the third category of threat. They do not aim at destroying hardware or software, but they manipulate the perception of reality. In doing so semantic attacks distort the decision-making process, whether of an autonomous agent or of a human user of a computer system. On some occasion – particularly violent – semantic attacks may even lead to information warfare. We argue that bullshitting as chance-faking can be considered as a form of semantic attacks. As already pointed

out, a bullshit appears to be a good chance when it is not. And even when there is no fraudulent intent from the bullshitter to deceive, a bullshit actually manipulates the way a reasoner or a decision-maker interprets his reality.

The idea of bullshit as a semantic attack has important connections especially in computer mediated communication. In such a context the main victim is a user acting accordingly to those information that have been received through information technologies. An interesting example of bullshitting as a semantic attack carried out by an artificial agent is reported in [15]. At the end of August 2000, a company press release was posted saying that “Emulex was suffering the corporate version of a nuclear holocaust”. The press release was immediately picked up by a number of business news services offering a potential cognitive chance for their readers. Now, the problem is that the report was widely distributed without any serious investigation whether it was true or false. It took a couple of hours to discover that the press release was fraudulently posted by a 23 years old guy on a website called Internet Wire. Indeed, the bogus press release brought about some real consequences in market share, as Emulex shares went from \$113 per share down to \$43. About 3 millions shares were affected by that bullshit. In this example, everything started with a manipulation, that is, a precise intent of deceiving. However, the bullshitting element was not that. The bullshitting element was related to the particular communication context mediated by computer. More precisely, the use of computational tools facilitated the bullshit to spread all over intoxicating the stock market. Traders picked up the news as a chance that, however, was merely a fake one, as it was unwarily manufactured as a bullshit by the carelessness of those business news services that contributed to distributing it.

4 Bullshitter as Chance-Confabulator

We have been illustrating so far the main elements characterizing bullshitting as chance-faking. In the present section we will shed light on the confabulating dimension of the agent whose behavior is affected by a fake chance. More precisely, we will present the bullshitter as a chance-confabulator.

In four words confabulations are: false reports about memories [16]. Quite recently, a number of studies have been conducted in order to shed light on the very nature of confabulation. One of the most interesting conclusions worth mentioning is that confabulation would be due to a reality monitoring deficit [17,18]. This deficit is explained at neurological level by the effects of focal lesions to the medial orbitofrontal cortex [19]. Basically, confabulating patients lack those mechanisms enabling them to inhibit information that are irrelevant or out of date. In our terminology, they lack a sort of fake chance inhibiting mechanism. For instance, they are not able to distinguish between previously and relevant stimuli [17]. As a consequence of this deficit, they are simply unable to control and assess the plausibility of their beliefs. The self-deceptive dimension can be also explained with relation to chance-seeking. They are basically unable to differentiate between those chances that have a certain degree of ecological validity and those that are simply apparent due to their deficit.

Some confabulation are extremely puzzling for their weirdness, and they have already recognized as symptoms of particular syndromes¹. For instance, patient affected by Anton's syndrome deny to be blind when they are; those who being affected by Capgras' syndrome think that their relatives have been replaced by impostors. Amazingly, patients that have been diagnosed with Cotard syndrome think that they are dead. As one can easily see, confabulators manufacture chances as the result of their impairment. We will come back to this issue in a few paragraphs.

More generally – and beyond its pathological dimension – confabulation emerges as human beings have a natural tendency of “coherencing” and filling in gaps. Confabulating patients have to face up to memories or perceptions that are basically false, but that they have been accepted as true because of the reality monitoring deficit they are affected by. Then, the need of coherencing creates chances that result to be completely implausible and unacceptable. In this sense, like in the case of bullshit, confabulating is not lying. But they are affected by what has been called *pathological certainty* [20]. Basically, confabulators do not doubt, when they should doubt.

Our main take is that in bullshitting all these effects we have briefly surveyed are present, although they are not resulting from a pathological deficit: they are brought about by the mindless attitude bullshitters have about truth-value. Drawing upon Hirstein's phenomenological definition of confabulation [16] we define bullshitting as follows. A person is bullshitting when:

1. He believes that a chance P is present.
2. His thought that P is ill-grounded.
3. He does not know that P is ill-grounded.
4. He should know that P is ill-grounded.

As to capture the difference between confabulation and bullshit, items (3) and (4) are very important. In confabulating patients a person holds a ill-grounded belief, because of his neurological deficit. As already mentioned, he has a reality monitoring deficit that impairs those mechanisms inhibiting irrelevant or out of date information that, in turn, may inform about the presence of a chance. If he were a normal person, he should know that P is ill-grounded. The case of bullshitter is quite different. The two cases partly overlap: both of them should know that P is ill-grounded, when they do not. However, what does not overlap is the reason behind that. Bullshitters have no lesions preventing them to meet any “epistemic” standard of truth. P is ill-grounded because of their careless attitude. So, in the case of confabulators they simply get things wrong, because they cannot discern relevant and up to date information from those that are not. Whereas bullshitters do not get things wrong, but, as Frankfurt put it, they are not even trying. Once we distinguish between these two states of the mind, then we analogically claim that bullshitter is confabulating, as he detects a chance that he would not consider as such, if he tried to get things right. This confabulatory dimension is clear in the example mentioned above. As the business news services distributed the fake press release, they started confabulating about the chance that Emulex was really going to collapse. That is, they immediately activated their usual routines acting as if they were in presence of an authentic chance.

¹ An complete list can be found in [20].

There is another element which illustrates the confabulatory dimension of the bullshitter as a chance confabulator. It follows from our definition that confabulating patients are indeed chance confabulators. However, it should be noted that their confabulation has no effect on other people. In fact, their confabulations are easily to debunk or dismiss as implausible explanations. For instance, the blind patient who claims he is not blind, or the one who thinks he is dead. That leads us to introduce our main contention: that their confabulating does not produce any bullshit. By contrast, when it is normal people confabulating, then confabulations may be regarded as bullshit. This is so because in the case of normal people confabulations are not so easy to dismiss – as we have already argued – and so they can be passed on intoxicating our cognitive niches. In the example of Emulex, the confabulatory dimension made the news services act according to their bullshit. In fact, they promptly distributed the fake report just like any other, and thus transmitted it socially.

5 Conclusion and Future Trends

In this paper we have introduced the idea of chance-faking as a possible outcome of the activity of chance-seeking. We have argued that chance-faking regards all those situations in which a person detects a chance when there is no such a chance. We have illustrated the idea of bullshit as that activity promoting fake chances. In the last part of the paper we have remarked the confabulating character of the bullshitter as chance-faker. The idea of chance faking delineated so far can be further developed by opening up to several other deceiving dimensions threatening the curation of chance. Chance faking can promote other phenomena causing cognitive niche impoverishment. For instance, *chance canceling*, *chance hiding*, and *chance inhibiting*, which are all dimensions strictly related to chance faking and worth further investigation.

References

1. Oshawa, Y., McBurney, P. (eds.): *Chance Discovery*. Springer, Berlin (2003)
2. Magnani, L.: Chance discovery and the disembodiment of mind. In: Oehlmann, R., Abe, A., Ohsawa, Y. (eds.) *Proceedings of the Workshop on Chance Discovery: from Data Interaction to Scenario Creation, International Conference on Machine Learning (ICML 2005)*, pp. 53–59 (2005)
3. Magnani, L., Bardone, E.: Sharing representations and creating chances through cognitive niche construction. The role of affordances and abduction. In: Iwata, S., Oshawa, Y., Tsumoto, S., Zhong, N., Shi, Y., Magnani, L. (eds.) *Communications and Discoveries from Multidisciplinary Data*, pp. 3–40. Springer, Berlin (2008)
4. Magnani, L.: *Abductive Cognition. The Eco-Cognitive Dimension of Hypothetical Reasoning*. Springer, Heidelberg (2009)
5. Abe, A.: Cognitive chance discovery. In: S, C. (ed.) *Universal Access in HCI, Part I, HCII2009, LNCS*, vol. 5614, pp. 315–323. Springer, Heidelberg (2009)
6. Peirce, C.S.: *Collected Papers of Charles Sanders Peirce*. Harvard University Press, Cambridge (1931-1958); Hartshorne, C., Weiss, P. (eds.), vols. 1-6, Burks, A. W. (ed.), vols. 7-8
7. Thomas, H.J.: Are theories of imagery theories of imagination? An active perception approach to conscious mental content. *Cognitive Science* 23(2), 207–245 (1999)

8. Odling-Smee, F.J., Laland, K.N., Feldman, M.W.: *Niche Construction. The Neglected Process in Evolution*. Princeton University Press, Princeton (2003)
9. Bardone, E.: *Seeking Chances. From Biased Rationality to Distributed Cognition* (2010) (forthcoming)
10. Frankfurt, H.: *On Bullshit*. Princeton University Press, Princeton (2005)
11. Misak, C.: Pragmatism and solidarity, bullshit, and other deformities of truth. *Midwest Studies in Philosophy* 32, 111–121 (2008)
12. Vrij, A.: *Detecting Lies and Deceit Pitfalls and Opportunities*. Wiley, New York (2008)
13. Simon, H.A.: Altruism and economics. *American Economic Review* 83(2), 157–161 (1993)
14. Jackson, S.: Black box arguments. *Argumentation* 22, 437–446 (2008)
15. Thompson, P.: Deception as a semantic attack. In: Kott, A., McEneaney, W. (eds.) *Adversarial Reasoning: Computational Approaches to Reading the Opponent's Mind*, pp. 125–144. Chapman & Hall/CRC, Boca Raton (2007)
16. Hirstein, W.: Introduction. what is confabulation? In: Hirstein, W. (ed.) *Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy*, pp. 1–12. Oxford University Press, Oxford (2009)
17. Schnider, A.: Spontaneous confabulation, reality monitoring, and the limbic system: A review. *Brain Research Reviews* 36, 150–160 (2001)
18. Fotopoulou, A., Conway, M., Solms, M.: Confabulation: Motivated reality monitoring. *Neuropsychologia* 45, 2180–2190 (2007)
19. Szatkowska, I., Szymajska, O., Bojarski, P., Grabowska, A.: Cognitive inhibition in patients with medial orbitofrontal damage. *Experimental Brain Research* 181(1), 109–115 (2007)
20. Hirstein, W.: *Brain Fiction. Self-Deception and the Riddle of Confabulation*. The MIT Press, Cambridge (2005)

Summarization for Geographically Distributed Data Streams

Anna Ciampi, Annalisa Appice, and Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari
via Orabona, 4 - 70126 Bari, Italy
{aciampi,appice,malerba}@di.uniba.it

Abstract. We consider distributed computing environments where geo-referenced sensors feed a unique central server with numeric and uni-dimensional data streams. Knowledge discovery from these geographically distributed data streams poses several challenges including the requirement of data summarization in order to store the streamed data in a central server with a limited memory. We propose an enhanced segmentation algorithm in order to group data sources in the same spatial cluster if they stream data which evolve according to a close trajectory over the time. A trajectory is constructed by tracking only data points which represent a change of trend in the associated spatial cluster. Clusters of trajectories are discovered on-the-fly and stored in the database. Experiments prove effectiveness and accuracy of our approach.

1 Introduction

Recent trends in pervasive computing together with advantages in sensor technologies and wireless communication have raised new research challenges due to the huge number of geo-referenced data streams to be managed continuously and at a very high rate. Although, several systems have been already designed to mine data streams on-the-fly, many real life applications need to keep permanent track of streamed data in a central server with a limited memory. To address storage requirements, we need a compact and informative representation of the unbounded amount of geo-referenced data streams in order to discard series of sensor readings and store only their compact representation in the database. In this paper, we consider the geographical dimension of sensor data sources as information bearing to address the task of spatio-temporal data summarization.

Several algorithms have been already investigated in the literature to address the task of summarization in general data streams [2,11,4] as well as in OLAP systems [3]. In this paper, we consider the case of distributed streams of uni-dimensional numeric data, where each data source is a *geo-referenced* remote sensor which periodically records measures for a specific numeric theme (e.g., temperature, humidity). Sensor data sources are geo-referenced by recording their latitude and longitude on the Earth. Due the geo-referencing of the data sources, the summarization task is addressed by resorting to spatial data mining

in addition to stream data mining. This way, we are able to capitalize on the assumption of positive spatial autocorrelation among spatially referenced readings which is common in ecological and environmental environments [9] where sensor data sources are largely distributed. Formally, the *positive spatial autocorrelation* [6] is the property of attributes (themes) taking values at pairs of locations a certain distance apart (neighborhood) to be more similar than expected for randomly associated pairs of observations. Based upon this consideration, we define a novel segmentation algorithm, called SUMMA, which groups sensor data sources in the same cluster if they stream readings which result to be spatially autocorrelated across the space and to evolve with a close trajectory over the time. This trajectory is a compact representation of the readings in the cluster. To be processed, the stream is broken into equally-sized windows of sensor readings which arrive at consecutive time points. After a window goes through the framework, the window is discarded while both clusters and trajectories discovered over the window are permanently stored in the database for future analysis. The trajectory associated to a cluster is obtained as a piecewise linear interpolation [5] of the readings falling in the cluster over the window. Interpolation takes into account a subset of the window time-points, i.e., the trend change points which track a change in the evolution trend (e.g., slope change) of the trajectory.

The paper is organized as follows. In the next section, we define data and patterns we consider in this study. Section 3 presents the summarization framework. Experiments are reported in Section 4 and conclusions are drawn.

2 Data and Pattern Definition

Remote sensor readings are modeled according to a *snapshot* spatio-temporal data model, in which geographical information is structured in subsequent time-stamped thematic layers. A thematic layer is a collection of numeric values measured for an attribute (theme) at the same time. Values vary over a continuous space surface. In a field-based data model [8], this spatial variation over each layer is modeled as a function ($f: \mathbb{R}^2 \times T \mapsto \text{Attribute domain}$) where \mathbb{R}^2 is the layer space and T is the time. In this work, we assume that the theme is measured at sites (sensor data sources) whose point position is fixed on the space surface. In the stream perspective, the sequence of time-stamped thematic layers arrive continuously at consecutive time points. The stream is broken down into windows of w layers arriving in series. Each time a window flows in, summarization patterns are generated locally to the window and stored in the database.

As summarization patterns we discover spatial clusters which are associated to piecewise linear trajectories over the window time. Spatial clusters group sites, such that sites in the same cluster model the continuity (similarity) in space of sensor readings, while sites in separate clusters model the variation over space [7]. Continuity detects the positive spatial autocorrelation of readings over the spatial organization arising in data. Spatial organization is here expressed by closeness between separate sensor sites. By considering consecutive layers, we observe that the sensor data sources, which are repeatedly grouped in a spatial

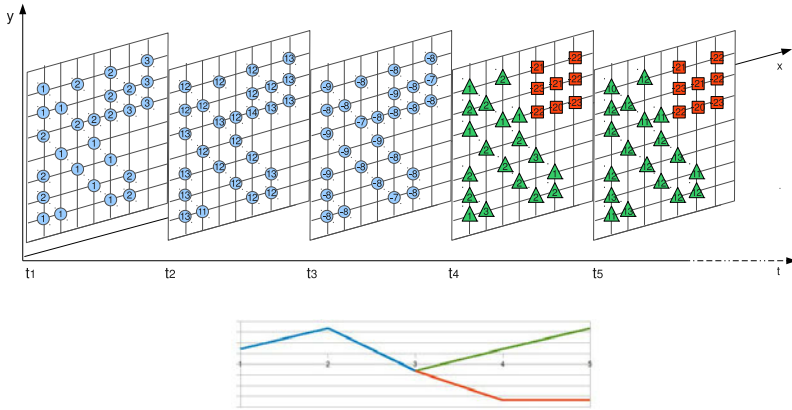


Fig. 1. Clusters and trajectories: the blue cluster groups circle sites which evolve according to the blue trajectory over t_1 , t_2 and t_3 . The red (green) cluster groups squared (triangular) sites which evolve according to the red (green) trajectory over t_4 and t_5 .

cluster, measure value series which evolve with a close trajectory over window time. Fore instance, in Figure 1, the blue cluster summarizes $27(\text{nodes}) \times 3(\text{time points})$ sensor readings in a three-points trajectory; the green (red) cluster summarizes $19 \times 2 (8 \times 2)$ sensor readings in a two-points trajectory.

3 The Algorithm

SUMMA operations consist in: (1) buffering window layers in a graph-based synopsis data structure, (2) discovering clusters of trajectories (trajectory-clusters) over the window, and then (3) identifying trend change points in each trajectory. After a window goes through SUMMA, the window is discarded, while the trajectories for each discovered cluster are stored in the database. Window size w , autocorrelation threshold δ are given before SUMMA starts.

3.1 Buffer Synopsis Structure

The spatial organization of sensor data sources is modeled by means of a graph $G(N, E)$ where N is a set of nodes (sensor sources or sites), and E is a binary (spatial) relation between nodes, $E \subseteq \{\langle u, v \rangle | u, v \in N\}$. The graph G provides the model of the synopsis data structure where layers are buffered window-by-window. In this synopsis data structure, a node $u \in N$ stores a w -sized bucket, denoted as B_u , which is in charge of buffering w readings streamed by the sensor data source over the window (see Figure 2). This way, the slot $B_u[l]$ is the value buffered by the node u at the layer l of the window. Each edge $\langle u, v \rangle \in E$ will be used in clustering to identify spatially close autocorrelated readings which are the only candidates to be grouped in the same cluster.

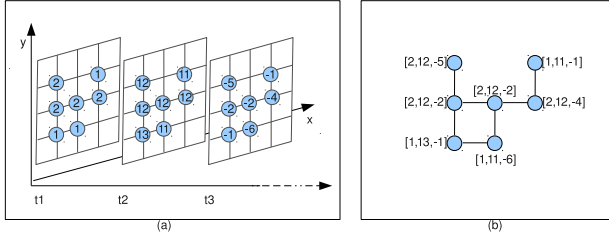


Fig. 2. The graph synopsis data structure (b) where layers (a) are buffered with $w = 3$

3.2 Trajectory Based Spatial Cluster Discovery

Before presenting how the cluster discovery is performed, we introduce some preliminary definitions. We define the δ -close measure, which is a measure of the dissimilarity between two values measured for a numeric theme.

Definition 1 (δ -close measure ψ_δ). Let X be a numeric theme with domain $[\alpha, \beta]$ and δ be a real value in $[0, 1]$. The δ -close measure is a function $\psi_\delta: X \times X \mapsto \{0, 1\}$ such that $\psi_\delta(x_1, x_2) = \begin{cases} 1 & \frac{\|x_1 - x_2\|_1}{\beta - \alpha} \leq \delta \\ 0 & \text{otherwise} \end{cases}$.

Based on Definition 1, the E_δ close relation between edged nodes is defined.

Definition 2 (E_δ close relation). Let δ be a real value in $[0, 1]$; E_δ is the binary relation between the edged nodes of G ($E_\delta \subseteq E$) which is defined as $E_\delta = \{\langle u, v \rangle \in E \mid \sum_{i=1}^w \psi_\delta(B_u[i], B_v[i]) = w\}$ where $B_u[i]$ ($B_v[i]$) is the value stored in the i^{th} slot of the bucket B_u (B_v).

We use the E_δ close relation to define the property of E_δ -connectivity in G .

Definition 3 (E_δ -connectivity). A node u is E_δ -connected to a node v ($u, v \in N$), with respect to E_δ , iff $\langle u, v \rangle \in E_\delta$ (direct connectivity) or $\exists w \in N$ such that $\langle u, w \rangle \in E_\delta$ and w is E_δ -connected from v (undirect connectivity).

Finally, we define the function of δ -homogeneity, which estimates the similarity over a set of nodes of G .

Definition 4 (δ -homogeneity). Let δ be a real value in $[0, 1]$. The function δ -homogeneity $: 2^N \mapsto \{true, false\}$ is defined as:

$$\delta\text{-homogeneity}(N_i) = \begin{cases} true & \forall u, v \in N_i, \frac{\|\eta(B_u) - \eta(B_v)\|_1}{\beta - \alpha} \leq \delta \\ false & \text{otherwise} \end{cases}$$

where $\eta: 2^X \mapsto X$ returns the median over a sequence of values in X 1.

¹ The choice of the median is motivated by the fact that it is robust, while the mean would be influenced by outliers.

Algorithm 1. Trajectory-cluster discovery in SUMMA: $G(N, E) \mapsto \Gamma$

– *Main routine*

```

1:  $\Gamma = \emptyset$ 
2: for all  $u \in N$  do
3:   if  $u$  is UNCLASSIFIED then
4:      $C \leftarrow \{u\} \cup \text{expandCluster}(C, u)$ 
5:      $T \leftarrow \langle 1, \eta(C, 1) \rangle, \langle 2, \eta(C, 2) \rangle, \dots, \langle w, \eta(C, w) \rangle$ 
6:      $\Gamma \leftarrow \Gamma \cup \{C \Leftrightarrow T\}$ 
7:   end if
8: end for

```

– *growCluster* (C, u)

```

1: if  $\delta$ -homogeneity( $C \cup N_\delta(u)$ ) then
2:    $C \leftarrow C \cup N_\delta(u)$ 
3:   for all  $n \in N_\delta(u)$  do
4:      $C \leftarrow \text{growCluster}(C, n)$ 
5:   end for
6: else
7:   for all  $n \in N_\delta(u)$  do
8:     if  $\delta$ -homogeneity( $C \cup \{n\}$ ) then
9:        $C \leftarrow C \cup \{n\}$ 
10:     $C \leftarrow \text{growCluster}(C, n)$ 
11:   end if
12: end for
13: end if

```

Both E_δ -connectivity and δ -homogeneity (see Definitions 4.3) are employed in the graph partitioning algorithm which is reported in Algorithm 1. An unclassified node is chosen as seed and labeled as new cluster. This cluster is recursively grown by merging the E_δ -neighborhoods (see Definition 5) which are constructed for each clustered node. The merge is performed only if the resulting cluster is a δ -homogeneous node set.

Definition 5 (E_δ -neighborhood). Let u be a node, the neighborhood $N_\delta(u)$ of u is defined as $N_\delta(u) = \{v | \langle u, v \rangle \in E_\delta \wedge \text{UNCLASSIFIED}(v)\}$, where *UNCLASSIFIED*(v) means that v is not yet assigned to any cluster.

In the *main routine* of Algorithm 1, a novel empty cluster C is created for a node $u \in N$ which is currently *UNCLASSIFIED*. C is firstly grown with u , then *growCluster*(C, u) grows C by using u as seed. In particular, *growCluster*(C, u) evaluates the property of δ – homogeneity for the node set $C \cup N_\delta(u)$ (see the call of δ – homogeneity(\cdot) function in Algorithm 1). If $C \cup N_\delta(u)$ is a δ -homogeneous node set, C is grown with $N_\delta(u)$. Otherwise, the addition of each neighbor $n \in N_\delta(u)$ is evaluated node-by-node. A neighbor is individually added to C only if the output cluster is a δ – homogeneous node set. Each time C changes, *growCluster*(\cdot, \cdot) is recursively called to further grow C by considering the new clustered nodes as candidate seed of the expansion (see the recursive call of *growCluster*(\cdot, \cdot) in Algorithm 1). When a cluster C is completely constructed

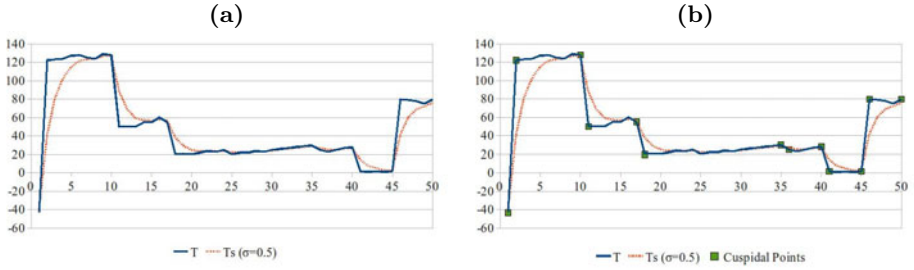


Fig. 3. (a) EMA ($\sigma = 0.5$) to smooth T and (b) cuspid points detected over T

(no further node can be added to), the trajectory T is built and associated to C . T is the sequence of w time-stamped points which track how nodes clustered in C evolve over time points of the window, that is, $T = \langle 1, \eta(C, 1) \rangle, \langle 2, \eta(C, 2) \rangle, \dots, \langle w, \eta(C, w) \rangle$. The time-stamped point $\langle l, \eta(C, l) \rangle$ is the representative of the behavior of the cluster C over the l^{th} layer in the window. In Algorithm 1, the function $\eta(C, l)$ (with $l = 1 \dots w$) returns the median of the values stored in the l^{th} slots of the buckets associated to the nodes grouped in C . The trajectory-cluster $C \Leftrightarrow T$ is added to Γ . Once all nodes are classified, bucket values are discarded and the algorithm stops by returning Γ as output.

3.3 Trajectory Compression

Before storing Γ in database, SUMMA performs a trend change point based compression of trajectories in Γ in order to reduce the number of time-stamped values to be stored in database. Let $T = \langle 1, v_1 \rangle, \langle 1, v_1 \rangle, \dots, \langle w, v_w \rangle$ be a w -sized trajectory that is included in Γ , SUMMA firstly smooths out short-term fluctuations in T and then identify trend change points over T . As smoothing function, the exponentially moving average (EMA) is adopted. EMA smoothies time series by applying a weighting factor that decreases exponentially for older data points. This way, EMA gives much more importance to recent readings, while still not discarding older observations. EMA is recursively formulated in the followings.

$$EMA(T, \langle 1, v_1 \rangle) = v_1$$

$$EMA(T, \langle l, v_l \rangle) = EMA(T, \langle l - 1, v_{l-1} \rangle) + \sigma(v_l - EMA(T, \langle l - 1, v_{l-1} \rangle))$$

where σ is the degree of weighting decrease that is expressed as a constant smoothing factor, i.e., a number between 0 and 1. By applying EMA, short-term fluctuations are smoothed out, as reported in Figure 3a.

A method of 2D-graphics is applied to identify the cuspid points in T which are labeled as trend change points. The angles of incidence w_l is computed as follows:

$$w_l = \arctan \frac{l - (l - 1)}{EMA(T, \langle l, v_l \rangle) - EMA(T, \langle l - 1, v_{l-1} \rangle)} \text{ with } l = 2, \dots, w. \quad (1)$$

w_l ranges in $[-90^\circ, 90^\circ]$ The differences Δw_l are calculated as $\Delta w_l = w_{l+1} - w_l$ with $l = 1, \dots, w - 1$. The time point l is a cuspid if the difference $|\Delta w_l - \Delta w_{l-1}|$ exceeds a given threshold τ_C . This way, only the trajectory points of T , whose time-stamp is labeled as cuspid point are stored in database as a compact representation of the trajectory T (see Figure 3b).

4 Experimental Study

We evaluate SUMMA with two geographically distributed data streams, namely, Berkley data stream and South American data stream.

Evaluation measures. Experiments are performed in order to evaluate both the degree of data summarization and the accuracy of trajectories in interpolating series of values streamed in the associated spatial clusters. As a measure of the summarization degree we consider both the number of clusters and the number of data points permanently stored for each window in database. As a measure of the accuracy of the summarization Γ , we use the average absolute percentage error (*MAPE*), which is the error that is performed when trajectories stored in database are used to fit data windowed in W . Formally,

$$MAPE(\Gamma, W) = \left(\sum_{u \in S} mape(\Gamma, u|_W) \right) / |S|,$$

where S is the set of sites over which the geographically distributed stream is transmitted and $u|_W$ is the series of w values which are streamed from the site $u \in S$ over the window W . $mape(\Gamma, u|_W)$ is the mean absolute percentage error which measures how trajectories fit real values streamed in the $u|_W$ (with $u \in C$). It expresses accuracy as a percentage, and it is defined by

$$mape(\Gamma, u|_W) = \frac{1}{w} \sum_{i=1}^w \left\| \frac{(u[i|_W]) - T[i]}{(u[i|_W])} \right\|_1 \text{ with } (C \Leftrightarrow T) \in \Gamma \text{ and } u \in C.$$

Due to the summarization, a trajectory T interpolates the trend change points detected over W . This means, that $T[i]$ is not stored in database in the case i does not correspond to a trend change point. Anyway we are able to predict the value of $T[i]$. Let (j, v_j) and (k, v_k) be trend change points over T such that $j < i < k$, $\exists(j', v_{j'}) \in T$ with $j' < i$ and $\exists(k', v_{k'}) \in T$ with $i < k'$. Let $T[X] = \alpha + \beta X$ be the straight-line which interpolates trend change points (j, v_j) and (k, v_k) with $\alpha = \frac{v_k - v_j}{k - j}$ and $\beta = v_j - \frac{v_k - v_j}{k - j} \times j$, then $T[i] = \alpha + \beta \times i$.

Parameters setting. To automatically derive a reasonable representation of the range $[\alpha, \beta]$, we use a box plot computed over a portion of the stream. Streamed values are depicted through 5 summaries: the smallest observation, lower quartile (Q_1), median (Q_2), upper quartile (Q_3) and largest observation. $\alpha = Q_1 - 1.5 * (Q_3 - Q_1)$ and $\beta = Q_3 + 1.5 * (Q_3 - Q_1)$. Experiments are run by setting $\delta = p\%[\beta - \alpha]$. Effectiveness and accuracy of SUMMA by tuning parameters' values is investigated in Berkley data stream.

Berkley data stream. This data stream (<http://db.csail.mit.edu/labdata/labdata.html>) contains temperature values streamed every 31 seconds (layers) from

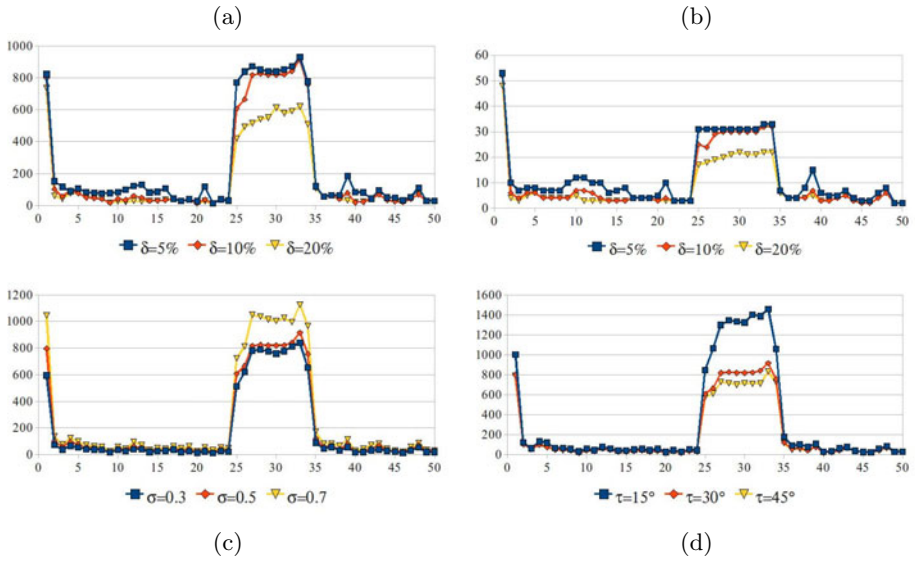


Fig. 4. Berkeley data $w = 100$: (a-b) Number of trend change values (discovered clusters) by varying δ ($\sigma = 0.5$ and $\tau = 30^\circ$). (c) Number of trend change values by varying σ ($\delta = 10\%[\beta - \alpha]$ and $\tau = 30^\circ$). (d) Number of trend change values by varying τ ($\delta = 10\%[\beta - \alpha]$ and $\sigma = 0.5$).

54 sensors deployed in the Berkeley Research lab between February 28th and April 5th. A sensor is considered to be close to sensors into the range of 6 meters. For each layer, measurements are not available for the entire sensor set: unobserved values are simulated with the lastly observed value for the same sensor. Firstly, we set $w = 100$ and study how SUMMA summarization capability (number of clusters and number of points) is affected by δ ($\delta = 5\%[\beta - \alpha]$, $\delta = 10\%[\beta - \alpha]$ and $\delta = 20\%[\beta - \alpha]$), σ (0.3, 0.5, 0.8) and τ (15° , 30° and 45°). Number of trend change point values and number of clusters are plotted, window by window, in Figure 4. We do not plot the number of clusters by varying δ and τ , since it does not change by varying these parameters. Results confirm that by decreasing autocorrelation threshold, the number of discovered clusters as well as the number of trend change points increase. By decreasing σ , EMA flattens the curvilinear trajectory to a linear trajectory, hence the number of detected trend change values decreases. Finally, as expected, by decreasing τ , the number of detected trend change points increases. Additionally, by analyzing the MAPE, we observe only a slight improvement of the accuracy if increasing the number of detected clusters/trend change points. This is due to the fact that, in this data stream, readings for several sensor data sources are missing on each layer and they are simulated by using the lastly observed values. Based on these considerations, intermediate values for δ , σ and τ are chosen to run SUMMA in the remaining experiments reported in the paper, i.e. $\delta = 10\%[\beta - \alpha]$, $\sigma = 0.5$ and $\tau = 30^\circ$. By using this parameter setting, experiments are performed by varying

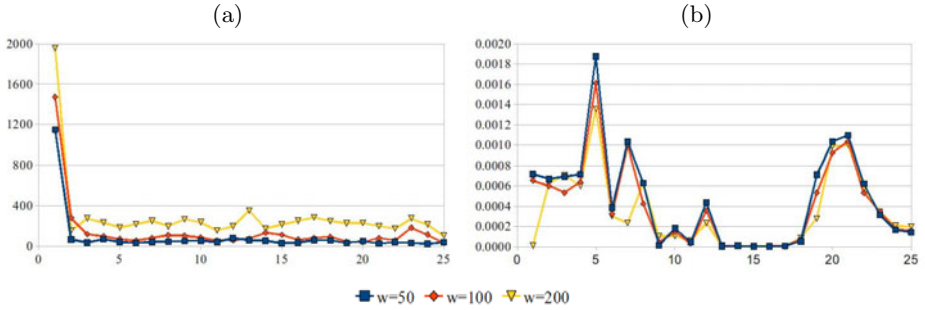


Fig. 5. Berkley data: (a) Elapsed time (ms) and (b) MAPE by varying w

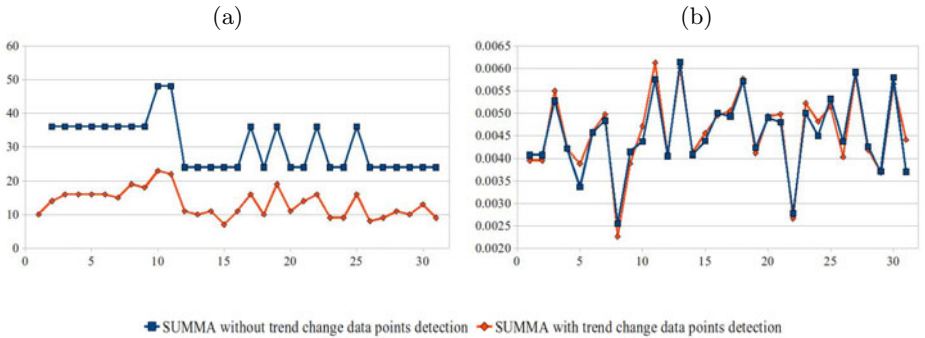


Fig. 6. South American data $w = 12$: (a) Number of points stored in database and (b) MAPE. SUMMA stores all the trajectory data points for the clusters (red diamonds) wrt. SUMMA stores trend change points (blue squares) only.

the window size ($w = 50, 100$ and 200). For the comparison, we consider the same portion of data streams, that is, one window with $w = 200$ is compared with 4 corresponding windows with $w = 50$ and 2 windows with $w = 100$. Elapsed time and MAPE are plotted in Figure 5 by varying w . The analysis of larger ($w = 200$) windows leads to higher time cost of the learning phase. At the same time, we observe that by processing smaller windows, we improve the accuracy of the piecewise linear interpolation for the trend change values stored in database.

South American Climate data stream. This data stream (http://climate.geog.udel.edu/~climate/html_pages/archive.html) contains monthly-mean air temperature values recorded between 1960 and 1990 over a 0.5 degree by 0.5 degree of latitude/longitude grid of South America, where the grid nodes are centered on 0.25 degree, for a total of 6447 node sites. A site is close to the neighbor sites which are located in the grid cells around the site. We run experiments with $w = 12$. This way, data streamed in each window are 77364. Results reported in Figure 6 confirm that SUMMA ($\delta = 10\%[\beta - \alpha]$, $\sigma = 0.5$ and $\tau = 30^\circ$) reduces the number of data values stored in database (less than

50 for window), but, at the same time, data stored in database allow to accurately reconstruct the originally streamed data. MAPE is always lower than 0.01 and, additionally, it does not significantly decrease by storing only trend change points in the database. We also perform experiments by varying w . We do not report results due to space limitation, they confirm considerations suggested by Berkeley data stream.

5 Conclusions

This paper presents an algorithm, called SUMMA, to analyze geographically distributed data streams and discover trajectories according to cluster of sensor readings evolve in time. Trajectories are permanently stored in databases as a summarization of streamed data. Experiments with real data streams demonstrate that SUMMA is effective and accurate in summarizing windowed data.

Acknowledgments

This work is supported by the Strategic Project PS121: “Telecommunication Facilities and Wireless Sensor Networks in Emergency Management”.

References

1. Chiky, R., Hébrail, G.: Summarizing distributed data streams for storage in data warehouses. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 65–74. Springer, Heidelberg (2008)
2. Cormode, G., Muthukrishnan, S.: Summarizing and mining skewed data streams. In: SDM (2005)
3. Cuzzocrea, A.: Cams: Olaping multidimensional data streams efficiently. In: Pedersen, T.B., Mohania, M.K., Tjoa, A.M. (eds.) DaWaK 2009. LNCS, vol. 5691, pp. 48–62. Springer, Heidelberg (2009)
4. Guha, S.: Tight results for clustering and summarizing data streams. In: ICDT 2009, pp. 268–275. ACM, New York (2009)
5. Kontaki, M., Papadopoulos, A.N., Manolopoulos, Y.: Continuous trend-based clustering in data streams. In: Song, I.-Y., et al. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 251–262. Springer, Heidelberg (2008)
6. Legendre, P.: Spatial autocorrelation: Trouble or new paradigm? *Ecology* 74, 1659–1673 (1993)
7. Malerba, D., Appice, A., Varlaro, A., Lanza, A.: Spatial clustering of structured objects. In: Kramer, S., Pfahringer, B. (eds.) ILP 2005. LNCS (LNAI), vol. 3625, pp. 227–245. Springer, Heidelberg (2005)
8. Shekhar, S., Chawla, S.: Spatial databases: A tour. Prentice Hall, Englewood Cliffs (2003)
9. Tobler, W.: Cellular geography. In: *Philosophy in Geography* (1979)

Gradual Data Aggregation in Multi-granular Fact Tables on Resource-Constrained Systems

Nadeem Iftikhar and Torben Bach Pedersen

Aalborg University, Department of Computer Science, Selma Lagerløfs Vej 300,
9220 Aalborg Ø, Denmark
{nadeem, tbp}@cs.aau.dk

Abstract. Multi-granular fact tables are used to store and query data at different levels of granularity. In order to collect data in multi-granular fact tables on a resource-constrained system and to keep it for a long time, we gradually aggregate data to save space, however, still allowing analysis queries, for example, for maintenance purposes etc. to work and generate valid results even after aggregation. However, ineffective means of data aggregation is one of the main factors that can not only reduce performance of the queries but also leads to erroneous reporting. This paper presents effective methods for data reduction that are developed to perform gradual data aggregation in multi-granular fact tables on resource-constrained systems. With the gradual data aggregation mechanism, older data can be made coarse-grained while keeping the newest data fine-grained. This paper also evaluates the proposed methods based on a real world farming case study.

Keywords: Data aggregation, multi-granular fact tables, gradual data aggregation.

1 Introduction

In order to have both fine-grained data and to store data for as long time periods as possible, we need to aggregate data in an intelligent way. As the detail data grows older, it slowly loses its value or may not have the same value as before. Therefore to save disk space and to perform efficient query processing there may be two options of data reduction, either to delete older data or to aggregate it gradually. However the major problem with deleting the older data could be organizational or governmental level data retention laws, therefore it may not be a feasible solution. Alternatively, data aggregation seems reasonable. The aggregated data could be quite useful for analysis purposes. For example, in the farming business, maintenance of the farming machinery such as when the service of the equipment is due or to determine the working life of different components of the equipment and future planning on efficient task management may be done.

In this paper, we present multi-granular fact table that is based on very low-end systems with limited storage and query processing capabilities. In particular, we specifically target data aggregation over a single fact table, even without joins. The proposed multi-granular fact table is the table without any dimensions for analysis, reporting and maintenance purposes. The key objective of the multi-granular fact

tables is to store both detail and aggregated data at different levels of granularity for as long time periods as possible. The main reason to gradually aggregate data at different levels of granularity is to make data available for analysis and reporting purposes at different detail levels. For example, the data at 10 minutes granularity level is more detailed than the data at day granularity level and so on. Furthermore, to gradually aggregate data, three aggregation methods are presented in this paper, namely *the interval-based aggregation method*, *the row-based aggregation method* and *the time granularity-based aggregation method*. The main research question presented in this paper is *how to design a multi-granular fact table and perform gradual aggregation on it on a resource-constrained system*.

Previously, other studies on data aggregation have been done. A comprehensive survey of most relevant techniques for the evaluation of aggregate queries on spatiotemporal data is presented by [1]. Efficient aggregation algorithms for compressed data warehouses are proposed by [2]. Techniques such as pattern identification, categorization, feature extraction, drift calculation and generalization for the aggregation of information are summarized in [3]. Multi-dimensional extension of the ER model in order to use aggregated data in complex analysis contexts is proposed by [4]. However, the main objective of these approaches is to perform one time aggregation of data rather than gradual aggregation, as presented in this paper. In the context of gradual data aggregation work has been reported. An efficient tree based indexing schemes for gradually maintaining aggregates is presented in [5]. The focal point of this work is on presenting effective indexing schemes for storing aggregated data. A data reduction system based on gradual data aggregation is also given in [6]. The main aim of this work is to propose a language for archiving data. Further, the semantic foundation for data reduction in data warehouses that permits the gradual aggregation of detailed data as the data gets older is provided by [7]. The work also described a query execution mechanism on the aggregated data. The work is highly theoretical and the main direction of this work is on querying multi-dimensional data that is being aggregated gradually. An algorithm for gradual data aggregation in multi-granular fact tables has previously been described in [8]. The algorithm is based on a single time hierarchy-based aggregation technique. In the context of data management on resource-constrained systems, work has also been reported. The concept of user-defined aggregation queries to reduce the quantity of data that has to be transmitted through the sensor network is presented in [9]. Finally, [10] and [11] highlighted that existing data management solutions cannot be reused and adapted appropriately in resource-constrained embedded systems.

To the best of our knowledge, this work is the first to propose and evaluate gradual data aggregation methods for multi-granular fact tables on resource-constrained systems. The rest of the paper is organized as follows: Section 2 describes the multi-granular fact tables and explains the motivation behind the proposed methods. Section 3 presents the aggregation methods. Section 4 provides evaluation of the proposed methods. Finally, Section 5 is devoted to conclusions and future work.

2 Multi-granular Fact Tables

This section defines the aggregation problems in multi-granular fact tables that relate to define the methods that enable gradual data aggregation in real world application

domains, such as farming. The multi-granular fact table (Table 1) has been specially designed to store data at different levels of granularity and to perform gradual aggregation on the stored data. Furthermore, it quite naturally captures the data described in the farming example that is very important for clearly presenting the methods. The two most significant attributes of the fact table are the *Granularity* and the *Timestamp*. The Granularity has dual purpose. First, it represents the initial time granularity of the captured Value and second, it represents the aggregated time granularity of the stored Value. For example, from row number 8000 to 11002 it represents initial time granularity of the captured Values and in row number 12000 and 13000 it represents aggregated time granularity of the stored Values. The Timestamp represents logging time in case of detailed data, whereas, it represents the beginning of each time interval or each set of rows (explained later in Section 3) in case of aggregated data. The snapshot of example data (Table 1) concerns farm equipments in fields. The example consists of a single task and four parameters with different levels of time granularities. The most important attributes in this case study are: *Task*, *Parameter*, *Timestamp*, *Granularity* and *Value*. The Task represents activities to distinguish all the work that is carried out in a particular field of a farm. The Parameter represents a variable code for which a data value is recorded. Each Parameter has a different data logging frequency that is represented by the Granularity, for example the logging frequency of Parameter 247 (amount of chemical sprayed in liters) is every 30 second, the logging frequency of Parameter 1 (tractor speed in km/h) is every 60 seconds and so on. The Timestamp represents an instance of time when a data value is recorded and it is in UTC [12] format. Lastly, the Value represents a numeric measure.

In Table 1, *Rid* represents Rowid that is only an abstract attribute used for row identification purposes, *T* represents Task, *P* represents Parameter, *TS* represents Timestamp, *G* represents Granularity and *V* represents Value. In addition to different

Table 1. Snapshot of the multi-granular fact table (Data_log)

Rid	T	P	TS	G	V
8000	8	41	1155208060	20	19.12
8001	8	248	1155208070	30	31.44
8002	8	247	1155208073	30	84.20
8003	8	41	1155208084	20	19.45
8004	8	1	1155208100	60	11.60
8005	8	248	1155208101	30	31.52
8006	8	41	1155208104	20	19.46
8007	8	247	1155208133	30	84.23
..
11000	8	41	1155211720	20	34.56
11001	8	1	1155211732	60	10.55
11002	8	247	1155211733	60	88.80
..
12000	8	247	1140446710	600	45.21
..
13000	8	247	1148359800	1200	63.05

logging frequencies there are situations that have to be taken special care while aggregating otherwise erroneous results may occur. The aggregation methods presented in this paper are considered with two possible aggregation viewpoints: data “without holes” and data “with holes”. Data without holes means that there is no missing data or rows. Whereas, data with holes means it is possible to have missing data or rows. Moreover, these are also other circumstances that have to be taken care of such as, *delay in capturing data*, *change of Granularity value for the same Parameter value (it may be due to the change in logging frequency or due to the aggregating process)* and *higher levels of data aggregation due to the presence of aggregated data over different time intervals*. In Table 1, row number 8002 represents a delay in capturing data (Timestamp=1155208070 should be the correct value), row number 8007 represents data with holes or missing data (Timestamp=1155208100 is missing), row number 11002 represents a change of Granularity value for the same Parameter value (Granularity value for Parameter 247 has been changed from 30 to 60 seconds due to the change in logging frequency) and finally row number 12000 and 13000 represents aggregated data over different time intervals, 10 minutes and 20 minutes granularity levels, respectively.

3 Aggregation Methods

In this section, we describe three methods for aggregating data in multi-granular fact tables on resource-constrained systems. The methods aggregate data with a user-defined *ratio* and/or *time intervals*. These methods are capable of handling most of the aggregation problems mentioned in the previous Section. In fact, three methods are presented for gradual aggregation, namely *the interval-based aggregation method*, *the row-based aggregation method* and *the time granularity-based aggregation method*. The interval-based aggregation method aggregates time intervals based on the aggregation ratio. The row-based aggregation method always aggregate rows based on the aggregation ratio. Finally, the time granularity-based aggregation method aggregates time intervals based on a single time hierarchy. Moreover, all of the proposed aggregation methods are able to aggregate data gradually. These methods are further described in the following subsections.

3.1 The Interval-Based Aggregation Method

This method aggregates time intervals with a user-defined aggregation ratio. It has further two variants. First, the source time period length is the same for all data in the query and we know it. Second, the source time period may vary, which means it is unknown. We now proceed to describe how the overall data aggregation is achieved by using the second variant of this method, which is most relevant to this paper. The aggregation process that enables gradual aggregation is a four-step process, out-lined in the following pseudo-code. 1) *Aggregate data that is more than m months old with a ratio of $1:r$* ; 2) *Get the aggregated data and insert it back into the fact table with changed granularity levels that represent higher level of granularity*; 3) *Delete all the rows that are just been aggregated from the fact table*; and 4) *Go to Step 1 for higher level of aggregation*. Step 1) relates to data aggregation. Step 2) and 3) correspond to

the overall aggregation process that consists of inserting the aggregated data and deleting the detailed data, which has just been aggregated. Step 4) relates to the concept of gradual data aggregation. Only the details of Step 1) and 4) are further described below, whereas, Step 2) and 3) are trivial. The following SQL statement represents the data aggregation part of Step 1).

```
SELECT Task, Parameter, Timestamp, Granularity*
       aggregation_ratio, MAX(Value)
FROM   Data_log
WHERE  Parameter = 247
AND    Timestamp < STRFTIME('%s', 'NOW', '-3MONTH')
GROUP BY Task, ROUND((Timestamp)/(Granularity*aggregation
                       _ratio), 0) * (Granularity*aggregation_ratio);
```

The most significant part of the statement is the GROUP BY clause. The GROUP BY clause aggregates time intervals by using MAX() function. The aggregation ratio is the user-defined ratio, such as, 2 (1:2), 5 (1:5) and so on. The Granularity initially represents the current level of time granularity of the data to be aggregated and later represents the aggregated level of time granularity. The Timestamp represents the beginning of each time span. In the GROUPBY clause, the Timestamp field is first divided by (*Granularity * aggregation_ratio*) and then the ROUND() function is used to round the result to the nearest integer. Further, the rounded result is multiplied with the same denominator in order to get the time intervals based on the frequency and the aggregation ratio. Moreover, Step 4) represents gradual aggregation and it can be achieved by repeating the Step 1) to 3) with higher aggregation ratio. For example, if the data is initially aggregated with a ratio of 2 (1:2) in that case it can be aggregation with a ratio of 5 (1:5) followed by 10 (1:10) and so on, as the data grows older.

3.2 The Row-Based Aggregation Method

This method always aggregate rows with a user defined aggregation ratio. For example, if we like to aggregate rows in Table 1 with an aggregation ratio of 5 (1:5) for parameter 247, this method will always aggregate five rows, based on the same level of granularity. One of the major differences in this method and the method presented in subsection 3.1 is the GROUP BY clause. The row-based aggregation *cannot* be done using a SQL GROUP BY clause instead it is done using a program (C, etc). This is because the decision about what rows to put into a group depends on the previous rows, not on any values in the rows themselves. The overall aggregation process using this method is a five-step process, out-lined in the following pseudo-code. 1) *Get data that is more than m month old in a array (record set) for further processing;* 2) *Aggregate the data by manipulating the array;* 3) *Get the aggregated data and insert it back into the fact table with changed granularity levels that represent higher aggregated from the fact table;* and 5) *Go to Step 1 for higher level of aggregation. level of granularity;* 4) *Delete all the rows that are just been* The details of Step 2) and 5) are further described below, whereas, Step 1), 3) and 4) are trivial. The following C code represents the data aggregation part of Step 2). It aggregate rows (by returning the largest value) based on the aggregation ratio, for

instance, if aggregation ratio is 5 (1:5) in that case it will always aggregate five rows based on the same level of granularity.

```

...for (i=17; i < columns*(rows+1); i=i+6){
    if ((count < aggregation_ratio)&&
        (atoi(result[i-2]==granularity)){
        if (max < atoi(result[i])
            {max=atoi(result[i]); count++;}}
        else if((count < aggregation_ratio)&&
            (atoi(result[i-2]!=granularity))
            {insert into fact table{...} count=1;
            granularity=result[i-2];}
        else{insert into fact table{...} count=1;
            max=atoi(result[i])}}};...

```

In the above mentioned code, *i* starts with 17 as we need to start from the third row (first row contains the attribute names) of the single dimensional array since the Value attribute of the second row is selected as maximum, *i* must be less than the length of the result set and the value of *i* is incremented by 6 as there are 6 columns in Table 1. For each increment in the *loop* multiple *if else* statements are executed to check the aggregation ratio and level of granularity, while applying the aggregate MAX. If *count* exceeds the *aggregation_ratio* or level of granularity changes in both cases the aggregated data is inserted into the fact table. Similarly, after the insertion of the first set of aggregated data, the next value is chosen as MAX and process goes on until the end of the *for loop*. Furthermore, Step 5) represents gradual aggregation and it can be achieved by repeating the Step 1) to 4) with higher aggregation ratio, as the data grows older.

3.3 The Time Granularity-Based Aggregation Method

This method aggregates time intervals based on single time hierarchy, such as, *Second, Minute, 2 Minutes, 10 Minutes, 20 Minutes, Hour, Day, Quarter* and *Year*. The overall aggregation process related to this method is similar to interval-based aggregation method (subsection 3.1); however, Step 1) of the aggregation process differs in a sense that it does not consider the Granularity (attribute) and the aggregation ratio. Instead, it makes use of time granularity to aggregate data. As an example, let us assume this method depends on the following flexible rules. 1) when data is 3 months old aggregate to 1 minute level from 20 seconds level or 30 seconds level, 2) when data is 6 months old aggregate to 2 minutes level from 1 minute level, 3) when data is 9 months old aggregate to 10 minutes level from 2 minutes level etc.

When compared with the ratio-based and row-based aggregation methods this method has a limitation. The limitation is due to the fixed time hierarchical structure. If we do not apply the fixed time hierarchical structure to gradually aggregate data from the lower level of granularity to higher levels of granularity in that case the aggregation becomes erroneous due to the varying nature of granularity levels. The following SQL statement represents the data aggregation part of Step 1).

```

SELECT Task, Parameter, Timestamp, Time_granularity,
       MAX(Value)
FROM   Data_log
WHERE  Parameter = 247
AND    Timestamp < STRFTIME('%s', 'NOW', '-3MONTH')
GROUP BY Task, ROUND((Timestamp/(Time_granularity), 0) *
                    (Time_granularity));

```

The most significant part of the SQL statement is the GROUP BY clause. The SQL GROUP BY clause calculates the beginning of each time interval based on the time granularity and aggregates several different levels of granularity into a single higher level of granularity. The *Timestamp* field is first divided by *Time_granularity* such as, 60 (60 seconds), 120 (120 seconds) and so on, next the ROUND() function is used to round the result to the nearest integer. Further, the rounded result is multiplied with the same denominator in order to get the single higher level of granularity.

4 Evaluation

An evaluation of the proposed aggregation methods has been done. Performance tests have been carried out for both single-level and multi-level aggregation queries. The single-level queries aggregate data from a single level of granularity to a higher level of granularity. The multi-level queries aggregate data from several different levels of granularity to a single higher level of granularity. The tests were designed to evaluate the effectiveness and efficiency of the proposed methods to handle the aggregation problems mentioned in Section 2.

The first test was performed on a fact table (Table 2) having data at a single level of granularity and without any aggregation problems (mentioned in Section 2). The results show that methods 1 and 2 (interval-based and row-based aggregation methods) produce exactly same output (1011, 7, 247, 1155208190, 150, 34.30) and (1012, 7, 247, 1155208340, 150, 34.45) with aggregation ratio of (1:5) and function MAX(). Both these methods produced two aggregated rows with changed Granularity and Timestamps. The changed Granularity value represents the higher granularity level at 150 seconds. Moreover, method 3 (time granularity-based aggregation method) produced the following result (1011, 7, 247, 1155208160, 120, 34.27), (1012, 7, 247, 1155208280, 120, 34.40) and (1013, 7, 247, 1155208340, 120, 34.45) with time granularity at 2 minutes level and function MAX(). This method produced three aggregated rows with changed Granularity and Timestamps. The changed Granularity value represents the higher granularity level at 120 seconds.

The second test was performed on a fact table (Table 3) having data at a different levels of granularity and with all the aggregation problems (mentioned in Section 2). In Table 3, row number 1002 represents data with holes (Timestamp = 1155208100 is missing), row number 1004 represents a delay in capturing data (Timestamp = 1155208190 should be the correct value), row number 1010 and 1011 represent a change of granularity level (Granularity value has been changed from 30 to 60 seconds due to the change in logging frequency) and finally row number 1012 represents already aggregated data at 10 minutes granularity level. The results show that methods 1 and 3 (ratio-based and time granularity based aggregation methods)

Table 2. Single granular fact table

Rid	T	P	TS	G	V
1001	7	247	1155208070	30	34.20
1002	7	247	1155208100	30	34.23
1003	7	247	1155208130	30	34.25
1004	7	247	1155208160	30	34.27
1005	7	247	1155208190	30	34.30
1006	7	247	1155208220	30	34.33
1007	7	247	1155208250	30	34.37
1008	7	247	1155208280	30	34.40
1009	7	247	1155208310	30	34.42
1010	7	247	1155208340	30	34.45

Table 3. Multi-granular fact table

Rid	T	P	TS	G	V
1001	7	247	1155208070	30	34.20
1002	7	247	1155208130	30	34.25
1003	7	247	1155208160	30	34.28
1004	7	247	1155208194	30	34.30
1005	7	247	1155208220	30	34.33
1006	7	247	1155208252	30	34.35
1007	7	247	1155208281	30	34.38
1008	7	247	1155208310	30	34.40
1009	7	247	1155208343	30	34.43
1010	7	247	1155208403	60	34.50
1011	7	247	1155208460	60	34.56
1012	7	247	1142561210	600	05.55

Table 4. 1st level aggregation (methods 1 & 3)

Rid	T	P	TS	G	V
1010	7	247	1155208403	60	34.50
1011	7	247	1155208460	60	34.56
1012	7	247	1142561210	600	05.55
1013	7	247	1155208070	60	34.20
1014	7	247	1155208130	60	34.25
1015	7	247	1155208194	60	34.30
1016	7	247	1155208252	60	34.35
1017	7	247	1155208310	60	34.40
1018	7	247	1155208343	60	34.43

Table 5. 1st level aggregation (method 2)

Rid	T	P	TS	G	V
1010	7	247	1155208403	60	34.50
1011	7	247	1155208460	60	34.56
1012	7	247	1142561210	600	05.55
1013	7	247	1155208130	60	34.25
1014	7	247	1155208194	60	34.30
1015	7	247	1155208252	60	34.35
1016	7	247	1155208310	60	34.40
1017	7	247	1155208343	60	34.43

produced the above output (Table 4) with an aggregation ratio of (1:2) and time granularity equal to 60 seconds, respectively. Both these methods use aggregate function MAX() to produce aggregated rows with changed Granularity. In Table 4, row number 1010 and 1011 are unchanged as they were already at granularity level equal to 60 seconds. Similarly, row number 1012 is also unchanged as it was at a higher level of granularity, whereas, the rest of the rows are aggregated based on the given ratio and time granularity. Method 2 (row-based aggregation method) on the other hand, produced the above results (Table 5) with an aggregation ratio of (1:2) and function MAX(). In Table 5, row number 1010, 1011 and 1012 are unchanged (due to the similar reasons as for Table 4), while rest of the rows are aggregated with a user-defined ratio. The aggregated results are quite similar to methods 1 and 3 except for row number 1013 in Table 4. The difference is due to the row-based aggregation characteristics of the method 2. The subsequent gradual aggregation based on methods 1 and 3 produced the following output (Table 6) with an aggregation ratio of (1:10) and time granularity equal to 10 minutes, respectively. The row number 1012 remained unchanged due to higher level of granularity, whereas, row number 1019 and 1020 contain aggregated data based on two different time spans. In comparison, method 2 with an aggregation ratio of (1:10) produced the following output (Table 7). Similar to Methods 1 and 3, the row number 1012 remained

unchanged and row number 1019 hold the aggregated data for all the remaining rows. Further aggregation with higher aggregation ratio and time granularity may be performed to reduce the number of rows.

Table 6. 2nd level aggregation (methods 1 & 3)

Rid	T	P	TS	G	V
1012	7	247	1142561210	600	05.55
1019	7	247	1155208194	600	34.30
1020	7	247	1155208460	600	34.56

Table 7. 2nd level aggregation (method 2)

Rid	T	P	TS	G	V
1012	7	247	1142561210	600	05.55
1019	7	247	1155208460	600	34.56

The tests were also designed to measure the aggregation speed and the overall aggregation process speed in seconds. The overall aggregation process consists of 1) aggregating the existing rows based on different levels of granularity; 2) inserting the newly aggregated rows in the fact table; and 3) deleting the previous rows from the fact table. The tests were performed on a 2.0 GHz Intel® Core Duo with 512 MB RAM, running Ubuntu 8.04 (hardy) and MySQL 5.0.5. From the tests, it is observed that methods 1 and 2 have performed well with respect to the aggregation speed that is approximately 300,000 rows per second and the overall aggregation process speed that is approximately 550 seconds to process 10,000,000 rows. Method 2 has not performed reasonably well as compared to the rest of the methods in terms of the aggregation speed that is approximately 35,000 rows per second and the overall aggregation process speed that is approximately 815 seconds to process 10,000,000 rows for the reason that it requires complex programming in a language such as C, instead of SQL. Based on the above mentioned facts, methods 1 and 2 ensure high degree of flexibility in terms of data aggregation because of user defined ratios, in contrast to method 3 that it is based on fixed time hierarchy structure. Furthermore, methods 1 and 3 are quite fast in terms of aggregation speed as compared to method 2 that depends on programming language instead of SQL. Finally, based on the results, the proposed methods have demonstrated their ability to handle data “without holes” and “with holes” along with other aggregation problems (mentioned in Section 2).

5 Conclusion

Inspired by the growing needs of data aggregation techniques to reduce data volumes over time, this paper presents three gradual data aggregation methods in multi-granular fact tables on resource-constrained systems. We proposed the notion of interval-based, row-based and time granularity-based aggregation methods, which are capable of aggregating data both “without holes” and “with holes” in addition to delay in capturing data, change of Granularity value and different time intervals. We also presented a real-world case study from the farming business, where we store and query multi-granular data.

To save storage space is important not only for the farming business but for any other type of industry in which significant amounts of data are generated for this reason the methods presented in this paper are general.

Acknowledgments. This work is supported by the LandIT [13] project and is funded by the Danish Ministry of Science, Technology and Innovation.

References

1. Lopez, I.F.V., Moon, B., Snodgrass, R.T.: Spatiotemporal Aggregate Computation: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 17(2), 271–286 (2005)
2. Li, J., Srivastava, J.: Efficient Aggregation Algorithms for Compressed Data Warehouses. *IEEE Transactions on Knowledge and Data Engineering* 14(3), 515–529 (2002)
3. Rasheed, F., Lee, Y.K., Lee, S.: Towards using Data Aggregation Techniques in Ubiquitous Computing Environments. In: 4th IEEE International Conference on Pervasive Computing and Communication Workshops, pp. 36–392. IEEE Press, New York (2006)
4. Schulze, C., Spilke, J., Lehner, W.: Data Modeling for Precision Dairy Farming within the Competitive Field of Operational and Analytical Tasks. *Computers and Electronics in Agriculture* 59(1-2), 39–55 (2007)
5. Zhang, D., Gunopulos, D., Tsotras, V.J., Seeger, B.: Temporal and Spatio-Temporal Aggregations over Data Streams using Multiple Time Granularities. *Information Systems* 28(1-2), 61–84 (2003)
6. Boly, A., Hébrail, G., Goutier, S.: Forgetting Data Intelligently in Data Warehouses. In: IEEE International Conference on Research, Innovation and Vision for the Future, pp. 220–227. IEEE Press, New York (2007)
7. Skyt, J., Jensen, C.S., Pedersen, T.B.: Specification-based Data Reduction in Dimensional Data Warehouses. *Information Systems* 33(1), 36–63 (2008)
8. Iftikhar, N.: Integration, Aggregation and Exchange of Farming Device Data: A High Level Perspective. In: 2nd IEEE Conference on the Applications of Digital Information and Web Technologies, pp. 14–19. IEEE Press, New York (2009)
9. Madden, S.R., Franklin, M.J., Hellerstein, J.M., Hong, W.: TinyDB: An Acquisitional Query Processing System for Sensor Networks. *ACM Transaction on Database Systems* 30(1), 122–173 (2005)
10. Rosenmuller, M., Siegmund, N., Schirmeier, H., Sincero, J., Apel, S., Leich, T., Spinczyk, O., Saake, G.: FAME-DBMS: Tailor-made Data Management Solutions for Embedded Systems. In: EDBT Workshop on Software Engineering for Tailor-made Data Management, pp. 1–6. ACM Press, New York (2008)
11. Kim, G.J., Baek, S.C., Lee, H.S., Lee, H.D., Joe, M.J.: LGeDBMS: A Small DBMS for Embedded System with Flash Memory. In: 32nd International Conference on Very Large Data Bases, pp. 1255–1258 (2006)
12. UTC, http://en.wikipedia.org/wiki/Coordinated_Universal_Time
13. LandIT, <http://www.tekkva.dk/page326.aspx>

A Refinement Operator Based Method for Semantic Grouping of Conjunctive Query Results

Agnieszka Lawrynowicz¹, Claudia d'Amato², and Nicola Fanizzi²

¹ Institute of Computing Science, Poznan University of Technology, Poland
alawrynowicz@cs.put.poznan.pl

² Dipartimento di Informatica, Università degli Studi di Bari, Italy
{fanizzi,claudia.damato}@di.uniba.it

Abstract. The methods proposed for aggregating results of structured queries are typically grounded on syntactic approaches. This may be inconvenient for an exploratory data retrieval, with often overwhelming number of the returned answers, requiring their further analysis and categorization. For example, if the values instantiating a grouping criterion are all different, a separate group for each answer would be created, providing no added value. In our recent work, we proposed a new approach, coined semantic grouping, where the results of conjunctive queries were grouped based on the semantics of knowledge bases (ontologies) of reference. Specifically, a user defined grouping criterion was expressed as a concept from a given ontology, and results grouped based on the concept subsumption hierarchy. In this work, we propose a novel method for the task of semantic grouping, that is based on an application of a concept refinement operator. This novel method is able to deal with some cases not handled by the initially proposed one, where, for example, a grouping criterion is a primitive concept thus not allowing for further semantic grouping of the results. In such a way, we achieve a solution able to deal with both problems: of too large and of too small number of groups.

1 Introduction

Given a source of knowledge, the most common task consists in querying such a source of knowledge. An example is given by the (Semantic) Web which represents the most comprehensive source of knowledge, currently available. The users of the (Semantic) Web often perform interactive, and exploratory data retrieval, where queries often result in an overwhelming number of the returned answers. However typically, only a small part of the result set is relevant to the user. Manually separating the interesting items from the uninteresting ones is a tedious and time consuming job. For this reason, various services have been set up to manage the information overload. A typical example is given by the well known search engines that typically adopt a keyword-based search approach. However, when the number of the results is huge, even despite their ranked order, manually investigating them is a big effort without any additional navigation tools.

In order to facilitate browsing and managing results of a Web search/query, methods for grouping answers on the ground of a user defined criteria would be exploited. Classically, aggregation abilities are provided by SQL-like GROUP BY clause, which will be also supported in a new version of a standard Semantic Web query language – SPARQL¹. However, for some scenarios, the classical GROUP BY semantics, which is to partition the results by identical values, is not proper. Consider, for example, a scenario in which a user searches for weekend break offers, and would like to have them grouped by a destination. In the case, when destination is represented in a database by a town name, one group for each town name would be created, which could result in too many groups to be easily managed by the user.

In such cases, an alternative way of grouping could be used, that we coined *semantic grouping* [1]. In [1], some preliminary results have been presented on a method performing semantic categorization of the results. Specifically, given on input a query in the form of a conjunctive query [2,3,4,5,6], and a background ontology, the method returns a dynamic, *runtime* categorization over ranked query results. The key features of this method are: (a) the exploitation of the *semantics* of knowledge bases of reference (in the form of ontologies) for grouping results with respect to a user defined criterion; (b) the exploitation of deductive reasoning involving background knowledge during the query results aggregation. Particularly, given a certain grouping criterion, expressed as a (complex) concept from a knowledge base of reference, results are grouped in agreement with (part of) the subsumption hierarchy deductively obtained by considering the specified concept and the given ontology. Consider, for example, that the destination criterion from our discussed scenario, would be represented by a concept from an ontology, being a root of a concept hierarchy which we would like to exploit for grouping the results. In case, the towns would be annotated by subconcepts of concept *Destination* such as *City*, *EuropeanDestination*, *ItalianDestination*, *PolishDestination*, it would be possible to aggregate the results into a smaller number of ‘semantic’ groups e.g. into Polish destinations or Italian destinations.

The problems for this approach arise when the criteria used for the aggregation correspond to leaf concepts of the reference ontology. Indeed, in this case no other information is available for performing semantic aggregation. In this paper, we propose an extension of the method proposed in [1] in order to overcome the aforementioned limitation. It is based on the exploitation of a concept refinement operator. The overall idea is that: when leaf concepts are used as aggregation criteria, this concepts could be specialized (the use of the proposed refinement operator) in order to be able to build a concept hierarchy as in the case in which the concepts are not leaves in the reference ontology, thus allowing for further semantic grouping of the results. Furthermore, we propose a technique to handle the opposite case, where the grouping criterion is a top (most general) concept in the ontology, what may lead to impractically large number of semantic groups. In such a way, we achieve a solution able to deal with the two main problems not solved in [1]: managing too large and too small number of (semantic) groups.

¹ <http://www.w3.org/TR/rdf-sparql-query/>

The rest of the paper is organized as follows. In Sect. 2 the basics of the knowledge representation formalism of choice (namely description logics and of conjunctive queries) and the basics of the method proposed in [1] are presented. In Sect. 3 the extension of the method for performing semantic group by able to overcome the limitation present in [1] is illustrated. Sect. 4 discusses the work related to ours. In Sect. 5 conclusions are drawn.

2 Preliminaries

2.1 Language of Knowledge Representation

We assume standard notation for the syntax and semantics of the description logics knowledge bases [7]. The main building blocks of DL knowledge bases are *atomic concepts* (denoted by A), and *atomic roles* (denoted by R, S). *Complex descriptions* (denoted by C and D) are inductively built by using concept and role *constructors* (such as \sqcap, \sqcup, \neg).

A DL *knowledge base*, KB , is formally defined as: $KB = (\mathcal{T}, \mathcal{A})$, where \mathcal{T} is called a TBox, and it contains axioms dealing with how concepts and roles are related to each other, and where \mathcal{A} is called an ABox, and it contains assertions of individuals to concepts and roles. A DL knowledge base can be given semantics by translating it into first-order logic with equality. Atomic concepts are translated into unary predicates, complex concepts into formulae with one free variable, and roles into binary predicates. For the details on the DL formalism please refer to [7]. An example of a DL knowledge base, describing weekend break offers, is shown below.

Example 1 (Description logic KB).

$$\begin{aligned} \mathcal{T} = \{ & \text{City} \sqsubseteq \text{Destination}, \text{EuropeanDestination} \sqsubseteq \text{Destination}, \text{ItalianDestination} \sqsubseteq \text{EuropeanDestination}, \\ & \text{PolishDestination} \sqsubseteq \text{EuropeanDestination}, \text{Hotel} \sqsubseteq \text{Accommodation}, \text{BudgetAccommodation} \sqsubseteq \\ & \text{Accommodation}, \text{B\&B} \sqsubseteq \text{BudgetAccommodation}, \text{Hostel} \sqsubseteq \text{BudgetAccommodation}, \text{SkiingSite} \sqsubseteq \text{Site}, \\ & \text{SightSeeingSite} \sqsubseteq \text{Site}, \top \sqsubseteq \forall \text{hasDestination} \neg .\text{WeekendBreakOffer}, \\ & \top \sqsubseteq \forall \text{hasAccommodation}.\text{Accommodation}, \top \sqsubseteq \forall \text{hasSite}.\text{Site} \}. \end{aligned}$$

$$\begin{aligned} \mathcal{A} = \{ & \text{locatedIn}(\text{ZAKOPANE}, \text{TATRA}), \text{hasSite}(\text{ZAKOPANE}, \text{SKILIFTS_NOSAL}), \text{SkiingSite}(\text{SKILIFTS_NOSAL}), \\ & \text{locatedIn}(\text{CHOCHOLOW}, \text{TATRA}), \text{hasSite}(\text{CHOCHOLOW}, \text{HIGHLANDERS_WOODEN_HOUSES}), \\ & \text{SightSeeingSite}(\text{HIGHLANDERS_WOODEN_HOUSES}), \text{Mountains}(\text{TATRA}), \text{WeekendBreakOffer}(\text{O1}), \\ & \text{hasAccommodation}(\text{O1}, \text{A1}), \text{B\&B}(\text{A1}), \text{hasDestination}(\text{O1}, \text{ZAKOPANE}), \text{PolishDestination}(\text{ZAKOPANE}), \\ & \text{WeekendBreakOffer}(\text{O2}), \text{hasAccommodation}(\text{O2}, \text{A2}), \text{B\&B}(\text{A2}), \text{hasDestination}(\text{O2}, \text{CHOCHOLOW}), \\ & \text{PolishDestination}(\text{CHOCHOLOW}), \text{WeekendBreakOffer}(\text{O3}), \text{hasDestination}(\text{O3}, \text{ROME}) \}. \end{aligned}$$

Queries admitted in this work are *conjunctive queries* over DL knowledge bases [2,3,6]. Let N_C, N_R, N_I be the sets of *concept names*, *role names* and *individual names* respectively and let N_V be a countably infinite set of variables disjoint from N_C, N_R , and N_I . Let by \mathbf{x} and \mathbf{y} denote the sets of distinguished and nondistinguished variables, respectively, where $\mathbf{x}, \mathbf{y} \subseteq N_V$. A conjunctive query, denoted with $Q(\mathbf{x}, \mathbf{y})$, is the finite conjunction of a non-empty set of *atoms*.

An *atom* is an expression of kind $A(t_1)$ (concept atom) or $R(t_1, t_2)$ (role atom), where A is a concept name, R is a role name, and t_1, t_2 are individuals from N_I or variables from \mathbf{x} or \mathbf{y} . An answer to a query $Q(\mathbf{x}, \mathbf{y})$ w.r.t. KB is an assignment θ of individuals to distinguished variables such that $KB \models \exists \mathbf{y} : Q(\mathbf{x}\theta, \mathbf{y})$.

2.2 Task of Semantic Grouping of Query Results

Let us consider the knowledge base from Example [1](#), and suppose that the user submits a query on weekend break offers. In case many offers are available, the user would be interested in grouping the results with respect to different destinations. As we already discussed in the introduction, a merely syntactic approach, as the one used by the *group by* clause in the database context, could not be of great help in cases in which many destinations are found or all results refer to the same destination. To manage cases like this, we proposed *semantic group by* [\[1\]](#): an approach to group query results on the ground of concepts of a reference ontology. For instance, looking at the knowledge base in Example [1](#), results could be grouped on the ground of the pertaining country of a destination (ItalianDestination, PolishDestination).

The general idea behind *semantic group by* is to categorize the results with regard to concept hierarchies inferred for each variable in a grouping condition. This is formalized by means of a special second order predicate `categorize_by`.

Definition 1 (`categorize_by`). *A conjunctive query with a semantic aggregate subgoal is of the form*

$$\text{categorize_by}([X_1, X_2, \dots, X_m]) : Q(\mathbf{x}, \mathbf{y})$$

where $[X_1, X_2, \dots, X_m]$ is a grouping list of variables appearing in \mathbf{x} .

Example 2 (Example queries with the `categorize_by` clause).

$$Q_1(x, y) = \text{categorize_by}(y) : \text{WeekendBreakOffer}(x) \wedge \text{hasDestination}(x, y)$$

$$Q_2(x, y, z) = \text{categorize_by}(y, z) : \text{WeekendBreakOffer}(x) \wedge \text{hasDestination}(x, y) \wedge \\ \text{PolishDestination}(y) \wedge \text{hasAccommodation}(y, z)$$

$$Q_3(x, y, z) = \text{categorize_by}(y, z) : \text{WeekendBreakOffer}(x) \wedge \text{hasDestination}(x, y) \wedge \\ \text{locatedIn}(y, v) \wedge \text{Mountains}(v) \wedge \text{hasAccommodation}(y, z) \wedge \text{BudgetAccommodation}(z)$$

Since there may be more than one grouping variable, each answer may be described by a tuple of concepts named *semantic category*.

Definition 2 (Semantic category). *Given a query*

$$Q = \text{categorize_by}([X_1, X_2, \dots, X_m]) : Q(\mathbf{x}, \mathbf{y})$$

a semantic category is a tuple of concepts $\langle C^1, C^2, \dots, C^m \rangle$, where each C^i corresponds to X_i in the grouping variables list of Q .

Semantic categories form a hierarchy \mathcal{H} induced by the subsumption relation (drawn from the reference ontology) among concepts appearing in the same place in tuples. Then, *the task of semantic grouping* (the operational semantics for `categorize_by` clause) consists in creating the hierarchy \mathcal{H} , which is then used as

a multi-valued classification of the query results. This multi-valued classification is performed by assigning tuples from the input relation to the semantic categories on the ground of the inferred semantic types of the grouping variables.

3 A New Method for Semantic Grouping

Below we will briefly discuss the basic algorithm proposed for semantic grouping in [1]. Let X_i denote a grouping variable. Assume that, in the first step (i) for each such variable a (complex) concept C^i , representing type of X_i , may be derived. This may be done by exploiting the semantics of the query atoms and the background ontology (e.g. as proposed in [1]). Then, in the next step (ii), C^i is classified in the subsumption hierarchy of the concepts from the ontology produced by a DL reasoner, and the sub-hierarchy rooted in C^i is considered. This can be done for each concept C^i representing the type of each variable X_i in the set of grouping variables. All sub-hierarchies are then further used as input for (iii) a tree product operator [1] that outputs a final hierarchy of semantic categories (multi-valued/faceted classification). The last step (iv) is to assign answers to the proper categories on the ground of an instance checking inference procedure that is performed by a standard deductive reasoner [7].

In some cases, however, an application of such a top-down approach may not be enough to obtain a meaningful hierarchy of groups. Consider for example the following cases: (a) it is not possible to derive any specific typing for the grouping variable, without looking at the actual instance data (an ABox), as in case of variable y from query Q_1 (Example 2) (b) the concept C^i is a leaf in the classified subsumption hierarchy, e.g. `PolishDestination` in the knowledge base from Example 1, as happens for query Q_2 ; (c) all the retrieved results for variable X_i , e.g. all destinations, fall under the same type, e.g. again `PolishDestination` as happens in case of query Q_3 for variable y .

In the first mentioned case, the type of variable y of query Q_1 will be inferred as the top concept (\top). Since the top concept subsumes all the other concepts in an ontology, the number of generated semantic groups may be too large. In turn, in both last cases all the results will be assigned to one "semantic" group.

To overcome the above problems, in this paper, we propose a new method for generation of semantic groups based on: (i) an application of a *concept refinement operator* when a hierarchy of concepts is too shallow, and (ii) an application of a bottom-up, instance-driven strategy for determining grouping variable typing, in case there is no available information on typing in the query atoms themselves.

The application of the refinement operator may introduce further, unnamed, complex concepts into the sub-hierarchy rooted at C^i . For example, the concept `PolishDestination` could be then specialized into concepts: `PolishDestination \sqcap \exists hasSite.SkiingSite` and `PolishDestination \sqcap \exists hasSite.SightSeeingSite` to differentiate two Polish destinations from the knowledge base from Example 1, ZAKOPANE, and CHOCHOLOW, based on the sites they offer.

In case there is no typing available for a grouping variable X_i , a solution could be to apply an additional scan on the retrieved results to determine the typing of X_i based on assertions from an ABox. In the following subsections we formalize the above propositions.

3.1 Downward Refinement Operator

For the proposed setting, that is grouping query results in a real-time during information retrieval on possibly many types of knowledge bases, it is desirable that a refinement operator of interest: (i) traverses the space of concept descriptions efficiently, (ii) is independent of any particular DL language. Therefore, drawing from the characteristics of the operator proposed in [8], we introduce a specialization operator which is not complete, but designed with the goal of efficiency, and which is independent of a particular DL language.

Definition 3 (Downward refinement operator ρ). $\rho = (\rho_{\sqcup}, \rho_{\sqcap})$, where:

$[\rho_{\sqcup}]$ given a description in normal form $D = D_1 \sqcup \dots \sqcup D_n$:

- $D' \in \rho_{\sqcup}(D)$ if $D' = \bigsqcup_{1 \leq i, j \leq n} D_k$ for some $j \neq i, 1 \leq k \leq n$,
- $D' \in \rho_{\sqcup}(D)$ if $D' = D'_i \sqcup \bigsqcup_{1 \leq i, j \leq n}^{j \neq i} D_k$ for some $D'_i \in \rho_{\sqcap}(D_i)$

$[\rho_{\sqcap}]$ given a conjunctive description $C = C_1 \sqcap \dots \sqcap C_m$, and a set of descriptions in a form of either an atomic description A , or an existential restriction $\exists R.D_{j+1}$, $\text{Domain}(R, D)$, $C_i \sqsubseteq D$ for some $i \in \{1, \dots, m\}$:

- $C' \in \rho_{\sqcap}(C)$ if $C' = C \sqcap C_{j+1}$, where $C_{j+1} = A$, and $KB \not\models C \sqsubseteq A$,
- $C' \in \rho_{\sqcap}(C)$ if $C' = C \sqcap C_{j+1}$, where $C_{j+1} = \exists R.D_{j+1}$,
- $C' \in \rho_{\sqcap}(C)$ if $C' = (C \sqcup \neg C_j) \sqcap C'_j$, for some $j \in \{1, \dots, m\}$, where $C'_j = \exists R.D'_j$, $C_j = \exists R.D_j$, $D'_j \in \rho_{\sqcup}(D_j)$.

ρ_{\sqcup} drops one top-level disjunct or replaces it with a downward refinement obtained with ρ_{\sqcap} . ρ_{\sqcap} adds new conjuncts in the form of an atomic description, or an existential restriction involving a role whose domain subsumes one of the concepts in the conjunction, or replaces one conjunct with a refinement obtained by specializing concepts in the range of an existential restriction by ρ_{\sqcup} .

Please note, that we do not specialize concepts by adding expressions with \forall quantifier. This is due to the open world assumption, where even though every instance in a KB of interest possesses certain property, it cannot be deduced by a reasoner, which always assumes incomplete knowledge, and possible existence of a counterexample. Figure 1 presents a sample semantic category hierarchy, that may be obtained by an application of a refinement operator to concept `PolishDestination`.

3.2 Building Hierarchies of Semantic Categories

In this paper, we propose the following extensions of the basic method described in this section: (i) for cases when there is no typing available for a grouping variable basing only on a TBox, to replace a top-down strategy, by an instance-driven, bottom-up one, (ii) for cases where there is too specific concept obtained

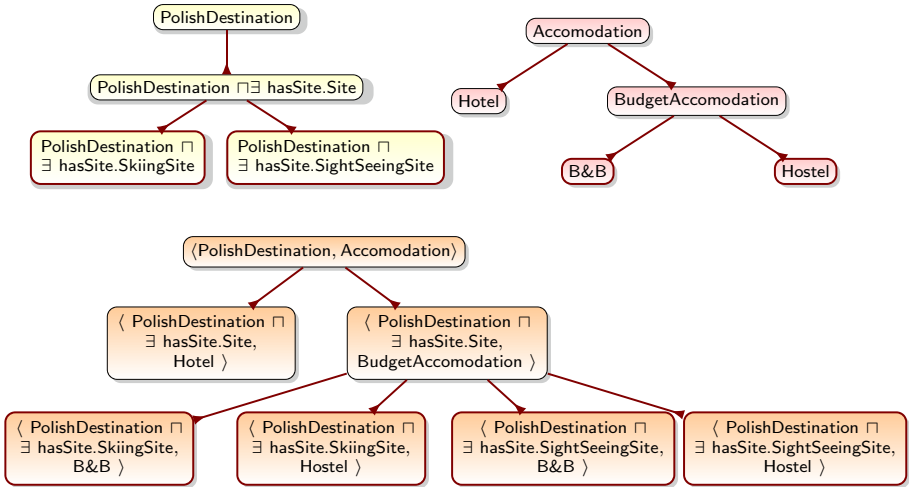


Fig. 1. Concept hierarchies, and a final hierarchy of semantic categories (obtained by an application of a levelwise tree product operator [11] for query Q_2 from Example 2)

as a grouping criterion, to apply a concept refinement operator for generation of new semantic groups. The new algorithm is presented in Figure 2.

For each grouping variable X_i , the algorithm first tries to infer a type of a variable (INFERTYPEOFGROUPINGVARIABLE) by exploiting the information from the query atoms. Firstly, the explicit typing represented by those concepts explicitly mentioned in the query atoms $C(X_i)$ is considered. Additionally, the implicit types inferred by the role atoms $R(X_i, \cdot)$ or $R(\cdot, X_i)$, respectively the domain and range of role R are added. Now, let $\mathcal{B}_i := \{C_1^i, \dots, C_{n_i}^i\}$ be the set of concepts describing individuals the variable X_i can be bound to according to the query atoms. The final concept determined for each query variable is $C^i := \sqcap_{C_p^i \in \mathcal{B}_i} C_p^i$. If any C^i is a top concept, then the algorithm performs one scan on the results (SCANRESULTSTODETERMINETYPEOFGROUPINGVARIABLES) to derive the variable typing. In the next step, a KB is classified, and each inferred concept C^i is placed in the proper place in the subsumption hierarchy. If a hierarchy for some C^i has less levels than a user-specified $MAXDEPTH$ threshold then the refinement operator is applied to further expand the hierarchy (SPECIALIZE). Then, all the hierarchies obtained for grouping variables are fed into a tree product operator (COMPUTETREEPRODUCT). A tree product is a binary operation on trees, which takes two trees T_1 and T_2 on input, and produces a tree T whose vertex set is a subset of the Cartesian product, and two product vertices (u_1, u_2) and (v_1, v_2) are connected in T , iff u_1, u_2, v_1, v_2 satisfy conditions of a certain type in T_1 and T_2 . Multiple variables are handled by an iterative application of a tree product. For some proposed tree product operators see [11]. Finally, the results (tuples) are assigned to the most specific semantic categories (POPULATE).

```

SEMANTICGROUPBY( $T, KB, Q, MAXDEPTH$ )
input:
 $T = (a_{ki})_{k=1,r}^{i=1,n}$ : query answer table
 $KB$ : knowledge base
 $Q$ : query
 $MAXDEPTH$ : maximum depth of semantic category hierarchy  $\mathcal{H}$ 
output:
 $\mathcal{H}$ : populated semantic category hierarchy
begin
allTyped  $\leftarrow$  true;
 $S \leftarrow \emptyset$ ; // set of (complex) concepts
 $\mathcal{HT} \leftarrow \emptyset$ ; // hash table, key - a concept from  $S$ , value - list of associated answer tuples
for  $X_i \in \text{GroupingVars}(Q)$  do
   $C^i \leftarrow \text{INFERTYPEOFGROUPINGVARIABLE}(Q, X_i, KB)$ ;
  if  $C^i = \top$  do
    allTyped  $\leftarrow$  false;
if  $\neg$ allTyped do
   $S, \mathcal{HT} \leftarrow \text{SCANRESULTSTOCOMPUTETYPEOFGROUPINGVARIABLES}(Q, T, KB)$ ;
   $KB \leftarrow KB \cup S$ ;
  CLASSIFY( $KB$ );
for  $X_i \in \text{GroupingVars}(Q)$  do
   $l \leftarrow \text{DepthOfConceptHierarchy}(Tax_{X_i} \in Tax)$ ;
  while  $l < MAXDEPTH$  do
    foreach concept  $C^{li}$  do
      set of  $C^{(l+1)i} \leftarrow \text{SPECIALIZE}(C^{li})$ ;
       $l \leftarrow l + 1$ ;
      UPDATE( $Tax_{X_i}$ );
 $\mathcal{H} \leftarrow \text{COMPUTETREEPRODUCT}(Tax)$ ;
 $\mathcal{H} \leftarrow \text{POPULATE}(\mathcal{H}, T, \mathcal{HT}, KB)$ ;
return  $\mathcal{H}$ 
end

SCANRESULTSTODETERMINETYPEOFGROUPINGVARIABLES( $Q, T = (a_{ki})_{k=1,r}^{i=1,n}, KB$ )
output:
 $S$ : set of (complex) concepts
 $\mathcal{HT}$ : hash table, key - a concept from  $S$ , value - list of associated answer tuples
begin
 $S \leftarrow \emptyset$ ;  $\mathcal{HT} \leftarrow \emptyset$ ;
for  $k \in \{1, \dots, r\}$  do
   $a_k \leftarrow T.\text{GETANSWER}(k)$ ;
  for  $i \in \{1, \dots, n\}$  do
    if  $C^i = \top$  do
       $a_{ki} \leftarrow a_k.\text{GETBINDING}(i)$ ;
       $C^{ki} \leftarrow \prod_{D \in \text{realization}(a_{ki})}$ ; // (conjunction of most specific
      atomic concepts in the realization
      [7] of  $a_{ki}$ .
       $C^i \leftarrow C^i \sqcap C^{ki}$ ;
       $S \leftarrow S \cup C^i$ ;
      update  $\mathcal{HT}$  with  $S$  and  $a_{ki}$ ;
end

```

Fig. 2. An algorithm for semantic grouping of query results

4 Related Work

To the best of our knowledge, the task of grouping conjunctive query results on the ground of semantics of the underlying knowledge base had not been addressed before our proposal presented in [1], and discussed throughout this paper.

The topic of aggregate queries was extensively studied for relational databases, but only few results for aggregate queries over ontologies may be found, especially such that target the peculiarities of the KR formalisms of the Semantic Web. In [9] the syntax and semantics for epistemic aggregate queries over ontologies were proposed, motivated by the non-adequacy of certain answer semantics for conditional aggregate queries over ontologies due to the open world assumption. In [10] the grouping and aggregate queries over RDF graphs were studied, motivated by the drawback of the previous works, not addressing the graph structure of the base data during aggregation. None of the above approaches exploited the peculiar feature of the Semantic Web datasets, that is possible availability of the background ontologies expressing semantics of the data.

Relevant work was presented in [11], where the results of SQL queries are automatically categorized, and a dynamic, labeled, hierarchical category structure generated. Since this was proposed for relational database model, the constructed categories were built based on the values in the retrieved tuples, and not on any semantic information linked to them, due to the lack of domain knowledge.

Relevant to ours are the works on concept refinement operators for description logic, in particular [8] from which we draw in this work, but also the other ones [12,13]. Since, differently to the cited works, we are not employing the operators for an inductive task, our basic aim is to have a very simple, but efficient operator, which is less complete than the ones proposed in the mentioned works. Importantly, our operator is independent of any DL language.

5 Conclusions and Future Work

This work addresses a new task of *semantic grouping*, which has been recently introduced in our previous work [1]. In this work, we have identified the cases where the basic techniques, initially proposed in [1] would provide poor solutions. Subsequently, we proposed a new method, based on the application of a concept refinement operator, and instance-driven strategy for deriving grouping variable typing, that is able to handle the identified cases. In the result, we extended our initial method, with capabilities of handling the cases when a grouping criterion is a primitive concept, or just the inverse - it is a top concept. This provides more complete solution for the task of semantic grouping.

In the future work we plan to perform an extensive investigation of our proposed techniques. Besides of testing the scalability, we will investigate the following research questions: *At which level of generality should the groups be? How to determine the depth of semantic categories (dynamically? user-defined?) How many groups is optimal? How many is too much?*. This will allow us to carry out an evaluation of the quality of the output of the algorithms, which will be twofold: by objective measures, by the subjective user assessment.

References

1. d'Amato, C., Fanizzi, N., Lawrynowicz, A.: Categorize by: Deductive aggregation of Semantic Web query results. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) *ESWC 2010*. LNCS, vol. 6088, pp. 91–105. Springer, Heidelberg (2010)
2. Calvanese, D., De Giacomo, G., Lenzerini, M.: On the decidability of query containment under constraints. In: *Proc. of the 17th ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS 1998)*, pp. 149–158 (1998)
3. Horrocks, I., Tessaris, S.: Querying the Semantic Web: a formal approach. In: Horrocks, I., Hendler, J. (eds.) *ISWC 2002*. LNCS, vol. 2342, pp. 177–191. Springer, Heidelberg (2002)
4. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R.: Data complexity of query answering in description logics. In: Doherty, P., Mylopoulos, J., Welty, C. (eds.) *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2006)*. AAAI Press, Menlo Park (2006)
5. Ortiz, M., Calvanese, D., Eiter, T.: Data complexity of answering unions of conjunctive queries in *SHIQ*. In: Parsia, B., Sattler, U., Toman, D. (eds.) *Proceedings of the International Workshop on Description Logics (DL 2006)*. CEUR-WS, vol. 189, CEUR (2006)
6. Glimm, B., Horrocks, I., Lutz, C., Sattler, U.: Conjunctive query answering for the description logic *SHIQ*. *J. Artif. Int. Res.* 31(1), 157–204 (2008)
7. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (eds.): *The Description Logic Handbook*. Cambridge University Press, Cambridge (2003)
8. Fanizzi, N., d'Amato, C., Esposito, F.: DL-Foil: Concept learning in Description Logics. In: Železný, F., Lavrač, N. (eds.) *ILP 2008*. LNCS (LNAI), vol. 5194, pp. 107–121. Springer, Heidelberg (2008)
9. Calvanese, D., Kharlamov, E., Nutt, W., Thorne, C.: Aggregate queries over ontologies. In: *ONISW 2008: Proc. of the 2nd international Workshop on Ontologies and Information Systems for the Semantic Web*, pp. 97–104. ACM, New York (2008)
10. Seid, D., Mehrotra, S.: Grouping and aggregate queries over Semantic Web databases. In: *International Conference on Semantic Computing*, pp. 775–782. IEEE Computer Society, Los Alamitos (2007)
11. Chakrabarti, K., Chaudhuri, S., Hwang, S.w.: Automatic categorization of query results. In: *SIGMOD 2004: Proc. of the 2004 ACM SIGMOD International Conference on Management of Data*, pp. 755–766. ACM, New York (2004)
12. Iannone, L., Palmisano, I., Fanizzi, N.: An algorithm based on counterfactuals for concept learning in the Semantic Web. *Appl. Intell.* 26(2), 139–159 (2007)
13. Lehmann, J.: DL-learner: Learning concepts in description logics. *Journal of Machine Learning Research (JMLR)* 10, 2639–2642 (2009)

Semantic Network of Ground Station-Satellite Communication System

Katarzyna Dąbrowska-Kubik

Warsaw University of Technology, Faculty of Electronics and Information Technology,
The Institute of Computer Science, Warsaw, Poland
K.Dabrowska@ii.pw.edu.pl

Abstract. This paper describes Semantic Network Methods of the Ground Station-Satellite Communication System (abbrev. GS-SCS), in the context of its further development and future needs. The abstract data model for the GS-SCP protocol is formulated with ASN.1. It supports a designer in reconfiguring the GS-SCS system. The paper shows how to create the ontology, and how to implement a software agent which uses this ontology. One of the aims is to apply the ASN.1 notation for GS-SCP protocol by this software Agent of the semantic network of the GS-SCS system. This tool has chance to be used in practice in the PW-SAT satellite project, developed by the students of Warsaw University of Technology.

1 Introduction

The Ground Station – Satellite Communication System is designed to provide communication via the Internet between satellite and ground stations located all around the world via the Internet [2]. It has client-server architecture (Fig. 1). Ground station uses a client application to communicate with the GS-SCS Server through the protocol layer based on TCP. The data stream of this protocol has binary format. This architecture allows the centralization of the data received from a satellite by various GSs in the Data Base. The client applications are able to communicate with the server only. People who do not have their own GS are able to communicate with a satellite using the limited version of GS-SCS Client.

Owing this architecture, the centralization of the data received from a satellite by various GSs is possible. The packets from the satellite are received by all GSs, which are available at a given moment. On the other hand, the data is sent only by one station, which is chosen using Packet Voting System [3] [4]. This solution allows to increase the amount of time during which a satellite is in the range of GS. The second advantage of the system is the possibility of distributed receipt of information through a couple of stations at the same time from the *CubeSats*.

In the last years *CubeSats* [2] are more and more common especially among the student groups. Owing this architecture the better use of the power of single station is possible. The GS-SCS system's users, which don't have their own GS-es, may use any available station at given moment. In addition, the high quality of receiving data from satellite is obtained using Packet Voting System [3].

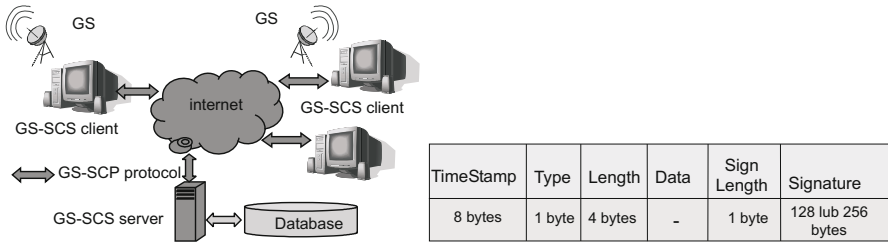


Fig. 1. The architecture schema of the GS-SCS system (left) and the GS-SCP protocol packet (right)

Up to now, the GS-SCS system serves only one type of satellite called PW-SAT, which uses GS-SCP protocol for transmitting data in two way directions [2][4]. The PW-SAT¹ is a CubeSat project that has chance to be the first Polish satellite. It is currently developed by students at the Warsaw University of Technology by the Faculty of Electronics and Information Technology and the Faculty of Power and Aeronautical Engineering. The main goal of this project is to develop an innovative and inexpensive method of satellite deorbitation based on sail. This method could be used in the future for removing various payloads from LEO orbits. The single packet of GS-SCP protocol is shown in figure (Fig. 1). It's worth creating an abstract model of the data for GS-SCP protocol. Thanks to it, GS-SCS system will be able to serves different types of satellites, which can appear soon in the polish space technology. Regardless of people's addresses and the satellites' positions, everyone can communicate with every satellite in the GS-SCS system in any moment. This solution requires an automatic reconfiguration of GS-SCS Client and Server [1].

2 Semantic Network of the GS-SCS

In order to achieve the idea, described in a previous section, a new semantic network for the GS-SCS system has been created. At the beginning the ontolgy has been built. It describes satellite in the context of the GS-SCP protocol. Moreover, it expresses semantic meaning of data structures and relations between them. At last, the ontology of the GS-SCS system describes the GS-SCS Server's and Client's answers to a receiving type of packet of the GS-SCP protocol. The second part of this semantic network is a *Library Generator*, which implements a packet flow between GS-SCS Client and Server using knowledge from the ontology of the GS-SCS system. The main advantage of this platform is possibility of saving a lot of developers' time and also improving the quality of a program's source code. Additionally, the ontology of the GS-SCS system has very precise information about this system, especially about the GS-SCP protocol.

¹ <http://www.pw-sat.pl/>, the PW-SAT project home page.

2.1 Abstract Syntax Notation Number One (ASN.1)

ASN.1 (Abstract Syntax Notation One) is an international standard, which aims at specifying of data used in telecommunication protocols. It is a computing language that is both powerful and complex. It was designed for efficient modeling communications between heterogeneous systems [5].

ASN.1 notation seems to be very good tool to formulate an universal data model for the GS-SCS system. A description of the GS-SCP protocol packet in the ASN.1 notation is given in figure (Fig.2). This packet is the same independently of the satellite type. The *data* structure may be one of the following type: *DataAP*, *DataACK-AP*, *DataCKP*, *DataIOP*, *DataROP*, *DataACK-SOP*, *DataCP*, *DataDOP*, *DataSG-SP*, *DataSRMP*, *DataQP*, *DataPP*, depending on the *Packet* type. In order to translate the abstract syntax notation of the GS-SCP protocol into Java library a tool called *Snacc For Java*² was used. The author suggests to apply the ASN.1 notation for the GS-SCP protocol by the software agent of semantic network of the GS-SCS system.

```

Packet ::= SEQUENCE
{
    type          Type,
    timeStamp     Timestamp,
    length        NumericString (SIZE(4)),
    data          Data,
    signLen      NumericString (SIZE(1)) OPTIONAL,
    sign         Signature OPTIONAL,
    satId        NULL OPTIONAL
}

```

Fig. 2. A description of the GS-SCP protocol packet in the ASN.1 notation

2.2 Building OWL Ontology for the GS-SCS

An ontology describes the concepts in the domain and also the relationships that hold between those concepts. Different ontology languages provide different facilities. The most recent development in standard ontology languages is the *Ontology Web Language (OWL)* from the World Wide Web Consortium.

For creating an OWL ontology, the Protege-OWL plug-in has been used [6]. This tool has been chosen because of supporting all stages ontology creating process (Fig.3) and fulfilling all requirements. The first stage is *determining domain*, which ontology is describing. In the GS-SCS the main domains are *Ground Station Network (GSN)* systems and protocols, which are used for communication between them [2]. The next stage is *defining classes* in a separating domain range. Things, which were managed to determining in this area, has been shown in figure (Fig.4). The third step is *defining datatype and object property* between individuals, which have been created in previous step. Each *object property* may have corresponding *inverse property* i.e. $isTypeOf(Type, Data) \Rightarrow hasType(Data, Type)$. Each property concerns the same individuals. Only domains and ranges has been exchanged. If a property is *functional* for a given

² <http://galera.ii.pw.edu.pl/docs/Snacc4Java/classdoc/>, Snacc4Java: ASN.1 Runtime Library.

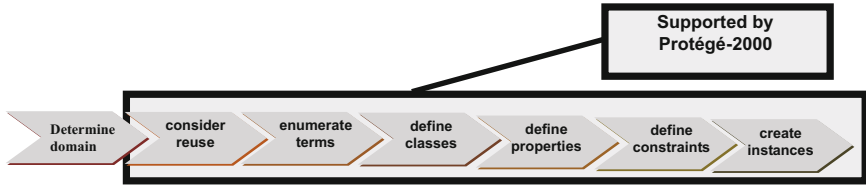


Fig. 3. The stages of the defining ontology process in Protégé-2000

individual, there can be at most one individual that is related to the individual via the property. The example of functional property is $hasType(Data, Type)$. If a property is *inverse functional*, then it means, that inverse property is functional. Protégé-2000 sets inverse functional property for $isTypeOf(Type, Data)$ automatically using inverse property and functional property defining before. In the OWL properties are used to create restrictions. Constraints are used to restrict the individuals that belongs to a class. The GS-SCS ontology needs quantifier restrictions. For example, for a Packet class the following restrictions has been created: $hasFields\ only((Data\ and\ Type\ and\ Length\ and\ Timestamp\ and\ Signature\ and\ SignLength)\ or\ (Data\ and\ Type\ and\ Length\ and\ Timestamp))$. Finally, the instances of the classes, called *individuals*, have been created.

What is important, the GS-SCS ontology has been used for automatically reasoning. In order to do this, some rules have been defined using *SWRL language*³. Rules describing a data packet flow in the GS-SCS system between the GS-SCS client and server have been created. Thanks to it, automatic generating of source code is possible both on the client and server side. The author assumes that the given rule is valid for all types of satellites, when there is no satellite individual itemized in this rule.

However, the SWRL language has a few disadvantages. One of them is lack of a negation operator. In the Protégé-2000 version 3.4 beta the only way to express negation is defining the second property, which has prefix -no in this name. For example, rule describing the GS-SCS server's answer on the receiving packet of a IOP type occurs in two versions: $receiveIOPyes$ and $receiveIOPno$. The main difference between them is that, the first one is executed when user has privileges to send a given command in a IOP packet type. In other words, the property $hasPrivileges(?y, ?z)$ has the positive value in this situation. The second rule is executed in the opposite case. Because of the lack of negation operator, the author had to defined additional property, called $hasNoPrivileges(?y, ?z)$. In comparison to the $hasPrivileges(?y, ?z)$ property, the only difference is a prefix -no in its name. The $receiveIOPno$ rule has been shown below:

$$sameAs(Packet, ?x) \wedge sameAs(UserDGSS, ?y) \wedge sends(?y, ?x) \wedge sameAs(Role, ?z) \wedge hasNoPrivilege(?y, ?z) \wedge recives(ServerDGSS, ?x) \wedge hasType(?x, IOP) \wedge sameAs(Packet, ?w) \Rightarrow sends(ServerDGSS, ?w) \wedge hasType(?w, RIOP)$$

³ <http://protege.cim3.net/cgi-bin/wiki.pl?SWRLLanguageFAQ#nid9KP>, the SWRL language description.

2.3 Agent’s Prototype of the GS-SCS Semantic Network

The semantic network of the GS-SCS system consists of the following modules: the GS-SCS ontology, the *LibGen* application, which plays agent’s role in the whole system, the GS-SCS client, the GS-SCS server and the knowledge database, which is in database form.

In a semantic network an agent is a software, which executes complex tasks given by the user automatically. The most popular activity for agents is selection and processing of information from the internet. In the GS-SCS system the agent plays similar role. The schema of the semantic network of the GS-SCS system has been shown in figure (Fig 4). As it has been mentioned before, the *LibGen* application is the agent of the semantic network. This software is used both on the client and server side. The agent gets all information about the current functional requirements from the internet in an ontology form. There is also a possibility to set input parameters for *LibGen* application by the user on the GS-SCS client side. The application generates a distinct block of a source code for different types of a satellite. The most important task of the *LibGen* application is to generate of library’s source code (Fig 4). After that, a recompilation of all parts of the GS-SCS system is done. If all of operations have been executed without any errors, the restart of the GS-SCS system is done.

On the server side all procedures of the GS-SCS system, which are modified by the *LibGen* application, are in the third library. The *Library3* consist of one file called *ServerPacketFlow.cpp*. These procedures, which are on the client side,

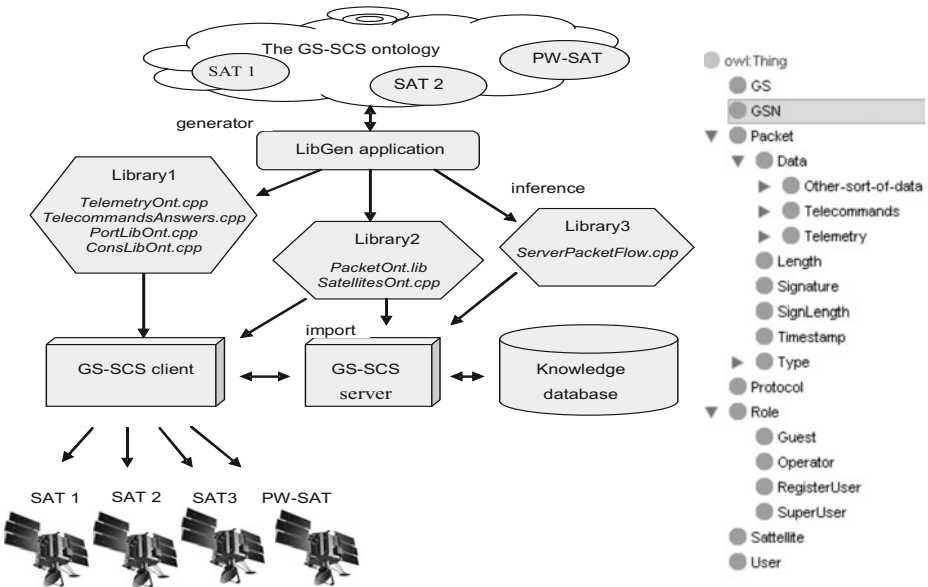


Fig. 4. Semantic network of the GS-SCS (left) and the individuals of the GS-SCS defined in Protégé-2000 (right)

are in the first library. Library1 consist of the following files: *TelemetryOnt.cpp*, *TelecommandsAnswers.cpp*, *PortLibOnt.cpp* and *ConsLibOnt.cpp*. This library is different depends on a satellite type. *PacketOnt.lib* and *SatellitesOnt.cpp* (Library2) are used by both the client and server.

Extended in this way the GS-SCS system is able to service any type of satellite. A company, which produces a new type of satellite, has to do only one thing - define the ontology for the data structures of its satellite. For a proper executing of the LibGen application each company had to restrict a few constraints during developing the GS-SCS ontology [1].

The author has assumed, that only one field of the GS-SCP protocol packet, called *Data* has been changed depending on the satellite type (Fig. 1). The second difference is a packet type, which is sent by the GS-SCS server in answer on receiving data form the GS-SCS client. At least, taking into consideration the difference of the data structures, the database has a distinct mutation for each satellite type.

2.4 Implementation of the Agent of the GS-SCS Semantic Network

For agent's implementation *Jena RDF framework*⁴ has been used, because it gives methods for processing an ontology defining in the OWL language. Jena storages data in the RDF format by default.

The database can be modified in two ways. The first one assumes, that the ontology is storages in form of a database schema. The second option is to write the whole ontology model to one table in the relational database, whatever the structure is. From a software developer point of view this solution is much easier. However, it may cause a trouble in the future. The main reason is that, the database will be denormalized. Apart from that, there will be a necessary to reconstruct the way of saving data in the database by the GS-SCS server.

That is way; the first solution has been chosen for the agent's implementation. In other words, classes have been transformed into tables in a database. The datatype properties of a given class has been rows in the table mentioned above. Taking into consideration the fact, that classes may be organized into a superclass-subclass hierarchy, which is also known as a taxonomy, the rules for setting constraints on tables have been formulated.

First of all, for each class from the ontology, a table has been created with one column, called *nameOfClassId*. This column has set a primary key with an auto-increment option. In the next step, the columns with a foreign key are created. For each subclasses of a given class, the column with a foreign key, called *nameOfSubclassId*, is created. The second group of columns is created from the datatype properties of a given class.

As mentioned above, that the main goal of the semantic network of the GS-SCS system is to provide "knowledge" for automatic reconfiguration of the system by using the files created by LibGen application. As can be seen in Fig. 4, we distinguish three groups of the files. The first group has files used by both

⁴ <http://jena.sourceforge.net/ontology/>, Jena Ontology API.

the GS-SCS server an client: *SatelliteOnt.cpp* and *PacketOnt.cpp*. They contain interface to a given kind of a satellite and packet of the GS-SCP protocol respectively. In the second group there is one file *ServerPacketFlow.cpp* used by the GS-SCS server only, which contains different procedures for receiving and sending packets of the GS-SCP protocol. In the third group are *PortLibOnt.cpp* and *ConsLibOnt.cpp*, which are used by the GS-SCS client only. The *PortLibOnt.cpp* contains an interface to sending packets to the ground station through a COM port. The interface for a data consumption is in the *ConsLibOnt.cpp*. It is worth to notice, that in the third group of files, the classes, which have been built using SnacForJava tool on base of ASN.1 notation for the GS-SCP protocol, are used. For generating *ServerPacketFlow.cpp* library, the agent LibGen uses rules created in Protégé-2000 tool, describing a data packet flow in the GS-SCS system. Jena framework gives plug-in to automatic reasoning, but it doesn't support rules created in Protégé-2000. That is way, the plugin *edu.stanford.smi.protege.owl* has been used. The one disadvantages of this plug-in is the lack of good documentation.

3 Experiments

In this section we present the experiments with the Semantic Network of the GS-SCS system. Their main goal is to prove the correctness of the implemented solutions. We have performed 4 experiments. The first and the second have been run for the GS-SCS server side, whereas the third and the fourth ones by the client side. Experiments have been done with two satellites: the PW-SAT and a theoretical one, called SAT-1.

In Experiment 1 a database schema generated using the GS-SCS ontology, according to implementation rules described in 2.4 section of this article, has been tested. The correctness of the source code, which has been generated for the GS-SCS server needs (Fig 4), has been demonstrated in Experiment 2.

In Experiment 3 the *PacketOnt.lib* library, using by both the GS-SCS server and client side and the *ServerPacketFlow.cpp* file, using by the server side, generated by the agent of the GS-SCS Semantic Network, has been tested. Additionally, the correctness of the automatic reconfiguration of this part of the GS-SCS system, which answers for receiving data from satellite, has been proved. This experiment has been done for two different satellites: PW-SAT and SAT-1. They have different types of receiving data called *DataCP_PWSAT* and *DataCP_SAT1* respectively. Examples of this types of packets is given in (Fig 5).

The fragment of the GS-SCS client application has been shown in figure (Fig 6). It is a buffer on packets received from the satellites. The client has been received both *DataCP_SAT1* and *DataCP_PWSAT* packets and sent them to the server successfully. The proof for this can be seen in the figure - packets has changed status from *not sent (NS)* to *sent (S)*. The satellites (*sat_name*), which have sent this packets, the time of sending given packet (*satellite_send_time*) and the number of packet, for which, the given packet is an satellite's answer (*telecommand_id*), has been circled. It is worth to notice, that for SAT-1 the field

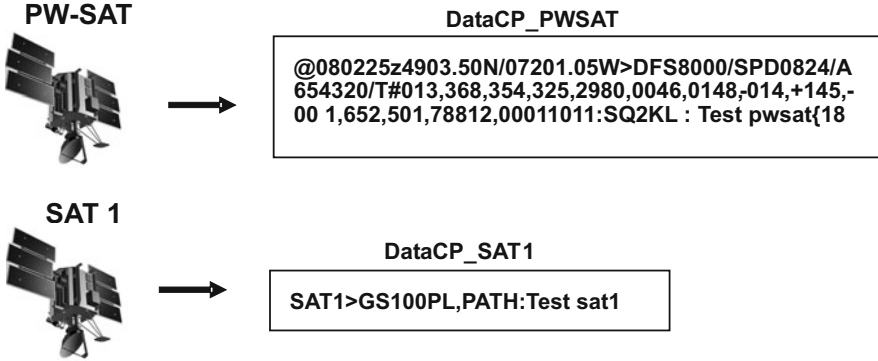


Fig. 5. The format of the *DataCP_SAT1* and *DataCP_PWSAT* packets received from the satellites SAT-1 and PW-SAT respectively

Telemetry						
sat_name	status	satellite_send_time	telecommand_id	receive_client_time	groundstation_id	data
sat-1	5		0	Tue Jan 27 13:20:01 200...	7	SAT1>GS100PL,PATH:Test sat1{yyyyy*****t
pw-sat	5	080225	18	Tue Jan 27 13:19:04 200...	7	@080225z4903.50N/07201.05W>DFS8000/SPD0824/A6543...

Fig. 6. The *DataCP_SAT1* and *DataCP_PWSAT* packets received from the satellites SAT-1 and PW-SAT respectively by the GS-SCS client

```
semweb=# select * from communicate;
communicate_id | @satellite_id | name |
contents
correct | (telecommand_id) | receive_server_time | received_client_time | data_id | (satellite_
send time
-----+-----+-----+-----+-----+-----+
25 | (1) | | @080225z4903.50N/07201.05W>DFS8000/SPD0824/A654320/T#013,368,3
54,325,2980,0046,0148,-014,+145,-001,652,501,78812,00011011:SQ2KL : Test pwsat{18 | 7 |
f | (18) | | Tue Jan 27 13:19:13 2009
| Tue Jan 27 13:19:04 2009
| | (080225)
| 26 | (2) | | SAT1>GS100PL,PATH:Test sat1 | 7 |
f | | | Tue Jan 27 13:20:11 2009
| Tue Jan 27 13:20:01 2009
| |
(2 rows)

semweb=#
semweb=# █
```

Fig. 7. The *Communicate* table of the database of the GS-SCS system at the end of the experiment no.3

satellite_send_time is empty and the field *telecommand_id* is equal to 0, because a packet, sent by SAT-1, doesn't have this information in it. In the next step of this experiment, the GS-SCS server has received these packets and stored in the database in the *Communique* table of the database of the GS-SCS system. The result of this operation has been given in the figure (Fig. 7), where the parameters *satellite_send_time* and *telecommand_id* have been highlighted too. Finally, in the experiment no.4 the correctness of the *ConsLibOnt.cpp* file, generated by the LibGen agent, has been checked.

4 The Assessment of Chosen Solution

The ASN.1 notation for the GS-SCP protocol has been converted to Java classes using *Snacc4Java* tool. Objects, created in this way, have been managed to use in the process of implementation of the agent of the GS-SCS Semantic Network. Indeed, a lot of time has been saved, because the author hasn't had to define the same classes in the *Jena API* framework. Unfortunately, the author hasn't found the way to import the abstract data model of the GS-SCP protocol in the ASN.1 notation to the Protégé-2000 framework, where ontology has been created. The classes and their properties have been created on the base of the data structure formulated in the ASN.1 notation.

The ontology of the semantic network of the GS-SCS system has been created in the *Protégé-2000* framework. This tool is free, still developing, and the most important, implemented in Java. That is way; this framework can be run under many different operating systems, such as MacOSX, Windows, AIX, Linux, Solaris and others. Its interface is quite intuitive and comfortable.

The most important part of the GS-SCS system is the agent, called *LibGen*. It generates files, which are used during the reconfiguration of the GS-SCS system. This application has been implemented in Java using *Eclipse*⁵ for two reasons. First of all, there are many free plugins in Java for creating and developing applications from semantic network area.

For the *LibGen* implementation, the author has used two plugins: *Jena API* and *Protégé-2000*. Both of them offer a rich set of features, which turn out to be sufficient for implementing an agent of semantic network. The plug-in *Protégé-2000* has been used, because Jena API doesn't support rules created in the SWRL language.

5 Conclusions

In this article the methods for improvement of the GS-SCS system have been presented. The abstract data model in ASN.1 notation and its use in the GS-SCS Semantic Network have been described. The project and implementation of the GS-SCS Semantic Network has been shown, especially the agent application, called *LibGen*. From many available solutions the author has chosen this one,

⁵ <http://www.eclipse.org/>, the Eclipse home page.

which has permitted to use created ontology by the *LibGen* agent. Finally, four experiments of the GS-SCS Semantic Network have been described [1].

The tool, presented in this article, has chance to be used in practice in the PW-SAT satellite project. The PW-SAT satellite project team has to change a data format of this satellite, because the engineering model and technical requirements have changed. The PW-SAT satellite launching time depends on the time when ESA will have been finished work at the Vega rocket. There will be a small amount of time for testing communication between the satellite and the GS-SCS system. That is way; the automatic reconfiguration of the GS-SCS system is very important and will make a success of the PW-SAT project possible.

References

1. Dąbrowska-Kubik, K.: Master thesis. Development of GS-SCS a satellite communication system (in polish: Praca magisterska. Koncepcje rozwoju systemu komunikacji satelitarnej GS-SCS), Warsaw (2009)
2. Dąbrowska, K.: Bachelor thesis. Ground Station - Satellite Communication System (in polish: Praca inżynierska. System komunikacji z satelitą poprzez internet i stacje naziemne), Warsaw (2007)
3. Stolarski, M.: The Use of Distributed Ground Station System for very low power communication. In: CD-ROM Conference Materials from The 1st International Workshop on Ground Station Network, Tokyo, Japan (2006)
4. Dąbrowska, K., Stolarski, M.: Ground segment of Distributed Ground Station System. In: The IEEE Region 8 Eurocon 2007 Conference, Warsaw, September 9-12 (2007)
5. Dubuisson, O.: ASN.1 Communication between heterogeneous systems. Morgan Kaufmann Publisher, San Francisco (September 2000)
6. Horridge, M., Knublauch, H., Rector, A., Stevens, R., Wroe, C.: A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools, 1st edn. The University Of Manchester (August 27, 2004)

W-kmeans: Clustering News Articles Using WordNet

Christos Bouras^{1,2} and Vassilis Tsogkas¹

¹ Computer Engineering and Informatics Department, University of Patras, Greece

² Research Academic Computer Technology Institute
N. Kazantzaki, Panepistimioupoli Patras, 26500 Greece
bouras@cti.gr, tsogkas@ceid.upatras.gr

Abstract. Document clustering is a powerful technique that has been widely used for organizing data into smaller and manageable information kernels. Several approaches have been proposed suffering however from problems like synonymy, ambiguity and lack of a descriptive content marking of the generated clusters. We are proposing the enhancement of standard kmeans algorithm using the external knowledge from WordNet hypernyms in a twofold manner: enriching the “bag of words” used prior to the clustering process and assisting the label generation procedure following it. Our experimentation revealed a significant improvement over standard kmeans for a corpus of news articles derived from major news portals. Moreover, the cluster labeling process generates useful and of high quality cluster tags.

Keywords: News clustering, k-means, Cluster Labeling, Partitional Clustering.

1 Introduction

While the amount of online information sources is rapidly increasing, so does the available online news content. One of the commonest approaches for organizing this immense amount of data is the use of clustering techniques. However, there are several challenges that clustering techniques normally have to overcome. Among them is efficiency: generated clusters have to be well connected from a notional point of view, despite the diversity in content and size that the original documents might have. For example, it is frequent for some news articles to belong to the same notional cluster, even though they do not share common words. The vice-versa is also possible: news articles sharing common words, while being completely unrelated to each other. Ambiguity and synonymy are thus two of the major problems that document clustering techniques regularly fail to tackle.

Furthermore, having IR systems simply generate clusters of documents is not enough per se. The reason is that it’s virtually impossible for humans to conceptualize information by merely browsing though hundreds of documents belonging to the same cluster. However, assigning meaningful labels to the generated clusters can help users conveniently recognize the content of each generated set and thus easily analyze the results.

Two generic categories of the various clustering methods exist: agglomerative hierarchical and partitional. Typical hierarchical techniques generate a series of partitions

over the data, which may run from a single cluster containing all objects to n clusters each containing a single object, and are widely visualized through a tree-like structure. On the other hand, partitional algorithms typically determine all clusters at once. For partitional techniques, a global criterion is most commonly used, the optimization of which drives the entire process, producing thus a single-level division of the data. Given the number of desired clusters, let k , partitional algorithms find all k clusters of the data at once, such that the sum of distances over the items to their cluster centers is minimal. Moreover, for a clustering result to be accurate, besides the low intra-cluster distance, high inter-cluster distances, i.e. well separated clusters, is desired. A typical partitional algorithm is k -means which is based on the notion of the cluster center, a point in the data space, usually not existent in the data themselves, which represents a cluster.

The family of k -means partitional clustering algorithms [1] usually tries to minimize the average squared distance between points in the same cluster, i.e. if d_1, d_2, \dots, d_n are the n documents and c_1, c_2, \dots, c_k are the k clusters centroids, k -means tries to minimize the global criterion function:

$$\sum_{i=1}^k \sum_{j=1}^n sim(d_j, c_i) \quad (1)$$

Several improvements have been proposed over this simple scheme, like bisecting k -means [2], k -means++ [3] and many more.

WordNet is one of the most widely used thesauri for English. It attempts to model the lexical knowledge of a native English speaker. Containing over 150,000 terms, it groups nouns, verbs, adjectives and adverbs into sets of synonyms called synsets. The synsets are organized into senses, giving thus the synonyms of each word, and also into hyponym / hypernym (i.e., Is-A), and meronym / holonym (i.e., Part-Of) relationships, providing a hierarchical tree-like structure for each term. The applications of WordNet to various IR techniques have been widely researched concerning finding the semantic similarity of retrieved terms [4], or their association with clustering techniques. For example in [5] they combine the WordNet knowledge with fuzzy association rules and in [7], they extend the bisecting k -means using WordNet; their methodology however is rather unclear.

Regarding cluster labeling, techniques frequently evaluate labels using information from the cluster themselves [8], while existing approaches that utilize other external databases, like Wikipedia [6] are only good for the labeling process and not the clustering one. Recently in [9], WordNet hypernyms were used for the labeling process; we found however that their weighting scheme didn't scale well with the number of documents.

In this paper we are presenting a novel algorithmic approach towards document clustering, and in particular, clustering of news articles deriving from the Web, that combines regular k -means with external information extracted from the WordNet database. We are also incorporating the proposed algorithm in our existing system [10], evaluating the clustering results compared to regular k -means using a large pool of Web news articles existing in the system's database.

2 Information Flow

The flow of information as handled by our approach is depicted in Fig. 1. At its input stage, our system crawls and fetches news articles from major or minor news portals from around the world. This is an offline procedure and once articles as well as meta-data information are fetched, they are stored in the centralized database from where they are picked up by the following procedures.

A key procedure of the system as a whole, which is probably as least as important as the clustering algorithm that follows it, is text preprocessing on the fetched article’s content, that results to the extraction of the keywords each article consists of. Analyzed in [10], keyword extraction handles the cleaning of articles, the extraction of the nouns [11], the stemming as well as the stopword removal process. Following, it applies several heuristics to come up with a weighting scheme that appropriately weights the keywords of each article based on information about the rest of the documents in our database. Pruning of words, appearing with low frequency throughout the corpus, which are unlikely to appear in more than small number of articles, comes next. Keyword extraction, utilizing the vector space model [12], generates the term-frequency vector, describing each article that will be used by the clustering approach technique that follows, as a ‘bag of words’ (words – frequencies).

Our aim towards increasing the efficiency of the used clustering algorithm is to enhance this ‘bag of words’ with the use of external databases, and in particular, WordNet (dashed box). This enhanced feature list, feeds the kmeans clustering procedure that follows. In this work, clustering is achieved via regular kmeans using the cosine similarity distance measure:

$$d(a,b) = \cos(\theta) = \frac{a \cdot b}{|a| |b|} \tag{2}$$

Where $|a|, |b|$ are the lengths of the vectors a, b respectively and the similarity between the two data points is viewed by means of their angle in the n -dimensional space. It is important to note however that the clustering process is independent of the rest of the steps, meaning that it can easily be replaced by any other clustering approach operating on a word-level of the input documents.

The generated clusters are finally forwarded for labeling, taking also advantage of the WordNet database. The labeling subprocess outputs suggested tags for the given cluster. Cluster assignments, as well as labels are the output of the proposed approach.

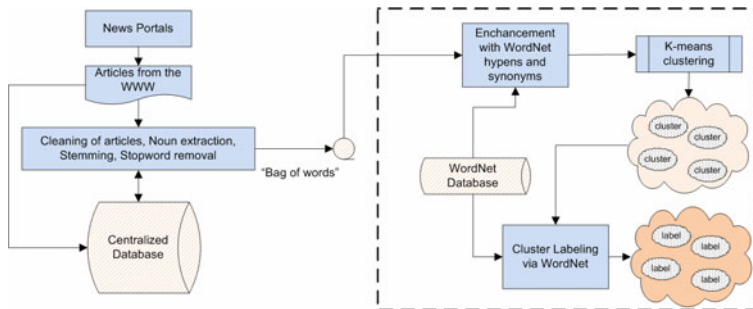


Fig. 1. Information flow for clustering news articles

3 Algorithm Approach

The WordNet lexical reference system, organizes different linguistic relations into hierarchies. Most importantly, given any noun, verb, adjective and adverb, WordNet can provide results regarding hypernyms, hyponyms, meronyms or holonyms. Using these graph-like structures, we can search the WordNet database for all the hypernyms of a given set of words, then weight them appropriately, and finally chose representative hypernyms that seem to extend the overall meaning of the set of given words. This intuitive approach, however, depends entirely on the weighting formula that will be used during the process. It is important that weighting only introduces “new knowledge” to the list of given words that will make the clustering result less fuzzy and more accurate.

3.1 Enriching Articles Using WordNet

Initially, for each given keyword of the article, we generate its graphs of hypernyms leading to the root hypernym (commonly being ‘entity’ for nouns). Following, we combine each individual hypernym graph to an aggregated one. There are practically two parameters that need to be taken into consideration for each hypernym of the aggregate tree-like structure in order to determine its importance: the depth and the frequency of appearance. For example, Fig. 2 depicts the aggregated hypernym graph for three terms: ‘pie’, ‘apple’, ‘orange’.

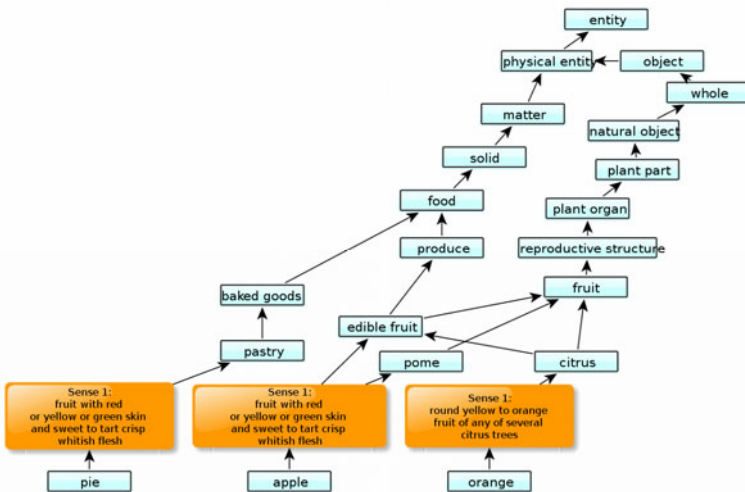


Fig. 2. Aggregate hypernym graph for three words: ‘pie’, ‘apple’, ‘orange’

It is observed that the higher (i.e. less deep, walking from the root node downwards) the hypernym is in the graph, the more generic it is. However, the lower the hypernym is in the graph, the less chances it has to occur in many graph paths, i.e. its frequency of appearance is low. In our approach, those two contradicting parameters are weighted using the formula (3).

$$W(d, f) = 2 \cdot \frac{1}{1 + e^{-0.125(d^3 \frac{f}{TW})}} - 0.5 \quad (3)$$

where d stands for the node's depth in the graph (starting from root and moving downwards), f is the frequency of appearance of the node to the multiple graph paths and TW is the number of total words that were used for generating the graph (i.e. total article's keywords). Function (3) is a sigmoid one with a steepness value including both the frequency and the depth of the hypernym. For large depth · frequency combinations, the weight of the hypernym reaches closer and closer to 1 (neither f nor d can be negative), whereas for low depth · frequency combinations the weight is close to 0. A keyword having no hypernym or not being in WordNet is omitted both from the graph and the TW sum. Furthermore, a hypernym may have multiple paths to the root, but is counted only once for each given keyword. Note also that the depth has a predominant role in the weighting process, much greater than frequency does. Frequency, however, acts as a selective factor when the graph expands with more and more keywords being added. We concluded to this weighting scheme after observations of hypernym graphs generated over hundreds of keywords because it scales well with real data. Given the aggregate hypernym graph in Fig. 2, we can compute the weight of the various hypernyms. For example for 'fruit': $d = 9$, $f = 2$ and $W = 0.9954$, where for 'edible fruit': $W = 0.8915$, and for 'food': $W = 0.6534$.

The enriching algorithm using WordNet hypernyms, as outlined in Algorithm 1, operates on the articles keywords generating a hypernym graph for each. We use only 20% of the article's most important keywords reducing, thus, dimensionality and noise as explained in [10]. Following, an aggregate graph is generated from which the weight of each hypernym is calculated using formula (3). The graph is sorted based on the nodes' weights and a list of the top keywords – hypernyms is returned, containing the suggested ones for enriching the article. We take into consideration a total size of a quarter of the article's hypernyms for the enriching ones.

Algorithm wordnet_enrich

Input: article a

Output: enriched list of keywords

```

total_hyphen_tree = NULL
kws = fetch 20% most frequent k/ws for a
for each keyword kw in kws
  htree = wordnet_hyphen_tree(kw)
  for each hyphen h in htree
    if (h not in total_hyphen_tree)
      h.frequency=1
      total_hyphen_tree ->append(h)
    else
      total_hyphen_tree ->at(h)->freq++
for each h in total_hyphen_tree
  calculate_depth(h)
  weight = 2 ((1/(1+ exp(-0.0125 * (h->depth ^3 * h->freq/
    kws_in_wn->size)))) - 0.5))
sort_weights(total_hyphen_tree)
important_hypens = (kws ->size/4)*top(total_hyphen_tree)
return kws += important_hypens

```

Alg. 1. Enriching news articles using WordNet hypernyms

3.2 Labeling Clusters Using WordNet

In order to generate suggested labels for each resulting cluster, we are also utilizing the WordNet hypernyms information as presented in Algorithm 2. Cluster labeling operates on each cluster, fetching initially 10% of the most important keywords belonging to each article of the cluster. We have found that this percentage is enough for the process to maintain a high quality level for the resulting labels by not introducing much noise. For each cluster's keyword we generate the hypernym graph and append it to the aggregate one. The resulting nodes are weighted, sorted and the top 5 hypernyms are returned as suggested labeling tags for the cluster. Using Algorithm 1 and 2, we can describe the algorithmic steps of W-kmeans as presented in Algorithm 3.

Algorithm wordnet_cl_labeling

```

Input: clusters
Output: cluster_labels
for each cluster c
  total_hyphen_tree = NULL
  for each article a in c
    cluster_kws += fetch 10% most frequent k/ws for a
  for each keyword kw in cluster_kws
    hypens_tree = wordnet_hyphen_tree(kw)
    for each hyphen h in hypens_tree
      if (h not in total_hyphen_tree)
        h.frequency=1
        total_hyphen_tree->append_child(h)
      else
        total_hyphen_tree->at(h)->frequency++
  for each hyphen h in total_hyphen_tree
    calculate_depth(h)
    weight = 2 ((1/(1+ exp(-0.0125 * (h->depth ^3 * h->frequency/
      kws_in_wordnet ->size)))) - 0.5))
  sort_weights(total_hyphen_tree)
  cluster_labels+=5*top(total_hyphen_tree)
return cluster_labels

```

Alg. 2. Labeling clusters using WordNet hypernyms

Algorithm W-kmeans

```

Input: articles, number of clusters
Output: cluster assignments
for each article a
  fetch 20% most frequent k/ws for a
  wordnet_enrich(a)
  clusters = kmeans()
return wordnet_cl_labeling (clusters)

```

Alg. 3. News article's clustering using W-kmeans

4 Experimental Procedure

For our experiments we used a set of 8000 news articles obtained from major news portals like BBC, CNN, etc. over a period of 2 months. Those articles were evenly shared among the 8 base categories that our system features. In order to determine the

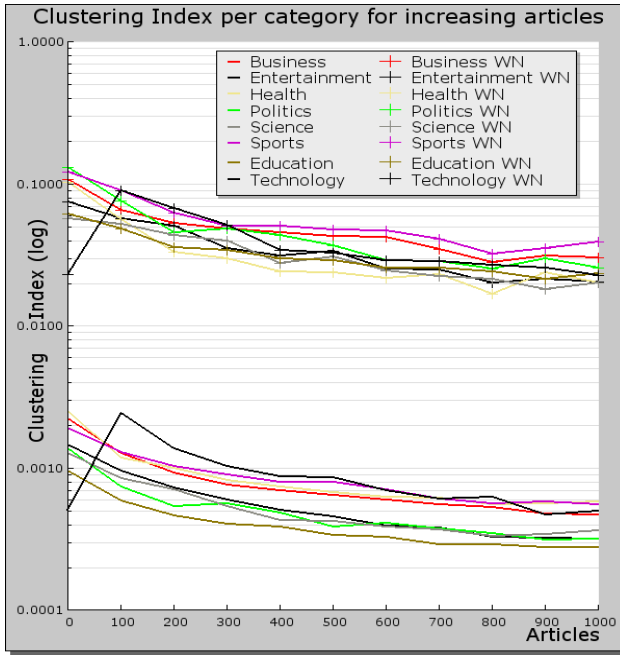


Fig. 3. Evaluating W-kmeans clustering over articles belong to various categories

efficiency of each clustering method, we used the evaluative criterion of Clustering Index (CI) as explained in [13], defined as: $CI = \bar{\sigma}^2 / (\bar{\sigma} + \bar{\delta})$, where $\bar{\sigma}$ is the average intra-cluster similarity and $\bar{\delta}$ is the average inter-cluster similarity. For our first experimentation set, we run both of the kmeans and W-kmeans algorithms on the dataset and observed the CI scores over varying categories, number of articles and number of clusters. For the results presented in Fig. 3, the top set of lines gives the CI for the case of WordNet enriched executions of the kmeans algorithm, compared to the non enriched ones (bottom set). It is clearly depicted that the quality of the kmeans algorithm has improved significantly when applied in our data set regardless the number of articles or the category they belong.

This provides a confirmation for the initial hypothesis that using outside features from the English language, apart from only textual - extracted features can be particularly useful. Another observation is that as the number of articles increase, the CI difference of W-kmeans compared to kmeans gets wider. We believe that this is because of the fact that while our experimentation data set grows larger the probability of hypernyms occurring also increases. Therefore, our clustering approach has a better chance of selecting clusters with improved connectivity. Fig. 4 presents the CI results for a variety of cluster numbers as averaged over all the categories (i.e. over all 8000 articles). The improvement, as before, is more than ten times over CI scores obtained with normal k-means (logarithmic scales in both Fig. 3 and 4). We also pinpointed that for the case of 50 clusters, the results are slightly improved over the rest of the cases which can be interpreted as a viable indication of the actual number of clusters our data set seems to have.

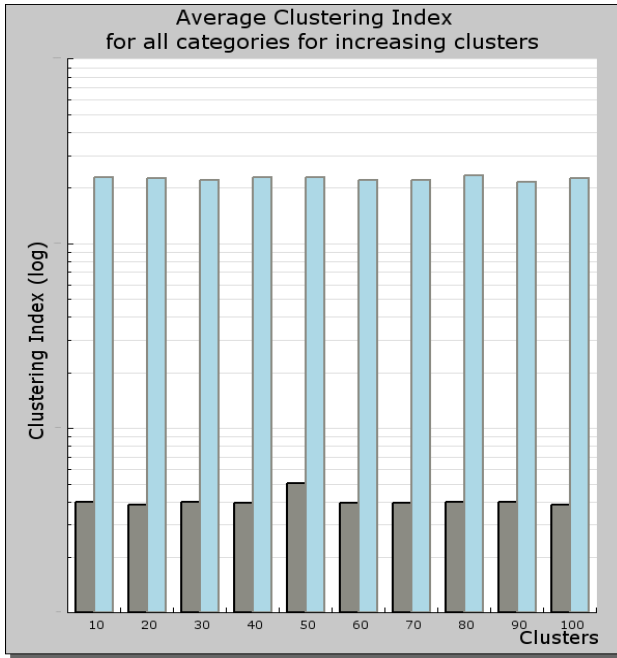


Fig. 4. Averaging Clustering Index over categories for various cluster numbers

For our second experimentation set, we evaluated the labeling results of the proposed algorithm. In order to do so, we applied W-kmeans over our data set using a total number of 8 clusters. Since the articles of the data set are pre-categorized to one of the 8 categories used, we compared the resulting cluster labels to aggregate lists created for each category containing: a) the 10 most frequent keywords of each category b) the category name itself. Labels getting ‘close’ (i.e. synonyms or derivatives) to the contents of the aggregate list are considered as representative ones. In addition, the category’s aggregate list to which a cluster has the most labels belonging to is accepted as the representative category for this cluster. We evaluated the accuracy of the labeling process using the precision of the suggested cluster labels against the aggregate list of the category that the respective cluster belongs to. Precision for labeling i and its belonging category j is defined as:

$$precision (label_i, category_j) = avg_rank (i, j) \cdot \frac{a}{a + b} \tag{4}$$

where $avg_rank(i, j)$ is the average rank that labeling i has in the aggregate list of category j , a is the number of terms labeling i has for category j and b is the number of terms that labeling i has but are not in the j^{th} ’s category aggregate list. The precision results per category presented in Table 1 show an overall precision rate of 75% for our labeling approach which would have been even better if the ‘technology’ and ‘science’ categories were not so closely related to each other.

Table 1. Precision results for cluster labeling over various categories using W-kmeans

Category	W-kmeans Precision
Business	85%
Entertainment	78%
Health	90%
Politics	88%
Science	65%

5 Conclusion

We have presented a novel algorithmic approach towards enhancing the kmeans algorithm using knowledge from an external database, WordNet, in a twofold manner. W-kmeans firstly enriches the clustering process itself by utilizing hypernyms and secondly, generates useful labels for the resulting clusters. We have measured a 10-times improvement over the standard kmeans algorithm in terms of high intra-cluster similarity and low inter-cluster similarity. Furthermore, the resulting labels are with high precision the correct ones as compared with their category tagging counterparts. As a future enhancement, we will be evaluating W-kmeans with regards to time efficiency using more clustering algorithms and larger document sets.

References

- [1] Zhao., Y., Karypi, G.: Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. *Machine Learning* 55(3), 311–331 (2004)
- [2] Yanjun, L., Soon, C.: Parallel bisecting k-means with prediction clustering algorithm. *The Journal of Supercomputing* 39, 19–37 (2007)
- [3] Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proc. of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035 (2007)
- [4] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., Miliotis, E.: Semantic similarity methods in wordNet and their application to information retrieval on the web. In: Workshop On Web Information And Data Management, Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 10–16 (2005)
- [5] Chen, C.-L., Frank, S., Tseng, C., Liang, T.: An integration of fuzzy association rules and wordNet for document clustering. In: Theeramunkong, T., Kijssirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 147–159. Springer, Heidelberg (2009)
- [6] Carmel, D., Roitman, H., Zwerdling, N.: Enhancing cluster labeling using wikipedia. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 139–146 (2009)
- [7] Sedding, J., Kazakov, D.: WordNet-based text document clustering. In: Proc. of COLING-Workshop on Robust Methods in Analysis of Natural Language Data (2004)
- [8] Treeratpituk, P., Callan, J.: Automatically labeling hierarchical clusters. In: Proceedings of the 2006 international conference on Digital government research, San Diego, California, May 21-24 (2006)

- [9] Tseng, Y.H.: Generic title labeling for clustered documents. In: *Expert Systems With Applications*, vol. 37(3), pp. 2247–2254. Elsevier, Amsterdam (2009)
- [10] Bouras, C., Pouloupoulos, V., Tsogkas, V.: PeRSSonal’s core functionality evaluation: Enhancing text labeling through personalized summaries. *Data and Knowledge Engineering Journal*, Elsevier Science 64(1), 330–345 (2008)
- [11] Bouras, C., Tsogkas, V.: Improving text summarization using noun retrieval techniques. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) *KES 2008, Part II. LNCS (LNAI)*, vol. 5178, pp. 593–600. Springer, Heidelberg (2008)
- [12] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Communications of the ACM* 18(11), 613–620 (1975)
- [13] Taeho, J., Malrey, L.: The Evaluation Measure of Text Clustering for the Variable Number of Clusters. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) *ISSN 2007. LNCS*, vol. 4492, pp. 871–879. Springer, Heidelberg (2007)

An Efficient Mechanism for Stemming and Tagging: The Case of Greek Language

Giorgos Adam, Konstantinos Asimakis, Christos Bouras, and Vassilis Pouloupoulos

Research Academic Computer Technology Institute, Greece
and

Computer Engineering and Informatics Department, University of Patras, Greece
{adam, asimakis, bouras, poulop}@ceid.upatras.gr

Abstract. In an era that, searching the WWW for information becomes a tedious task, it is obvious that mainly search engines and other data mining mechanisms need to be enhanced with characteristics such as NLP in order to better analyze and recognize user queries and fetch data. We present an efficient mechanism for stemming and tagging for the Greek language. Our system is constructed in such a way that can be easily adapted to any existing system and support it with recognition and analysis of Greek words. We examine the accuracy of the system and its ability to support personal a medium constructed for offering meta-portal news services to internet users. We present experimental evaluation of the system compared to already existing stemmers and taggers of the Greek language and we prove the higher efficiency and quality of results of our system.

Keywords: Greek stemmer, natural language processing, information tagging, context, keyword extraction.

1 Introduction

As the Internet expands dramatically more and more people are concerned about the difficulty in searching information on the WWW. The task of searching for useful information involves a large amount of procedures that are transparent to the end-user, who just needs to locate the information needed each time. One of the many procedures that are executed during the data processing is the stemming of words and root extraction. It remains an open issue of the academia whether an NLP procedure of a specific language should contain a stemmer or a lemmatizer. It seems that the solution is not as simple as it seems to be from a first analysis. In languages with large morphological linguistic variety a lemmatizer is more accurate while a stemmer may categorize words with extremely different meaning.

Another huge problem that exists, and is produced by the expansion of the web, is the fact that nowadays almost everybody is searching for information to their own language and not in English as it used to be some years ago. This implies that the data that exist on the Internet are written in many different languages. The Greek language, like many others, uses extensive inflection and therefore semantic relations between words cannot be detected unless those words are stemmed. Stemming Greek words is much harder than stemming words of other European languages because of the large

number of possible suffixes, many of which cannot be properly separated from the word stem without knowing the grammatical type of the word. Moreover, stemming of words with different meaning may lead to the same stem.

Many steps have been made towards the issue of construction of stemmers for many languages. The first stemmer that was ever presented was the effort made by Lovins. The Lovins stemming algorithm [1] was first presented in 1968 by Julie Beth Lovins. It is a single pass, context sensitive stemmer, which removes endings based on the longest-match principle. The stemmer was the first to be published and was extremely well developed considering the date of its release and has been the main influence on a large amount of the future work in the area. One of the most important efforts that became the basis of many other stemmers is Porter Stemmer [2]. The Porter Stemmer is a conflation Stemmer developed by Martin Porter at the University of Cambridge in 1980. The Stemmer is based on the idea that the suffixes in the English language are mostly made up of a combination of smaller and simpler suffixes. This Stemmer is a linear step Stemmer. Specifically it has five steps applying rules within each step. Within each step, if a suffix rule matched to a word, then the conditions attached to that rule are tested on what would be the resulting stem, if that suffix was removed, in the way defined by the rule. For example such a condition may be, the number of vowel characters, which are followed by a consonant character in the stem (Measure), must be greater than one for the rule to be applied. Finally, another commonly used stemmer is Paice-Husk stemmer [3] implemented at the University of Lancaster. The stemmer is a conflation based iterative stemmer. The stemmer, although remaining efficient and easily implemented, is known to be very strong and aggressive. A very common usage of the stemmers is the search engines and their query analysis and enhancement subsystems. The stemming procedure is applied to the input query in order to increase the recall rate, when it is necessary. It is important to note that the major search engines (such as Google¹ or Yahoo²) utilize an analyzer for the languages that they support. Despite the fact that such service is not referenced to, as it is transparent to the end user, it is obvious that it exists from the results presented to the end user. A Greek search engine that was constructed recently does utilize a stemmer; though, no information is publicly available except for a reference to the complete work that was done for the specific search engine [11].

The porter stemmer is the root of many stemmers that were produced for many European languages. The similarities on the construction of words in many European languages make it easy to construct stemmers starting from the basic Porter Stemmer. The basic categories to which we can possibly put a stemmer are three: Stemmers that are based (a) on dictionaries, (b) on algorithms, or (c) on hybrid algorithms that are based on both (a) and (b). We are putting the focus on the stemmers constructed for the Greek language. The first two are the TZK algorithm [4] by Kalamboukis and Nikolaidis in 1995 and the Automated Morphological Processor (AMP) [5] by Tambouratzis and Carayanis in 2001. The latest system presented was the work of George Ntais [6] while Spyridon Saroukos [7] has presented in 2008 an enhanced version of Ntais' stemmer.

¹ <http://www.google.com>. Google Search Engine

² <http://www.yahoo.com>. Yahoo! Search Engine

The first suffix stripping algorithm for the Greek language was presented in 1995 from Kalamboukis and Niloaidis. Design for information retrieval from Greek corpora the algorithm deals with inflections of the Greek language. They make extensive usage of suffix lists for inflection while in parallel; at a second level they remove derivational suffixes. Only 6 years later Tambouratzis and Carayannis presented a system for automated morphological categorization (AMP). Their work is based on extraction of stems through matching and masking with an initial set of correct stems and suffixes. A basic assumption for the usage of the system is that each word consists of a stem and a suffix. Every other word cannot be stemmed by their system. Nevertheless, because of the fact that in both the aforementioned effort no code or corpus was ever revealed it is impossible to make a direct comparison of a new stemmer.

The latest stemming algorithm developed for the Greek language is presented by George Ntais in 2006. The algorithm is based on Porter stemmer and an online implementation is available on the web. According to the author, the algorithm can handle a large number of suffixes of the Greek language, clearly outperforming the first two algorithms presented and aforementioned. A clear disadvantage of the system is the inability to manipulate with any other form of a word apart from the word in capital letters. Due to the fact that the morphology of the Greek language implies that a minor change to the accent mark of a word can change its meaning Ntais algorithm does have some weak points.

We propose a novel effort towards stemming for the Greek language. Our system is a hybrid system that is able to apply stemming on branches of texts without any limitation on the way text is written. The novelty of the system compared to the older efforts for the creation of a Greek stemmer is the fact that we are enhancing the stemming procedure by word tagging techniques. Tagging is as hard as stemming unless words are viewed in context. Greek syntax may be almost free but still some few useful rules apply which can be used to increase the accuracy of the grammatical tagging. Additionally, proper tagging enables us to ignore words based on their grammatical type instead of ignoring words using word length thresholds or stop-words lists.

The rest of the paper is structured as follows. In the next section we present our motivation: peRSSonal, a meta-portal; while in section 3 we present the architecture of our system. In section 4 we describe the algorithmic aspects of our system while the following section contains the results of the experimental evaluation of our system as well as a comparison to already existing mechanisms. The paper concludes with remarks on our system and what feature enhancements we consider interesting and perhaps useful.

2 Motivation: peRSSonal Meta-portal

The prevalent idea that motivated the peRSSonal [10] mechanism has its roots in the change of our web life, which has turned every single corner of the Web in a potential source of valuable information. However, benefit never comes without cost: locating the desired piece of information among irrelevant data has become a difficult and tedious task, even for the more experienced users, as everyone has different special needs from the medium that is called World Wide Web. Major search engines

(e.g. Google¹, Yahoo²) are trying to refine search. In parallel, attempts for the creation of WWW ontologies (e.g. DMOZ) look forward to resolve these issues. Our experience made us realize that among these facts lies another huge problem that is of primary concern for millions of users all over the world. As the Internet expands and acts as a form of “digital newspaper”, more and more people come to realize that they are able to read and stay informed by articles in real time. This leads to a problematic situation where users have to visit a big number of news portals to read the news from the categories they are concerned. The problem is partially answered by RSS feeds and personalized micro sites. In the first case, the user does not have to browse to every single website but, as a forfeit, they must undergo data filtering due to the fact that there is no specialization of these feeds on his needs. In the second case, focus on each user’s preferences can be guaranteed but he or she still has to visit each one or several of these sites in order to track down all the information on a specific subject.

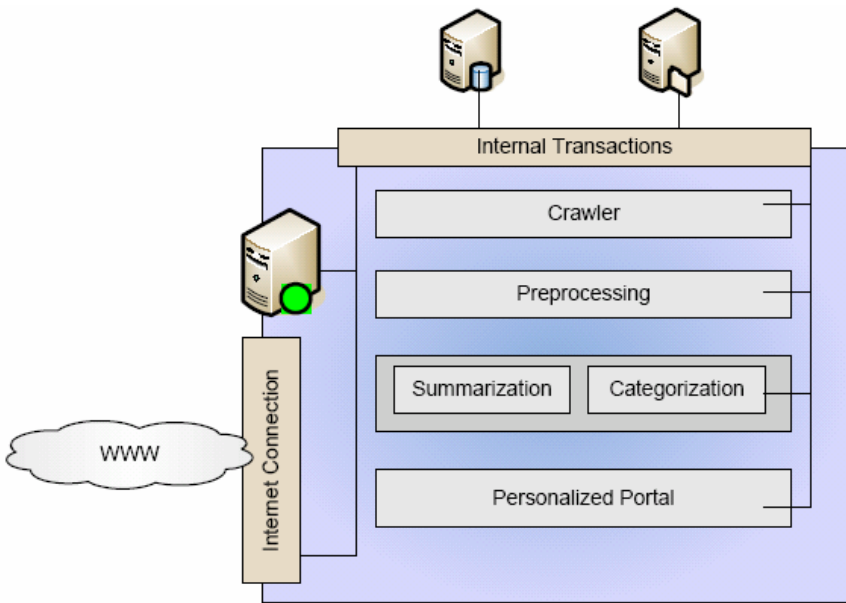


Fig. 1. peRSSonal architecture

3 Architecture and Algorithm Analysis

The architecture of the proposed system is similar to a context sensitive stemmer that applies a suffix stripping algorithm to produce the stem. It takes as input complete sentences and takes advantage of the POS tagging process in order to limit the possible suffixes that are going to be removed. The suffix stripping algorithm is rule-based and utilizes a table with possible suffixes for every part of speech tag. The output is a list of the input words, their stem and their grammatical type.

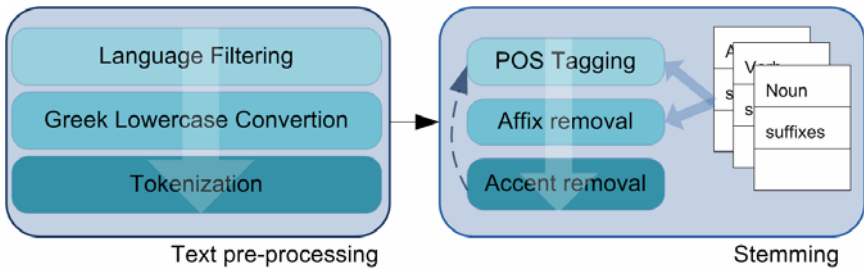


Fig. 2. G.I.C.S. architecture

The stemmer was initially designed in order to support the Greek language in the peRSSonal system. The procedure of this system is: (a) capture pages from the www and extract the useful text, (b) parse the extracted text, (c) summarize and categorize the text, and (d) present the personalized results to the end user. The capture is done using a crawler that takes as input a list of RSS feeds from news portals. The crawler extracts the articles and stores the html pages without any other element of the web page, like CSS and JavaScript files. At the next step it triggers a useful text extractor mechanism in order to be able to get the real text of the article and to omit undesired content (advertisements, etc.). The useful text can be defined as the title and the main body of the article.

The stemmer is invoked at the second analysis level, parsing the whole useful text and extracting the keywords (stems) of the article. This level receives as input XML files that include the title and body of articles. Its main scope is to apply preprocessing algorithms on this text and provide as output keywords, their location into the text and the frequency of their appearance in the text. These results are necessary in order to proceed to the third analysis level. The core of peRSSonal mechanism is located in the third analysis level, where the summarization and categorization subsystems are located. Their main scope is to characterize the article with a label (category) and produce a summary of it. All these results are then presented back to the end users of our personalized portal. The portal can feed each user only with articles that the user may find them interesting, according to his/her dynamically created profile.

The proposed context sensitive stemmer has a time and space complexity of $O(n)$ and it can be considered as the combination of two different procedures: stemming and tagging. Although the main scope of the system is to provide stemming functionality, we incorporated a grammatical tagger in the algorithm for two reasons. First of all, viewing words in the context of the grammatical types of nearby words helps resolve some ambiguous cases where different suffixes could be removed. Secondly, because discarding words belonging to certain grammatical categories (that usually carry no significant meaning for our purposes, e.g. pronouns) produces better results than using stop-word lists or discarding words based on their size.

Firstly, the algorithm accepts an array encoded in ISO-8859-7 and discards all characters except for English and Greek letters. The remaining characters are converted to lower case and any dieresis diacritics are removed. Accenting is left intact since it provides extra information that can be used in tagging and stemming. The last step of the preprocessing is the tokenization process which creates a linked list of structures, each of which hold a word along with space for storing the tag, the stem,

and information used internally by the algorithm. After the preprocessing step, the tagging process starts which consists of two rounds.

In the first round each word is tested for fitness in different grammatical categories, using tables with known suffixes and words of the Greek language [8]. Some of the categories cause the testing process to stop if the word is found fit for them while other let it continue. In cases where the word fits in more than one category, the grammatical type of the previous word is used to determine which category is chosen. While the words are matched to categories, information about the size of the affixes is stored in the linked structure list. During the second round the same steps are followed for the words that haven't been recognized during the first round. This time the accenting is ignored as a last resort, since in many cases words are either mis-accented or unaccented.

Next, the stemming procedure starts. During that, words are trimmed based on information stored in the previous step. For verbs, additional conversions are applied in an attempt to make the stems of different tenses match each other.

Finally, after the stemming process is finished. The algorithm will return the linked structure list, which now contains the stems and the tags, to the caller.

4 Experimental Evaluation

In contrast to other stemming algorithms, G.I.C.S. is not trying to extract the grammatically correct stem of words, although in most cases it does so. Instead we consider a stem correct as long as grammatically similar words are assigned to the same stem.

Specifically, for verbs we consider a stem correct if it is matching that of the first person singular of the same verb in present tense. For nouns/adjectives and pronouns we consider a stem correct if it is matching that of the masculine singular of the nominative case of the same noun. If there is no masculine singular then any singular of the nominative case is used instead. Additionally, although not necessary for considering the stemming correct, we tried to assign the noun to the same stem of the verb from which it derives (if any). For verb-derived proverbs, we consider the stem correct if it matches that of the verb which the proverb derives from. For other type of words stemming is trivial and not of any importance so we only rated the tagging.

We evaluated the algorithm using two different sets of text. The first set was composed of news articles like those that will be used by perSSonal, which totaled in around one thousand words. The second similarly sized set was composed of both formal and informal emails. The execution time of the mechanism for these two sets was insignificant, as the average stemming speed has been estimated, through extra testing, to be eabout 163 thousand characters per second. We first evaluated G.I.C.S tagging precision on all the words in the dataset. Finally we evaluated the stemming precision of both G.I.C.S and Ntais stemmer ignoring all words except nouns, adjectives, verbs and proverbs. Words in other categories are known, though our experiments, to carry no significant meaning. Additionally these words appear frequently and their stem can be easily extracted so they affect the results positively, making the useful precision of the stemmer harder to be measured.

98.6% of the words of the first set were tagged correctly by G.I.C.S.; while 96.7% of the useful words were stemmed correctly (Fig.3) 12.5% of the erroneous stems

were the result of over-stemming, meaning that the algorithm removed more letters than it should. The rest 87.5% of the words were either under-stemmed or they were irregular and the stemmer wasn't able to convert the stem correctly to match the desired one, based on our criteria. Specifically, 75% of the errors were made when an irregular verb was stemmed. (Fig.5) By comparison, using the same words and the same criteria, the Ntais stemmer stemmed correctly 91.1% of the words of the first set (the articles).

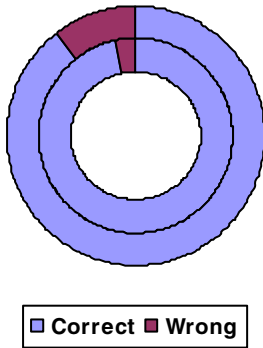


Fig. 3. G.I.C.S. and Ntais stemmer comparison. G.I.C.S. stemming precision on the inside. Ntais stemming precision on the outside. (articles dataset)

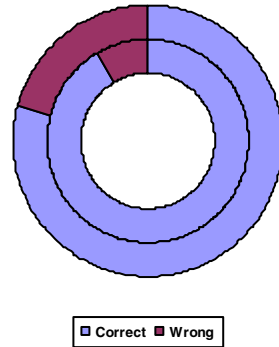


Fig. 4. G.I.C.S. and Ntais stemmer comparison. G.I.C.S. stemming precision on the inside. Ntais stemming precision on the outside.(emails dataset)

Regarding the second set which included emails, our algorithm achieved 96.7% correctly stemmed words while the tagging was successful for 91.7% of the useful words (Fig.4). Of the errors, 6.65% was because of under-stemmed words while 93.3% was of over-stemmed words including the irregular verbs. 80% of the errors were made during the stemming of irregular verbs (Fig. 5). By comparison, the Ntais stemmer achieved 88.15% of correctly stemmed words.

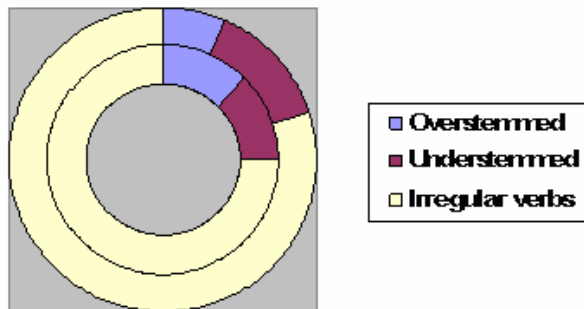


Fig. 5. G.I.C.S. types of error. Articles dataset (inside), Emails dataset (outside)

At the next experiment we utilize three different metrics [9] in order to compare the proposed stemming algorithm to Ntais implementation. The first metric (index compression factor) indicates the index reduction that can be achieved through stemming.

$$ICS = \frac{n - s}{n} . \quad (1)$$

Where n is the number of words in the corpus and s is the number of produced stems. The other two metric are the mean and median Hamming distance. The Hamming distance is defined as the number of characters in the two strings that are different at the same position. For unequal length strings, the difference of their length is added. The dataset for this experiment consists of 10981 words that have been retrieved from major Greek news portals. The results are shown in table 1.

Table 1. Comparison of G.I.C.S. and Ntais stemming algorithms

	Ntais stemmer	G.I.C.S.
Index compression factor	76.8%	80.9%
Mean modified Hamming Distance	1.95	2.73
Median modified Hamming Distance	2	2

5 Conclusion and Future Work

In this paper we have presented an efficient stemmer and tagger for the Greek language. The stemmer and tagger is created in order to support an existing meta-portal peRSSonal which intends to present news articles collected from the WWW to the users in a personalized manner. The stemming is performed using a rule-based algorithm that removes suffixes. It bases most of its procedures on the fact that all the words of the text are grammatically tagged by a tagger which works as part of the stemmer. The mechanism was compared to the latest known Greek stemmer implementation and the experimental evaluation showed that it can achieve higher stemming precision. Despite the fact that we just wanted to create groups of similar words as input for the peRSSonal system we managed to construct a Greek stemmer that can possibly achieve very high scores of stemming on the text that we have as input. Our input texts are news articles which are usually short in length and they use compact language with many forms of the same word used across the text.

For the future we would like to enhance the current system in order to improve the tagging process using the punctuation information. Moreover, when processing a specific word, the mechanism could take into account the tags of more words that it currently does (e.g. the tags of all the words in the current sentence). This information will lead to better POS tagging and thus, better stemming precision.

References

1. Lovins, J.B.: Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11, 22–31 (1968)
2. Porter, M.F.: An algorithm for suffix stripping. *Program; automated library and information systems* 14(3), 130–137 (1980)

3. Paice, D.: Another stemmer. *ACM SIGIR Forum* 24(3), 56–61 (1990)
4. Kalamboukis, T.Z.: Suffix stripping with modern Greek. *Program* 29(3), 313–321 (1995)
5. Tambouratzis, G., Carayannis, C.: Automatic corpora-based stemming in Greek. *Literacy and Linguistic Computing* 16, 445–466 (2001)
6. Ntais, G.: Development of a stemmer for the Greek language, MSc Thesis, Stockholm University (2006)
7. Saroukos, S.: Enhancing a Greek Language Stemmer, MSc Thesis, University of Tampere (2008)
8. Triantafillidis, M.: *Modern Greek Grammar (Dimotiki)* (in Greek). Reprint with corrections 1978. Institute of Modern Greek Studies, Thessaloniki (1941)
9. Frakes, W.B., Fox, C.J.: Strength and similarity of affix removal stemming algorithms. *SIGIR Forum* 37(1), 26–30 (2003)
10. Bouras, C., Pouloupoulos, V., Tsogkas, V.: PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries. *Data and Knowledge Engineering Journal* 64(1), 330–345 (2008)
11. Papadakos, P., Vasiliadis, G., Theoharis, Y., Armenatzoglou, N., Kopidaki, S., Marketakis, Y., Daskalakis, M., Karamaroudis, K., Linardakis, G., Makrydakis, G., Papathanasiou, V., Sardis, L., Tsialiamanis, P., Troullinou, G., Vandikas, K., Velegrakis, D., Tzitzikas, Y.: *The Anatomy of Mitos Web Search Engine*. CoRR, Information Retrieval, abs/0803.2220. CoRR Technical Report (March 2008)

Co-clustering Analysis of Weblogs Using Bipartite Spectral Projection Approach

Guandong Xu^{1,2}, Yu Zong², Peter Dolog¹, and Yanchun Zhang²

¹ IWIS - Intelligent Web and Information Systems, Aalborg University, Computer Science Department Selma Lagerlofs Vej 300 DK-9220 Aalborg, Denmark

² Center for Applied Informatics, School of Engineering & Science, Victoria University, P.O. Box 14428, Vic 8001, Australia

Abstract. Web clustering is an approach for aggregating Web objects into various groups according to underlying relationships among them. Finding co-clusters of Web objects is an interesting topic in the context of Web usage mining, which is able to capture the underlying user navigational interest and content preference simultaneously. In this paper we will present an algorithm using bipartite spectral clustering to co-cluster Web users and pages. The usage data of users visiting Web sites is modeled as a bipartite graph and the spectral clustering is then applied to the graph representation of usage data. The proposed approach is evaluated by experiments performed on real datasets, and the impact of using various clustering algorithms is also investigated. Experimental results have demonstrated the employed method can effectively reveal the subset aggregates of Web users and pages which are closely related.

1 Introduction

Recently, the Web becomes an important and popular platform for distributing and acquiring information and knowledge due to its rapid evolution of Web technology and the influx of data sources available over the Internet in last decades [1]. Web clustering is emerging as an effective and efficient approach to organize the data circulated over the Web into groups/collections in order to re-structure the data into more meaningful blocks and to facilitate information retrieval and representation, and at the same time to meet user preferences [2,3,4].

Web clustering could be performed on either Web pages or user sessions in the context of Web usage mining. Web page clustering is one of popular topics in Web clustering, which aims to discover Web page groups sharing similar functionality or semantics. For example, [5] proposed a technique LSH (Local Sensitive Hash) for clustering the entire Web, concentrating on the scalability of clustering. Snippet-based clustering is well studied in [6]. [3] reported using a hierarchical monothetic document clustering for summarizing the search results. [7] proposed a Web page clustering algorithm based on measuring page similarity in terms of correlation.

In addition to Web page clustering, Web usage clustering is proposed to discover Web user behavior patterns and associations between Web pages and users

from the perspective of Web user. The motivation behind is that each user visiting session could be expressed by a pageview vector with weights, and the clustering on user sessions leads to finding the user session aggregates, which could be viewed as user access patterns. [8] combined user transaction and pageview clustering techniques, which was to employ the traditional k-means clustering algorithm to characterize user access patterns for Web personalization based on mining Web usage data. In [9] Xu et al attempted to discover user access patterns and Web page segments from Web log files by utilizing a so-called Probabilistic Semantic Latent Analysis (PLSA) model.

The clustering algorithms described above are mainly manipulated on one dimension/attribute of the Web usage data only, i.e. user or page solely, rather than taking into account the correlation between Web users and pages. However, in most cases, the Web object clusters do often exist in the forms of co-occurrence of pages and users - the users from the same group are particularly interested in one subset of Web pages. For example, in the context of customer behavior analysis in e-commerce, this observation could correspond to the phenomenon that one specific group of customers show strong interest to one particular category of goods. In this scenario, Web co-clustering is probably an effective means to address the mentioned challenge. The study of co-clustering is firstly proposed to deal with co-clustering of documents and words in digital library [10]. And it has been widely utilized in many studies which involved in multiple attribute analysis, such as social tagging system [11] and genetic representation [12] etc.

In this paper, we aim to address finding Web object co-clusters by employing a bipartite spectral projection approach. The main strengths of the proposed approach is the capability of projecting the original usage matrix (i.e. row and column vectors) simultaneously into a single unified spectral space in lower dimension based on graph partition algorithm. Experimental study is carried out to validate the clustering performance.

The main contribution of this paper are as follows:

- The bipartite spectral clustering algorithm is employed into Web usage mining for co-clustering user sessions and pageviews. To our best knowledge, finding the session and page co-clusters is rarely studied in Web usage mining although it achieved great success in text mining.
- We conduct experiments to evaluate the effectiveness of proposed technique.
- We also experimentally investigate the selection of co-cluster number and the impact of different clustering algorithms on the performance of co-clustering.

The rest of the paper is organized as follows: Section 2 presents the problem formulation and the introduced mathematical models. In section 3, we describe the details of bipartite spectral clustering algorithm. Section 4 discusses the experiment design in terms of datasets and evaluation measures. And experimental results and comparisons are presented in this section as well. Eventually we conclude the paper and indicate possible future research directions in section 5.

2 Problem Formulation

2.1 Web Usage Data

Web usage data is mainly sourced from Web log files, which include Web server access logs [8]. The log data collected at Web access reflects the navigational behavior knowledge of users in terms of access pattern. Web page is a basic unit of Web site organization, which contains a number of meaningful units serving for the main functionality of the page. User session is a sequence of Web pages clicked by a single user during a specific period. In general, the exhibited user access interests may be reflected by the varying degrees of visits on different Web pages during one user session. Thus, we can represent a user session as a weighted vector of pageviews visited by the user during a particular period. In this manner, Web usage data is modeled as a Web page collection and a user session collection as well as a session-pageview matrix. In this paper, we use the following notations to model the co-occurrence activities of usage data:

- $\mathcal{S} = \{s_i, i = 1, \dots, m\}$: a set of m user sessions.
- $\mathcal{P} = \{p_j, j = 1, \dots, n\}$: a set of n Web pages.
- For each user, a user session is represented as a vector of visited pages with corresponding weights: $s_i = \{a_{ij}, j = 1, \dots, n\}$, where a_{ij} denotes the weight for page p_j visited in s_i user session. The corresponding weight is usually determined by the number of hit or the amount time spent on the specific page.
- $A_{m \times n} = \{a_{ij}, i = 1, \dots, m, j = 1, \dots, n\}$: the ultimate usage data in the form of a weight matrix with a dimensionality of $m \times n$.

Generally, the element in the session-page matrix, a_{ij} , associated with the page p_j in the user session s_i , should be normalized across pages in same user session in order to eliminate the influence caused by the amount difference of visiting time durations or hit numbers and to capture the relative significance of a page within one user session with respect to others pages accessed by the same user.

2.2 Bipartite Graph Model

As the nature of Web usage data is a reflection of a set of Web users visiting a number of Web pages, it is intuitive to introduce a graph model to represent the visiting relationship between them. In particular, here we use the Bipartite Graph Model to illustrate it.

Definition 1: Given a graph $G = (\mathcal{V}, E)$, where \mathcal{V} is a set of vertices $\mathcal{V} = \{v_1, \dots, v_n\}$ and E is a set of edges $\{i, j\}$ with edge weight E_{ij} , the adjacency matrix M of the graph G is defined by

$$M_{ij} = \begin{cases} E_{ij}, & \text{if there is an edge (i,j)} \\ 0, & \text{otherwise} \end{cases}$$

Definition 2 (Cut of Graph): Given a partition of the vertex set \mathcal{V} into multiple subsets $\mathcal{V}_1, \dots, \mathcal{V}_k$, the cut of the graph is the sum of edge weights whose vertices are assigned to two different subsets of vertices:

$$\text{cut}(\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_k) = \sum_{i \in \mathcal{V}_i, j \in \mathcal{V}_j} M_{ij}$$

As discussed above, the usage data is indeed demonstrated by the visits of Web users on various Web pages. In this case, there are NO edges between user sessions or between Web pages, instead there are only edges between user sessions and Web pages. Thus it is essential that the bipartite graph model is an appropriate graphic representation to characterize their mutual relationships.

Definition 3 (Bipartite Graph Representation): Consider a graph $G = (\mathcal{S}, \mathcal{P}; E)$ consisting of a set of vertices $\mathcal{V} = \{s_i, p_j : s_i \in \mathcal{S}, p_j \in \mathcal{P}; i = 1, \dots, m, j = 1, \dots, n\}$, where \mathcal{S} and \mathcal{P} are the user session collection and Web page collection, respectively, and a set of edges $\{s_i, p_j\}$ each with its weight a_{ij} , where $s_i \in \mathcal{S}$ and $p_j \in \mathcal{P}$, the links between user sessions and Web pages represent the visits of users on specific Web pages, whose weights indicate the visit preference or significance on respective pages.

Furthermore, given the $m \times n$ session-by-pageview matrix A such that a_{ij} equals to the edge weight E_{ij} . It is easy to formulate the adjacency matrix M of the bipartite graph G as

$$M = \begin{bmatrix} 0 & A \\ A^t & 0 \end{bmatrix}$$

In this manner, the first m rows in the reconstructed matrix M denote the co-occurrence of user sessions while the last n rows index the Web pages. The element value of M is determined by click times or duration period. Because our ultimate goal is to extract subsets of user sessions and Web pageviews to construct a variety of co-clusters of them such that they possess the closer cohesion within the same cluster but the stronger disjointness from other clusters, it is necessary to model the user session and Web page vectors in a same single unified space. In the coming section, we will discuss how to perform co-clustering on them.

2.3 An Example of Usage Bipartite Graph

To better illustrate the bipartite graphic representation of Web usage data, here we present a very simple example. Consider a bipartite graph $G = (\mathcal{S}, \mathcal{P}; E)$ depicted in Figure 1, where the set of user sessions $\mathcal{S} = \{s_1, s_2, s_3\}$, the set of Web pages $\mathcal{P} = \{p_1, p_2, p_3, p_4\}$ and the set of edges connecting user sessions and Web pages $E = \{a_{ij} : s_i \in \mathcal{S}, p_j \in \mathcal{P}\}$. Each of edges indicates the relation between the connected user session s_i and Web page p_j , and its weight is set to be the hit number of the page in the user session. According to the definition 2, the cut of the bipartite graph into two subsets is illustrated in Figure 2, in which the dash line of s_3 and p_1 indicates the cut degree between them.

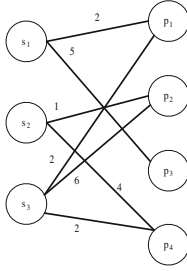


Fig. 1. Bipartite graph representation of usage data

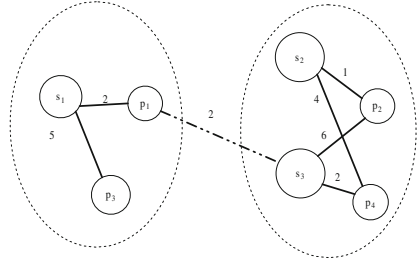


Fig. 2. Cut of Bipartite graph

3 Clustering User Sessions and Web Pages

3.1 Dual Subset Clustering

Upon the introduction of bipartite graph expression of usage data, we aim to perform the co-clustering of user sessions and Web pageviews. Essentially, the co-clustering could be considered as a problem of Dual Subset Clustering (*DSC*) of user sessions and Web pageviews, that is, the session clustering results in the pageview clustering while the pageview clustering reinforces the session clustering.

Consider a set of disjoint session clusters $\mathcal{S}_1, \dots, \mathcal{S}_k$ and their associated set of pageview clusters $\mathcal{P}_1, \dots, \mathcal{P}_k$. Since the usage data is modeled as a bipartite graph of sessions and pageviews, there are no connections between sessions or between pageviews themselves. Thus for a co-clustering operation it is intuitive to assign respective sessions and pageviews into various clusters $\mathcal{C}_j = (\mathcal{S}_j, \mathcal{P}_j), \mathcal{S}_j \subset \mathcal{S}, \mathcal{P}_j \subset \mathcal{P}$ such that the sessions have visited the pageviews that are in the same cluster (or the pageviews are visited by the sessions that are in same cluster) but the sessions have less visited the pageviews that from the different clusters (or the pageviews are almost seldom visited by the sessions that from the other clusters). In other words, according to definition 2 of cut of graph, we can further formulate the dual subset clustering of sessions and pageviews as a solution of minimization of graph cut:

$$DSC(\mathcal{C}_1, \dots, \mathcal{C}_k) = \min_{\mathcal{V}_1, \dots, \mathcal{V}_k} cut(\mathcal{V}_1, \dots, \mathcal{V}_k)$$

where $\mathcal{V}_1, \dots, \mathcal{V}_k$ is a k -partitioning of the bipartite graph.

3.2 Spectral Co-clustering Algorithm

Modeling the usage data with a bipartite graph representation motivates us to employ the spectral graph theory to induce the co-clusters, which has been successfully in graph partitioning problem [10]. Spectral graph clustering indeed utilizes the eigenvalues of the adjacency matrix of usage data to map the original

inherent relationships of co-occurrence onto a new spectral space, on which the new session or pageview vector is projected. After the projection, the sessions and pageviews are simultaneously partitioned into disjoint clusters with minimum cut optimization.

According to the spectral graph theory proposed in [10], the k left and right singular vectors of the reformed matrix $RA = D_s^{-1/2}AD_p^{-1/2}$ present a best approximation to the projection of row and column vectors on the new spectral space. The D_s and D_p are the diagonal matrices of sessions and pageviews respectively, and are defined as:

$$D_s(i, i) = \sum_{j=1}^n a_{ij}, i = 1, \dots, m, D_p(j, j) = \sum_{i=1}^m a_{ij}, j = 1, \dots, n$$

Let L_s denote the $m \times k$ matrix of the left k singular vectors and R_p the $n \times k$ matrix of the right k singular vectors of RA . As our aim is to conduct a dual subset clustering on both session and pageview attributes, we create a new $(m + n) \times k$ matrix PV to reflect the projection of the row and column vectors on the new spectral space in lower dimension as:

$$PV = \begin{bmatrix} D_s^{-1/2}L_s \\ D_p^{-1/2}R_p \end{bmatrix}$$

The full steps of co-clustering algorithm is summarized in the below Algorithm 1.

Algorithm: Spectral Co-Clustering Algorithm

Input: The user session collection \mathcal{S} and pageview set \mathcal{P} , and the Web log file

Output: A set $C = \{C_1, \dots, C_k\}$ of k subsets of sessions and pageviews such that the cut of k -partitioning of the bipartite graph is minimized.

1. Construct the usage matrix A from the Web usage log, whose element is determined by the visit number or duration of one user on a specific page;
2. Calculate the two diagonal matrices D_s and D_p of A ;
3. Form a new matrix $NA = D_s^{-1/2}AD_p^{-1/2}$;
4. Perform SVD operation on NA , and obtain the left and right k singular vectors L_s and R_p , and combine the transformed row and column vectors to create a new projection matrix PV ;
5. Execute a clustering algorithm on PV and return the co-clusters of subsets of \mathcal{S} and \mathcal{P} , $C_j = (\mathcal{S}_j, \mathcal{P}_j)$.

Meanwhile, the selection of cluster number k is another concern in the context of clustering, which is commonly encountered. The selection of k value has a straight impact on the performance of clustering: the bigger number of k results in the over-separation of sessions and pageviews (e.g. too trivial clusters), while the smaller number of it prevents the data from being sufficiently partitioned. Thus it is necessary before performing the clustering to select an appropriate value of k to achieve a better clustering performance. We study this issue in the following section of experiments.

4 Experiments and Discussions

4.1 Datasets

We take two Web log files, which are public to access on the Internet for the purpose of research, as the usage data for experiments. These data sets are in either a raw data format or a pre-processed format. The first dataset used in this study is downloaded from KDDCUP (www.ecn.purdue.edu/kddcup/). After data preparation, we have setup a data set including 9308 user sessions and 69 pages. We refer this data set to “KDDCUP data”. In this data set, the entries in session-pageview matrix associated with the specific page in the given session are determined by the hit numbers of Web pages by a given user.

The second data set is from a university website log files and was made available by the authors of [13]. The data is based on a random collection of users visiting this site for a 2-week period during April of 2002. After data pre-processing, the filtered data contains 13745 sessions and 683 pages. This data file is expressed as a session-pageview matrix where each column is a page and each row is a session represented as a vector. The entry in the table corresponds to the amount of time (in seconds) spent on a page during a given session. For convenience, we refer this data as “CTI data”.

4.2 Clustering Evaluation and Experiment Design

In this section, the overall evaluation of the obtained clusters is examined in terms of clustering quality. Our aim is to validate how strong the correlation between the user sessions and pageviews within the co-cluster is. Here we assume a better co-clustering representing the fact that most of the user sessions visited the pages which are from the same co-cluster whereas the pages were largely clicked by the user sessions within the same subset of sessions and pageviews. In particular we use precision and recall measures to show whether the users with similar visiting preference are grouped together with the related Web pages within the same cluster. For each $\mathcal{C}_j = (S_j, P_j)$, its precision and recall measures are defined as below.

Definition 4: Given a cluster $\mathcal{C}_j = (S_j, P_j)$, its precision and recall measure are defined as the linear combination of the row (session) precision and column (pageview) precision, and the linear combination of the row and column recall.

$$precision(\mathcal{C}_j) = \alpha * precision(R_{\mathcal{C}_j}) + (1 - \alpha) * precision(C_{\mathcal{C}_j})$$

$$recall(\mathcal{C}_j) = \alpha * recall(R_{\mathcal{C}_j}) + (1 - \alpha) * recall(C_{\mathcal{C}_j})$$

where the row and column precision, and the row and column recall are defined as follows, respectively

$$precision(R_{\mathcal{C}_j}) = \frac{count(s_i):s_i \cap P_j \neq \emptyset, \forall s_i \in S_j}{|S_j|}, \quad precision(C_{\mathcal{C}_j}) = \frac{count(p_i):p_i \cap S_j \neq \emptyset, \forall p_i \in P_j}{|P_j|}$$

$$recall(R_{\mathcal{C}_j}) = \frac{count(s_i):s_i \cap P_j \neq \emptyset, \forall s_i \in S_j}{count(s_x):s_x \cap P_j \neq \emptyset, \forall s_x \in S}, \quad recall(C_{\mathcal{C}_j}) = \frac{count(p_i):p_i \cap S_j \neq \emptyset, \forall p_i \in P_j}{count(p_y):p_y \cap S_j \neq \emptyset, \forall p_y \in P}$$

where *count* denotes the count number of pageviews (or sessions) from one co-cluster appearing simultaneously in the same col-cluster of sessions (or pageviews). And α is the combination factor.

Because we consider that the sessions and pageviews have the equal contribution on the cluster conformation, in the experiment, the combination factor α is set as 0.5.

The overall precision and recall of the obtained clusters are defined as the average value of precision and recall of each cluster respectively. Given the precision of each cluster, $precision(C_j), j = 1, \dots, k$ and the recall of each cluster, $recall(C_j), j = 1, \dots, k$, the overall F-score of the obtained clustering is defined as: $F - score = \frac{2 * precision * recall}{precision + recall}$. It is clear that, the higher value of precision and recall denotes a better clustering being executed.

Experiments are conducted on a Pentium 4 machine with 2.66GHz CPU and 2GB RAM, running Windows XP. The algorithms are implemented by using Matlab 7.5.

4.3 Experimental Results and Discussions

Selecting an Optimal Cluster Number k. First we need to determine an appropriate number of clusters which results in a clustering with optimal performance. Here we use the assumption that an ideal clustering operation is in relation to a best F-score, and observe the change of F-score curve to select the optimal number. Figure 3 depicts the curve of F-score with different *k* settings for CTI dataset. From the figure, we can clearly see that when *k* changes from 2 to 30 the F-score climbs greatly until it reach the highest climax of *k*=5 then it decreases down monotonously and slowly to the value of 0.2. Thus we choose *k*=5 as the cluster number and use it in later experiments. Similarly, we select *k*=3 for KDD Cup dataset.

4.4 Evaluation Results

We conducted evaluations on these two datasets in terms of precision and recall measures. We run 30 times of the spectral co-clustering algorithm and presented

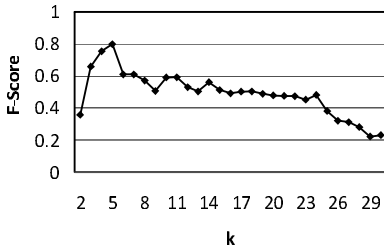


Fig. 3. Determining the optimal cluster number k of CTI dataset

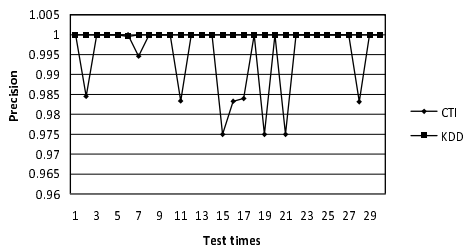


Fig. 4. Precision results of co-clustering

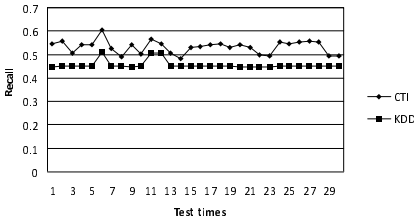


Fig. 5. Recall results of co-clustering

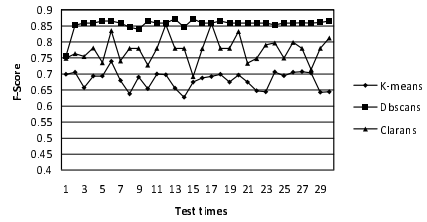


Fig. 6. F-score comparisons of various clustering

the results in Figure 4 and Figure 5. From these figures, it is seen that the precisions of co-clustering for CTI dataset are between 0.97 to 1 while the recall values range from 0.5 to 0.65. For the KDD dataset, the calculated values of precision are very close to 1 but the values of recall look a bit worse, fluctuating between 0.45 - 0.5. This is likely because that the dataset is relative small in column dimension (only 69 attributes) and the small cluster number was experimentally set. Figure 6 depicts the comparisons of using various clustering algorithms in co-clustering in terms of F-score. It is seen that using various clustering will result in variation of results. Overall, we claim that the quality of obtained co-clusters is quite good and satisfactory with F-score varying between 0.65 -0.75 for CTI dataset and 0.62-0.67 for KDD dataset.

5 Conclusion and Future Work

Web clustering is an approach for aggregating Web objects into various categories according to underlying relationships among them. The conventional clustering algorithms are mainly manipulated on one dimension/attribute of the Web usage data only, i.e. user or page solely, rather than taking into account the correlation between Web users and pages. Thus co-clustering of users and pages proposes a promising way to better reveal the relationships between these two kinds of Web objects and capture the user navigational behavior. In this paper, we have addressed the study of co-clustering Web usage data by using a bipartite spectral clustering approach. After modeling the usage data into a bipartite graph, a SVD-based projection is executed to transform the row and column vectors of original usage space for spectral clustering. Experiments have been conducted on two real world usage datasets to evaluate the performance of co-clustering by using precision, recall and F-score measures. Furthermore, we have also investigated the selection of the number of co-clusters and the impact of different clustering algorithms on clustering quality. One of possible future research direction is to compare the proposed co-clustering algorithm with other existing co-clustering algorithms.

Acknowledgment

This work has been partially supported by FP7 ICT project M-Eco: Medical Ecosystem Personalized Event-Based Surveillance under grant number 247829 and “KIWI - Knowledge in a Wiki” under grant number No. 211932.

References

1. Zhang, Y., Yu, J.X., Hou, J.: *Web Communities: Analysis and Construction*. Springer, Berlin (2006)
2. Wang, X., Zhai, C.: Learn from web search logs to organize search results. In: *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 87–94. ACM, New York (2007)
3. Kummamuru, K., Lotlikar, R., Roy, S., Singal, K., Krishnapuram, R.: A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In: *WWW 2004: Proceedings of the 13th International Conference on World Wide Web*, pp. 658–665. ACM, New York (2004)
4. Flesca, S., Greco, S., Tagarelli, A., Zumpano, E.: Mining user preferences, page content and usage to personalize website navigation. *World Wide Web Journal* 8(3), 317–345 (2005)
5. Haveliwala, T.H., Gionis, A., Indyk, P.: Scalable techniques for clustering the web (extended abstract). In: *WebDB 2000, Third International Workshop on the Web and Databases in Conjunction with ACM SIGMOD (2000)*
6. Ferragina, P., Gulli, A.: A personalized search engine based on web-snippet hierarchical clustering. In: *WWW 2005: Special Interest Tracks and Posters of The 14th International Conference on World Wide Web*, pp. 801–810. ACM, New York (2005)
7. Hou, J., Zhang, Y.: Utilizing hyperlink transitivity to improve web page clustering. In: *Proceedings of the 14th Australasian Database Conferences (ADC 2003)*, vol. 37, pp. 49–57. ACS Inc, Adelaide (2003)
8. Mobasher, B., Dai, H., Nakagawa, M., Luo, T.: Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery* 6(1), 61–82 (2002)
9. Xu, G., Zhang, Y., Zhou, X.: A web recommendation technique based on probabilistic latent semantic analysis. In: Ngu, A.H.H., Kitsuregawa, M., Neuhold, E.J., Chung, J.-Y., Sheng, Q.Z. (eds.) *WISE 2005*. LNCS, vol. 3806, pp. 15–28. Springer, Heidelberg (2005)
10. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: *KDD 2001: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 269–274. ACM, New York (2001)
11. Giannakidou, E., Koutsonikola, V.A., Vakali, A., Kompatsiaris, Y.: Co-clustering tags and social data sources. In: *WAIM*, pp. 317–324 (2008)
12. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T.: Co-clustering of biological networks and gene expression data. In: *ISMB*, pp. 145–154 (2002)
13. Jin, X., Zhou, Y., Mobasher, B.: A maximum entropy web recommendation system: Combining collaborative and content features. In: *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2005)*, Chicago, pp. 612–617 (2005)

Talking Biology in Logic, and Back*

Hasan Jamil

Department of Computer Science
Wayne State University, USA
jamil@cs.wayne.edu

Abstract. While computation of biological information takes center stage in today's research, knowledge-based query processing in biological databases has not gained much attention. The complex relationship among biological entities, and the numerous possibilities make it extremely difficult for a user to query data repositories in traditional ways since the information users seek are often not stored directly as extents. Thus, to meaningfully query such repositories, users often write complex applications to compute the needed response based on universal knowledge, or user specific hypothesis. In this paper, we propose a new data model based on object-oriented deductive databases and ontological concepts to allow knowledge-based querying. We propose a new query language called *ConLog* (for Concept Logic) in the direction of the acclaimed F-Logic and show that ConLog is able to express biological information and support queries in intuitive and user specific ways.

1 Introduction

Intelligent response to queries has been a subject of intense research for a long-time since such a querying paradigm closely parallels the real world environment within which we interact with each other and with other systems. In our human world, we use our knowledge and deductive ability to draw analogies, reason, comprehend, and recognize the intent of queries to respond to the best of our abilities. This is partly true because in our queries, we assume that the responder knows the context, has the knowledge and the ability to respond by reasoning. In computer based systems, especially in databases, queries are concrete and mostly based on syntax and structure and have little to do with the meaning of the data in contains. Whatever little semantics is understood, is buried in the syntax and in the users mind.

Recently, ontology based querying is being proposed as the vehicle for semantic querying [179]. While most proposals focus on modeling and querying applications using languages such as RDF or OWL, very few have addressed the issues related to rich knowledge modeling and answering queries using entailed knowledge. This is partly due to the fact that RDF and OWL are not very suitable for expressing knowledge using rich rules as in object-oriented systems. The deficiencies of OWL, in particular, led to its extension to include rules [10], and inheritance and encapsulation [1112]. Despite similar extensions, querying

* Research supported in part by National Science Foundation grants CNS 0521454 and IIS 0612203.

based on the knowledge ontologies entail remain an illusive goal. It was observed that in areas such as Life Sciences, simple use of ontologies can improve the answering capabilities of application manifold [20,18]. Our goal in this paper is to show that incorporation of ontological structure in rule processing systems and their proper interpretation can support knowledge-based querying in quite powerful ways.

Our work is primarily motivated by a pioneering effort at Virginia Commonwealth University in which the BioBike system [7] tries to respond to queries using background knowledge and database intension. However, in BioBike, the application logic is embedded into the query processing engine written in LISP which guides users to make the right choices through a graphical user interface. Though limited in its capability, BioBike is able to respond to queries that are biologically similar if the response is not available in the extensional database. BioBike is limited by the fact that it is custom designed to function for specific data repositories in which specific functions have specialized meanings. It is hard to imagine its use in other data repositories without extensive customization. In contrast, the ConceptBase system we propose, is a general purpose data modeling and querying system based on an F-Logic [16] like rule language called *ConLog* (short for Concept Logic) that is capable of capturing all the functionalities of BioBike, and can be used to model applications not specific to any domain. In a companion work [6], we are working on a natural language interface for ConceptBase in which we plan to develop a mapping for all natural language queries to ConLog so that the limitations posed by logical formulae are eliminated and a near NLP querying capability is supported.

The rest of the paper is organized as follows. Given the limited scope and length of the paper, we present our language and its salient features mostly using examples rather than a formal model and theory. We plan to discuss its formal foundation in an expanded version of this paper. In section 2, we introduce ConLog on intuitive grounds. We discuss an example and ConLog's implementation in SelfLog [1] (using SWI Prolog) in section 3 based on a transformation procedure. We present a discussion on related research in section 4 before we summarize in section 5.

2 The Language of Biology – ConLog

The language ConLog essentially supports the description of basic and composite concepts, their associations and derivation rules. ConLog is derived from ORLog [14], a derivative of F-Logic [16]. But unlike ORLog, ConLog supports schema description, schema compliance and schema queries somewhat similar to F-Logic, albeit in a restricted way. Since a full description of ConLog is beyond the scope of this paper, we will use illustrative examples to discuss the salient features of our language.

2.1 Ontological Structures

The model of ConLog is actually quite simple. It consists of a set of basic concepts, and a set of associations involving these concepts that give rise to a set

of composite concepts. A ConLog database is a set of modules [19,3] of composite concepts organized in a hierarchy (more precisely a DAG) of modules with overriding [1]. Basic concepts in this model are any concrete or abstract terms in the domain of discourse, such as *gene*, *regulation*, *organism*, etc. An association, on the other hand, is a relationship between two or more basic concepts. Hence, *gene mapk1 regulates mitosis* is an association between gene *mapk1* and function *mitosis*. Similarly, *mapk1 is a protein coding gene*, *fbxw2 is also known as UniProt q9ukt8*, and *mapk1 is a gene found in human, monkey and chicken*, are all associations. All concepts participate in a special association, called the specialization-generalization hierarchy, or is-a hierarchy. An is-a hierarchy is a directed acyclic graph by definition. A composite concept is a set of associations for a given concept.

The language \mathcal{L} of ConLog consists of an infinite set of id terms or concepts \mathcal{C} , an infinite set of function symbols \mathcal{F} for attribute or association names, infinite set of variable names \mathcal{V} , a set of terms \mathcal{T} such that $\mathcal{T} = \mathcal{C} \cup \mathcal{V}$, and a set of predefined polymorphic predicates \mathcal{P} , called the associations or relationships that are a subset and adaptation of Gruber's [8] and Bunge's [4,5] ontological relationships. The set \mathcal{P} includes *isa*, *partof*, *property*, and *alias*. Each of these are directional binary predicates with polymorphic behavior such as, *unique*, *multiple* and *open*. A *unique* relationship entails a one-to-one mapping, while *multiple* entails one-to-many, and finally *open* means many-to-many. The meanings of each of these predicates coincide with Gruber's and Bunge's definition of associations as follows:

- *A isa B*: *B isa A* is always false when *A isa B* holds.
- *A partof B*: for every instance of *A*, *B* is always true.
- *A property of B*: $A \not\rightleftharpoons B$.
- *A alias of B*: $A \not\rightleftharpoons B$, but if *A* alias of *B*, *B* alias of *A* holds.

Figure 1 shows an example of a ConLog ontological representation of an application. In this example, each box represents a basic concept, and each arrow represents an association. The thick arrows represent *isa* relationships, while thin dashed lines depict *part of*, *property* and *alias* type relationships. The arrows in thin lines capture relationship cardinalities much the same way ER model does, i.e., arrow at the box side is meant for one, and a no arrow is meant for many. Therefore, a unique relationship will be depicted using arrows at both ends, and open an relationship with no arrows. A multiple relationship from *A* to *B* will be shown with no arrow at *A* and an arrow at *B* to depict a single *A* being related with multiple *Bs*. Finally, a mandatory or required relationship is shown with a star on the line. Instances of a *isa* hierarchy can participate in instances of the relationships shown. For example, gene *mapk1* is part of $\{human, monkey\}$, gene *gnrhr2* has a mandatory *function*, i.e., *reception*, etc.

2.2 Syntax and Semantics of ConLog

Basic concepts in ConLog are just ID terms and hence, do not play much of a useful role by themselves. But in association with other concepts, they form

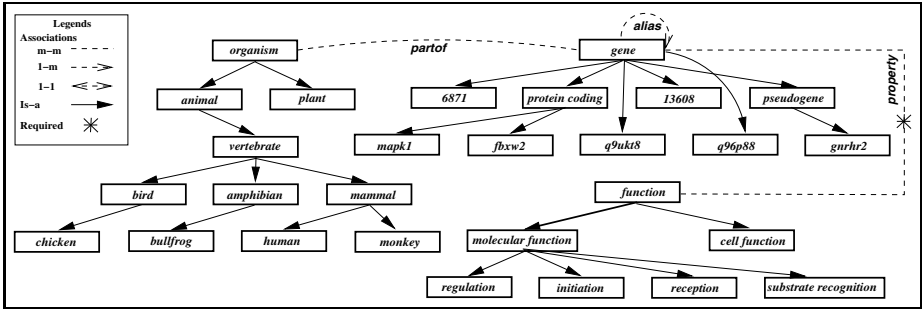


Fig. 1. Ontological representation of a gene function database in ConceptBase

complex concepts to play a vital role in modeling real world objects and concepts. Although we use only five types of associations from thousands possible [15], together with the cardinality constraints we have introduced, they capture a large set of structures sufficient to model practical applications.

Atomic Formulas. In the language of \mathcal{L} , if $c \in \mathcal{T}$, $c()$ and $c\{\}$ are c -terms to capture basic concepts, where c is the root of a concept isa hierarchy, $c()$ represents a node in the hierarchy for which at most one descendant exists, and multiple descendants $c\{\}$ may exist for c . For two concepts $\{c, v\} \in \mathcal{T}$, $v : c$ is an i-term or isa term that captures the fact that v is a descendant of c , or v is also a concept c . Given concepts $\{c, v\} \in \mathcal{T}$, an attribute name $f \in \mathcal{F}$, and an association type $p \in \mathcal{P}$, $c[f :: p \leftrightarrow v]$ is an a-term for association, in which \leftrightarrow is one of \Rightarrow (multiple), \Leftrightarrow (unique), or \Leftrightarrow^* (open [2]). An association type declares what kinds of associations are possible in a way similar to relation schemes or signatures in object-oriented languages such as F-Logic and ORLog. The corresponding instances, or concrete associations, called the v-terms (for value terms) are the instances of these concepts that satisfy the specified cardinality constraints. For example, given an a-term $c[f :: p \leftrightarrow v]$, and concepts $\{o, u\} \in \mathcal{T}$, $o[p = u]$ is a v-term, such that $o : c$, and $u : v$ hold.

Complex Formulas. Given v-terms of the form $o[p_i = u_i]$, and i-term of the form $o : c$, $o : c[p_1 = u_1, p_2 = u_2, \dots, p_n = u_n]$ is a complex formula which is a shortcut and logical equivalent of the formula $o : c \wedge o[p_1 = u_1] \wedge o[p_2 = u_2] \wedge \dots \wedge o[p_n = u_n]$. As customary in standard logic programming, given a set of basic and complex formulas in \mathcal{L} , a formula of the form $A \leftarrow B_1, B_2, \dots, B_m$. is a Horn clause such that either A is an i-term, or a-term and $m = 0$, or A is a v-term, B_i s are any basic or complex formula in \mathcal{L} and $m \geq 0$. A is called

¹ The boldface italic terms represent syntactic variables that could either be a variable or a constant in the language of \mathcal{L} .

² Mandatory associations are depicted using a star above the arrows, e.g., $\overset{*}{\Rightarrow}$. The corresponding keyboard notations we adopt are as follows: \Rightarrow , \Leftrightarrow , \Leftrightarrow^* , \Rightarrow^* , \Leftarrow^* , \Leftrightarrow^* . Interestingly, the numerous possible interpretations of associations that created substantial ambiguity and complexity in [15] is given a limited set of possible interpretations in ConLog using this set of definite possible incarnations of an association.

the consequent and B_i s are called the antecedent of the rule. Notice that a rule is called a fact, when $m = 0$, and we do not allow a-terms or i-terms in the consequent of a rule other than when it is a fact due to computational safety reasons vis-a-vis in ORLog. A rule is called a query, when it is of the form $\leftarrow B_1, B_2 \dots, B_m, m \geq 1$. The clauses are called i-, a- or v-clauses when A is a i-, a- or v-term respectively, basic or otherwise.

Modules and Databases. A collection of ConLog rules and facts f_i can be grouped in a named module called M , such that M is a pair of the form $\langle S, V \rangle$ where $S = I \cup A$ is the set of i- and a-clauses, and V is the set of v-clauses. A ConLog database has at least one module \top which is the default module in which all executions begin. A ConLog database is a pair $\langle \mathcal{M}, \preceq \rangle$ such that \mathcal{M} is the set of all modules, and \preceq is a partial order among the modules in \mathcal{M} . The purpose of the partial order relation is to organize knowledge modules in an inheritance hierarchy so that knowledge can be refined in a way similar to object-oriented hierarchies in languages such as F-Logic, ORLog or SelfLog. However, in ConLog we adopt a parametric inheritance model as in [13] and a module composition approach as in SelfLog that essentially captures knowledge refinement using program rewriting and overriding. In our model, knowledge refinement is captured via inflation (union) inheritance of S [13] and overriding inheritance of V in M . Finally, to be able to evaluate a formula in a module as a query goal, we introduce a new term, called the m-term, as $M \triangleright A$ (keyboard notation \triangleright) where A is any basic or complex formula, and M is the module name.

An Illustrative Example. The instance of the database shown in figure 1 can be described in ConLog as the program shown in figure 2 using the keyboard syntax. In this example, we have three modules, *mgene*, *ygene* and the default top module *top*. The partial order among the modules (in this example a linear hierarchy), can be described using the isa declaration of the form *module* : {*parents*}. For example, in figure 2, its written as *ygene* : *mgene*, and so on at the beginning of each module definition.

Interpretation of this program can be explained using the two queries in the module *top*. Let us consider query q1 first. As mentioned before, the hierarchy of modules determine the semantics of the program in which each module participate. In *top*, all the signatures and isa definitions are inherited monotonically. But, *mapk1* in *top* overrides the organism value ($\{human, monkey, chimp\}$) from *ygene*, and hence in response to the query we expect to get a binding $Z/\{human, chimp\}$. In query q2, we expect to get the binding $Y/\{human, monkey, chimp\}$ since the goal is evaluated in the module *ygene*. So, rules and facts corresponding to each concept are overridden in a lower module in a way similar to SelfLog's predicate replacement using program composition and rewriting at the module level.

The more interesting query is the *functions* part of query q1. Notice that in module *top*, the way *functions* are computed for the concept *gene* is introduced through rule (r2). Hence in addition to all the definitions of gene functions in facts (f1) through (f3), we can compute additional functions using rule (r2) which states that a gene has function F , if gene X is similar to another gene Z , and Z

```

mgene {
gene[partof::organisms <<=>> organism; property::functions <<*>> function].
gene{.
  organism{.
    function{.
proteinencoding : gene.      pseudogene : gene.      mapk1 : proteinencoding.
gnrhr2 : pseudogene.      fbw2 : proteinencoding.  q9ukt8 : proteinencoding.
animal : organism.        vertebrate : animal.    bird : vertebrate.
chicken : bird.          amphibian : vertebrate.  bullfrog : amphibian.
mammal : vertebrate.     human : mammal.        monkey : mammal.
molecularfunction : function.  geneticfunction : molecularfunction.
regulation : geneticfunction.  initiation : geneticfunction.
reception : geneticfunction.   substrate_recognition : geneticfunction.
mapk1[organisms = {human,monkey}, functions = {regulation,initiation}].      (f1)
fbw2[organisms = {human,chicken}, functions = {substrate_recognition}].      (f2)
gnrhr2[organisms = {monkey,bullfrog}, functions = {reception}].      (f3)
}

ygene : mgene {
gene[alias::aliases =>> gene].
chimp : mammal.
p28482 : gene.    6871 : gene.    13608 : gene.    q9ukt8 : gene.    q96p88 : gene.
mapk1[organisms = {human,monkey,chimp}, aliases = {6871}].      (f4)
fbw2[aliases = {13608,q9ukt8}].  gnrhr2[aliases = {q96p88}].  p28482[functions = {reception}].
}

top : ygene {
mapk1[organisms = {human,chimp}, aliases = {6871,p28482}].      (r1)
similar(X,Y) :- X : gene, Y : gene, ortholog(X,Y).      (r2)
X:gene[functions = Y] :- Z:gene[functions = Y], similar(X,Z).      (r3)
:- mapk1[functions = Y, organisms = Z].      (q1)
:- ygene |> mapk1[organisms = Y].      (q2)
}

```

Fig. 2. Partial ConLog implementation of the database in figure 1

has a function. Perhaps not as evident is the fact that a gene shares the functions of its aliases by definition, and so, for *mapk1*, the functions for genes 6871 and p28482 will also be returned, as well as for its orthologs³.

3 Translational Semantics and Implementation of ConLog

The semantics of ConLog programs are given using a translational semantics into a SelfLog [11] program as follows. All root concepts are represented as a predicate *concept*. The concept hierarchy is captured using the *isa* predicate, and their transitive relationship is captured using the axioms for *tisa*. The alias relationship is encoded into a rule for the predicate *similar* in module *common* using a transitivity of alias. Associations of the form *property* and *partof* are captured using two other predicates *property* and *partof*. Each rule involving *c*-, *a*-, *i*-, and *v*-terms are converted to corresponding predicates in a first-order Horn rule.

The difference with SelfLog is that in ConLog, the module composition function is much more complex and involves overriding based on concepts (objects) and not just predicate names as was done on SelfLog. We also do not override *c*-, *i* and *a*- clauses. Only *v*-clauses (their corresponding SelfLog clauses) are overridden in the module hierarchies. However, as mentioned before, a complete

³ We are assuming that there is a way to compute the orthologs of genes either from a table or using a computable function. In fact in ConceptBase, we support a set of computable functions for real life applications.

exposition is outside the scope of this paper due to space limitations and will be published elsewhere.

```

common{
tisa(X,Y) :- isa(X,Y).      tisa(X,X) :- isa(X,Y).
tisa(Y,Y) :- isa(X,Y).      tisa(X,Y) :- isa(X,Z), tisa(Z,Y).
talias(X,Y) :- alias(X,Y).  talias(X,Y) :- alias(Y,X).
talias(X,X) :- alias(X,Y).  talias(Y,Y) :- alias(X,Y).
talias(X,Y) :- alias(X,Z), talias(Z,Y).
similar(X,Y) :- talias(X,Y).
}

mgene : common {
concept(isa, gene, multiple).      concept(isa, organism, multiple).
concept(isa, function, multiple).
concept(alias, gene, aliases, multiple, optional).
concept(property, gene, functions, open, required).
concept(partof, gene, organisms, open, optional).
isa(gene, proteinencoding).  isa(gene, pseudogene).      isa(proteinencoding, mapk1).
isa(pseudogene, gnhr2).      isa(proteinencoding, fbw2).  isa(proteinencoding, q9ukt8).
isa(organism, animal).      isa(animal, vertebrate).  isa(vertebrate, bird).
isa(bird, chicken).        isa(vertebrate, amphibian). isa(amphibian, bullfrog).
isa(vertebrate, mammal).  isa(mammal, human).      isa(mammal, monkey).
isa(function, molecularfunction). isa(molecularfunction, geneticfunction).
isa(geneticfunction, regulation). isa(geneticfunction, initiation).
isa(geneticfunction, reception). isa(geneticfunction, substrate_recognition).
property(functions, mapk1, regulation).      property(functions, mapk1, initiation).
property(functions, fbw2, subsrtate_recognition).  property(functions, gnhr2, reception).
partof(organisms, human, mapk1).  partof(organisms, monkey, mapk1).
partof(organisms, human, fbw2).  partof(organisms, chicken, fbw2).
partof(organisms, monkey, gnhr2). partof(organisms, bullfrog, gnhr2).
}

ygene : mgene {
isa(chimp, mammal).  isa(gene, p28482).  isa(gene, 6871).  isa(gene, 13608).
isa(gene, q9ukt8).  isa(gene, q96p88).
property(functions, p28482, reception).
partof(organisms, human, mapk1).      partof(organisms, monkey, mapk1).
partof(organisms, chimp, mapk1).
alias(mapk1, 6871).  alias(fbwx2, 13608).  alias(fbwx2, q9ukt8).  alias(gnhr2, q96p88).
}

top : ygene {
alias(mapk1, 6871).  alias(mapk1, p28482).
partof(organisms, human, mapk1).      partof(organisms, chimp, mapk1).
similar(X,Y) :- tisa(gene, X), tisa(gene, Y), ortholog(X, Y).
property(functions, X, Y) :- tisa(gene, X), tisa(gene, Z), property(functions, Z, Y),
similar(X,Z).
:- property(functions, mapk1, Y), property(organisms, mapk1, Z)
:- ygene |> property(organisms, mapk1, Y).
}

```

4 Related Research and Novelty of ConLog

From a language standpoint, ConLog supports constructs to directly embed ontological concepts into the program. In ConLog, concepts are first class citizens whereas in F-Logic and SelfLog, objects and predicates respectively are primary entities. The concepts in ConLog have semantic interpretations very distinct from F-Logic and SelfLog. For example, *partof*, *property of* and *alias* have no counterpart in either of these languages, while they both support variants of *isa* associations. It can be easily shown that without such embedding and linguistic abstractions, languages such as F-Logic or SelfLog will have to simulate these

missing features through numerous explicit rule definitions. Within this context ConLog can be viewed as a higher level abstraction of F-Logic and SelfLog much the same way, these two languages can be viewed as higher level abstractions of Datalog or first-order logic.

Specifically, the *isa* relationship in F-Logic relates classes and objects, and inheritance of properties (values) is always non-monotonic. Furthermore, classes are used to group objects into collections where classes can define default properties for objects. However, objects in F-Logic can have their sub-objects because no distinction is made between classes and objects. But in ConLog, objects or concepts cannot be further specialized. In ConLog, *isa* relationship acts purely as a grouping and type safety mechanism. In SelfLog on the other hand, no explicit *isa* relationship is supported. A compositional semantics of modules in a module hierarchy is used instead to simulate non-monotonic inheritance of properties vis-a-vis inheritance in F-Logic's *isa* hierarchies.

The fundamental departure from hierarchies in F-Logic and SelfLog in ConLog is the way modules are organized in a static hierarchy in which only concepts (*v*-clauses) are inherited non-monotonically, while concept hierarchies are inherited monotonically. In other words, regardless of their scope, in module hierarchies, they are global. In F-Logic, objects are used to override properties in objects, while in ConLog modules override a group of concepts. Though it is not a serious weakness, in SelfLog, hierarchies are composed dynamically to determine inheritability, and the long chains of module hierarchies could be complicated and may lead to erroneous programs. It also does not allow modules to be organized in a tree or a graph like fashion because all hierarchies in SelfLog are linear chains. Finally, in ConLog, we support both dynamic module composition and static hierarchy of modules without the need for program transformation which SelfLog requires⁴.

Modules in ConLog serve as a mechanism to isolate knowledge that are potentially contradicting or interfering. The only way modules interact with others is through explicit module composition, which by design override conflicting concepts with the most specific ones⁵. Modules may be considered as user specific structuring of knowledge and user modules can be reused through refinement (module composition) using our module hierarchies and module compositions ($M \triangleright G$).

It is certainly possible to imagine such structuring using encapsulation. But there are several technical challenges. First encapsulation requires explicit interface definition for every property that needs to be exposed, which is quite burdensome when almost all the properties are actually public. In our applications

⁴ Although we do not discuss the proof theory and model theory of ConLog in this paper, ConLog do have a proof system that does not depend on program transformation, i.e., \oplus operator in SelfLog.

⁵ In [13], we have introduced multiple types and modes of inheritance some of which were later given a complete logical characterization by Yang and Kifer [21]. It is certainly possible to non-intrusively adapt such elaborate inheritance mechanism in ConLog as an orthogonal extension. However, for now, we will only consider non-monotonic inheritance for ConLog.

in Biology, we are just focused on avoiding conflicts, not hiding them. Even if we use encapsulation, we are again left with sorting out conflicting knowledge in a module and selecting the desired ones based on classical means such as negation or overriding. Furthermore, logical characterization of encapsulation is not a truly well understood subject and it is not clear enough how encapsulation interacts with inheritance in complex inheritance hierarchies although our own work on encapsulation has been shown to have significant promise [2].

In summary, ConLog has been designed to have specific language features to aid knowledge representation and knowledge structuring in Biology in particular, and science in general, although it can be used in other application domains requiring such features. Our claim is that the features included in ConLog are not directly supported in contemporary languages. In Biology, the knowledge scientists use have many facets: shared or universal knowledge, personal or user specific knowledge, and hypothetical knowledge or conjectures. Resources used are distributed, extremely large, heterogeneous and ever changing making it extremely difficult to create a unified and centralized repository. Hence, it is unlikely that a user will have complete knowledge about the structure and content of the resources she uses. In such an environment, it is helpful to form queries based on ontological concepts and available knowledge in prudent ways. In ConLog we have allowed ontological knowledge to be universal (*isa* and *a-terms*) and is thus shared across modules. Users are able to use their specific knowledge using overriding. They can also use selective overriding using dynamic module composition similar to SelfLog using \triangleright .

Finally, the constructs *partof*, *property of* and *alias* play important roles in representing knowledge specific to Biology. For example, the fact that *fbxw2* is also known as *UniProt q9ukt8* gene has complex biological implications. We can actually return everything relevant related to gene *q9ukt8* when a query about gene *fbxw2* is asked, and transitively, *q9ukt8* plays an important role in organisms that are composed of gene *fbxw2*. No contemporary logic other than ConLog is capable of capturing this knowledge directly into its semantics as explained earlier.

5 Summary

It was our goal to show that ontological constructs may be used in advanced object-oriented languages to support knowledge-based querying as a preliminary research. We have presented an early version of ConLog as a prototype that adequately demonstrates that intensional query answering is possible in unprecedented ways using a modular logic programming approach with inheritance. The advantage of module hierarchy is that we can organize the knowledge-base in a partial order and compute the entailment as a composition of modules as in SelfLog [1]. However, SelfLog does not support inheritance conflicts and multiple mode of inheritance, whereas ORLog and its parametric inheritance extension [13] do. The details of module composition still remains to be worked out which we hope to carry out as our future research.

References

1. Bugliesi, M.: A declarative view of inheritance in logic programming. In: JICSLP, pp. 113–127 (1992)
2. Bugliesi, M., Jamil, H.: A logic for encapsulation in object oriented languages. In: Penjam, J. (ed.) PLILP 1994. LNCS, vol. 844, pp. 215–229. Springer, Heidelberg (1994)
3. Bugliesi, M., Lamma, E., Mello, P.: Modularity in logic programming. *J. Log. Program.* 19/20, 443–502 (1994)
4. Bunge, M.: *Treatise on Basic Philosophy. Ontology I: The Furniture of the World*, vol. 3. D. Reidel Publishing Company, Inc., New York (1977)
5. Bunge, M.: *Treatise on Basic Philosophy. Ontology II: A World of Systems*, vol. 4. D. Reidel Publishing Company, Inc., New York (1979)
6. Dey, S., Hossain, S., Jamil, H.: A knowledge-based middleware to support cooperative response to natural language queries over structured scientific databases. Technical report, March 2010. Under review, ACM UIST (2010)
7. Elhai, J., Taton, A., Massar, J.P., Myers, J.K., Travers, M., Casey, J., Slupesky, M., Shrager, J.: Biobike: A web-based, programmable, integrated biological knowledge base. *Nucleic Acids Research* 37(Web-Server-Issue), 28–32 (2009)
8. Gruber, T.R.: A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 199–220 (1993)
9. Horrocks, I.: Scalable ontology-based information systems. In: EDBT, p. 2 (2010)
10. Horrocks, I., Patel-Schneider, P.F., Bechhofer, S., Tsarkov, D.: Owl rules: A proposal and prototype implementation. *J. Web Sem.* 3(1), 23–40 (2005)
11. Hosain, S., Jamil, H.: Empowering OWL with overriding inheritance, conflict resolution and non-monotonic reasoning. In: AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 meets Web 3.0 (2009)
12. Hosain, S., Jamil, H.: OWL that can choose to inherit and hide it too. In: IEEE International Conference on Semantic Computing, Berkeley, CA (September 2009)
13. Jamil, H.M.: A logic based language for parametric inheritance. In: *Knowledge Representation*, Breckenridge, CO, pp. 611–622 (April 2000)
14. Jamil, H.M., Lakshmanan, L.: A declarative semantics for behavioral inheritance and conflict resolution. In: ILPS (1995)
15. Kashyap, V., Borgida, A.: Representing the umls semantic network using owl (or "what's in a semantic web link?"). In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 1–16. Springer, Heidelberg (2003)
16. Kifer, M., Lausen, G., Wu, J.: Logical foundations of object-oriented and frame-based languages. *Journal of ACM* 42(4), 741–843 (1995)
17. Knappe, R., Bulskov, H., Andreassen, T.: Perspectives on ontology-based querying. *Int. J. Intell. Syst.* 22(7), 739–761 (2007)
18. Mabotuwana, T.D.S., Warren, J.: An ontology-based approach to enhance querying capabilities of general practice medicine for better management of hypertension. *Artificial Intelligence in Medicine* 47(2), 87–103 (2009)
19. Monteiro, L., Porto, A.: A transformational view of inheritance in logic programming. In: ICLP, pp. 481–494 (1990)
20. Wolstencroft, K., Brass, A., Horrocks, I., Lord, P.W., Sattler, U., Turi, D., Stevens, R.: A little semantic web goes a long way in biology. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 786–800. Springer, Heidelberg (2005)
21. Yang, G., Kifer, M.: Well-founded optimism: Inheritance in frame-based knowledge bases. In: Meersman, R., Tari, Z., et al. (eds.) CoopIS 2002, DOA 2002, and ODBASE 2002. LNCS, vol. 2519, pp. 1013–1032. Springer, Heidelberg (2002)

Analysis of Medical Pathways by Means of Frequent Closed Sequences

Elena Baralis, Giulia Bruno, Silvia Chiusano, Virna C. Domenici,
Naeem A. Mahoto, and Caterina Petrigni

Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy

Abstract. Analysing sequential medical data to detect hidden patterns has recently received great attention in a variety of applications. This paper addresses the analysis of patients' exam log data to rebuild from operational data an image of the steps of the medical treatment process. The analysis is performed on the medical treatment of diabetic patients provided by a Local Sanitary Agency in Italy. The extracted knowledge allows highlighting medical pathways typically adopted for specific diseases, as well as discovering deviations with respect to them, which can indicate alternative medical treatments, medical/patient negligence or incorrect procedures for data collection. Detected medical pathways include both the sets of exams which are frequently done together, and the sequences of exam sets frequently followed by patients. The proposed approach is based on the extraction of the frequent closed sequences, which provide, in a compact form, the medical pathways of interest.

Keywords: Pattern extraction, sequential pattern mining, closed sequences, medical pathways.

1 Introduction

The introduction of electronic medical records has made available a large amount of medical data, storing the medical history of patients. This large data collection can be profitably analyzed by using data mining techniques to extract a variety of information, for example the relationships between medical treatment and final patient condition, or the medical protocols usually adopted for patients with a given disease [4].

This paper addresses the problem of analysing patients' exam log data and performing reverse engineering of the steps of the medical treatment process. The proposed approach allows the identification of (i) the sets of exams which are frequently done together, and (ii) the sequences of exam sets frequently performed by patients. Both aspects are crucial for health care organizations, because they can significantly impact on the effectiveness of the medical treatments as well as on the costs incurred by the organizations.

Standard medical pathways have been defined as care guideline for a variety of chronic clinical conditions. They specify the sequence and timing of actions necessary to provide treatments to patients with optimal effectiveness and efficiency.

The adoption of these pathways allows health care organizations to control both their processes and costs [3] [8]. Medical pathways deviate from predefined guidelines when they include different or additional exams, or some exams are missing. When not justified by specific patient conditions, these non compliant pathways may cause additional costs without improving the effectiveness of the treatment.

Non compliant pathways may be due to patient negligence in following the prescribed treatments, or medical ignorance of the predefined guidelines, or incorrect procedures for data collection. For example, to speed up the registration process, the operator may only record a subset of the exams actually performed by the patient. On the other hand, available guidelines may not provide procedures to cover some particular (e.g., rare or new) diseases. By analysing the electronic records of patients, we can identify the medical pathways commonly followed by patients. This information may be exploited to improve the current treatment process in the organization and to assess new guidelines.

Sequential pattern mining [1] has been successfully exploited to analyze electronic medical data. Some works analyze the trend over time of a specific exam, to detect sequences which can potentially lead to a critical event (e.g., thrombosis [7]). By analysing time sequences of biochemical variables, in [2] frequent sequential patterns are extracted to individuate which variables can positively influence the effectiveness of the photopheretic therapy in liver transplant. The works in [5] and [10] focus on identifying patient flows among different hospitals or wards, to optimize health care resources. To analyze a collection of patients' exam log data and extract from it the medical pathways of interest, our approach relies on the extraction of frequent closed sequences. To the best of our knowledge, the analysis of this kind of data to extract frequent medical pathways was not addressed before.

The paper is organized as follows. Section 2 provides the formal definition of the problem by explaining how medical pathways are represented by means of frequent closed sequences. Section 3 presents the proposed approach to derive medical pathways. In Section 4 our approach is applied to diabetics patients' log data, while Section 5 draws conclusions and discusses future work.

2 Representing Medical Pathways by Means of Closed Sequences

Patients' exam log data store the sequences of medical exam¹ sets performed by patients in subsequent days. This data collection can be represented as a sequence database, i.e., a collection of tuples (p_{id}, S) , where p_{id} is the patient identifier and S is the temporal list of sets of exams (e_i) done by the patient. As an example, consider the following sequence database including two tuples: (patient p_1 , $\{e_1\}\{e_2, e_3\}\{e_1\}$), (patient p_2 , $\{e_1\}\{e_2, e_3\}\{e_4\}$).

A sequence S is said to contain a sequence S' if S' is a subsequence of S , i.e., S' contains a subset of the elements in S and preserves their order. S is called

¹ With the term "exam" we indicate each kind of medical test or service recorded in the patients' log data.

supersequence of S' . The *support* (or frequency) of sequence S' is the percentage of tuples in the database that contain S' . A sequence S' is called *frequent* if its support is above a specified support threshold. A sequence S' is called *frequent closed sequence* if there exists no proper supersequences of S' , having the same support as S' [11]. In this work, we are interested in mining frequent closed sequences, because they represent the medical pathways hidden in the data.

Consider the previous example database. When all sequences are extracted, both sequences $S'=\{e_1\}\{e_2\}$ and $S''=\{e_1\}\{e_2,e_3\}$ are generated. S' is a subsequence of S'' and both sequences have support 2. It follows that all patients that did first exam e_1 and then exam e_2 , also did exam e_3 together with e_2 . Sequence S' is thus redundant for our investigation, because its information is also included in S'' . Sequence S'' is a closed sequence, because all its supersequences (i.e., $\{e_1\}\{e_2,e_3\}$, $\{e_1\}$ and $\{e_1\}\{e_2,e_3\}\{e_4\}$) have lower support. Thus S'' is not redundant.

Considering only the subset of closed sequences allows us to considerably reduce the size of the solution set. A compression factor CF can be defined to evaluate the compactness achieved obtained by considering the frequent closed sequences (FCS) instead of all frequent sequences (FS), as

$$CF = (1 - \frac{\#FCS}{\#FS})\% \quad (1)$$

In the former example, when considering a support threshold equal to 1, 22 sequences are extracted, while the closed sequences are only 3. The achieved compression is thus $CF=86\%$.

3 Deriving Medical Pathways

Patients' exam log data are analyzed to identify hidden medical pathways by means of frequent closed sequence extraction. The meaning of extracted medical pathways is then assessed by exploiting medical domain knowledge. The main phases of the proposed approach are reported in Figure 1, while a more detailed description of each phase is presented in the following.

3.1 Data Collection and Preprocessing

In the data collection phase the raw medical exam logs provided by different medical units are collected and integrated into a common data structure. The resulting data collection contains information about patient id, exam name, exam date, and additional information as prescription code, prescription date, kind and location of the exam. In the preprocessing phase the irrelevant information is removed and only the following attributes are selected (patient id, exam date, exam name). Cleaned and integrated data are finally transformed into a sequence database on which the extraction process takes place.

3.2 Medical Pathway Extraction

The extracted medical pathways are of two types: (i) the sets of exams usually performed together (i.e., in the same day) and (ii) the frequent sequences of exam

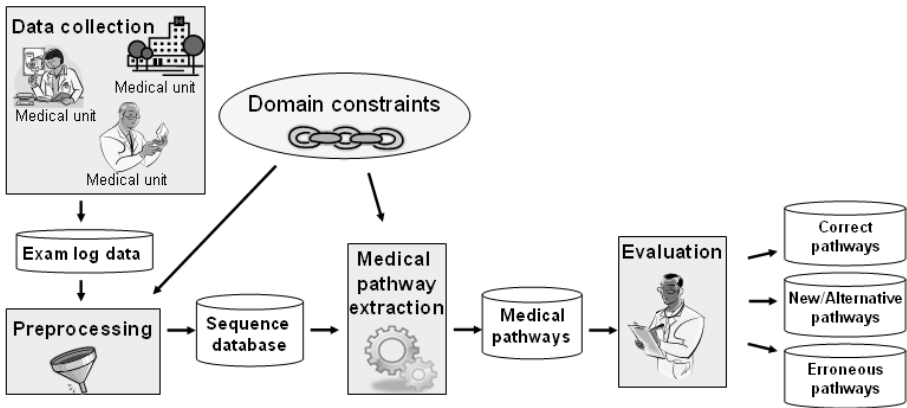


Fig. 1. Main phases of the proposed approach

sets. Each extracted pathway is a frequent closed sequence, characterized by a support value. Currently, the analysis exploits the BIDE algorithm [11], an effective state of the art method for frequent closed sequence extraction. It adopts the BI-Directional Extension closure checking scheme to prune the search space.

3.3 Domain Constraints

When additional domain information is available, our approach allows the inclusion of the following domain constraints to focus the analysis on specific data subsets.

Target exam set. This constraint is used to focus the analysis on a specific set of exams, such as the most expensive, or typical of a specific pathology. The constraint is exploited in the preprocessing phase to select only patients who did the target exams. This approach is particularly useful in case of infrequent exams usually associated to more critical pathologies.

Medical pathway length. Constraints on medical pathway length are used to focus the analysis on patients which have a particular clinical history (e.g., perform exams many times in a year). This constraint is exploited in the preprocessing phase to discard patients whose clinical history does not satisfy the length constraint.

As future work we plan to extend the medical pathways extraction phase to incorporate the management of temporal constraints on the exam sequences (e.g., the maximum time interval among sets of exams in the sequence, or the maximum time window of interest for the analysis).

3.4 Medical Pathway Evaluation

The extracted medical pathways are evaluated based on the available medical knowledge for the considered domain (e.g., clinical guidelines, medical experts). The following situations have been detected during the analysis.

Correct pathways. The extracted pathway is coherent with the medical knowledge. Thus, the treatment process followed by the supporting patients is correct.

New or alternative pathways. The extracted pathway is not included in the available medical knowledge. For example, the available guidelines do not cover a specific and rare disease. In this case the approach allows identifying the medical pathways commonly followed by patients. This information can be exploited for the assessment of new guidelines.

Erroneous pathways. The extracted pathway is not coherent with the medical knowledge. For example, it includes different or additional exams, or some exams are missing. Erroneous pathways can be for example imputed to incorrect procedures for data collection. The detection and analysis of data entry errors can be profitably exploited both to perform input data cleaning and to improve the medical data entry process.

4 Medical Pathway Analysis for Diabetic Patients

Medical pathways have been extracted from a dataset recording the medical treatments of diabetics patients. The dataset includes 95,788 records logging all the exams performed by 6,380 diabetic patients, which were collected in the year 2007 by the Local Sanitary Agency of the Asti province (Italy).

In the preprocessing phase the exam log data were transformed into a database of 6,380 sequences. This sequence database contains 159 distinct exams, and the maximum, minimum and average sequence length are 41, 1, and 3.58, respectively. We extracted the medical pathways from the sequence database and we investigated their meaning with the supervision of a domain expert. We also evaluated the performance of the proposed approach in terms of execution time, number of extracted sequences, and achieved compression factor.

4.1 Characterization of the Extracted Medical Pathways

Exam frequencies. The occurrence in the database of each exam is evaluated. The exam frequencies reflect the expected frequencies in diabetics treatments. Even if the data only covers one year of treatments, a wide spectrum of clinical situations is represented.

The most frequent exams are the base tests routinely repeated by diabetic patients to monitor the current concentration of sugar in blood. They include the glucose level exam (84.76% of patients), the venous blood sample (79.25%), the capillary blood sample (75.03%), and the urine test (74.87%).

Also exams typically prescribed in case of serious complications caused by diabetes appear in the database. These exams are (correctly) characterized by a lower frequency. Exams concerning cardiovascular diseases as the total cholesterol level (35.96%), and the triglycerides level (35.69%), or exams to determine the liver health as the measure of alanine aminotransferase enzyme (30.14%) and of aspartate aminotransferase enzymes (29.51%) are relevant examples of this situation.

Another disease degeneration is the damage to the eye retina (retinopathy). The database contains both exams to monitor the eye status (e.g., the examination of fundus oculi, 27.24%) and more specific exams for retina repairing (e.g., laser photocoagulation 2.24%). The frequency is significantly lower for the latter case, being the exam more specific.

Frequent exam sets. Medical pathways which represent the sets of exams frequently done together (i.e., in the same day) are analyzed. For the considered database, the extracted pathways are generally coherent with the medical knowledge, but also few anomalies exist in the data.

For example, since the glucose level is measured by analysing a sample of either venous blood, or capillary blood, or urine, at least one of these three exams is expected to be associated to the glucose level exam. Most of the extracted medical pathways verify this constraint. The following frequent sets have been detected: {glucose, urine} (74.86% of patients), {glucose, capillary blood} (74.40%), {glucose, venous blood} (70.99%).

On the other hand, for 5.56% of the patients the glucose level exam appears at least once not associated with any of the three exams. These sequences clearly highlight an error condition, being it impossible to evaluate the glucose level without any blood or urine sample. These errors may be due to an incorrect data entry process, in which only a subset of the exams actually performed by the patient have been recorded.

Frequent sequences of exam sets. Medical pathways which represent the sequences of exam sets frequently done by patients are analyzed. The extracted pathways are coherent with the usual diabetics treatments.

For example, patients are required to routinely repeat the glucose measure during the year to monitor the disease status. The most frequent sequences in the database reflect this behavior, showing that the glucose exam control is repeated two times (58.20% of patients), three times (31.83%), or four times (14.78%) during the considered year.

Medical pathways including target exams. To extract pathways related to a specific pathology from the sequence database, only patients who did at least one of the target exams are selected in the preprocessing phase. Then, medical pathways are extracted from the selected subset.

For example, a serious diabetes degeneration is the damage of eye retina (retinopathy). The retinal photocoagulation is a therapy used to repair retina lacerations, and a patient with proliferative retinopathy often requires multiple therapy treatments. To analyze the medical pathways including the retinal photocoagulation therapy, the exam data for the 143 patients (i.e., 2.24% of the total patients) performing at least once this treatment are selected. Evidence of multiple treatments emerge because pathways are available where the therapy is repeated two times (50.35% of the sample, but 1.13% of the total) or three times (25.17% of the sample, but 0.56% of the total).

4.2 Performance Evaluation

To verify the feasibility of our approach we analyzed the performance of the closed sequence extraction by means of the BIDE algorithm whose Java implementation has been kindly provided by Philippe Fournier-Vigier [6]. All experiments were performed on a Pentium 4 at 3.2 GHz with 2 GByte of RAM. The BIDE execution time ranges between few seconds for support threshold 70% and about 23 minutes for support threshold 30% (see Figure 2(a)). In Figure 2(b) we report the number of closed sequences of different length for support values ranging from 60% to 30%.

We also evaluate the compression achieved by mining only the *closed* frequent sequences instead of *all* sequences by computing the compression factor (see Section 2). The complete set of frequent sequences has been extracted using the PrefixSpan algorithm [9] (Java implementation kindly provided by [6]). Almost no compression is achieved for support thresholds between 70% and 60%, because few sequences are extracted and they are mainly closed. The compression factor increases from 11% to 48% for support thresholds decreasing from 50% to 30%. These results show that by focusing on the closed sequences only, the amount of extracted information is significantly reduced, with the twofold effect of (i) a more easily manageable result for pathway evaluation, and (ii) a more efficient sequence extraction process.

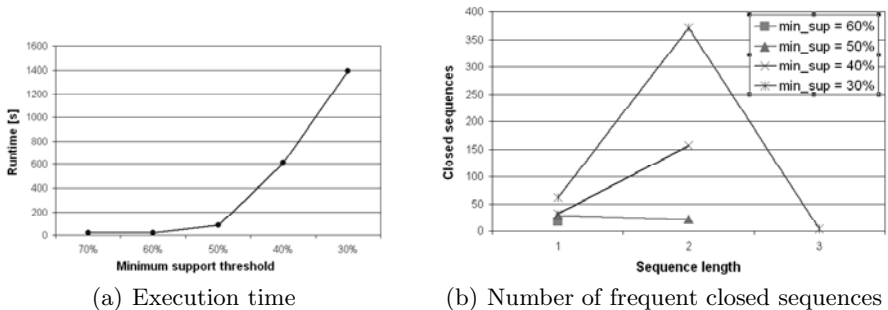


Fig. 2. Performance evaluation

5 Conclusion and Future Works

In this paper we presented an approach to analyse patients' exam log data to detect frequent medical pathways followed by patients. The approach is based on the extraction of frequent closed sequences and it allows setting domain constraints to drive the analysis.

Future extensions will address (i) the implementation of temporal constraints to further drive the analysis and (ii) the definition of input data cleaning procedures driven by the incoherences detected in the extracted pathways.

Acknowledgments

The authors would like to thank Prof. Baudolino Mussa (University of Torino Medical School, Italy) for his support and for serving as domain expert, and Prof. Dario Antonelli (Politecnico di Torino) for fruitful discussions.

References

1. Agrawal, R., Srikant, R.: Mining sequential patterns. In: ICDE 1995, pp. 3–14 (1995)
2. Berlingerio, M., Bonchi, F., Curcio, M., Giannotti, F.: Mining Clinical, Immunological, and Genetic Data of Solid Organ Transplantation Biomedical Data and Applications, pp. 211–236. Springer, Heidelberg (2009)
3. Calligaro, K.D., Dougherty, M.D., Raviola, C.A., Musser, D.J., DeLaurentis, D.A.: Impact of clinical pathways on hospital costs and early outcome after major vascular surgery. *Journal of Vascular Surgery* 22(6), 649–660 (1995)
4. Cerrito, P.B.: Mining the Electronic Medical Record to Examine. *Physician Decisions Studies in Computational Intelligence* 48, 113–126 (2007)
5. Dart, T., Cui, Y., Chatellier, G., Degoulet, P.: Analysis of hospitalized patient flows using data-mining. *Studies in Health Technology and Informatics* 95, 263–268 (2003)
6. Fournier-Viger, P.: <http://www.philippe-fournier-viger.com/spmf/>
7. Jensen, S.: Mining medical data for predictive and sequential patterns: PKDD 2001. PKDD Discovery Challenge on Thrombosis Data (2001)
8. Panella, M., Marchisio, S., Stanislao, F.: Reducing clinical variations with clinical pathways: do pathways work? *Int. J. for Quality in Health Care* 15(6), 509–521 (2003)
9. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, W.H., Chen, Q., Dayal, U., Hsu, M.C.: Mining sequential patterns by pattern-growth: the prefixspan approach. *IEEE Transaction on Knowledge and Data Engineering* 16(11), 1424–1440 (2001)
10. Rossille, D., Cuggia, M., Arnault, A., Bouget, J., Le Beux, P.: Managing an emergency department by analysing HIS medical data: a focus on elderly patient clinical pathways. *Health Care Manage Sci.* 11, 139–146 (2008)
11. Wang, J., Han, J.: BIDE: Efficient Mining of Frequent Closed Sequences. In: International Conference on Data Engineering, pp. 79–90 (1995)

Inheriting Access Control Rules from Large Relational Databases to Materialized Views Automatically

Alfredo Cuzzocrea¹, Mohand-Said Hacid², and Nicola Grillo³

¹ ICAR-CNR and University of Calabria, Italy

² University Claude Bernard Lyon 1 and LIRIS, France

³ DEIS Department, University of Calabria, Italy

cuzzocrea@si.deis.unical.it, mohand-said.hacid@univ-lyon1.fr,
ngrillo@deis.unical.it

Abstract. A novel approach for *automatically inheriting access control rules from large relational databases to materialized views* defined on such databases is proposed in this paper, along with main algorithm VSP-Bucket. Our proposal introduces a number of research innovations, ranging from a novel *Datalog*-based syntax, and related semantics, for modeling and expressing access control rules over relational databases to algorithm VSP-Bucket itself, which is a meaningfully adaptation of a well-know view-based query re-writing algorithm for database optimization purposes. A preliminary experimental evaluation and analysis of performance of algorithm VSP-Bucket completes our foremost analytical contribution made in this research.

1 Introduction

Database security is playing a major role in state-of-the-art security literature [5], as DBMS breaches represent the main reason of vulnerability for large corporate organizations. A well-recognized solution to this research challenge is represented by *security policies* [2] that determine the way database users can access and manage *database objects*. Several models have been proposed to this end. *Fine-grained access control models* (e.g., [1]) are among the most popular ones, and undoubtedly play a critical role in database security research, as confirmed by several studies in this scientific field. Basically, given a table T belonging to a database D , fine-grained access control models apply *restrictions* to individual rows, columns or even cells of T , being these restrictions expressed in terms of *access control rules* [23] based on a given language, e.g. query languages such as SQL (e.g., [21,1]), logical languages such as *First Order Logic* (FOL) [23], declarative languages such as XML [8,7,4], embedded database languages such as PL/SQL [16], and so forth. On the other hand, fine-grained access control models have been implemented within the core layers of a number of DBMS platforms such as *Oracle* [16], *SQL Server* [13] and *Sybase* [24]. This further confirms to us the great interest for such a database security solution by both the academic and industrial research community.

In today DBMS platforms, beside database security issues highlighted before, a critical role is also played by *query optimization issues* [12], which concerns with efficiently evaluating queries against the target database. Across years, this challenge has originated a wide and well-understood research literature. Indeed, although a wide spectrum

of data representation and related management/query technologies, more or less efficient, exist (e.g., *text databases*, *XML databases*, *multidimensional and OLAP databases*, and so forth), *the majority of information/knowledge of any corporate organization is still stored and processed within conventional relational databases*, whereas so-originated data sets are increasing in size, dimension and complexity continuously.

As a consequence, a well-known mechanism for taming the complexity of evaluating queries against large relational databases is represented by the so-called *view-based query answering techniques* [11]. Given a database D and an input query-workload QWL , which could even be a synthetic query-workload modeling real-life ones on the basis of a *probabilistic* or *data-mining-based* interpretation [3], view-based query answering techniques pursue the idea of effectively and efficiently computing a set of views V such that queries in QWL evaluated against V instead than D introduce a *lower complexity* with respect to the case of evaluating queries in QWL against D , thus reducing query response time drastically. Views can be either *virtual* or *materialized*, meaning that in the first case they are computed at query time whereas in the second case they are pre-computed and stored within the database as normal (database) tables. The first solution does not impose us to take into consideration *view update* and *synchronization* issues while it increases query response time, whereas the second solution decreases query response time while it poses several problems related to the issue of *managing views efficiently* [9].

View-based query answering techniques have already reached a sufficient maturity in the context of DBMS technology, so that almost all today corporate databases store views within their core layers, for query optimization purposes. At the same, commercial DBMS platforms include at now the view definition and management layer in their reference architectures (e.g., *Oracle*). As possible drawbacks that may limit the success of this database technology, view update, synchronization and management represent day-by-day serious problems for database administrators. Coming back to theoretical aspects, it has been already proven that, given a database D and an input query-workload QWL against D , computing the set of views V that optimize a given criterion C (e.g., minimization of the query error of queries in QWL , or minimization of the total occupancy of views in V) is an *NP-Hard problem* [10].

Given a database D having schema S , access control rules R are defined against S (by the database administrator, or granted users/applications) and they are not automatically inherited by *materialized* views in V that may be defined on D . As a consequence, tuples stored in views of V are *not* secured like tuples stored in tables of D , so that possible *security breaches* arise, even because views in V may also be used by external users/applications during the evaluation of resource-consuming queries. Therefore, *protecting tuples in V by inheriting access control rules from R is a big research challenge for database technology*, due to the fact that, as highlighted before, almost all today corporate databases store and manage views for query optimization purposes, without any effective neither efficient security solution. To the best of our knowledge, there is not any research work in literature that deals with this so-important challenge, despite its significativity and relevance.

Looking at practical implementations of commercial DBMS platforms, editing access control rules from R on V is still a *manual task* performed by the database administrator, like in *Oracle*, who determines a sub-set Z (of R) that is valid for views in V . Obviously, this solution exposes to a number of drawbacks, such as: (i) dealing with large databases

[2], (ii) dealing with a large number of access control rules [2], (iii) dealing with incomplete or uncertain databases [18], (iv) dealing with heterogeneity and inconsistency between the original database schema and the view schema [13]. Therefore, it clearly follows that it is necessary to make the *access control rule selection process* automatic or semi-automatic for database management efficiency purposes.

Starting from these considerations, *in this paper we propose and experimentally assess an innovative algorithm, called VSP-Bucket (View Security Policy by Bucket algorithm), for effectively and efficiently selecting access control rules on materialized views over relational databases.* Basically, given (i) a database D having schema S and access control rule set R , and (ii) a set of queries Q that define the materialized view set V on D , we select from R a (sub-)set of access control rules Z suitable to secure tuples stored in views of V . Summarizing, in this paper we make two main research contributions: (i) we introduce a *Datalog*-based syntax, and related semantics, to model and express access control rules over relational databases, thus taking advantages from the flexibility and the expressive power of such a modeling formalism; (ii) we introduce algorithm VSP-Bucket as a meaningfully adaptation of the well-known *view-based query re-writing bucket algorithm* by Halevy [11] to the problem of effectively and efficiently selecting the set of access control rules to be propagated on the target view set. To the best of our knowledge, our research is the first in proposing the usage of *Datalog* to model and reason on access control rules over relational databases, with no similar initiatives in previous studies.

2 Related Work

A significant amount of research has been devoted to the problem of modeling and effectively and efficiently reasoning on access control rules to define powerful database security policies [5], beyond limitations of native SQL statements [21]. In particular, this problem is directly related to the issue of *re-writing queries according to a pre-defined set of access control rules*, thus ensuring the security of tuples [23]. Among the literature devoted to access control rules for the security of relational databases, we recall: [15], which introduces the so-called *Access Control Matrix (ACM)* for effectively and efficiently implementing access control mechanisms over database objects by means of user grants defined on relational tables and portions of them; [21], which starts from limitations of SQL in capturing authorization mechanisms (due to progressively-added constructs like triggers, objects, and so forth) and proposes *abstracted authorization definitions* rather than syntax-related authorization definitions (like in traditional initiatives) in order to overcome drawbacks above; [19], which proposes the usage of so-called *SQL-based authorization views* to flexibly implement fine-grained access control models within DBMS platforms; [26], which proposes innovative access control models on *derived objects*, i.e. database objects which have a some relation with a set of primary database objects (e.g., views and tables, respectively); [1], which introduces novel SQL-based constructs, which can be directly integrated within the core layer of DBMS platforms, for defining restrictions over database tables, along with an algorithm able to enforce fine-grained access control during query evaluation by re-writing approaches; [13], which focuses the attention on possible redundancy and information loss issues arising in view-management-based

access control mechanisms over relational databases; [25], which defines correctness criteria to assess the quality of a fine-grained access control model over relational databases; [14], which focalizes the attention on a real-life application scenario represented by fine-grained access control mechanism over distributed bio-medical databases.

Starting from the initial context of relational databases, researchers have sequentially devoted attention to access-control-rule-based security policies in different and innovative contexts. Among these contexts, indoubtely XML databases play a leading role, due to the fact they are very popular in actual information systems. In line with this assertion, [8] investigates the problem of *flexibly* defining access control rules over XMLSchema schemas of XML documents by means of XQuery. [7] introduces *security views* over XML documents by means of a novel formalism for modeling *security constraints* over XML documents that control the access of multiple users to multiple portions of the target XML document corpus. [4] moves instead the attention on the issue of automatically adapting so-called *XPath-formatted access control views* over XML documents subjected to updates. Security policies over *Web databases* has also been of strident interest for database researchers. For instance, [20] studies acces control rules over such databases by means of so-called *parametrized views* able to capture the particular dynamicities of the Web. Finally, novel contexts like *uncertain databases* (e.g., RFID data sets) have also been addressed, like [18], which introduces a new access control language for uncertain data, called *UCAL*, which exploits a *multiple-instance-based probabilistic interpretation* of uncertain databases in order to define a novel semantics according to which access control over such databases is handled in terms of *partial information* releasement from the so-modeled probabilistic databases.

3 Expressing Access Control Rules over Relational Databases by Datalog

In this Section, we provide our novel Datalog-based notation for modeling and expressing access control rules over relational databases we use in our proposed security framework. First, some definitions are necessary in order to better understand our research.

Given a relational database $D = \{R_0, R_1, \dots, R_{|D|-1}\}$ having schema S , such that $R_i(A_{i,0}, A_{i,1}, \dots, A_{i,|R_i|-1})$, with $0 \leq i \leq |D|-1$, denotes a relation in S , and $A_{i,0}, A_{i,1}, \dots, A_{i,|R_i|-1}$ denote the attributes of R_i , a Datalog *conjunctive query* q over D is defined as follows:

$$\begin{aligned}
 q(X) \leftarrow & R_{k_0}(A_{k_0,0}, A_{k_0,1}, \dots, A_{k_0,|R_{k_0}|-1}) & \wedge \\
 & R_{k_1}(A_{k_1,0}, A_{k_1,1}, \dots, A_{k_1,|R_{k_1}|-1}) & \wedge \\
 & \dots & \wedge \\
 & R_{k_{U-1}}(A_{k_{U-1},0}, A_{k_{U-1},1}, \dots, A_{k_{U-1},|R_{k_{U-1}}|-1}) & \wedge \\
 & P_{C_0}(A_{h_0}) \wedge P_{C_1}(A_{h_1}) \wedge \dots \wedge P_{C_{W-1}}(A_{h_{W-1}}) & \wedge
 \end{aligned} \tag{1}$$

such that: (i) R_{k_j} , with $0 \leq k_j \leq U-1$, denotes a *standard predicate* that is derived from a database relation R_{k_j} in D ; (ii) $A_{k_j,0}, A_{k_j,1}, \dots, A_{k_j,|R_{k_j}|-1}$ denote attributes of the

predicate R_{k_j} (which correspond to attributes of the relation R_{k_j} in D) or costants; (iii) $P_{C_w}(A_{h_w})$, with $0 \leq w \leq W-1$, denotes a *comparison predicate* between a singleton attribute A_{h_w} in D (i.e., $A_{h_w} \in \{A_{0,0}, A_{0,1}, \dots, A_{|D|-1, |R_{|D|-1}|-1}\}$) and a constant γ , i.e. $P_{C_w}(A_{h_w}) = A_{h_w} \theta \gamma$, such that θ belongs to the following alphabet of operators: $\{<, >, \leq, \geq, =, \neq\}$ (it should be noted that the latter operator alphabeth is useful enough to cover a wide set of case studies of interest for database security applications); (iv) $X \subseteq \{A_{k_0,0}, A_{k_0,1}, \dots, A_{k_{U-1}, |R_{k_{U-1}}|-1}\}$ denotes a set of attributes in D or costants. Furthermore, we denote as $R(q)$ the set of predicates of the body of q , i.e. $R(q) = \{R_{k_0}, R_1, \dots, R_{k_{U-1}}\}$, and as $P(q)$ the set of comparison predicates of the body of q , i.e. $P(q) = \{P_{C_0}, P_{C_{0,1}}, \dots, P_{C_{W-1}}\}$, respectively. Evaluated against D , q retrieves a set of tuples that satisfy its body.

Containment and equivalence are two important properties one can find and check among two given Datalog queries. We next provide the formal definitions of these properties, by inheriting classical glossaries (e.g., [22]). Given two queries q_i and q_j , with $i \neq j$, we say that q_i is *contained* by q_j , and denote as $q_i \subseteq q_j$, iff, given a database D , evaluating q_i against D retrieves a sub-set of tuples of the tuple set retrieved by evaluating q_j against D . In other words, we can equivalently say that q_i is contained by q_j iff a *containment mapping* between q_j and q_i exists, i.e. (i) *each* predicate R_{k_j} in q_j matches with a predicate R_{k_i} in q_i , with $0 \leq k_j \leq |D|-1$ and $0 \leq k_i \leq |D|-1$, and (ii) X_j in the head of q_j matches with X_i in the head of q_i . Based on the containment concept, we can easily derive the equivalence concept. Given two queries q_i and q_j , with $i \neq j$, we say that q_i is *equivalent* to q_j , and denote as $q_i \equiv q_j$, iff $q_i \subseteq q_j$ and $q_j \subseteq q_i$. From active literature, it follows that, given two conjunctive queries q_i and q_j , with $i \neq j$, determining if q_i is contained in q_j is an *NP-complete problem* [6].

In our proposed security framework, we make use of conjunctive queries (1) to model both queries in q defining the set of materialized views V and the same access control rules over both the target relational database D and the view set V . Since the usage for generating materialized views is trivial, let us focus on more interesting access control rules. First, we introduce the *access control scheme* over relational databases supported by our proposed security framework. Given a relational database $D = \{R_0, R_1, \dots, R_{|D|-1}\}$ and an access control rule r over D , r can restrict the access to a *singleton* relation R_i in D or to *multiple* relations $R_{k_0}, R_{k_1}, \dots, R_{k_{N-1}}$ in D , with $0 \leq k_j \leq |D|-1$. Beyond to be innovative in the context of database security research, this specialized property supported by our framework allows us to gain flexibility extremely, hence a higher expressive power (thanks to the Datalog syntax) in editing complex access control rules over very large relational databases. As regards a singleton relation R_i in D , an access control rule r over R_i can restrict the access to the following database objects of R_i : (i) tuples (i.e., rows) $t_{i,h_0}, t_{i,h_1}, \dots, t_{i,h_{L-1}}$ in R_i , with $0 \leq h_l \leq \|R_i\|-1$, such that $\|R_i\|$ denotes the *cardinality* of R_i ; (ii) attributes (i.e., columns) $A_{i,u_0}, A_{i,u_1}, \dots, A_{i,u_{F-1}}$ in R_i ; with $0 \leq u_f \leq |R_i|-1$; (iii) cells $c_{i,h_0,u_0}, c_{i,h_1,u_1}, \dots, c_{i,h_{L-1},u_{F-1}}$ in R_i , with $0 \leq h_l \leq \|R_i\|-1$ and $0 \leq u_f \leq |R_i|-1$. As regards multiple relations $R_{k_0}, R_{k_1}, \dots, R_{k_{N-1}}$ in D , with $0 \leq k_j \leq |D|-1$, an access control rule r over $R_{k_0}, R_{k_1}, \dots, R_{k_{N-1}}$ can restrict the access to tuples, attributes and cells of

$R_{k_0} \bowtie R_{k_1} \bowtie \dots \bowtie R_{k_{N-1}}$ by combining the elementary access control scheme provided above of each relation R_i in $R_{k_0}, R_{k_1}, \dots, R_{k_{N-1}}$.

Upon the access control scheme above, we introduce some classes of access control rules, which we define next. First, we distinguish between *positive access control rules* and *negative access control rules*. Given a relational database D , a positive access control rule r over D defines X as a set of attributes in D that *can* be accessed by users of D . Symmetrically, a negative access control rule r over D defines X as a set of attributes in D that *cannot* be accessed by users of D . In our proposed security framework, we adopt negative access control rules, as the latter are suitable and targeted to algorithm VSP-Bucket (see Sect. 1), as we demonstrate in Sect. 4.

4 Effectively and Efficiently Selecting Access Control Rules on Materialized Views over Relational Databases: The VSP-Bucket Algorithm

Algorithm VSP-Bucket implements the main task of our proposed security framework by effectively and efficiently selecting access control rules from the set R defined over the target relational database D , thus generating the access control rule (sub-)set Z over the view set V . In this Section, we present in detail VSP-Bucket. Basically, VSP-Bucket is a meaningful adaptation of the well-known view-based query rewriting bucket algorithm by Halevy [11], which has been proposed for query optimization purposes, to the specific context of security of relational databases, which is investigated in our research.

Algorithm VSP-Bucket consists of three atomic routines, called *bucketize*, *partition*, and *select*, respectively. We next describe these routines in details. *bucketize* implements the *bucketization step* of VSP-Bucket, according to which a set of *buckets*, denoted as $B(Q)$, which is exploited by the following steps of algorithm VSP-Bucket, is generated from the set of queries Q . Given a query $q \leftarrow R_{k_0} \wedge \dots \wedge R_{k_{U-1}} \wedge P_{c_0} \wedge \dots \wedge P_{c_{W-1}}$ in Q , a bucket $b_{q,i}$ associated to q is a *container of access control rules* in R corresponding to a predicate R_{k_j} in $R(q)$. For the sake of simplicity, we again denote as R_{k_i} the predicate associated to a bucket $b_{q,i}$. Therefore, given a query q , the bucket set of q , denoted by $B(q)$, is obtained by generating one bucket for each predicate R_{k_j} in $R(q)$, thus finally generating $|R(q)|$ buckets (it should be noted that, for each query q in Q , $|B(q)| = |R(q)|$). This process is iterated for each query q in Q . As a consequence, $B(Q) = \bigcup_{k=0}^{|Q|-1} B(q_k)$.

partition implements the *partitioning step* of VSP-Bucket, according to which, for each query q in Q , access control rules in R are partitioned with respect to buckets of the bucket set $B(q)$, thus generating a partitioned representation of R , denoted by $\Phi(R,q)$. By iterating this process for each query q in Q , we finally obtain the partition set $\Phi(R,Q) = \bigcup_{k=0}^{|Q|-1} \Phi(R,q_k)$. Now, focus the attention on how a singleton partitioned representation of R $\Phi(R,q)$ is generated. To this end, we make use of the classical *unification procedure* proposed in logic programming. Here, we briefly recall the unification concept. Given two predicates R_{k_i} and R_{k_j} , with $i \neq j$, we say that R_{k_i}

and R_{k_j} can be unified, denoted as $R_{k_i} \oplus R_{k_j}$, if the following conditions are true: (i) predicate symbols match, i.e. $R_{k_i} \equiv R_{k_j}$; (ii) R_{k_i} and R_{k_j} have the same number of attributes, i.e. $|R_{k_i}| = |R_{k_j}|$; (iii) for each pair of attributes $A_{k_i,l}$ and $A_{k_j,l}$, with $0 \leq l \leq |R_{k_i}|-1$ (equivalently, $0 \leq l \leq |R_{k_j}|-1$), a matching between $A_{k_i,l}$ and $A_{k_j,l}$ exists. Furthermore, it is necessary to introduce the concept of *comparison predicate compatibility*. Given two comparison predicates P_{C_i} and P_{C_j} , with $i \neq j$, defined on the same attribute set Y (i.e., $P_{C_i} = P_{C_i}(Y)$ and $P_{C_j} = P_{C_j}(Y)$), we say that P_{C_i} and P_{C_j} are *compatible*, denoted as $P_{C_i} \approx P_{C_j}$, if conditions they express are *not disjointed*.

Based on the concepts/constructs introduced above, the partitioning step works as follows. Given a query $q(X^q) \leftarrow R_{k_0}^q(Y_{k_0}^q) \wedge \dots \wedge R_{k_{U-1}}^q(Y_{k_{U-1}}^q) \wedge P_{C_0}^q(A_{h_0}^q) \wedge \dots \wedge P_{C_{W-1}}^q(A_{h_{W-1}}^q)$ in Q , such that $Y_{k_j}^q$, with $0 \leq k_j \leq U-1$, denotes the set of attributes of $R_{k_j}^q$, its bucket set $B(q) = \{b_{q,0}, b_{q,1}, \dots, b_{q,|B(q)|-1}\}$, and the access control rule set R , the baseline goal of the partitioning step consists in partitioning rules in R with respect to buckets in $B(q)$, thus generating the set $\Phi(R,q)$. Therefore, at the end of this process, each bucket $b_{q,l}$ in $B(q)$ contains a set of rules from R , denoted by $R(b_{q,l})$. Note that, since predicates are defined from relations of the target database D , and bodies of rules in R can contain the *same* predicates, buckets in $B(q)$ can have non-null intersections (i.e., $R(b_{q,i}) \cap R(b_{q,j}) = \emptyset$, such that $i \neq j$). As a consequence, $\Phi(R,q)$ and $\Phi(R,Q)$ are, formally, *degenerate partitions*. In our partitioning approach, a rule $r(X^r) \leftarrow R_{k_0}^r(Y_{k_0}^r) \wedge \dots \wedge R_{k_{G-1}}^r(Y_{k_{G-1}}^r) \wedge P_{C_0}^r(A_{h_0}^r) \wedge \dots \wedge P_{C_{J-1}}^r(A_{h_{J-1}}^r)$ in R belongs to a bucket $b_{q,l}$ in $B(q)$ of the query q if the following conditions are true: (i) there exists a predicate $R_{k_g}^r$ in r that can be unified with the predicate $R_{k_l}^q$ associated to buckets in $b_{q,l}$; (ii) r and q are *compatible*. Specifically, condition (ii) must be tested only if condition (i) is true. We say that a rule r and a query q are compatible if the following conditions are true: (i) let $P^q(r)$ denote the sub-set of comparison predicates of r , $P(r)$, that match with a sub-set of comparison predicates of q , $P(q)$, denoted by $P^r(q)$ – predicates in $P^q(r)$ and predicates in $P^r(q)$ are compatible, i.e. $\nexists P_{C_w}^q \in P^q(r) \wedge P_{C_j}^r \in P^r(q) : P_{C_w}^q \neq P_{C_j}^r$; (ii) $X^r \subseteq X^q$; (iii) there exists a predicate $R_{k_g}^r$ in r such that its set of attributes $Y_{k_g}^r$ is a sub-set of X^r , i.e. $Y_{k_g}^r \subseteq X^r$. If conditions (i), (ii) and (iii) are true, then the rule r is finally “inserted” into the bucket $b_{q,l}$ in $B(q)$ (i.e., $r \in R(b_{q,l})$) and, as a consequence, $r \in \Phi(R,q)$ and $r \in \Phi(R,Q)$. By iterating the process above for each rule r in R with respect to the query q in Q , we obtain the set $\Phi(R,q)$ (and, as a consequence, the bucket set $B(q)$ is determined), and, by iterating the whole process for each query q in Q we finally obtain the set $\Phi(R,Q)$ (and, as a consequence, the bucket set $B(Q)$ is determined) that constitutes a partitioned representation of R with respect to Q .

select implements the *selection step* of VSP-Bucket, which is in charge of effectively selecting the set of access control rules Z (which is defined as a sub-set of R) over the view set V from access control rules contained in the bucket set $B(Q)$. This process is accomplished by selecting, from each bucket set $B(q)$ in $B(Q)$, a sub-set of rules suitable to V , denoted by $Z(q)$. By iterating this process for each bucket set $B(q)$

ALGORITHM VSP-Bucket

Input: The relational database D ; the access control rule set R ;
the query set Q .

Output: The set of access control rules over the
view set V defined by Q, Z .

Begin

```

 $Z \leftarrow \emptyset;$             $Z(q) \leftarrow \emptyset;$ 
 $q \leftarrow \emptyset;$       $\mathbb{C}_{B(q)} \leftarrow \emptyset;$ 
 $B(q) \leftarrow \emptyset;$    $\Omega_{B(q)} \leftarrow \emptyset;$ 
 $B(Q) \leftarrow \emptyset;$    $k \leftarrow 0;$ 
 $\Phi(R, q) \leftarrow \emptyset;$    $l \leftarrow 0;$ 
 $\Phi(R, Q) \leftarrow \emptyset;$ 
for  $k = 0..|Q| - 1$  begin
   $q \leftarrow Q.getQuery(k);$ 
   $B(q) \leftarrow buildBucketSet(q);$ 
   $B(Q).add(B(q));$ 
end;
for  $k = 0..|Q| - 1$  begin
   $B(q) \leftarrow B(Q).getBucketSet(k);$ 
   $\Phi(R, q).partition(R, B(q));$ 
   $\Phi(R, Q).add(\Phi(R, q));$ 
end;
for  $k = 0..|Q| - 1$  begin
   $q \leftarrow Q.getQuery(k);$ 
   $Z(q) \leftarrow \emptyset;$ 
   $B(q) \leftarrow B(Q).getBucketSet(k);$ 
   $\mathbb{C}_{B(q)} \leftarrow computeCartesianProduct(B(q));$ 
   $\Omega_{B(q)} \leftarrow computeCombinedConjRuleSet(\mathbb{C}_{B(q)});$ 
  for  $l = 0..|\Omega_{B(q)}| - 1$  begin
     $r_{c,l} \leftarrow \Omega_{B(q)}.getConjQuery(l);$ 
    if  $\{r_{c,l}.contains(q) == \text{TRUE}\}$  then
       $Z(q).add(r_{c,l});$ 
    endif;
  end;
   $Z.add(Z(q));$ 
end;
return  $Z;$ 

```

End

Fig. 1. Algorithm VSP-Bucket

in $B(Q)$, we finally obtain the access control rule set Z as $Z = \bigcup_{k=0}^{|Q|-1} Z(q_k)$. To this end, given a bucket set $B(q) = \{b_{q,0}, b_{q,1}, \dots, b_{q,|B(q)|-1}\}$ we introduce the concept of *combined conjunctive rule set*, denoted by $\Omega_{B(q)}$, which is obtained by means of the *Cartesian product* among buckets in $B(q)$, denoted by $\mathbb{C}_{B(q)} = b_{q,0} \times b_{q,1} \times \dots \times b_{q,|B(q)|-1}$. Specifically, each (conjunctive) rule $r_{c,l}$ in $\Omega_{B(q)}$ is defined by combining rules from buckets in $B(q)$ belonging to the *same* tuple in $\mathbb{C}_{B(q)}$, i.e. $r_{c,l} := r_{u_0} \wedge r_{u_1} \wedge \dots \wedge r_{u_{|B(q)|-1}}$ such that $\exists t_l \in \mathbb{C}_{B(q)} : t_l = (r_{u_0}, r_{u_1}, \dots, r_{u_{|B(q)|-1}})$. We name rules $r_{u_0}, r_{u_1}, \dots, r_{u_{|B(q)|-1}}$ as *primitive rules* of $r_{c,l}$. It should be noted that $|\Omega_{B(q)}| = |\mathbb{C}_{B(q)}|$. Based on the set $\Omega_{B(q)}$, we obtain the access control rule (sub-)set $Z(q)$ by checking, for each rule $r_{c,l}$ in $\Omega_{B(q)}$, the following condition: $q \subseteq r_{c,l}$. If this condition is true, then the primitive rules of $r_{c,l}$, $r_{u_0}, r_{u_1}, \dots, r_{u_{|B(q)|-1}}$, are finally selected and added to $Z(q)$. This process is then iterated for each query q in Q , so that the final set of access control rules Z over the view set V is obtained.

Finally, Fig. 1 shows the pseudo-code of algorithm VSP-Bucket. Based on previous detailed descriptions, procedures and routines embedded in the pseudo-code of Fig. 5 are self-describing, thus they do not deserve further details.

5 Experimental Evaluation and Analysis

In this Section, we provide a comprehensive experimental assessment and analysis of algorithm VSP-Bucket, which represents the core of our proposed security framework. Indeed, it is easy to understand how the efficiency of the framework largely depends on the efficiency of algorithm VSP-Bucket. Also, it should be noted that corporate databases are typically very large in size, and, in addition to this, we usually experience a large size in views and a large number of access control rules defined over such databases. The convergence of these three separate factors makes gaining efficiency a critical requirement for algorithm VSP-Bucket. Inspired by these considerations, in our experimental assessment we stressed both the *performance* and the *scalability* of VSP-Bucket, having recognized these factors as critical in the context of handling the security of very large databases.

First, our experimental infrastructure was composed by an *AMD Turion X2 Dual-Core RM-70* at 2.00 GHz and 4.00 GB RAM running *Ubuntu 9.10* at 64 bits. In order to achieve a reliable experimental assessment, we captured and formalized a number of parameters characterizing our security framework. In particular, we focused the attention on parameters modeling the access control rule set R defined on the target database D , and the query set Q that defines the view set V , respectively. Here, we outline these parameters. For what regards rules, the following parameters were introduced: (i) N_R^R , which models the number of access control rules defined over the target database D ; (ii) $N_{H,r}^R$, which models the number of attributes in the head of the rule r ; (iii) $N_{G,r}^R$, which models the number of (standard) predicates of the rule r ; (iv) $N_{P,r}^R$, which models the number of comparison predicates of the rule r ; (v) $N_{A,r}^R$, which models the total number of attributes of the rule r . For what regards queries, the following parameters were introduced: (i) N_R^Q , which models the number of queries in Q that define the view set V over D ; (ii) $N_{H,q}^Q$, which models the number of attributes in the head of the query q ; (iii) $N_{G,q}^Q$, which models the number of (standard) predicates of the query q ; (iv) $N_{P,q}^Q$, which models the number of comparison predicates of the query q ; (v) $N_{A,q}^Q$, which models the total number of attributes of the query q .

Given the target database D having schema S , we generated from S the set of access control rules R and the set of queries Q via *random sampling* over S by generating both (standard) predicates and comparison predicates directly from relations R_i in D , with $0 \leq i \leq |D|-1$, and from the domains of attributes and constants of the universe of D , denoted by $\mathbb{U}(D)$. In order to ensure an *effective* experimental evaluation, we introduced the concept of *confidence of access control rules* in R with respect to queries

$$\begin{array}{ll}
 N_{H,r}^R \in [1:3]; N_{G,r}^R \in [1:5]; N_{P,r}^R \in [1:3]; & N_{H,r}^R \in [1:3]; N_{G,r}^R \in [1:5]; N_{P,r}^R \in [1:3]; \\
 N_{A,r}^R \in [4:8]; N_R^Q \in [10:30]; N_{H,q}^Q \in [1:5]; & N_{A,r}^R \in [4:8]; N_R^Q \in [10:30]; N_{H,q}^Q \in [1:5]; \\
 N_{P,q}^Q \in [1:5]; N_{A,q}^Q \in [4:8]; & N_{P,q}^Q \in [1:5]; N_{A,q}^Q \in [4:8];
 \end{array}$$

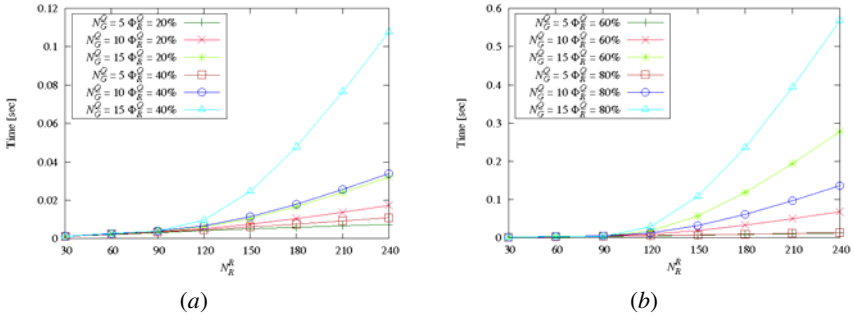


Fig. 2. VSP-Bucket performance for the base setting of the experimental assessment

in Q , denoted by Φ_R^Q . Formally, Φ_R^Q models the percentage of rules r in R related to queries q in Q , i.e. $R(r) \cap R(q) \neq \emptyset$ and $P(r) \cap P(q) \neq \emptyset$, where $R(y)$ and $P(y)$ denote the set of predicates and the set of comparison predicates of y , respectively, with y in $\{r, q\}$.

We conducted our experimental evaluation as follows. We executed a *large* number of (experimental) runs such that, for each run, we randomly ranged the number of queries N_R^Q on the interval $[10:30]$ and the number of rules N_R^R on the interval $[30:240]$, respectively. For each run, we measured the *absolute time* required by VSP-Bucket to select the access control rule set Z suitable to the view set V . The final metrics was obtained by *averaging* the so-determined amounts of time required by the whole collection of runs. This allowed us to assess the performance of VSP-Bucket. In order to assess the scalability of VSP-Bucket, we conducted the same experiments by ranging *all* the remaining experimental parameters modeling queries and rules (introduced above), thus finally obtaining a reliable and comprehensive experimental evaluation. Also, we considered two reference application scenarios that well describe scenarios coming from real-world relational database settings. The first one describes a scenario characterized by a *low confidence*, i.e. a low percentage of access control rules in R are related to queries in Q . In this case, we set Φ_R^Q as ranging over the interval $[20:40]$ %. The second one instead describes an application scenario characterized by an *high confidence*, i.e. an high percentage of access control rules in R are related to queries in Q . In this case, we set Φ_R^Q as ranging over the interval $[60:80]$ %.

Fig. 2 shows the performance of VSP-Bucket when ranging the number of access control rules N_R^R , for the two distinct cases of low confidence (a) and high confidence (b), under the random variation of the number of predicates of queries q in Q , $N_{G,q}^Q$, over the interval $[5:15]$, being the latter one a critical parameter in our proposed security framework. Similarly, the remaining experimental parameters were ranged randomly on given intervals. In our experimental assessment, the latter one is named as the *base setting*, meaning that the scalability of VSP-Bucket was stressed throughout suitable variations of the experimental parameters with respect to such a base setting. As shown in Fig. 2 (a) and Fig. 6 (b), VSP-Bucket allows us to select the desired access control rule set Z over the view set V very efficiently.

6 Conclusions and Future Work

Inspired by actual challenges of database security research, in this paper we have presented and experimentally assessed a novel framework for effectively and efficiently selecting access control rules on materialized views over very large relational databases. Indeed, looking at actual DBMS platform implementations (e.g., Oracle), the process of inheriting access control rules from the target database to the view set defined on such a database is still a manual task, which exposes to several security breaches that have been deeply discussed in this paper. We have also provided and experimentally assessed algorithm VSP-Bucket, which implements the main selection task of our proposed security framework. Our comprehensive experimental campaign has clearly confirmed to us the benefits deriving from applying our framework to the context of very large relational database systems, where security issues play a challenging role. Future work is oriented to a double-fold goal: (i) embedding *more complex* predicates in both queries and access control rules actually supported by our proposed framework, beyond the simple-yet-relevant ones considered in this research; (ii) integrating our proposed framework with a real-life DBMS platform, like Oracle, in order to study how the effectiveness and the efficiency of access control can be combined with *privacy preserving aspects* of relational databases.

References

1. Agrawal, R., Bird, P., Grandison, T., Kiernan, J., Logan, S., Rjaibi, W.: Extending Relational Database Systems to Automatically Enforce Privacy Policies. In: Proc. of ICDE 2005, pp. 1013–1022 (2005)
2. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Hippocratic Databases. In: Proc. of VLDB 2002, pp. 143–154 (2002)
3. Ahmad, M., Aboulnaga, A., Babu, S., Munagala, K.: Modeling and Exploiting Query Interactions in Database Systems. In: Proc. of CIKM 2008, pp. 183–192 (2008)
4. Ayyagari, P., Mitra, P., Lee, D., Liu, P., Lee, W.-C.: Incremental Adaptation of XPath Access Control Views. In: Proc. of ASIACCS 2007, pp. 105–116 (2007)
5. Castano, S., Fugini, M., Martella, G., Samarati, P.: Database Security. Addison Wesley, Reading (1995)
6. Chandra, A.K., Merlin, P.M.: Optimal Implementation of Conjunctive Queries in Relational Data Bases. In: Proc. of STOC 1977, pp. 77–90 (1977)
7. Fan, W., Chan, C.-Y., Garofalakis, M.: Secure XML Querying with Security Views. In: Proc. of SIGMOD 2004, pp. 587–598 (2004)
8. Goel, S.K., Clifton, C., Rosenthal, A.: Derived Access Control Specification for XML. In: Proc. of XMLSEC 2003, pp. 1–14 (2003)
9. Gupta, A., Mumick, I.S.: Materialized Views: Techniques, Implementations, and Applications. The MIT Press, Cambridge (1999)
10. Gupta, H.: Selection of Views to Materialize in a Data Warehouse. In: Afrati, F.N., Kolaitis, P.G. (eds.) ICDT 1997. LNCS, vol. 1186, pp. 98–112. Springer, Heidelberg (1996)
11. Halevy, A.: Answering Queries Using Views: A Survey. The VLDB Journal 10, 270–294 (2001)
12. Jarke, M., Koch, J.: Query Optimization in Database Systems. ACM Computing Surveys 16(2), 111–152 (1984)

13. Kabra, G., Ramamurthy, R., Sudarshan, S.: Redundancy and Information Leakage in Fine-Grained Access Control. In: Proc. of SIGMOD 2006, pp. 133–144 (2006)
14. Matthias, A., Onur, K., Yi, P.: Approaching Fine-grain Access Control for Distributed Biomedical Databases within Virtual Environments. In: Proc. of CGW 2009, pp. 311–319 (2009)
15. Olson, L.E., Gunter, C.A., Cook, W.R., Winslett, M.: Implementing Reflective Access Control in SQL. In: Gudes, E., Vaidya, J. (eds.) Data and Applications Security XXIII. LNCS, vol. 5645, pp. 17–32. Springer, Heidelberg (2009)
16. Oracle Corp.: The Virtual Private Database in Oracle9iR2: A Technical White Paper (2002), <http://www.cgisecurity.com/database/oracle/pdf/VPD9ir2twp.pdf>
17. Pottinger, R., Halevy, A.: MiniCon: A Scalable Algorithm For Answering Queries Using Views. The VLDB Journal 10, 182–198 (2001)
18. Rastogi, V., Suci, D., Welbourne, E.: Access Control over Uncertain Data. In: Proceedings of the VLDB Endowment, vol. 1, pp. 821–832 (2008)
19. Rizvi, S., Mendelzon, A., Sudarshan, S., Roy, P.: Extending Query Rewriting Techniques for Fine-Grained Access Control. In: Proc. of SIGMOD 2004, pp. 551–562 (2004)
20. Roichman, A., Gudes, E.: Fine-Grained Access Control to Web Databases. In: Proc. of SACMAT 2007, pp. 181–184 (2007)
21. Rosenthal, A., Sciore, E.: Abstracting and Refining Authorization in SQL. In: Jonker, W., Petković, M. (eds.) SDM 2004. LNCS, vol. 3178, pp. 148–162. Springer, Heidelberg (2004)
22. Sagiv, Y., Yannakakis, M.: Equivalences Among Relational Expressions with the Union and Difference Operators. Journal of the ACM 27, 633–655 (1980)
23. Stonebraker, M., Wong, E.: Access Control in a Relational Data Base Management System by Query Modification. In: Proc. of ACM 1974, vol. 1, pp. 180–186 (1974)
24. Sybase Corp.: New Security Features in Sybase Adaptive Server Enterprise. Sybase Technical White Paper (2003)
25. Wang, Q., Yu, T., Li, N., Lobo, J., Bertino, E., Irwin, K., Byun, J.-W.: On the Correctness Criteria of Fine-Grained Access Control in Relational Databases. In: Proc. of VLDB 2007, pp. 555–556 (2007)
26. Zannone, N., Jajodia, S., Massacci, F., Wijesekera, D.: Maintaining Privacy on Derived Objects. In: Proc. of WPES 2005, pp. 10–19 (2006)

MySQL Data Mining: Extending MySQL to Support Data Mining Primitives (Demo)

Alfredo Ferro, Rosalba Giugno, Piera Laura Puglisi, and Alfredo Pulvirenti

Dept. of Mathematics and Computer Sciences,
University of Catania
{ferro,giugno,lpuglisi,apulvirenti}@dmi.unict.it

Abstract. The development of predictive applications built on top of knowledge bases is rapidly growing, therefore database systems, especially the commercial ones, are boosting with native data mining analytical tools. In this paper, we present an integration of data mining primitives on top of MySQL 5.1. In particular, we extended MySQL to support frequent itemsets computation and classification based on C4.5 decision trees. These commands are recognized by the parser that has been properly extended to support new SQL statements. Moreover, the implemented algorithms were engineered and integrated in the source code of MySQL in order to allow large-scale applications and a fast response time. Finally, a graphical interface guides the user to explore the new data mining facilities.

Keywords: Data Mining, MySQL, APRIORI, Decision trees.

1 Introduction

Commercial database systems such as Oracle¹ and SQL Server² are equipped with a wide range of native data mining primitives. They provide predictive analytical tools equipped with graphical user interface allowing to access and explore data to find patterns, relations and hidden knowledge.

On the open source databases front, a widely used system, e.g. MySQL, lacks of such data mining primitives. Some basic mining tasks may be performed by facing complex SQL queries, others could be issued through well known stand alone suites such as WEKA [6] and RAPIDMINER³. However those suites are useful only for prototyping since they do not scale well on the size of the data and result unsuitable for most applications.

In this paper, we present MySQL Data Mining⁴, a web-based tool that performs an integration of *Frequent itemset computation* [1] and *Classification* based on C4.5 [3] algorithm on top of MySQL.

¹ http://www.oracle.com/lang/it/solutions/business_intelligence/data-mining.html

² <http://www.microsoft.com/italy/server/sql/default.msp>

³ <http://rapid-i.com/>

⁴ <http://kdd.dmi.unict.it/tedata/>

These algorithms were implemented in C++ and integrated in the standard distribution of MySQL version 5.1 on Linux OS. In order to execute these commands, the parser of MySQL server was modified and extended to support new SQL statements. Moreover, the implemented algorithms were engineered to allow large-scale applications and a fast response time. Finally, a graphical interface guides the user to explore the new data mining facilities.

The demo is organized as follows. Section 2 briefly reviews the data mining algorithms. Section 3 describes the new SQL statements and the main steps of the integration process. Section 4 shows the navigation of the graphical interface. Finally, section 5 reports conclusions and propose future extensions.

2 Data Mining Algorithms in MySQL Data Mining

This section briefly describes the data mining algorithms integrated in MySQL. **The Frequent Itemsets computation algorithm.** Mining frequent itemsets can support business decision-making processes such as cross-marketing or analyses on customer buying behavior. APRIORI is an algorithm proposed by [1] for finding all frequent itemsets in a transactional database. It uses an iterative *level-wise* approach based on candidates generation exploring $(k + 1)$ -itemsets from previously generated k -itemsets. Let L_k be a set of frequent k -itemsets and C_k a set of candidate k -itemsets. Our implementation consists of two steps:

1. *Join Step*: find C_k by joining L_{k-1} with itself [1].
2. *Find Step*: find L_k , i.e. a subset of C_k of frequent itemsets. This step is implemented following the strategy presented in MAFIA [2]. It uses a vertical bitmap representation of transactions and performs bitwise *AND* operations to determine the frequency of the itemsets.

The algorithm iterates until $L_k = \emptyset$.

Classification based on C4.5 algorithm. Classification allows to extract models describing important data classes that can be used for future predictions. A typical example is a classification model to categorize bank loan applications as either safe or risky.

Data classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing a training set consisting of a set of tuples and their associated class labels. In the second step, the model is used for classification. First, the predictive accuracy of the classifier is estimated, then if the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.

MySQL was extended to support data classification using the implementation of algorithm C4.5 of Quinlan [3]. This algorithm uses decision tree as classifiers, in which internal nodes are tests on an attribute and branches represent the outcomes of the test. Leaf nodes contain a class label.

3 Integrating Data Mining Algorithms into MySQL

MySQL architecture consists of five subsystems (the *query engine*, the *storage, buffer, transaction, and recovery managers*) that interact with each other in order to accomplish the user tasks. In particular, the query engine contains the *syntax parser*, the *query optimizer* and the *execution component*. The syntax parser decomposes the received SQL commands into a form that can be understood by the MySQL engine. The query optimizer prepares the execution plan and passes it to the execution component which interprets and retrieves the records by interacting with the storage manager.

The integration of new data mining procedures required the following steps:

1. implementation and optimization of the algorithms described in section 2;
2. definition of new SQL statements for the execution of 1. and extension of Bison grammar file (MySQL *syntax parser*);
3. integration of 1. in the MySQL server by modifying the *query engine*.

The first step is described in section 2. Next sections describe the other steps.

3.1 Extension of MySQL Syntax Parser

As an example of computation, Figure 2 shows the main phases for the integration of Apriori. In order to define the new command, we modified parser by:

- extending MySQL’s list of symbols with new keywords (e.g. APRIORI): lexical analysis recognizes new symbols after defining them as new MySQL keywords;
- adding to the parser (i.e. Bison grammar file) new grammar rules for the introduced primitive: the parser matches the grammar rules corresponding to SQL statements and executes the code associated to that rules.

The syntax of SQL statement for APRIORI is reported in Figure 1 (a). The *table_name* represents the input table for Apriori. Moreover, the user has to specify the minimum support (*threshold*) and the columns containing transaction ids (*col_name_tid*) and item ids (*col_name_item*). Other optional parameters can be used to (i) limit the size of the itemsets to compute, (ii) report other information related to items (such as the details of the related transactions) and

SQL APRIORI	SQL Generation Model
APRIORI table_name threshold [max_itemset_size] ON col_name_tid, col_name_item [optional_fields] [TYPE= storage_engine]	CREATE DTREE training_table BY class_name [GR]
	SQL Classification
	CLASSIFY new_table BY rules_table

(a)

(b)

Fig. 1. SQL statements syntax. (a) Apriori command. (b) Create model and classify commands.

(iii) specify the type of storage engine (myISAM, InnoDB etc). Default storage engine is MyISAM. This command is recognized and executed by MySQL Data Mining. The result is then stored into the database and will be available to the user for further analysis sessions.

Figure 1(b) reports also the syntax of SQL statements for training and classification phases. The integration of these SQL commands followed exactly the same steps of Apriori integration. Here, the generation of the model requires the specification of the training set (*training_table*) and the attribute representing the class (*class_name*). The model implements Information Gain (default one) and the Gain Ratio (*GR*) as splitting conditions. In this phase, classification rules are stored into the database. Next, the classification of data tuples (*new_table*) is performed by selecting a previously generated model (*rules_table*).

3.2 Extension of MySQL Execution Component

Figure 2 reports the modifications to MySQL Engine. Here, the MySQL Execution Component was modified in the following way: (i) main MySQL procedures (server side) were modified to support the execution of a new command Apriori; (ii) new C++ code (computing frequent itemsets as described in section 2) was added to the standard distribution. Moreover, CREATE and INSERT SQL statements were executed at the low level of the engine in order to store the frequent itemsets in a relational table (that will be available to the user for querying the result).

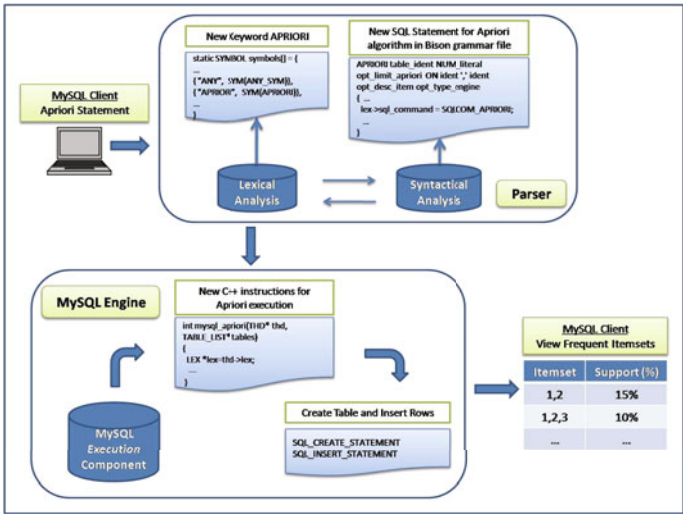


Fig. 2. Framework of MySQL: syntax parser and MySQL engine modifications

4 The User Interface

We equipped MySQL Data Mining with a web interface based on the LAMP⁵ framework. The user starts by logging into the system using his MySQL account. Then, he selects the database and the data mining algorithms to use. The web interface contains also a loader to upload data coming from external sources. The results of each task are stored in the database and will be available to the user in future sessions. This demo is available online⁶.

4.1 The Frequent Itemsets Computation Web Interface

APRIORI interface includes the following modules:

1. *data preparation module*: the user can load data from external sources or choose the data from tables stored in the database. The input table for Apriori must have at least two columns, in which the first one contains the transaction ID and the second one contains the item ID (see Fig. 3 (b));
2. *statement preparation module*: the user can set the input parameters to generate the frequent itemsets, that is the input table, the fields representing the transactions and the items, the support threshold and an optional limit to the size of the frequent itemsets (see Fig. 3 (a));

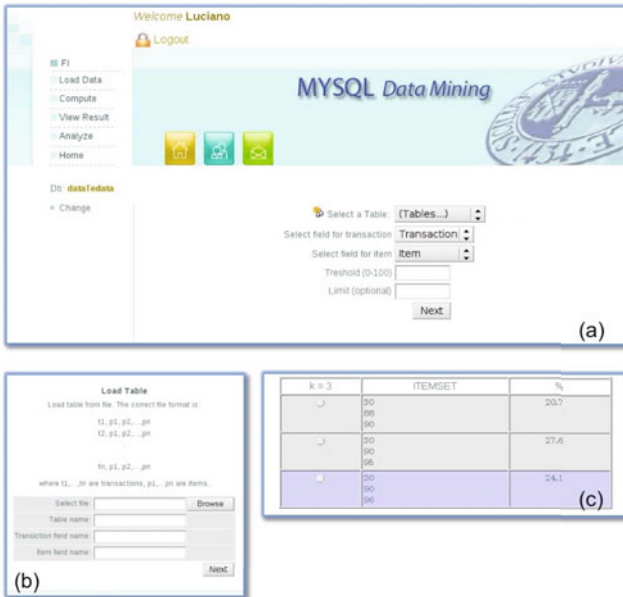


Fig. 3. (a) Statement preparation module. (b) Data preparation module. (c) Data analysis module.

⁵ <http://www.apachefriends.org/it/xampp-linux.html>

⁶ <http://kdd.dmi.unict.it/tedata/> use (guest/guest10) to access.

- 3. *data analysis module*: it is possible to visualize and query the tables containing the frequent itemsets.

4.2 The Decision Tree Algorithm

The generation of the model is supported by a simple interface which guides user to select the input table and the class from the list of possible attributes. Optionally, the user can set Gain Ratio as splitting condition. Moreover, the user can create an input table by getting the schema and the data from external files (.names, .data). The classification is performed by choosing a previously generated model and a set of tuples to be classified. Such tuples are provided by the user into a table which schema must be consistent with the selected model and must contain a column corresponding to the classifier attribute.

5 Performance Analysis

In this section, we report some preliminary experiments concerning the performances on FI computation of our MySQL Data Mining system. Experiments have been performed on a HP Proliant DL380 with 4GB RAM, equipped with Linux Debian Operating System. We used two different benchmark datasets, called *mushroom* and *chess* respectively, obtained from the FIMI (Frequent Itemset Mining Implementations) repository⁷. The mushroom dataset contains 8124 transactions and 23 items per transaction whereas the chess dataset contains 3196 transactions and 37 items per transaction. Fig. 4 reports the running time of our MySQL Apriori (Apriori-DM) without I/O operations and the total execution time needed by the SQL statement (Apriori-SQL). We show also a comparison of our standalone Apriori algorithm (Apriori-extern) with two freely available standalone Apriori implementations of Bodon⁸ [4] and Borgelt⁹ [5] respectively. Comparisons with Mafia are not reported since the algorithm is optimized for MFI. On the mushroom dataset (Fig. 4 (a)), the Apriori-DM

Mushroom Dataset					
Threshold	Time (sec.)				
	Apriori-DM	Apriori-SQL	Apriori-extern	Bodon	Borgelt
40	0,45	0,7	0,12	0,12	0,07
30	0,97	1	0,21	0,11	0,7
20	1,61	12	1,34	0,14	1
10	15,29	19	12,57	20,27	12,57
5	99	128	61,3	156,86	61,3

(a)

Chess Dataset					
Threshold	Time (sec.)				
	Apriori-DM	Apriori-SQL	Apriori-extern	Bodon	Borgelt
90	0,03	0,5	0,06	0,1	0,09
80	0,5	0,8	0,20	0,13	0,1
70	0,77	1	0,94	0,7	0,66
60	3,4	5	4,89	14,98	9,63
50	16,57	26	25,19	140	71,1
40	86	102	130,06	659	323

(b)

Fig. 4. Running times varying the threshold. (a) Mushroom dataset. (b) Chess dataset.

⁷ <http://fimi.cs.helsinki.fi/>
⁸ <http://www.cs.bme.hu/~bodon/en/apriori/>
⁹ <http://www.borgelt.net/apriori.html>

outperforms Bodon [4] because of the use of bitmaps during the verification phase [2]. However, Borgelt implementation yields the best results. On the chess dataset (Fig. 4 (b)), Apriori-DM and Apriori-SQL outperform all the standalone implementations. This is due to the fact that the number of generated FI is very high and the I/O operations in a DBMS are faster than I/O operations on text files.

6 Conclusions and Future Work

We have presented an integration of data mining algorithms on MySQL. Differently from other database systems, MySQL lacks of such features. Although user may overcome such unavailability, it could result unsuitable for most applications. The main advantages of this approach rely on fact that user can use simple SQL commands to perform complex data mining analysis. Future work includes integration of a wider range of data mining algorithms together with statistical primitives.

Acknowledgement

We thank all students that have collaborated on the development of the system, in particular Aurelio Giudice, Luciano Gusmano, Antonino Schembri, and Tiziana Zapperi.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proc. of the 20th VLDB Conf., pp. 487–499 (1994)
2. Doug Burdick, M.C., Gehrke, J.: Mafia: A maximal frequent itemset algorithm for transactional databases. In: Proc. of the 17th International Conference on Data Engineering, pp. 77–90 (April 2001)
3. Quinlan, J.: Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research* 4, 77–90 (1996)
4. Bodon, F.: Surprising results of trie-based FIM algorithms. In: 2nd Workshop of Frequent ItemSet Mining Implementations (FIMI 2004), Brighton, UK (2004)
5. Borgelt, C.: Recursion Pruning for the Apriori Algorithm. In: 2nd Workshop of Frequent ItemSet Mining Implementations (FIMI 2004), Brighton, UK (2004)
6. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco (2005)

A Genetic Algorithm to Design Industrial Materials

E. Tenorio, J. Gómez-Ruiz, J.I. Peláez, and J.M. Doña

Department of Languages and Computer Sciences
University of Malaga, Malaga 29071, Spain
etgil@hotmail.com, {janto,jignacio,jmdona}@lcc.uma.es

Abstract. At present, the development of our society is still marked by the need for lighter and stronger structures with a minimum manufacturing cost. The materials that are responding best to these needs are composite materials and as a result, these are replacing many traditional materials such as steel, wood or aluminium. Designing composite materials is difficult because it involves designing the geometry of the element and composition. Traditionally, due to the limited knowledge of these materials, these design tasks have been based on approximate methods; the possibilities for creating composite materials is almost unlimited, characterization by testing is very expensive and it is difficult to apply the results to other contexts. Due to this fact, the whole design task relies on the ability of an expert to select the best combination based on their knowledge and experience. This paper presents and compares a genetic Algorithm to design industrial materials.

Keywords: Genetic Algorithm, Heuristic, Composites, Soft Computing, Production.

1 Introduction

A composite material, or composite, is made by combining two or more materials to form another that is appropriate for a specific application. The agglomerate material is known as the matrix, and the rest are the reinforcement materials that may be made up of continuous fibers, short fibers or particles [2, 9]. When the design is good, the new material adopts the best properties of its constituents, and sometimes even some that none of these possess. The properties that the design of a composite material aims to improve on are strength, stiffness, toughness, lightness, thermal insulation, etc [12, 14]. Of course, not all of them can be improved simultaneously so the design objective is to obtain a new material that offers the best adaptation possible to the required specifications.

Although composite materials, such as the addition of straw to mud or the use of laminated wood, have been used since ancient times, the real boom has only happened recently with the development of materials based on a resin matrix reinforced by fibers [6]. The exceptional strength and lightness of these materials has led to the development of an enormous number of applications, particularly in the aeronautical and space industries due to the economic significance of these properties. Figure 1 shows the markets for composite materials.

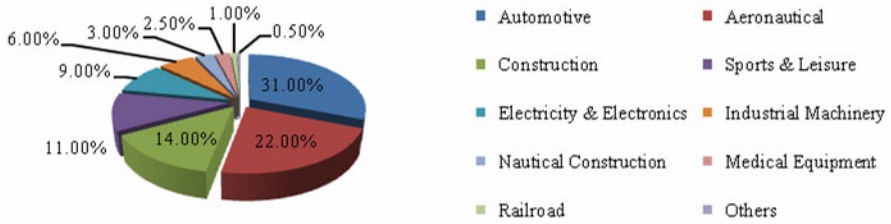


Fig. 1. Markets for composite materials

However, this type of material is difficult to develop, because the design of a new composite material involves designing both the geometry of the element and the configuration of the material itself so as to best exploit the qualities of the constituent materials. The deterioration of the properties over time due to stress, which can lead to unanticipated behaviour (breaks) or failure of the structural element in question, must also be evaluated [15,16].

The composites with unidirectional reinforcement of the fibers are particularly important [18]. These are made up of layers of the same composition, stacked and bound together by the same material that forms the matrix, but with distinct fiber orientation. From now on, this type of element will be referred to as a laminate [5]. Within this category, symmetric laminates are worthy of special attention, both those with geometrical symmetry and those with structural symmetrical relative to their mean surface.

The design process for a laminate starts with the definition of the problem to be solved and the specifications to be met by the element to be designed. From this information, a series of solutions are generated by a synthesis process that is usually, primarily, supported by the expertise of the designer. Potentially viable solutions are subsequently analysed to test their effectiveness. This analysis is made using finite element modelling based software packages, such as ANSYS [1]. This whole process is an iterative task that allows the proposed solution to be repeatedly improved until the final design is obtained.

Traditionally, the design work, both synthesis and analysis has been carried out using empirical knowledge based methods [8, 3, 11]. This is partly because the number of possible combinations of composites is almost unlimited and also because characterization by experimentation is very expensive.

Genetic Algorithms [10] are robust search techniques based on the principles of evolution that is proving rather efficient and easily to apply in a large number of situations [7, 4, 13, 17]. In this paper we propose a genetic algorithm to design symmetric laminate.

The paper is organized as follows: section 2 shows the symmetric laminate; section 3 proposes the model for designing laminates; in section 4, the model is applied to a real application and the results are compared to those obtained using different algorithms. Finally, conclusions are presented.

2 The Symmetric Laminate

A laminate is a set of several laminas tightly bound together, so that they act as a single structural element, where each lamina has its main axes forming different angles with the overall axes of the laminate (figure 2). In a symmetric laminate, layers are symmetrically arranged about the middle surface of the laminate, resulting in the lack of coupling between bending and extension, and consequent simplification of the design.

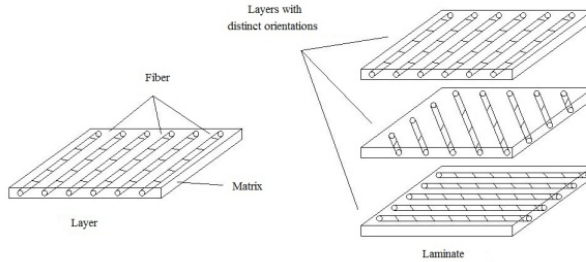


Fig. 2. Structure of a lamina and make up of a laminate

The characteristics of a laminate are calculated starting from the number of laminas, the stacking sequence and their geometric and mechanical characteristics.

For a single lamina where a plane stress state exists, the constitutive relation described in material coordinates 1, 2 is

$$\begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \tau_{12} \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} & 0 \\ Q_{12} & Q_{22} & 0 \\ 0 & 0 & Q_{66} \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \gamma_{12} \end{pmatrix}$$

In this equation, $\sigma_1, \sigma_2, \tau_{12}$ are the in-plane stresses and $\varepsilon_1, \varepsilon_2, \gamma_{12}$ the in-plane strains; the material stiffness matrix coefficients are

$$Q_{11} = \frac{E_1}{1 - \nu_{12} \cdot \nu_{21}} \quad Q_{22} = \frac{E_2}{1 - \nu_{12} \cdot \nu_{21}} \quad Q_{12} = \frac{\nu_{21} \cdot E_1}{1 - \nu_{12} \cdot \nu_{21}} = \frac{\nu_{12} \cdot E_2}{1 - \nu_{12} \cdot \nu_{21}} \quad Q_{66} = G_{12}$$

where E_1 and E_2 are the Young's modulus in the principal material 1 and 2 directions, ν_{12} and ν_{21} are the Poisson's ratios, and G_{12} is the shear modulus in the 1-2 plane.

In case that the main axes of the lamina, 1, 2, do not coincide with the natural axes for the solution of the problem, X, Y, the lamina stops behaving orthotropically and all the elements of the stiffness matrix [Q] take non-zero values.

In order to determine Hooke's law for the new coordinate system, the coordinate transformation matrix [T] is used, where:

$$[T] = \begin{pmatrix} \cos^2 \theta & \sin^2 \theta & 2 \cdot \sin \theta \cdot \cos \theta \\ \sin^2 \theta & \cos^2 \theta & -2 \cdot \sin \theta \cdot \cos \theta \\ -\sin \theta \cdot \cos \theta & \sin \theta \cdot \cos \theta & \cos^2 \theta - \sin^2 \theta \end{pmatrix}$$

The stiffness matrix in the new coordinate system, $[Q]$, is

$$[\bar{Q}] = \begin{pmatrix} \bar{Q}_{11} & \bar{Q}_{12} & \bar{Q}_{16} \\ \bar{Q}_{12} & \bar{Q}_{22} & \bar{Q}_{26} \\ \bar{Q}_{16} & \bar{Q}_{26} & \bar{Q}_{66} \end{pmatrix}$$

According to the classical plate theory, the strain at an arbitrary point of the laminate is

$$\begin{pmatrix} \epsilon_x \\ \epsilon_y \\ \gamma_{xy} \end{pmatrix} = \begin{pmatrix} \epsilon_x^0 \\ \epsilon_y^0 \\ \gamma_{xy}^0 \end{pmatrix} + z \cdot \begin{pmatrix} \kappa_x \\ \kappa_y \\ \kappa_{xy} \end{pmatrix}$$

Substituting these equations in the stress-strain relationships, Hooke's law allows the stresses at any point in the laminate to be calculated, providing the mid-plane deformations and curvatures at that point are known

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} = \begin{pmatrix} \bar{Q}_{11} & \bar{Q}_{12} & \bar{Q}_{13} \\ \bar{Q}_{21} & \bar{Q}_{22} & \bar{Q}_{23} \\ \bar{Q}_{31} & \bar{Q}_{32} & \bar{Q}_{33} \end{pmatrix} \cdot \left[\begin{pmatrix} \epsilon_x^0 \\ \epsilon_y^0 \\ \gamma_{xy}^0 \end{pmatrix} + z \cdot \begin{pmatrix} \kappa_x \\ \kappa_y \\ \kappa_{xy} \end{pmatrix} \right]$$

Once the stresses in each layer have been calculated, the membrane forces and moments per unit length, acting at the mid-plane of the laminate, are obtained by integration of the stresses in each layer through the laminate thickness,

$$\begin{pmatrix} N_x \\ N_y \\ N_{xy} \end{pmatrix} = \sum_{k=1}^N \int_{z_{k-1}}^{z_k} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix}_k \cdot dz; \quad \begin{pmatrix} M_x \\ M_y \\ M_{xy} \end{pmatrix} = \sum_{k=1}^N \int_{z_{k-1}}^{z_k} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix}_k \cdot z \cdot dz$$

where N is the number of layers of the laminate and z_k is the distance between the mid-plane and the k^{th} layer.

Finally, the constitutive equation of the in case of symmetric laminates, $B_{ij} = 0$, which gives

$$\begin{pmatrix} N_x \\ N_y \\ N_{xy} \\ M_x \\ M_y \\ M_{xy} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & A_{16} & 0 & 0 & 0 \\ A_{12} & A_{22} & A_{26} & 0 & 0 & 0 \\ A_{16} & A_{26} & A_{66} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & D_{11} & D_{12} & D_{16} \\ 0 & 0 & 0 & D_{12} & D_{22} & D_{26} \\ 0 & 0 & 0 & D_{16} & D_{26} & D_{66} \end{pmatrix} \begin{pmatrix} \epsilon_x^0 \\ \epsilon_y^0 \\ \gamma_{xy}^0 \\ \kappa_x \\ \kappa_y \\ \kappa_{xy} \end{pmatrix}$$

where it can be seen that there is no bending-extension coupling.

3 Proposed Genetic Algorithm

This subsection presents a genetic algorithm for the design of symmetric laminates. In order to develop this model an encoding method for the representation of the problem needs to be proposed first of all. Then, a fitness function is required to evaluate the suitability of solutions.

3.1 Coding of a Symmetrical Laminate

A string of genes or chromosome is used, which represents the structure of the laminate. This chain is made up of four genes that encode the fiber, matrix, volume fraction and geometry of the laminate. Figure 3 shows an individual, which is made up by the structure of the laminate (chromosome) and the fitness.

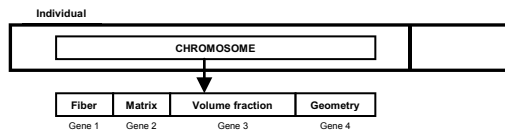


Fig. 3. Representation of an individual

Each of the four genes that make up the chromosome represents a different characteristic of the laminate's composition:

- Fiber. An integer number that identifies a fiber.
- Matrix. An integer numbers that identifies a matrix.
- Volume fraction. A real number belonging to the set $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. These values correspond to data that use the industries to design laminates.
- Geometry of the laminate. A sequence of integer numbers, one for each lamina, in the range $[-80^\circ, +90^\circ]$, with 10 increments.

3.2 Fitness of a Laminate

The design of a laminate is based on criteria that are fundamentally economic and safety related. From an economic perspective it is best that the laminate is comprised of the smallest number of layers and that the volume fraction is minimal; the fewer the number of laminas and the lower the volume fraction, the lower the cost. From a safety perspective, it is best that stresses are distributed along the direction of the fibers, as in this way the laminate will be less affected by any unanticipated stress. It is also important that residual strains, which appear as a result of stacking of layers with the same fiber orientation, do not occur during the material's curing process. In this example, no more than four sheets of this type are allowed to be stacked because it has been shown experimentally that a higher number can lead to residual stresses [2].

In this article, a fitness function is proposed that takes into account the economic and security criteria mentioned above. For this reason, whilst alignment of stresses along the direction of the fibers is favoured in the numerator, the denominator penalizes: volume fraction and high laminate thickness; lamina rupture and, finally, the

distribution of stress in directions other than that of the fibers. The higher the value of fitness function, the better the solution. The fitness function, F , proposed is as follows:

$$F = \frac{P_1^\alpha}{FV \cdot (n \cdot e)^\beta \cdot (R+1)^\gamma \cdot (P_2 + 1) \cdot (P_{12} + 1)}$$

where P_1 is the longitudinal coefficient along the direction of the fibers; P_2 is the longitudinal coefficient perpendicular to the direction of the fibers; P_{12} is the shear coefficient; FV is the laminate volume fraction; R indicates the number of layers that break and α, β, γ real value.

3.3 Crossover Operator

Crossover operator is different for each gene. In case of ‘fiber’, ‘matrix’ and ‘volume fraction’ genes, we use a classic one point crossover operator, adapted to the different cases. However, for the geometry gene, we don’t use the classic crossover operator, because it is very static in changing the number of layers. To avoid this problem, we propose a dynamic crossover operator with a different crossover point for each parent generated in a random way.

4 Experimental Results

The evolving model developed in this work is now going to be applied to a design problem. This problem has been chosen because, with minor variations, it covers a wide range of real problems.

The results obtained with the proposed model are compared with other metaheuristics in order to verify its advantages. These other models are a classic genetic algorithm with one point crossover, a simulated annealing algorithm and a tabu search algorithm. Their features are summarized below:

Classic Genetic algorithm

- Initial Population: one hundred randomly generated individuals.
- Selection of parents: roulette method.
- Crossover: randomly chosen, one point crossover.
- Mutation: random application to each child (with a probability of 1%) of one of nine, randomly chosen, operators.
- Selection of the next generation: elitist, select the ten best parents and the ninety best children to complete the one hundred individuals.
- Stopping criterion: after one hundred iterations.

Simulated annealing algorithm

- Initial solution: one hundred individuals are generated randomly and the best are selected.
- Control parameter ‘c’: initially equal to 10^{-2} , decreases quadratically with each iteration until it reaches 10^{-6} .
- Stopping criterion: After two hundred iterations.

Tabu search algorithm

- Initial solution: one hundred individuals are generated randomly and the best is selected.
- Tabu structure: the local search operators used in the last five iterations are classified as tabu.
- Stopping criterion: After one hundred iterations.

4.1 Application to a Real Problem

The problem is to correctly size the wall thickness, e , of a driveshaft, where e is small compared to the average radius, r , and subjected to an external torque T . Figure 4 shows the driveshaft along with the stresses on an element of it.

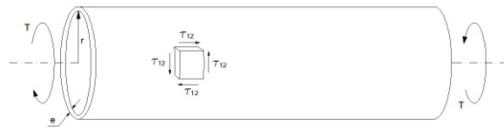


Fig. 4. Hollow driveshaft subjected to an external torque

Assuming that the stress is uniformly distributed throughout the thickness of the ring, on an element located far the edges there will only be a tangential stress with the following value:

$$\tau_{12} = \frac{T}{2 \cdot \pi \cdot r^2 \cdot e}$$

where N_{12} is the coplanar tangential force, resulting from the normal stress τ_{12} acting throughout the thickness, per unit length, calculated as follows:

$$N_{12} = \frac{T}{2 \cdot \pi \cdot r^2}$$

For example, if $T = 40000 \text{ N} \cdot \text{m}$ and $r = 0.05 \text{ m}$ then

$$N_{12} = \frac{T}{2 \cdot \pi \cdot r^2} = \frac{40000}{2 \cdot \pi \cdot 0.05^2} = 2.54648 \cdot 10^6 \text{ N/m}$$

Table 1 presents the statistics corresponding to the 300 tests completed. The results obtained with the proposed genetic algorithm are compared to the other algorithms described: a classic genetic algorithm with one point crossover operator, a simulated annealing algorithm and a tabu search algorithm. For each model, the results shown correspond to the least number of layers of the laminated material obtained, NL_{\min} , together with the number this solution is obtained, N_p , and its average, \overline{NL} . Also presented is the best (maximum) fitness obtained, F_{\max} , together with its average, \overline{F} . Comparing the results obtained in the simulations, one can observe that the proposed model offers better results because, in addition to obtaining a material with a lower number of layers in a higher proportion of trials, there is very little deviation in the solutions obtained and a much higher value for the fitness function.

Table 1. Statistics corresponding to the series of tests

Algorithm	$NL_{\min}(N_1)$	\overline{NL}	F_{max}	\overline{F}
Proposed Genetic	28 (47)	30.67	3.45	1.44
Classic Genetic	36 (2)	61.58	0.64	0.02
Simulated Annealing	44 (1)	71.48	0.0056	0.00016
Tabú Search	34 (1)	82.87	0.48	0.007

Figures 5 and 6 show the histograms corresponding to the number of laminas and the fitness of the tests made with the proposed genetic algorithm.

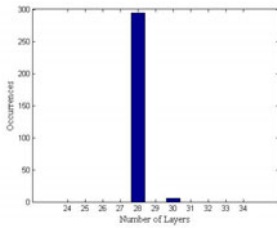


Fig. 5. Histogram of the number of laminas

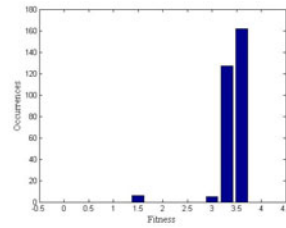


Fig. 6. Histogram of the fitness

Table 2 shows the data of the best laminate obtained with the new model: composition (fiber and matrix), volume fraction (V_f), fitness and thickness; then, they are compared with the best results obtained with the other techniques. The angles are measured relative to the circumferential direction.

Table 2. Characteristics of the best laminate obtained in a series of tests

Algorithm	Fiber Matrix	V_f	Fitness	thickness (mm)
Proposed Genetic	P-100 Peek	0.50	3.4525	5.04
Classic Genetic	P-100 Peek	0.30	0.6419	6.48
Simulated Annealing	P-100 Polyamide	0.60	0.0056	7.92
Tabú Search	P-100 Polyamide	0.40	0.4841	6.12

Finally, figures 7 to 10 show graphically the different coefficients (P_k or Tsai-Wu, P_1 , P_2 and P_{12}) corresponding to the layers of the best laminate obtained by the proposed Genetic model, and which are involved in the optimization of the laminate. In this example, a large degree of uniformity of the coefficient P_k around 1 can be observed. This indicates an excellent exploitation of all the laminas.

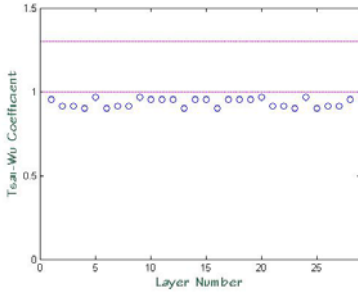


Fig. 7. P_k (Tsai-Wu) coefficient

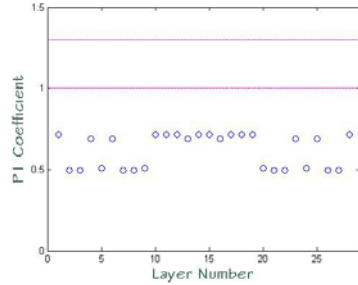


Fig. 8. P_1 coefficient

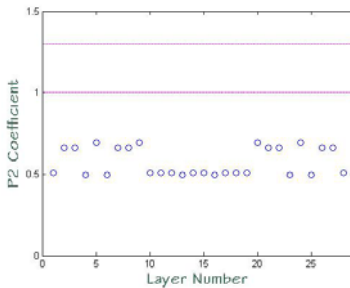


Fig. 9. P_2 coefficient

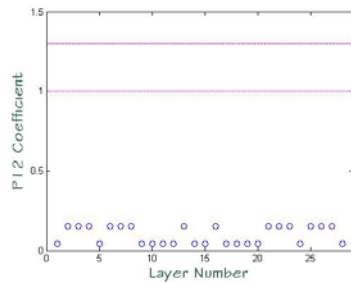


Fig. 10. P_{12} coefficient

5 Conclusions

This paper has proposed the use of a genetic algorithm with a dynamic crossover operator for the design of the geometry and composition of symmetric laminated composite materials.

The proposed model has been applied to a real situation, comparing the results obtained with different metaheuristics. The results obtained are favourable to the proposed model; the final solution it obtains is the best laminate. It should be noted that the adjustment coefficients obtained from P_k is close to 1, meaning that all laminas work near to their strength limits with the consequent highly efficient use of the material.

Considering the principal aspects of the approach, it can be concluded that the proposed genetic algorithm gives promising results.

References

1. ANSYS Multiphysics, Release 11.0. ANSYS, Inc.
2. Barbero, E.J.: Introduction to composite materials design. Taylor Francis, London (1999)
3. Conti, P., et al.: Layer thickness optimization in a laminated composite. Composites Part B, 28B, pp. 309–317. Elsevier Science Limited, Amsterdam (1997)

4. Davis, L. (ed.): *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York (1991)
5. Dietz, A.G.H.: *Composite Materials*. Edgar Marburg Lecture. American Society for Testing and Materials (1965)
6. Duratti, L., Salvo, L., Landru, D., Bréchet, Y.: Selecting the components of polymeric composites. *Advanced Engineering Materials* 4(6), 367–371 (2002)
7. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading (1989)
8. Grosset, L., Le Riche, R., Haftka, R.T.: A double-distribution statistical algorithm for composite laminate optimization. *Structural and Multidisciplinary Optimization* 31(1), 49–59 (2006)
9. Gürdal, Z.: *Design and optimization of laminated composite materials*. Wiley, New York (1999)
10. Holland, J.: *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor (1975)
11. Kim, C.W., et al.: Stacking sequence optimization of laminated plates. *Composite Structures* 39(3-4), 283–288 (1998)
12. Mallick, P.K.: *Composites engineering handbook*. University of Michigan, Dearborn (1997)
13. Michalewicz, Z.: *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, Heidelberg (1992)
14. Miravete, A.: *Materiales Compuestos*. Zaragoza (ed.), 1^a ed., Antonio Miravete (2000)
15. Puck, A., Schurmann., H.: Failure analysis of FRP laminates by means of physically based phenomenological models. *Compos. Sci. Technol.* 58, 1045–1068 (1998)
16. Puck, A., Schurmann., H.: Failure analysis of FRP laminates by means of physically based phenomenological models. *Compos. Sci. Technol., Part B* 62, 1633–1672 (2002)
17. Reeves, C.: *Modern Heuristic Techniques for Combinatorial Problems*. Blackwell Scientific Publications, Malden (1993)
18. Revuelta, D.: *Refuerzos y matrices*. *Materiales Compuestos Avanzados en la Construcción*. Madrid, Instituto de Ciencias de la Construcción Eduardo Torroja (2004)

A Proposal of P2P Content Retrieval System Using Access-Based Grouping Technique

Takuya Sasaki¹, Jun Sawamoto¹, Takashi Katoh¹, Yuji Wada²,
Noriyoshi Segawa¹, and Eiji Sugino¹

¹ Faculty of Software and Information Science, Iwate Prefectural University, Japan

² Department of Information Environment, Tokyo Denki University, Japan
g031e078@edu.soft.iwate-pu.ac.jp, sawamoto@iwate-pu.ac.jp

Abstract. In recent years, the diversification and ubiquitousness of information is rapidly advancing due to the development of the information and communication technology. As for the usage of the information, the attention of sharing of information using P2P network rises, and it is becoming to be used in many fields. The distributed hash table (DHT) is one of the typical overlay networks on P2P. However, there is a problem of lacking of flexibility in the retrieval on DHT. In this paper, to improve the convenience of the content access, we propose an efficient content access method based on the grouping technique of the contents using the content access history and frequency and the evaluation of grouping efficiency (precision, recall) is performed by the simple simulation.

Keywords: P2P network, distributed hash table, overlay network, content retrieval, grouping technique.

1 Introduction

In recent years, for the user who acts while moving around with a mobile device, the free mobile computing environment with an ad hoc communication like wireless LAN and the short distance wireless access, etc. are coming to be achieved and removing the restraint on user's movement [1].

As for the usage of the information, the attention of sharing of information using P2P network rises, and it is becoming to be used in many fields [2], [3]. P2P networks directly connect computers in an equal relationship and the load is distributed by exchanging information and services among computers, and high fault tolerance is also one of the advantages. P2P networks can be classified into two types; unstructured and structured overlay network, by the transfer method of the search messages. Flooding is a fundamental search method in unstructured P2P systems [4], [5]. However, the searching cost is high by generating a huge amount of network traffic. Structured P2P network employ a globally consistent protocol to ensure that any node can efficiently route a search to some peer that has the desired content. The most common type of structured P2P network is the DHT [6], [7], in which a variant of consistent hashing is used to assign ownership of each content to a particular peer. Therefore, DHT can search content with the number of very few messages in comparison with

the flooding method; however, the perfect matching with the search keyword is demanded in the hush table. The DHT has the drawbacks that the partial matching or the group matching retrieval cannot be achieved. It is difficult to do a flexible retrieval. This becomes a problem that should be solved when the search engine is constructed.

In this paper, to improve the flexibility of the content retrieval, we propose an efficient content access method based on the grouping technique of the contents, i.e., group matching or group recommendation, using the content access history and frequency. We consider mechanism of automatic generation and reformation of group structure based on the users' access records. When a user accesses data, we can consider a set of continuous data accesses as a candidate for grouping. When not only a single user but also multiple users are doing a similar set of accesses, they could be registered in the DHT as a group. Less frequently accessed groups are deleted from the table. In the present study, it is targeted to simulate and to evaluate whether the automatic grouping on the DHT is carried out as we assumed in P2P environment.

2 Content Grouping in P2P Network

P2P overlay networks are distributed systems without any hierarchical organization or centralized control. DHT-based systems can guarantee that any data object can be located in a small $O(\log N)$ overlay hops on average, where N is the number of peer nodes in the system.

2.1 Contents Grouping Method

In this paper, the object of the grouping is the contents that nodes of P2P network maintain. Grouping of nodes (contents) of P2P network is based on dynamic access history and access frequency by user nodes. We assume that it is possible to view those accessed contents as one group when the user accesses not only a single content but multiple contents continuously. The group of contents that a certain user used has the possibility to be used as a useful content group for other users. However, when a flow of time and a new kind of item appears, the item that comes off as a retrieval target is not permanently left as a member of a group but the group is updated in accordance with the access frequency etc. enabling the access efficient and suitable for tendencies such as the taste of the user or the fashion on the net in real time.

2.2 Relevance Index for Contents Grouping

A P2P node works as a content holder and at the same time a user node who executes accesses to the contents on other nodes. When a user accesses plural contents in succession within a certain time period, e.g., a day, we assume that those contents become to have high relevance indices with each other. Relevance index shows how much value the continuous access of two contents has for users. Relevance index is shown by equation (1). Here, $K_{s,i}$ shows the relevance index between contents s and i . $v_{s,i}^n$ has value 1 if contents s and i are continuously accessed by a user at P2P node n , otherwise has value 0.

$$K_{s,i} = \sum_{n \in \text{nodes in P2P}} v_{s,i}^n \tag{1}$$

2.3 A Small Example of Group Creation and Deletion (Example of PC Parts)

An example flow of the automatic creation of a new group that occurs when users actually access and contents are downloaded, and deletion of a group when it becomes not used, is described as follows and shown in Fig.1.

1. Contents registered in the DHT do not belong to a group in the initial state. Contents are shown in (a). "P6T" is a name of a mother board, "CD-552GA" is a CD drive and "Core i7 920" is a CPU respectively.
2. These three contents are retrieved and downloaded frequently by many users.
3. These three contents are registered in the DHT as a group accessed often at the same time. Grouped contents are shown in (b).
4. DVD drive appears. The DVD drive is a content named "DVR-216DBK" here. The state is shown in (c).
5. The user comes to do the retrieval of the combination of DVD drive, CPU, and mother boards by the appearance of the DVD drive in place of the group that contains the CD drive.
6. The group is newly generated because the combination of DVD drive, CPU, and the mother board frequently came to be retrieved by the user, and it is registered to the DHT. (d) shows the state in which a group is newly generated.

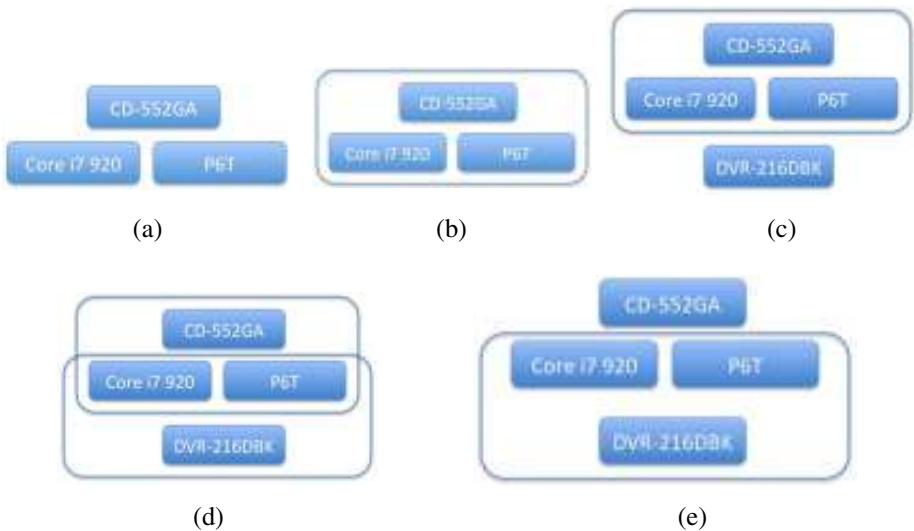


Fig. 1. A small example of group creation and deletion: (a) Initial state without a group, (b) A group is generated, (c) Appearance of DVD drive, (d) New contents group is generated, (e) The group that contains the CD drive is deleted.

7. The group including the DVD drive comes to be retrieved and downloaded by more users than the group of the CD drive. The CD drive after the DVD drive appears is retrieved and downloaded less frequently.
8. The group that contains the CD drive enters the state that is with less profitable information for the user. Then, the group that contains the CD drive is deleted as shown in (e). However, only the group of the CD drive is deleted, and CD drive itself is possible to be retrieved and downloaded as a single content.

3 System Configuration

P2P network is constructed with the DHT that uses Kademlia algorithm [6]. A P2P node usually participates to the network as a content holder (provider) and/or a user (consumer) of contents as shown in Fig. 2. In this system, we assume that one node holds a single content for simplicity.

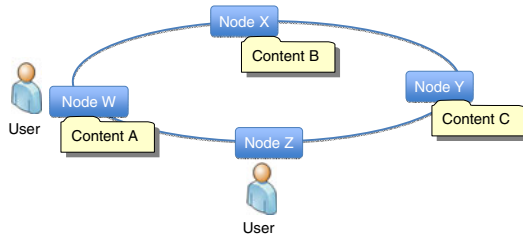


Fig. 2. P2P network configuration where a node usually participates to the network as a content holder (provider) and/or a user (consumer) of contents

3.1 P2P Node

In general, the node of P2P network holds hash table and the contents information. In this paper, a list of access history and a group list are hold at each node. Fig. 3 shows the state that keeps "a list of access history", "a group list" in a P2P node.

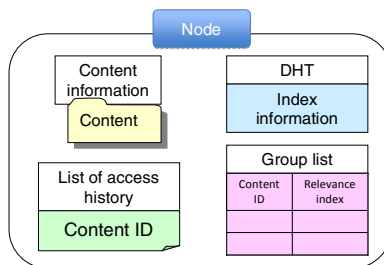


Fig. 3. P2P node keeps "a list of access history", "a group list" in addition to hash table and the contents information

Following four data structures are located in a P2P node.

- 1) *Contents information*: Content files which this node holds.
- 2) *Distributed hash table*: Index information of the node, i.e., hash value and node ID, are hold and updated.
- 3) *A list of access history*: We maintain a list of access history for a set of continuous access actions at a node. When there is a node which holds the object content that a user is accessing, the content ID is added to the list of access history, and it is erased when the acquisition of contents are completed.
- 4) *A group list*: We maintain a group list which consists of content ID of high relevance index with this hosting node, and it is sorted in descending order of the relevance index value. As time passes, the index values in the group list are maintained and decreases. When this hosting node is accessed from a user node, the nodes with high relevance indices are recommended back to the user node.

3.2 Contents Grouping by Content Access History and Frequency

The communication process between nodes for the grouping of the contents (or here, nodes which contain contents) is detailed. (Fig. 4)

- ① A user at node A requests to access content at node B after some retrieval action.
- ② Node A sends its own list of access history to node B. At node B, after receiving the list of access history, contents in the received list are merged to the group list of node B. If contents with the same IDs exist already in the B's group list, only relevance indices are increased by one.
- ③ Node B returns the requested own content to node A.
- ④ Node B sends back information of contents, which are members of B's group list and have relevance indices higher than a certain threshold set by the system parameter, to node A.
- ⑤ At node A, B's content IDs are added to the access history. By using information of contents from node B, A might access contents in the B's group list. This function is called the recommendation of contents in terms of the value of relevance index by node B to node A.

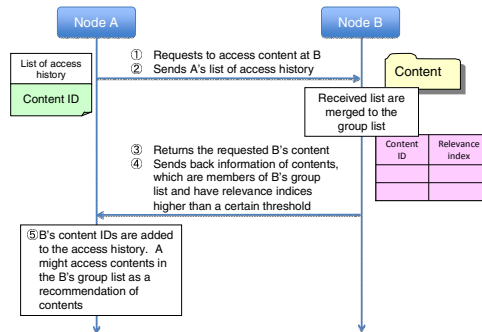


Fig. 4. The communication process between nodes for the grouping of the contents by content access history and frequency

3.3 System Parameters for Contents Grouping

Table 1 shows system parameters that are main factors for determining characteristics of contents grouping. Number of participating nodes and number of nodes which hold contents determine the size of P2P network. In this situation, nodes are classified into two types; node with or without content. A node without content mainly accesses contents as a user. Access tendency index and access count are considered to affect how quickly contents groups are formed in the process.

Table 1. System parameters for contents grouping

Name of parameter	Explanation
Number of nodes	Maximum number of participating nodes in P2P network
Number of content nodes	Number of nodes which hold contents
Max size of the list of access history	Maximum number of content items the list of access history holds
Max size of the group list	Maximum number of content items the group list hold. This parameter determines the size of the group that can be formed in the system.
Reduction rate of the relevance index	The rate the relevance index is reduced in one day period. If the value is $x\%$ then current relevance index S becomes $S \cdot x/100$.
Threshold of relevance index	The threshold to tell whether contents are considered in the same group or not by the relevance among contents.
Access tendency index	Retrieval tendency index indicates how frequently the user retrieves contents which are candidates for grouping in a short period, e.g. a day. The tendency is higher the group could be formed more quickly.
Averag. access frequency	The average number of access a user executes a day.
Recommendation function (on/off)	This function is a recommendation function of contents in terms of the value of relevance index. If recommendation function is on, it could be considered to accelerate formation of contents grouping.

4 Performance Evaluation

We use the Distributed Environment Emulator of "OverlayWeaver" [8], which is an overlay construction toolkit, and perform the experiment. We prepare the scenario file for the emulator which appointed the movement of the node such as retrieval/acquisition of contents.

The typical flow of the scenario file is shown in Fig. 5. For the content node we pre-set temporary groups beforehand in this scenario. The access of contents by user node is carried out from the pre-determined candidate groups based on the access tendency index. Steps D) to G) correspond to one day activity at each node. At the end of the day, the relevance index is reduced by the reduction rate given by the parameter. In this experiment, we iterate this activity for thirty days. We run experiments for the combination of parameter values and calculate performance measures for output data and then evaluate the performance of the grouping process.

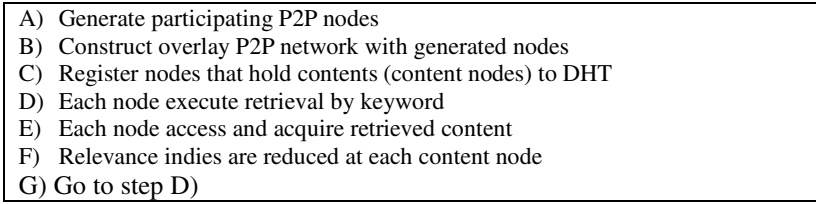


Fig. 5. The typical flow of the scenario file

4.1 Performance Measures

The measures we use to express the performance of contents grouping are:

- 1) *Number of content nodes that are in the formed groups:* Number of content nodes that have high relevance indices with intended candidate nodes and can be considered to be forming a group. Number of content nodes is 1000 and 1000 is the maximum value.
- 2) *Average number of contents that are in the group lists:* Average number of contents that are in the group lists of nodes which are members of the group. Maximum size of group list is 10 and 10 is the maximum number.
- 3) *Precision:* It is the fraction of the contents in the correctly (as intended) formed group relevant to contents in formed group. Precision is shown in equation (2), where, P: precision, k: number of contents in the correct groups, n: number of intended groups of contents, m: number of formed group, l: maximum size of group list = 10.

$$P = \frac{\sum_1^n \left(\frac{k}{l}\right)}{m} \tag{2}$$

- 4) *Recall:* It is the fraction of the contents in the correctly formed group relevant to all contents. Recall is shown in equation (3), where, R: recall, k: number of contents in the correct groups, n: number of intended groups of contents, l: maximum size of group list = 10.

$$R = \frac{\sum_1^n \left(\frac{k}{l}\right)}{n} \tag{3}$$

4.2 Experimental Results and Consideration

In this section, we explain some of the experimental results and discuss the feature of the proposed contents grouping method.

Effect of threshold of relevance index. This experiment inspects what kind of change occurs for the grouping by altering the setting of the threshold of relevance

Table 2. Parameters set for the experiment

Cases	No. of nodes	No. of content nodes	Max size of the access history	Max size of the group list	Reduc. rate of the relev. index	Threshold of relev. index	Access tendency index	Ave. access freq.	Recomm. function
A	1000	1000	10	10	10%	2	80%	3	on
B	1000	1000	10	10	10%	3	80%	3	on
C	10000	1000	10	10	5%	2	80%	3	on
D	10000	1000	10	10	5%	2	50%	3	on
E	2000	1000	10	10	5%	2	80%	3	on
F	2000	1000	10	10	5%	2	50%	3	on

index. We evaluate it using parameter values of Case A and B as show in Table 2. Fig. 6 is the graph which express precision and recall for thirty days of the experiments. We consider that we can perform correct grouping by setting an appropriate threshold of relevance index. However, when it is set high, even normal contents grouping would be filtered out. In this experiment, we set threshold relevance index to 2 and 3 and the result is that grouping is performed a little better in case of index 2. However, grouping is not well performed enough because of the small number of nodes, i.e., total 1000 participating nodes, compared to 1000 content nodes and limited access frequency.

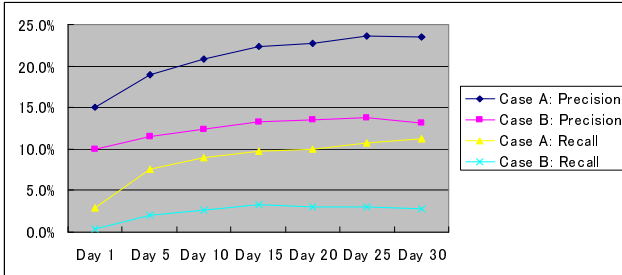


Fig. 6. This experiment shows the effect of threshold on grouping by altering the setting of the threshold of relevance index

Effect of number of nodes. This experiment is comparing how grouping changes by the number of participating nodes. We evaluate it using parameter values of Case C through F as show in Table 2. Fig.7 is the graphs which express precision and recall for thirty days of the experiments. We examine the group formation process by increasing the participating nodes to 2,000 and 10,000. Grouping is performed more rapidly by increasing nodes, and precision, recall show very high values. It can say number of participating nodes increases in P2P network, the proposed method for contents grouping shows a very high effect.

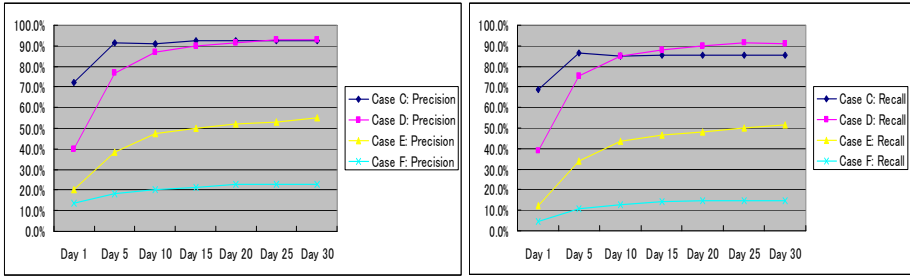


Fig. 7. This experiment is comparing how grouping changes (precision and recall) by the number of participating nodes

5 Conclusion

In this paper, we proposed access-based contents grouping technique aiming at an information retrieval system targeting contents in the P2P network, and performed evaluation by simulation. We intend to implement and evaluate the full function of grouping of contents in DHT. Based on the evaluation, improving the function of grouping of contents, it is scheduled to go to the system design of the efficient ubiquitous data retrieval system. In addition, the verification of the usability is scheduled to be covered in consideration of a concrete application system, and experimenting with the realistic applications.

Acknowledgments. This work was supported by Grant-in-Aid for Scientific Research (C) (20500095).

References

1. Chakraborty, D., et al.: Toward Distributed Service Discovery in Pervasive Computing Environments. *IEEE Trans. Mobile Computing* 5(2), 97–112 (2006)
2. Napster, <http://www.napster.jp/>
3. Gnutella, <http://www.gnutella.com>
4. Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. In: *Proceedings of the ACM SIGCOMM 2001 Conference*, San Diego California (2001)
5. Maymounkov, P., Mazières, D.: Kademlia: A Peer-to-peer Information System Based on the XOR Metric. In: *Proceedings of the 1st International Workshop on Peer-to-Peer Systems (IPTPS)* (2002)
6. Zhang, R., Charlie Hu, Y.: Assisted Peer-to-Peer Search with Partial Indexing. *IEEE Transactions On Parallel And Distributed Systems* 18(8), 1146–1158 (2007)
7. Sripanidkulchai, K., Maggs, B., Zhang, H.: Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems. In: *Proc. IEEE INFOCOM* (2003)
8. OverlayWeaver, <http://overlayweaver.sourceforge.net/>

The Effects of Individual Differences in Two Persons on the Distributed and Cooperative KJ Method in an Anonymous Environment

Takaya Yuizono and Zhe Jin

Japan Advanced Institute of Science and Technology,
1-1 Asahidai Nomi, Ishikawa 923-1292, Japan
yuizono@jaist.ac.jp

Abstract. The effects of three kinds of individual differences (a difference in the knowledge domain, a difference in the Big Five personality, and a difference in nationality) on a knowledge collaboration of the distributed and cooperative KJ method were examined with a groupware in an anonymous environment. The collaboration method has three steps: brainstorming, grouping, and writing. Twelve individuals participated in the experiment, eight of whom were Japanese and the other four were Chinese. The experimental results revealed the following: (1) Brainstorming by a pair in the interdisciplinary knowledge domain (humanities and sciences) led to the generation of more ideas than by a pair in the homo knowledge domain. (2) Persons possessing a factor called agreeableness in the Big Five factors of personality kept working until the final step.

Keywords: groupware, knowledge work, KJ method, individual differences, personality, knowledge domain, nationality.

1 Introduction

The infusion of the computer network into our daily lives has opened up several opportunities for communication and collaboration via the Internet. In order to support such work, researches about groupware have increased. Researches about an idea generation support system have been advanced in Japan. Such researches support the KJ method [1]. An anthropologist named Jiro Kawakita developed and diffused the KJ method in the 1960s. The groupware research realized a visual editor for the KJ method, for example, KJ-Editor [2] and D-Abductor [3]. Another research supports a distributed form of group work used the KJ method. In order to support the distributed and cooperative KJ method, the GUNGEN (groupware for a new idea generation support system) [4] has been developed and utilized in order to understand how the system and multimedia communication (voice and video) affects the distributed work [4, 5]. These researches aimed to develop and understand technology for idea generation [2–5].

On the other hand, personality is assumed to affect the usage of the groupware. If a person prefers individual work, he or she will display a dislike for group work, as a result of which the quality of the performance of the group work might deteriorate.

In this paper, the effects of individual differences on the distributed and cooperative KJ method with groupware technology were described. The differences were obtained from the five factors, the knowledge domain (sciences or humanities), and nationality.

2 Related Works

Some psychologists think that personality can be explained with the help of five factors; this explanation is called the Big Five [6]. For the Japanese, Murakami et al. classified extroversion, agreeableness, conscientiousness, emotional stability, and intellectual interests as the Big Five [7]. They also developed reliable questionnaires that eliminated the possibility of receiving ambiguous answers; these questionnaires comprised 70 questions for deciding the Big-Five factors.

In GUNGEN [5], video and voice, which form a part of multimedia conferencing tools, affected the text-chat communication to ask the state of the participants but did not affect the various factors, such as time and conclusion, that governed the output performance of the distributed and cooperative KJ method.

It is a well-known fact that computer mediated communication (CMC) has a more profound effect on our decision as compared with face-to-face communication [8]. The results showed that participants with CMC talk with more honesty and equality than those in face-to-face environments. Moreover, the participants tended to avoid a one-man show and produce constructive opinions. On the other hand, the participant tended to express excessive and angry opinions.

3 The Experimental Method

3.1 Procedure

First, we investigated individual properties in terms of three features—personality, knowledge domain, and nationality. Second, we formed pairs with the individual differences for group work. Third, each pair carried out the group work assigned. Finally, we compared these results with the three features.

The task and the environments for the group work are described in 3.3, and the results of these experiments are described and discussed in the next section.

3.2 Decision of Difference Pairs with Three Individual Features

3.2.1 Collection of Three Features

We developed web questionnaire systems to collect individual data for personality, knowledge domain, and nationality as a component of culture. The web system was developed with PHP, Apache Web Server, and MySQL database. For personality with the Big-Five test, there are 70 questions, with checkboxes provided wherein the answer is indicated by selecting either “yes” or “no.” Likewise, for the knowledge domain, checkboxes are provided in which the answer is indicated by choosing from either “humanities” or “sciences.” The same goes for nationality, wherein the options are “Japanese” or “Chinese.”

Table 1. Individual features of each person

Person	Big five factors						Knowledge domain	Nationality	
	Two points code		extroversion (EX)	agreeableness (AG)	conscientiousness (CO)	emotional stability (EM)			intellectual interests (IN)
A	IN+	CO+	3	6	11	7	12	Sciences	Chinese
B	IN+	AG+	6	9	5	6	10	Humanities	Chinese
C	EX-	CO-	2	10	4	4	3	Sciences	Japanese
D	EX-	CO-	2	8	3	8	3	Sciences	Japanese
E	AG+	CO+	3	12	11	10	7	Humanities	Chinese
F	AG+	CO+	9	11	11	9	10	Humanities	Japanese
G	ES-	CO-	2	4	1	0	1	Sciences	Japanese
H	EX-	IN+	2	10	3	7	1	Humanities	Japanese
I	EM-	AG+	3	8	7	0	5	Sciences	Japanese
J	IN+	EX+	10	11	10	11	12	Humanities	Chinese
K	EX+	AG+	12	12	11	10	5	Humanities	Japanese
L	EM-	EX-	3	9	5	0	4	Sciences	Japanese

The data collected through the Web questionnaire system are shown in Table 1. For personality, we scored each of the five factors and used “+” and “-” to indicate high-scoring and low-scoring factors, respectively. If a person had more than two scores with high or low scores, we selected the two factors with a striking difference. In cases of high scores for the extroversion dimension and low scores for the agreeableness dimension, the personality was coded as “EX+, AG-.”

3.2.2 Decision of Pairs

A pair was formed depending on the person’s personality, and a pair was judged as similar when the persons constituting the pair showed similarity in more than three of the Big Five factors; on the other hand, a pair was judged as different when the persons constituting the pair showed similarity in less than one of the Big Five factors. The result of the pairing is shown in Table 2.

With respect to personality, five similar and six different pairs were formed. With respect to the knowledge domain, six interdisciplinary and five homogeneous pairs were formed. With respect to nationality, six Japanese-Chinese and five same-nation pairs were formed.

3.3 Tasks and Environment

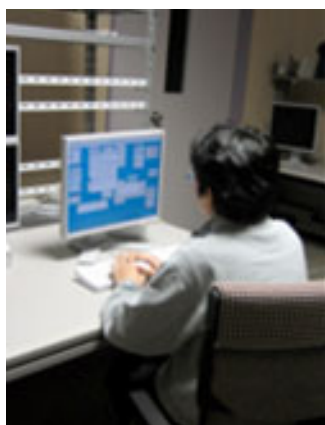
The task called the distributed and cooperative KJ method [3] was adapted to the groupware work by Munemori from the original KJ method. This task has three steps: generating ideas by brainstorming, grouping the ideas as per similarities for concept formation, and framing a concluding statement from the previous steps. With respect to creative style, we use divergent thinking in the first step, convergent thinking with gaps in the second step, and convergent thinking with linear sentences in the final step.

Table 2. Results of pairing

	Two persons	Personality	Knowledge domain	Nationality
Exp1	A-C	Different	S-S	J-C
Exp2	G-H	Similar	H-S	J-J
Exp3	J-K	Similar	H-H	J-C
Exp4	I-H	Different	H-S	J-J
Exp5	A-B	Similar	H-S	C-C
Exp6	J-L	Different	H-S	J-C
Exp7	G-F	Different	H-S	J-J
Exp8	D-E	Different	H-S	J-C
Exp9	D-C	Similar	S-S	J-J
Exp10	E-F	Similar	H-H	J-C
Exp11	B-K	Different	H-H	J-C

The discussion theme selected is entitled “How to become a president in a big company,” because the theme was appealing to students, and its contents would reflect aspects of the individual’s personality.

The experiment has been conducted in a distributed and anonymous environment, because the anonymity ensures that the user speaks frankly and honestly. The communication media is text-chat communication only for the purpose of maintaining anonymity. The environment setup is shown in Figure 1.

**Fig. 1.** A shot of an experimental setup

The KUSANAGI [9], which was groupware and supports the distributed and cooperative KJ method, was utilized as groupware software for the task. The KUSANAGI supports the concept of showing an opinion as a label and grouping labels, and naming each group on a shared screen. It also supports the framing of concluding statements on the shared window for writing. A sample screen of a result obtained from



Fig. 2. A screen shot of KUSANAGI

the collaboration task is shown in Figure 2. In the screen of the groupware KUSANAGI, the user name is inputted as “A” or “B” instead of a real name to preserve anonymity.

The participants in the experiments were twelve graduate students from our university, and they executed the cooperative task twice. However, the experimental data obtained covers only eleven out of the twelve participants, because two of the participants in one experiment accidentally met, and the experimental data did not come from an anonymous environment.

4 Results and Discussion

4.1 Results

The results of each experiment are shown in Table 3. The results data consist of the number of ideas, the number of groups and time required for grouping, the characters in a concluding statement and time required for writing, and chats for communication.

In other hands, the participants answered the seven-scale questionnaires about their interest in the theme, collaboration, and satisfaction rate for the result obtained in each experiment. The results are shown in Table 4, and the overall score is more than the neutral score 4. The participants felt that both oneself and his partner interested in the theme, and they had friendliness with the partner, a good cooperation, and the satisfied result.

Table 3. Results of each experiment

	Opinions	Groups	Time required for grouping (minutes)	Characters in each sentence	Time required for writing (minutes)	Chats
Exp1	23	2	12	364	33	16
Exp2	63	16	30	355	32	61
Exp3	34	7	32	515	49	7
Exp4	36	6	29	405	54	56
Exp5	43	2	13	352	63	34
Exp6	62	9	14	409	22	104
Exp7	80	14	35	351	13	32
Exp8	69	12	30	362	45	60
Exp9	42	9	24	395	52	45
Exp10	62	8	14	472	35	32
Exp11	50	14	27	257	24	84
Mean values	51.3	9.0	23.6	385.2	38.4	48.3

Table 4. Results of the seven-scale questionnaires

Questions	Score
Your interest in the theme	5.3
Partner's interest in the theme	5.2
Friendliness with a partner	5.1
Good cooperation	5.7
Satisfaction with the result	5.2

4.2 Comparison by Three Individual Differences

Three kinds of individual differences of personality, knowledge domain, and nationality are compared in table 5 in order from left to right.

As the first step, the results showed that the difference of knowledge domain affect the idea generation step. The number of ideas in the interdisciplinary pair, which consists of a person in humanities and a person in sciences, was more than the number of ideas in the homogeneous pair, which consists of two persons in humanities or two persons in sciences.

In other hands, there is no difference in the second step of forming groups on the basis of ideas and the last step of writing concluding statements.

To examine the tendency to generate more ideas in the case of interdisciplinary pairs, we investigated the number of ideas obtained per person.

First, the learning effects for the tenth–twelfth persons were investigated, because they executed the cooperative KJ method twice, so it was predicted that the number of ideas in the second execution would be more than that of the ideas in the first. The results indicated no difference between the number of ideas: $N = 24.9$ in the first

Table 5. Comparison by three kinds of individual differences

Items	Personality		Knowledge domain		Nationality	
	Similar	Different	Interdisciplinary	Homogeneous	Japanese-Chinese	Same nationality
Ideas	53.3	48.8	58.8	42.2	50.0	52.8
Groups	9.5	8.4	9.8	8.0	8.7	9.4
Time required for grouping (minutes)	24.5	22.6	22.6	24.5	21.5	26.2
Characters in a each sentence	358.0	417.8	372.3	400.6	396.5	371.6
Time required for writing (minutes)	31.8	46.2	38.2	38.6	34.7	42.8
Chats	58.7	35.8	57.8	36.8	50.5	45.6
N	5	6	6	5	6	5

execution and $N = 26.0$ in second execution (T-test, $p = 0.79 > 0.10$). Therefore, the experience did not affect the number of ideas in the task.

Next, we compared the number of ideas ($N = 29.0$) by the participants in humanities with the number of ideas ($N = 21.6$) by the participants in sciences. The result indicated the possibility of differences between the two (T-test, $p = 0.064 < 0.10$). Moreover, we compared the number of ideas ($N = 29.4$) by the participants in the interdisciplinary pair with the number of ideas ($N = 21.1$) by participants in the homogeneous pair. The result revealed a meaningful difference between the two (T-test, $p = 0.034 < 0.05$).

In addition, the number of ideas per a person, who participated both as the interdisciplinary pair and as the homogeneous pairs, was investigated, the result of which is shown in Table 6. It showed that all the persons offered more ideas in the case of the interdisciplinary pair than those in the case of homogeneous pair.

Table 6. Increase in ideas by interdisciplinary participants

Person	Interdisciplinary pair	Homogeneous pair
A	19	12
B	24	22
D	34	23
E	35	30
F	48	32
J	34	21

It was assumed that the number of ideas increased in the interdisciplinary pair, because the overlap in knowledge was less, and the persons in the pair could be more knowledgeable than those in the homogeneous pair. When the contents of the ideas were examined, the difference in opinions increased to some extent. In future, the metrics for clarifying such differences by the concept dictionary will be developed and could lead to the revelation of some differences in quality.

4.3 Personality and Work Process

To consider some effects of personality, the amount of cooperative work executed per participant was investigated, the results of which are presented in Table 7.

Participants who completed less than 20 percent of the work were judged as less cooperative in the task. For example, the rates of person G for writing operations in Ex2 was low—2.8 percent (12/427)—and this person did not work cooperatively in the last step. Looking over table 7, it is evident that the persons with “AG+” personality, which indicates agreeableness with others, were persons B, E, F, L, and, K, and they worked cooperatively from the first step to the last.

In other hands, the effects of the other four factors were not very evident. For example, the person with the trait of “extroversion” did not chat more than the other persons. In this research, the number of cases was too small to confirm the different varieties of personality.

Table 7. Personality and amount of cooperative work

	Person	Personality	Number of ideas	Number of group- ing operations	Number of writing operations
Ex1	A	IN+, CO+	12	28	110
	C	EX-, CO-	11	77	79
Ex2	G	ES-, CO-	25	168	12
	H	EX-, IN+	38	109	415
Ex3	J	IN+, EX+	21	54	61
	K	EX+, AG+	13	39	549
Ex4	I	EM-, AG+	13	47	117
	H	EX-, IN+	23	83	169
Ex5	A	IN+, CO+	19	25	38
	B	IN+, AG+	24	20	703
Ex6	J	IN+, EX+	34	91	547
	L	EM-, EX-	28	64	51
Ex7	G	ES-, CO-	32	114	301
	F	AG+, CO+	48	109	224
Ex8	D	EX-, CO-	34	124	606
	E	AG+, CO+	35	106	208
Ex9	D	EX-, CO-	23	98	840
	C	EX-, CO-	19	29	3
Ex10	E	AG+, CO+	30	34	317
	F	AG+, CO+	32	52	442
Ex11	B	IN+, AG+	22	80	400
	K	EX+, AG+	28	88	124

5 Conclusions

The effects of the individual differences paired with the distributed and cooperative KJ method were investigated using three factors—personality, the knowledge domain, and nationality. The results showed that the pair in the interdisciplinary knowledge domain had produced more ideas than the pair in the homogeneous knowledge domain. The persons who possess the agreeable factor within the Big Five personality maintained a cooperative working style throughout the three steps of the distributed and cooperative KJ method.

In future, the metrics for clarifying knowledge differences by the concept dictionary will be developed and the making an effect of the increased ideas on the post-collaboration steps will be considered.

Acknowledgement

This research was partially supported by The Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Grant-in-Aid for Young Scientists (B) 21700133, 2010.

References

1. Kawakita, J.: *Idea Generation Method*, Chuokoron-sha, Tokyo (1967) (in Japanese)
2. Yuizono, T., Kayano, A., Shigenobu, T., Yoshino, T., Munemori, J.: Groupware for a New Idea Generation with the Semantic Chat Conversation Data. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) *KES 2005. LNCS (LNAI)*, vol. 3681, pp. 1044–1050. Springer, Heidelberg (2005)
3. Munemori, J., Nagasawa, Y.: GUNGEN: Groupware for a New Idea Generation Support System. *Inf. and Soft. Technology* 38(3), 213–220 (1996)
4. Misue, K., et al.: Enhancing D-ABDUCTOR towards a Diagrammatic User Interface Platform. In: *Proc. KES 1998*, pp. 359–368 (1998)
5. Munemori, J., Yuizono, T., Nagasawa, Y.: Effects of Multimedia Communication on GUNGEN (Groupware for an Idea Generation Support System). In: *Proc. of 1999 IEEE International Conference on Systems, Man, and Cybernetics*, vol. II, pp. 196–201 (1999)
6. Ohiwa, H., Takeda, N., Kawai, K., Shimomi, A.: KJ editor: A Card-Handling Tool for Creative Work Support. *Knowledge-Based Systems* 10, 43–50 (1997)
7. Sproull, L., Kiesler, S.: Computers, Networks, and Work. *Scientific American* 265(3), 84–91 (1991)
8. In Wikipedia, *Big Five Personality Traits*, http://en.wikipedia.org/wiki/Big_Five_personality_traits (access on March 19, 2010)
9. Murakami, Y., Murakami, T.: *Big Five Handbook*, Gakugei-tosho, Tokyo (2001) (in Japanese)

Pictograph Chat Communicator III: A Chat System That Embodies Cross-Cultural Communication

Jun Munemori, Taro Fukuda, Moonyati Binti Mohd Yatid,
Tadashi Nishide, and Junko Itou

Wakayama University, Faculty of Systems Engineering, 930 Sakaedani,
Wakayama, Japan
{munemori, s095045, s105054, s125040, itou}@sys.wakayama-u.ac.jp

Abstract. The Pictograph Chat Communicator III is a modified version of the conventional system, where pictograph selection and the taxonomy of pictographs were taken a closer look. We added new pictographs (subjects, verbs, adjectives, 5WH, symbols, and alphabets) and deleted unnecessary pictographs. We also changed the taxonomy of pictographs. Experiments were carried out 9 times in international conference halls of America, Vietnam, and Portugal, between people who do not share the same spoken language (or native language). As a result, the level of understanding is 91.1%. We supposed that pictograph communication was enough in the ice-breaking communications.

Keywords: Cross-cultural communication, Groupware, Pictograph, Chat.

1 Introduction

The basic interaction tool that enables two or more people to communicate and collaborate with each other is the common language. Without sharing the same spoken language, we become handicapped in handling shared tasks.

Thus, the usage of widely spread text-based chat communications like MSN and Yahoo Messenger become limited to people, who only share the same spoken language. Language barrier occurs between people who don't share the same spoken language, and learning a new language takes time and effort. Therefore, we focused on the usage of pictographs [1] to express emotions and slight nuances, which may allow communication. Pictographs are used for instructions in the instruction books of medicine for foreigners in Japan [2]. The interpretation of the pictograph is thought to be approximately the same all over the world. So, we should prepare only one language sets using pictographs. On the other hand, the approach using language translation requires same number of language translation functions as the number of pairs of different languages.

Since the usage of pictographs have spread almost everywhere around the globe, we believe that pictographs may be the answer to break language barrier. Therefore, chats based on only pictographs were conducted and Pictograph Chat Communicator II was developed to assist communications between people who do not share the same spoken language [3]-[5]. More than 500 pictographs were prepared in the system.

We believe that communication between two parties who do not share the same spoken language could happen when the meaning of the pictographs could be understood. However, the understanding of pictographs may differ caused by differences of native languages and culture. When this happens, misinterpretations and misunderstandings occur.

In this paper, we first explained experiments using the conventional system (the Pictograph Chat Communicator II), and extracted the problems in the pictograph chat communication. Besides misinterpretations, the whole process of searching for pictographs and sentence making, everything until a chat user expresses himself, takes quite some time. As solutions to the problems, we proposed a new design selection of pictographs (newly designed pictographs of 5W1H, adjectives, verb etc, addition of important pictographs like alphabets, modification of current pictographs and deletion of unnecessary pictographs) and re-group the pictographs in the system. Experiments using Pictograph Chat Communicator III were conducted abroad (America, Vietnam, and Portugal) for 9 times. We examined how far communications using the pictograph chat could go.

2 Related Work

Zlango [6] is a pictograph-based system built for web and mobile messaging. The system has about 200 pictographs, which are changed from time to time, depending on its usage. Unused pictographs are deleted and new ones are being added to the system. The pictographs are divided into groups such as “People”, “Actions”, “Places”, and “Feelings”.

Pictograph Chat Communicator II [3] –[5] consisted of two PCs via LAN. Pictographs are in color but only a small part is in monochrome. There are 547 pictographs in the system. This system has 9 tabs (including the History Tab), and pictographs are divided into 8 tabs. The experiments were carried out 8 times [3][4]. Each experiment consists of 2 people, a Japanese and non-Japanese chatting using the system. Based on the results, a few problems regarding pictograph chat communication could be given. The problems and solutions are explained below.

- 1) Without the appropriate grouping (of pictographs), it is difficult to search for pictographs that one needs to use. Therefore, taxonomy of pictographs is extremely important especially to save time in pictograph chat.
- 2) There are a lot of unnecessary pictographs in the system. Pictographs that could not be understood or not used in daily conversations need to be deleted and changed to more important ones.
- 3) Pictographs that are not properly designed and monochrome pictographs whose meaning could not be identified lead to misunderstandings. Therefore, modification of pictograph design is important.

Lack of important pictographs like 5W1H (who, what, when etc), subjects (me, you, family etc) adjectives and verbs lead to difficulty in making pictograph sentences. New pictographs that express these need to be added to the system.

3 Pictograph Chat Communicator III

3.1 Design Policy

Based on experimental results of previous experiments and questionnaire results on the pictograph knowledge (answered by 12 Japanese and 10 Chinese people in Beijing), design selection of pictographs was applied. For examples, 5W1H (who, what, when etc), subjects (me, you, family etc), adjectives, and verbs to use well every day, are added. Pictographs that could not be understood or not used in daily conversations are to be deleted. Pictographs that are not properly designed and monochrome pictographs whose meaning could not be identified lead to misunderstandings are modified. New pictograph designs were made based on the understanding of two people, a Malaysian and Japanese. We guarantee the versatility of the pictographs by this approach. The taxonomy of the conventional system was based on the frequency of the usage of pictographs in mobile phones. Most of the languages in the world are in the order of Subject+Verb+Object (SVO), and so the subjects mostly ordered the pictographs in the same order. The Japanese people, however, could use both Subject+Object+Verb (SOV) and SVO when pictographs were selected in the pictograph communication [3]. We decided to group the pictographs in the order of the English language (SVO). That means it starts with the subject tab, verb tabs then the object tabs.

3.2 Composition of System

A modified version of the conventional system (the Pictograph Chat Communicator II), named Pictograph Chat Communicator III was developed (Fig. 1). This system focused on the design selection of pictographs and their taxonomy. More than 500 pictographs are used in the system. The hardware of the system is SONY VAIO type-U (OS: Windows XP) computers. Two PCs are linked by LAN. Software was developed by FLASH Professional 8 (Macromedia). It is a program of about 1100 lines. Pictographs are in colour but a part is PIC-DIC monochrome symbols [7]. The pictographs are sized 54*54 pixels. Each is represented about 5mm*5mm on the screen. All operations are performed by 'pen'.

3.3 Function of System

This system has 9 tabs (including the History Tab), and pictographs are divided into 8 tabs. Different pictographs appear when different Pictograph Tab (A) is tapped. The pictographs will be on display in the Pictograph Selection Screen (B). When a selected pictograph is tapped, it will be added in the Pictograph Input Screen (C). When the Chat Input Button (D) is tapped, pictographs will be sent to the Chat Log Screen (E) for both chat users to see. Displaying only user's icons (ex. penguin and fish) and pictograph sentences caused by the small size of the Chat Log Screen.

This system has a History Tab. The top tab is the History Tab (F), which displays the pictographs that have been used by both chat users. The maximum of 77 used pictographs could be on display. Pictographs could also be tapped in here. The part (B) of Fig. 1 shows the content of the History tab.

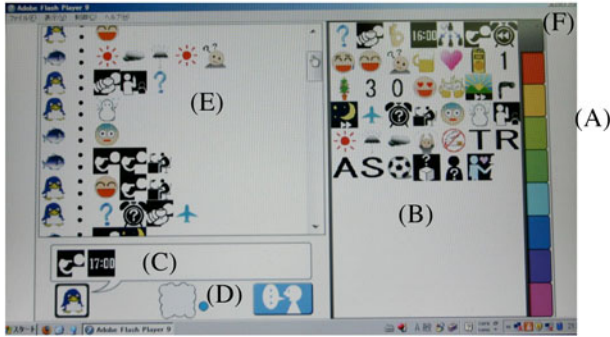


Fig. 1. The screen of the Pictograph Chat Communicator III

There were 551 pictographs in the system of Pictograph Chat Communicator III in the experiments in America and 564 pictographs in the experiments in Vietnam and Portugal. The modifications done are explained below.

3.4 Pictographs

Fig. 2 shows examples of the newly designed pictographs added in the system. [🗑️] is deleted from system in the experiments in Vietnam and Portugal.

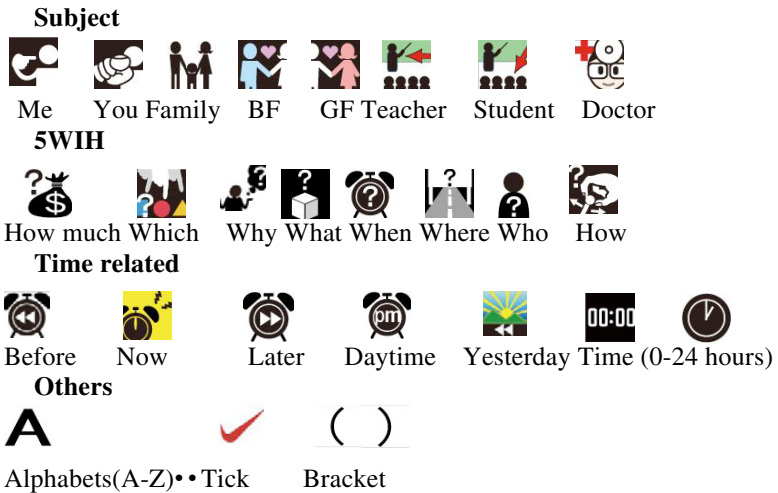


Fig. 2. Newly designed pictographs

There were a few pictographs that created misunderstandings. Examples of the modified and colored pictographs are shown in Fig. 3.



Fig. 3. Modified and colorized pictographs

Unnecessary pictographs shown in Fig. 4 were examples of deleted from the system.

- 1) Usage of Kanji Characters in pictographs



- 2) Similar pictographs (one is removed)



- 3) Pictographs only understood in Japan



- 4) Unidentified meaning of pictographs



Fig. 4. Unnecessary Pictographs Deleted from the System

In reference of Dongba Characters [8] pictographs that express verbs shown in Fig. 5) were added in the system (part). These pictographs were added only in the experiments in Vietnam and Portugal.



Fig. 5. Addition of Pictographs in Reference of Dongba Characters

3.5 Taxonomy of Pictographs

The contents of each tab are shown below. Therefore, the tabs are arranged in order of SVO. New pictographs are underlined. Approximately 100 new pictographs were added to the system. Fig. 6 shows the 2nd tab, 4th, and the 9th tab.

The 1st tab is the History Tab.

The 2nd tab contains pictographs that express subjects, 5W1H, time related, weather, face emoticons.

The 3rd tab contains pictographs that express verbs.

The 4th tab contains pictographs that express verbs, adjectives, place and transports.

The 5th till 8th tab contains pictographs that express nouns (food, animals, games).

The 9th tab contains pictographs of alphabets, numbers, marks and time.

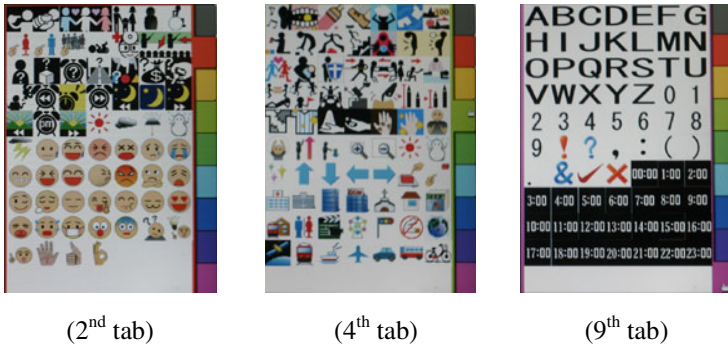


Fig. 6. Example of tabs (2nd, 4th, and 9th)

4 Experiments

Experiments were conducted 9 times. There were 18 subjects, 8 Japanese (7 males and 1 female) and 10 non-Japanese (1 American female, 1 Canadian male, 5 Thai males, 1 Vietnamese male, 1 Cambodian male, and 1 Luxembourg male).

The subjects were given no chat topic and were told to have a free chat communication for 15 minutes (except Exp. 2 which was for 10 minutes). They were only taught the way to use the system (how to choose and delete pictographs etc), and were told that the purpose of the experiment is to study the possibility of using pictograph chat for communication. After 15 minutes of chatting, both chat users were told to write down the meaning of each line of pictographs they made and the meaning of lines they think their partners said. After that, they were asked to answer a questionnaire of five-point scale evaluation on the experiment and the system.

Experiments in America (CSCW 2008);

[Experiment 1] Canadian (male) – Japanese (male)

[Experiment 2] American (female) – Japanese (female)

Experiments in Vietnam (KICSS 2008);

[Experiment 3] Vietnamese (male) – Japanese (male)

[Experiment 4] Thai (male) – Japanese (male)

[Experiment 5] Thai (male) – Japanese (male)

[Experiment 6] Thai (male) – Japanese (male)

[Experiment 7] Thai (male) – Japanese (male)

[Experiment 8] Cambodian (male) – Thai (male)

Experiments in Portugal (WEBIST 2009);

[Experiment 9] Luxembourg (male) – Japanese (male)

5 Experimental Results

An example of chat and the meaning of their pictographs expressed by both chat users (Exp. 9) are shown in Fig. 7. Chat user A (user's icon is 'snail') is a Japanese male in his 50's and chat user B (user's icon was 'dog') is a Luxembourg male in his 30's.



- (1) A: I am coming from Japan by plane.
B: I am coming from Japan (via plane).
- (2) A: I like wine.
B: I like wine.
- (3) B: Nice to meet you.
A: I like you.
- (4) A: I like you.
B: I like you.
- (5) B: I'm coming from Lux (via plane).
A: I am coming from Luxembourg.
- (6) B: How long was your flight?
A: When you arrived at the airport? or How long was your flight?
- (7) A: 15 hours.
B: 15 hours.
- (8) B: Long flight.
A: Long flight
- (9) A: Cold sweat.
B: Agree.
- (10) B: Indeed 3 hours.
A: 3hours (from Lux).
- (11) A: Did you present a paper?
B: Did you present a paper?
- (12) B: Yes
A: Yes
- (13) A: I listen only.
B: I listen only.
- (14) A: Are you taking the bus tonight?
B: Are you taking the bus tonight?

Fig. 7. Chat log and the meaning of their pictographs expressed by both chat users. A: Interpretation of user A, B: Interpretation of user B (A part of Exp. 9).

The pictograph chat was performed an average of 2.4 lines for one minute. About 2.8 pictographs on an average line (one remark) were used. Total number of pictographs was 822. The pictograph, which expressed [Me], was 11.3% (the most). The pictograph, which expressed [You], was 7.5%. All linage is 307 lines. Table 1 shows chat lines, numbers of pictographs per line, understanding level (later description), usage of new pictographs, and use of Alphabets (26 pictographs).

For a conversation of N lines, if a line is completely understood, it gets a score of $(1/N)*100\%$; if the interpretation is very different, it gets 0%. In N line of M pictographs, if there is one non-understood pictograph, the understanding level is $(M-1/M)*1/N\%$. In N line of M pictographs, if a pictograph is partially understood but not exactly right interpretation, the understanding level is $\{(M-1/M)+1/2*1/M\}*1/N\%$.

When an error of the interpretation of the pictographs occurred, you or a partner should notice and repair it.

Most chat contents were for “ice-breaking”. Because we carried it out in the meeting place of the international conference, there were some topics about the international conference.

Table 1. Results of experiments

Experiment	Average
Chat lines (line/min)	2.4
Pictographs (numbers/line)	2.8
Understanding level (%)	91.1
Use of new pictographs (%)	46.0
Use of Alphabets (%)	12.5

Table 2. Questionnaire Result (5-point scale)

Questions	Average
1. In one tap, a chosen pictograph is added into the input field. That operation was convenient.	4.2
2. Sentence making (using pictographs) was easy.	2.6
3. I was able to understand the meaning of all pictographs in the system.	3.4
4. There were targeted pictographs (pictographs that I wanted to use).	3.1
5. I was able to look for the targeted pictographs smoothly.	2.4
6. I was able to understand the things my partner was trying to say.	3.9
7. I think communication went well.	3.8
8. I think that a chat system based on only pictographs could allow communication.	3.8
9. This experiment was interesting.	4.7

1: I disagree strongly, 2: I disagree,

3: I don't agree nor disagree, 4: I agree, 5: I agree strongly

The contents of chats are shown below: sports, their families, and liking the cities in Japan (Exp. 1), their interests, families, their work, and their hometowns (Exp.2), coming by airplane, the time they woke up in the morning, when and with whom he goes back to his country, and their interests (Exp. 3), topics on cigarettes, weather, about liking beer and presentations (Exp. 4), liking dogs, swimming, karaoke and beer, and Christmas (Exp. 5), each other's hometown, likes and dislikes in food, and presentations during the conference (Exp. 6), soccer, favorite fruits, and beer (Exp. 7), having flu, hometown, their families, the conference, liking soccer, and beer (Exp. 8), and coming by airplane, the conference, and their profession (Exp. 9). They used alphabets to express the names of those cities.

After completing the chat experiment, the chat users answered a questionnaire about the system and the chat experiment they took part in. The result of the questionnaire is shown in Table 2.

6 Discussion

The average level of understanding using Pictograph Chat Communicator III is 91.1%. Although the 5% significance level shows no significance difference, this is a big increase compared to the average level of understanding using the conventional system [3], which is 84%. The pictograph design selection is believed to be the main factor of this increase.

There are about 100 new pictographs added to the Pictograph Chat Communicator III. This is about 20% of all pictographs in the system. Researches and questionnaires have been conducted, and as a result, necessary pictographs are detected. These are pictographs that express 5WIH (What, Who, When etc), subjects like [You] and [Me], verbs and adjectives. The average usage of new pictographs is 46.0%, shows that these pictographs are used frequently. Alphabets were used mainly to express names of places. The average usage of alphabets is about 12.5%. [When] was used 8 times. [Where] was used 4 times. [How much] was used 3 times.

Based on the questionnaire results, upon the question [There were targeted pictographs (pictographs that I wanted to use)], points increased from 2.6/5.0 [3] to 3.1/5.0(new system). The new design selection of pictographs is believed to have caused this, which means that chat users feel that the pictographs that they want to use could be found in the system.

The average chat lines of the experiments using Pictograph Chat Communicator III are 2.4 lines/min. This is almost twice the number of chat lines using the conventional system [3], which is 1.3 lines/min. The examined data shows that the significance level is 1%, which shows significance difference between the data.

Many sentences were written in subject + α . There are a lot of sentences without the verb, too. The 9th tab (see Fig. 6) was used frequently (Total 25.8%). The 9th tab contains pictographs of alphabets, numbers, marks and time. We should re-arrange the order of the tabs (SVO order). We would like to change thirdly the ninth tab.

7 Conclusion

We have developed a pictograph chat communicator, named Pictograph Chat Communicator III. Experiments were conducted in America, Vietnam, and Portugal using the system. We show the results of our experiments.

- (1) The average understanding level is 91%. This is a big increase compared to the understanding level using the conventional system, which is 84%. The main factor of misunderstandings is the usage of pictograph. It is necessary to analyze which pictograph is easy to be wrong.
- (2) The average usage of newly added pictographs (subjects, 5WIH, alphabets, adjectives etc) is 46.0%. The pictograph, which expressed [Me], was 11.3% (the most). The pictograph, which expressed [You], was 7.5%. The usage of alphabets is about 12.5%.

(3) About 2.8 pictographs on an average line (one remark) were used. The average chat lines are 2.4 lines/min. The average chat lines of experiments using the conventional system are 1.3 lines/min, almost half the productivity of the modified system. Significance level is 1%, which shows statistical significance between the data.

More experiments and modifications of system will be done in the future to enhance the usability of pictograph chat communication.

Acknowledgments. This research was partially supported by Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Scientific Research (B) 20300047, 2008.

References

1. Ota, Y.: Pictogram Design. Kashiwa Shobou, Tokyo (1993) (in Japanese)
2. RAD-AR Council, Japan, <https://www.rad-ar.or.jp/english/index.html>
3. Mohd Yatid, M.B., Fukuda, T., Itou, J., Munemori, J.: Proposal and Evaluation of Pictograph Chat Communicator II. In: Proceedings of The Fourth International Conference on Collaboration Technologies, pp. 66–71 (2008)
4. Mohd Yatid, M.B., Fukuda, T., Itou, J., Munemori, J.: Pictograph Chat Communicator II: A Chat System that Embodies Cross-cultural Communication. In: CSCW 2008, Poster paper (2008) (CD-ROM)
5. Munemori, J., Fukuda, T., Mohd Yatid, M.B., Itou, J.: The Pictograph Chat Communicator II. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 167–174. Springer, Heidelberg (2008)
6. Pic-Talk, Zlango, <http://www.zlango.com>
7. Pic-Talk, Zlango, <http://www.zlango.com>
8. Tonpa de asobou kai. Dongba Characters: Tonpa!! Personal Media, Tokyo (2001) (in Japanese)

Distance Learning Support System for Game Programming with Java

Kouji Yoshida¹, Takumu Yaoi¹, Isao Miyaji²,
Kunihiro Yamada³, and Satoru Fujii⁴

¹ Shonan Institute of Technology

1-1-25, Tsujido Nishikaigan, Fujisawa, Kanagawa 251-8511, Japan

² Okayama University of Science, 1-1 Ridaicho, Okayama, Okayama 700-0005, Japan

³ Tokai University, 1117 Kitakinme, Hiratsuka, Kanagawa 259-1292, Japan

⁴ Matsue National College of Technology, 14-4 Nishi-Ikuma, Matsue,
Shimane 690-8518, Japan

{yoshidak,a063124}@info.shonan-it.ac.jp,
miyaji@mis.ous.ac.jp, yamadaku@tokai.ac.jp,
fujii@matsue-ct.ac.jp

Abstract. In recent years, students have confronted the importance of developing intermediate level programming skills. Many students are weak not only in programming; they are also not good at cooperation and collaboration when designing a program. Therefore, we produced a distance-learning support system using Java.

Generally, students can acquire knowledge that is constant in distance learning, but they find that it is difficult to finish a program independently. In distance learning, it is important to use mutual teaching and regular communication among students. When interesting programming of contents exists, games can make more effectively using Java software. This paper presents Java programming techniques that are necessary for game programming. Moreover this system supports upper level functions of game programming.

Keywords: E-Learning, Java language, Game programming, Group work.

1 Introduction

It is not easy for a person who understands the C language to a certain degree to create a game program that can run on the internet in Java language. It would be very helpful to create net games and increase enjoyment of programming in Java if there were some distance learning support system available for such a person who understands the C language but not Java language. Such a system could support the creation of a game program. Such a system would also contribute comprehensively to training for intermediate level programmers.

Herein, we first describe an environment with a Java learning system to support its execution in addition to how to use Eclipse for Java program creation, editing, and debugging. Subsequently, this paper presents techniques of class and inheritance of Java programming that are necessary for game programming, followed by instruction

of an actual program creation of a shooting game or match-play game, for example, using mouse and the keyboard input.

2 System Overview

This support system comprises learning basic Java language concepts, learning how to use Eclipse and how to debug a program with it, learning how to do programming for game creation using Java classes and their combination for game programming, and actual game program creation. The system also includes the capability of integrating the necessary pieces of information for game creation as well as guidelines to produce a complete game program at the end of the process.

Using this system, a student can learn all the way from basic Java programming to actual creation using it. During the course of learning, some optional problem exercises are given. After solving them, it is possible to progress further in learning programming skills. Fig.1 presents a system outline.

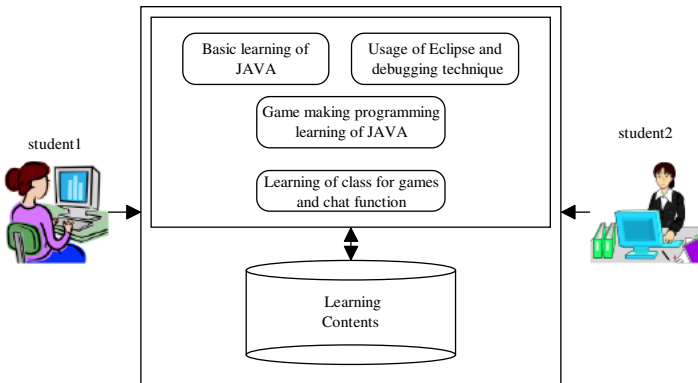


Fig. 1. System outline

3 Learning Contents

3.1 Learning Java Basic

A student will be given Web pages that are available on the world-wide web (WWW) when learning Java.

The student can then browse through those Web pages to learn basic information related to Java. First, it is presumed that the person understands C language, based on which the user is going to learn characteristics of the Java language as well as its difference from C. The learner will then understand that Java runs on a virtual machine as it runs in the environment in which the system functions.

During this learning process of Java basic in terms of software development, the student will learn from the command-level programming on the console screen to the class file creation including the execution of the program completed.

Listed in the following are items that are explained in this basic learning step.

- Java basic environment
- Source code creation
- Programming
- Class file creation
- Program execution
- Programming and debugging

This system showing the Java basic learning process is presented in Fig. 2.

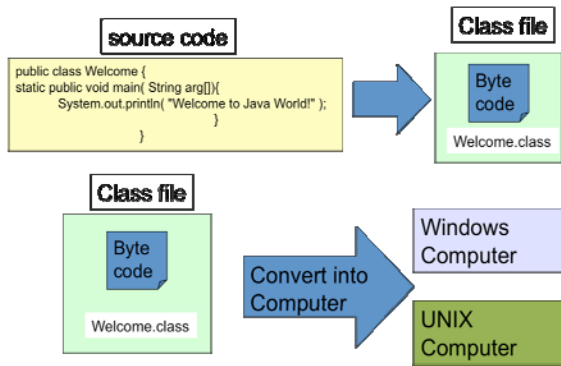


Fig. 2. Java basic learning process

3.2 Using Eclipse

A student who learns Eclipse will be given the contents available on the WWW the same way when learning Java basic. The contents are created on the HTML base given with the pictures of every associated Eclipse execution screen. Consequently, they are designed for easier understanding for the student and are uploaded to the server to be available.

The use of Eclipse is classified into three categories. The first is about how to create and execute Java basic classes. The second is how to operate the keyboard and the mouse for the input and output uses. The third one is how to input or output image pictures along with the applet manipulation. Through this learning process, the objective is for the student to acquire the ability to create a basic program and debug it as well. An image picture input can be accomplished using the import function available with Eclipse by setting up necessary files and folders for that.

3.3 Learning Game Programming in Java

Any game program can be more easily understood if it provides the necessary displays on the screen or if it is accompanied with actual game behavior. However, such a game-like program would generally be complex in terms of how it works, and would tend to be large; it would be somewhat inappropriate for beginners.

Considering those factors described above, our system is provided in a way in which every part of the available functionality is explained as necessary so that the student can learn the game-associated functionalities and behaviors in a practical sense while learning Java programming.

(1) Classes and methods

The things that are displayed on the screen when creating a basic class have been explained above already when the basic operation of Eclipse was explained. Here some examples of Java method in the calculation of subtraction will be given using the necessary classes that are unique characteristics of Java.

(2) Applet

Defining an image picture data in a class, the data are displayed on the screen using an applet. Understanding this functionality supports the processing of character images of various kinds in addition to background images.

(3) Input operation

For input processing using a keyboard or a mouse, the MouseListner and KeyListner methods are used. When judging what was input from the keyboard, the arg information of the method getKeyCode is used.

(4) Controlling

As a basic game control technique, determination of whether objects hit each other is important. Figure 4 shows how to accomplish that. When an object A is crossing another object B, we assume the coordinates of the object, A (x1,y1) at its top left and (x2,y2) at its bottom right, and of the object B, (x3,y3) at its top left and (x4,y4) at its bottom right. Under these circumstances, the judgment of whether those two objects collide or not is made by the IF statement shown at the lower part of Fig. 3. This mode of judging a hit should be convenient if remembered when needed to determine collisions happening in various games or particularly in a shooting game.

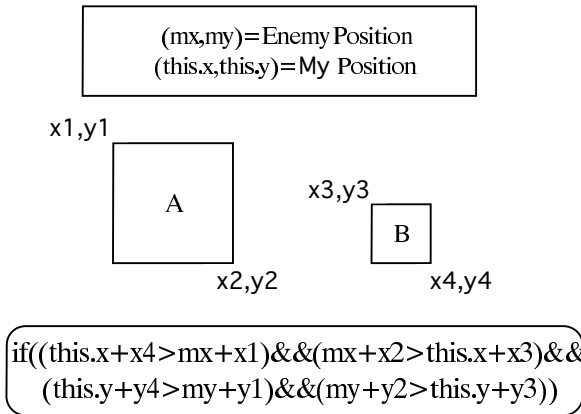


Fig. 3. Judgment of whether objects are hit or not

3.4 Learning Classes and Their Combination for Game

Even if a learner comes to understand a series of Java functionalities for game creation in Java, it does not necessarily mean that a game can be created easily. What is important is that learner learns through the available screens while combining the necessary functionalities that are also available. Then the learner can understand what every game is like and can create it accordingly. Even being provided with merely a description and/or an explanation of programming a game is insufficient for a learner's understanding. Using existing programs and combining them in a practical sense to verify how they would behave is also required. Executing some applied programs as well would be of great help in increasing the level of the programming technique step-by-step.

(1) Window operations

To create a Java program, "Java application", "Java applet" and "Java servlet" are available for use. For an online game, either "Java application" or "Java applet" is useful. However, because "Java applet" is restricted in many ways in terms of functionality, "Java application" is used here. Screen operation using the various parts of windows and buttons combined is called a Graphical User Interface (GUI). Here, we use Swing as the GUI.

Using Swing, the class `Jframe` is provided for window creation so that a display of a window can be accomplished merely by creating the instance of that class. Here, we will learn how to input a background screen image as well.

(2) Classes, methods, and associated input and output

An object can be displayed using a class. Using a method, an object's movement can be given: the enemy character can be displayed, for example. Furthermore, the output and input from a keyboard or by a mouse-click enables a user to control the game screen to advance the game. With the added object-hitting judgment technique, the creation of games of various types can be accomplished.

(3) Error processing

Error processing in Java must be done differently from the way it is done in C language, in which the programmer is responsible for keeping track of any errors if they occur and in which the programmer is responsible for describing them all (using the `try` statement). Depending on the error type that is caused, a user must describe in advance how to process the exception for the error (using the `catch` statement). In game programming, for example, an exception would occur when an interface must be done with an external device. Therefore, it would occur if no preparatory setting is made when an initial screen image data or any text data were to be input from a file.

Fig.4 presents an example in which an error message, "data file read error", is posted when the input data file "data.txt" does not exist in the recognizable range by Eclipse.

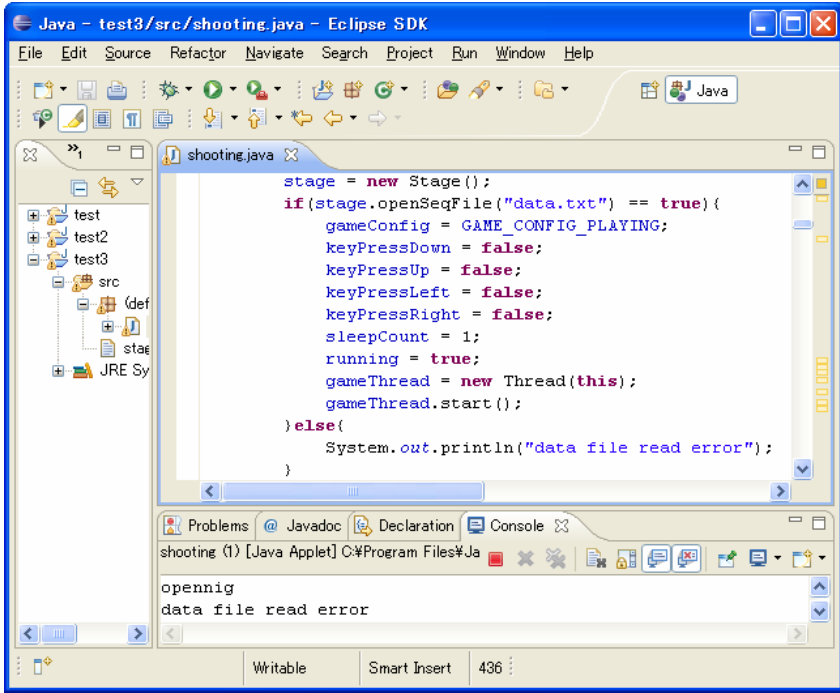


Fig. 4. Example of Java error processing

4 Game Integration

4.1 For a Shooting Game

In a shooting game[1], for example, after an initial screen is set, some objects are moved intentionally on the screen at first(Example Fig.5). Then the game progresses with some other objects or the mouse set on them, or responding to the input data from the keyboard after a series of such movements is processed or when the game control normally returns to the initial screen.

The game integration function is available to describe all necessary and complementary functions to finalize the game creation based on the information so far known during the game creation using each available function and combining them.

(1) Checking keyboard or mouse input

Using that method, Listner, available to handle the input from a keyboard or a mouse, the necessary next-step processing is moved to proceed after the data are input.

(2) Hitting

The decision of whether an input is hit or objects hit each other, for example, can be checked using the judgment function of hitting. The results are counted or the associated information is saved.

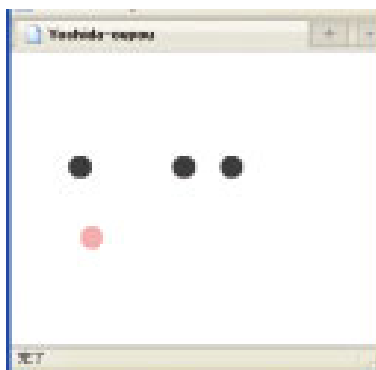


Fig. 5. Execution screen for a shooting game

(3) Registration processing

As a registration process, every individual process done during the game as well as the ranking information, for example, can be registered. Consequently, the entire system operation is managed. Then the associated information is saved in an external file.

4.2 For a Match-Play Game

For such a static match-play game as Othello, for example, a component for grid preparation must be created to show on the initial screen. Other requirements are as follows.

(1) Winner or loser determination in a match-play game

It is a fairly complex process to determine who is the winner or loser of the game. The determination logic might differ depending on the type of the game. For example in an Othello game(Fig.6), if a player catches the opponent's piece or pieces between the player's own pieces, then the opponent's pieces, either black or white, are reversed in

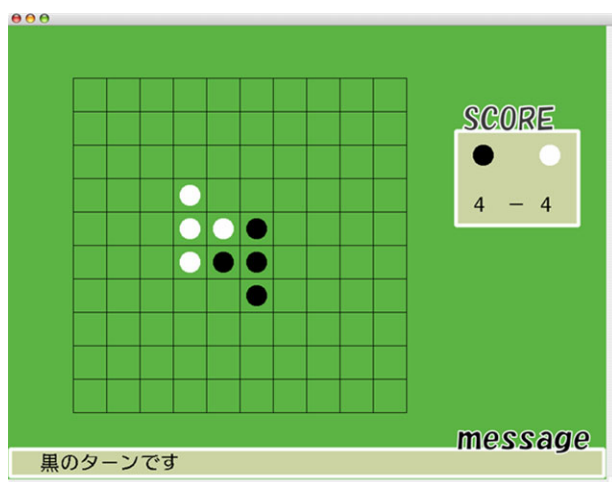


Fig. 6. Othello execution screen

color. In a Go-bang play, for example, one would need a search process to perform along the vertical, horizontal, and diagonal direction every time a piece is placed. If a five-piece alignment is found to be valid after doing search work and checking in all four directions, then the winner can be determined.

(2) Score ranking

In any game, if the winners' ranking is displayed, then the players would be more excited about the game. In an Othello or Gomoku game, the ranking display associated with the maximum score taken by every game or the rapidity of each player's action would be helpful. It might be interesting to display not only the ranking with respect to the game results, either win or lose, but also to display the difference in the number of black or white pieces possessed by each player.

5 Questionnaire Results and Evaluation

From Java language beginners, we received a good evaluation; they stated repeatedly that the way to learn Java and the examples given along with Eclipse were easy to understand. Furthermore, the evaluation we received of the learning program creation for the game was that it was suitable for self-learning because the program to run it was as compact and appropriate in size as to be understandable. Moreover, the problem exercises that were given were helpful for understanding.

However, learning related to classes was evaluated; their combination for game programming was not easy to understand because the associated program was complex and the complex relation was becoming involved with other programs, too. The distance-learning support system[2] was evaluated well overall because the students were able to progress at their own pace as they checked the results and thought about what would be the necessary next step to take.

Some responses included the explanation of a program: when it was complex, it was not given in an interactive way. For that reason, it was hard to understand the explanation. Additionally, it was hard to know how well each student had comprehended the material because no good way was available to check every player's step-by-step progress through the tests to take. Especially, the evaluation point as to whether or not it is fit for intermediate programmers was low. That is probably true mainly because the contents supported now are not suited for complex programming and no functionality is available for communication among the different games using a chat function or similar facility to that of a typical online game, for example.

- Evaluation and consideration

These results are expected to raise awareness of room for improvement in peer-to-peer communications. Additionally, participants in this study collaborated to produce a program through an exercise. They learned cooperation with others and the difficulty of producing a program. Furthermore, they understood that documentation quality closely reflected this communication and connection of a program. The students recognized the importance of documentation. The system improved motivation. It was particularly remarkable in terms of developing documentation, communication and listening skills. Students reached the goal through mutual assistance because we used both electronic media and communication through meetings.

Table 1. Questionnaire item and the result

No.	questionnaire item	Averaged evaluation pts. (1–5)
1	Do you think that the Web page is easy to see from a visual perspective?	4.25
2	Do you think that the Web page is easy to manipulate intuitively?	4.25
3	Do you think that the number of the Web pages is appropriate from the learning perspective?	4.5
4	Do you think that the use of Eclipse is easy to understand?	4
5	Are contents of "learning of a basic function of a game" of 3 chapters of titles appropriate?	4.1
6	Are contents of "learning of an online function" of 4 chapters of titles appropriate?	4.2
7	Do you think you can create a game program for yourself?	4
8	Do you think what you created is useful?	4.75
9	Do you agree that this training is useful for the intermediate level?	3
10	Do you think that this system is helpful for you?	4
11	Overall evaluation —	4

6 Conclusion and Potential

The distance-learning support system[3] introduced herein is intended to support a student learning basic concepts of Java. This study demonstrated how to complete a game program to create. Using this system and based on knowledge of C language, the student can learn Java basic and use the classes and the class library, which are typical characteristics of object-oriented programming. Moreover, a learner can start creating a program and complete it using Eclipse, which is a Java development tool.

The game programming[4] can be accomplished visually. Therefore, it is expected to attract users' interest and hold their willingness to perform at a higher level longer. Additionally, game development in Java is accompanied with a full class library, which means that sufficient functionality is available to run the resulting program in the www environment. Once it is understood, it is easy to use and an easy way to create a program as well. In summary, results show that a beginner can join in game programming without much hesitation. However, beginners who are not used to any complex processes might encounter some obstacles that they cannot overcome. We included some trials to help with those obstacles in an easy way, but we recognize that many problems remain to be solved beforehand.

Programming[5] in general has two aspects: it must be easy to understand, and it must be compact. This system would be harder to understand if we made it too compact. It would fail to be compact if we made it easy to understand. This was explained in the description of the evaluation. Therefore, we are considering implementing such functionality to enable a match-play game to be played in the network with some communication capability to allow for chatting, for example, for intermediate-level Java learners. This study has received assistance from a scientific research grant "18500737", along with the above.

References

1. Ootsuki, Y.: Programming classroom for JAVA online game, Rutles (2009)
2. Lopez, N., Nunez, M., Rodriguez, I., Rubio, F.: Including Malicious Agents into a Collaborative Learning Environment. In: Cerri, S.A., Gouardères, G., Paraguaçu, F. (eds.) ITS 2002. LNCS, vol. 2363, pp. 51–60. Springer, Heidelberg (2002)
3. NSW Department of Education and Training, Blended Learning, <http://www.schools.nsw.edu.au/learning/yrk12focusareas/learntech/blended/index.php>
4. Yoshida, K., Miyaji, I., Yamada, K., Ichimura, H.: Distance Learning System for Programming and Software Engineering. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS (LNAI), vol. 4694, pp. 460–468. Springer, Heidelberg (2007)
5. Scott, A.: Wallace and Andrew Nierman. Addressing the need for a JAVA-based game curriculum. *Journal of Computing Sciences in Colleges* 22(2), 20–26 (2006) ISSN 19374771

Evidence Analysis Method Using Bloom Filter for MANET Forensics

Takashi Mishina¹, Yoh Shiraishi², and Osamu Takahashi²

¹ Graduate School of Systems Information Science, Future University Hakodate,
116-2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041-8655, Japan

² School of Systems Information Science, Future University Hakodate
116-2 Kamedanakano-cho, Hakodate-shi, Hokkaido, 041-8655, Japan

{g2109042,siraishi,osamu}@fun.ac.jp

Abstract. Various security weaknesses have been identified in mobile ad-hoc networks (MANET). The paper focuses on MANET forensics whereby a third party can prove there was attack by collecting and analyzing evidence about it. The paper describes such a MANET forensics analysis method using a Bloom filter.

Keywords: Bloom filter, MANET, forensics, security.

1 Introduction

Mobile ad-hoc networks (MANET) can be built without relying on existing infrastructures. MANETs, however, have various security weaknesses, especially, regarding broadcast nodes and communication attacks (for example, disposal of received packets). We should be able to confirm whether a broadcast node has relayed packets correctly or not. A broadcast node that has suffered from some sort of damage may not be able to conduct a relay. Moreover, it may be possible for a normal broadcast node to be mistaken for a broadcast node of a malicious actor.

Here, we focus on the problem of forensics [1] whereby a third party can prove there was an attack by collecting communication logs as evidence and analyzing them. The task of MANET forensics is to prove whether a normal node was mistaken for a malicious one. For this to be possible, when each node relays data in MANET, each node generates evidence that can be subject to objective investigation, and each node should be able to collect this evidence, save it, and analyze it and each node should be able to collect and save this evidence and pass it on to a third party who will analyze it.

So far, we have suggested a method of collecting evidence in MANET forensics [2]. The model that we suggested, however, did not include a method of analysis.

In this paper, we propose a forensics analysis method that takes into account the limited calculation resources of MANET nodes.

2 Related Studies

2.1 Evidence Collection Method

A. Otaka, et al. [2,3] suggested a method for collect evidence in MANET forensics. This method limits the object of investigation to a broadcast node and collects evidence from neighboring nodes, or ‘eyewitnesses’, as proof that it relayed. Figure 1 shows the basic model of evidence collection. If a problem is detected when a broadcast node relays a data packet, the node that the data packet reached in 0 hops starts evidence collection. While the upstream node collects evidence, the broadcast node does not throw away the packets which it received.

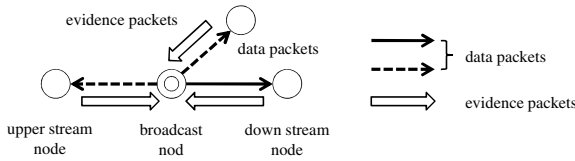


Fig. 1. Basic model of evidence collection

2.2 Bloom Filter

A Bloom filter [4] is a kind of data structure. It is used to determine whether the collection contains elements that may be targeted. It has much better space efficiency compared with other data structures, and it is time efficient in terms of collection and manipulation of data. Table 1 compares the characteristics of a Bloom filter(BF) with those of data structures (binary search tree, trie, hash table, simple array, linear list).

Table 1. Comparison of data structures (⊙:excellent,○:good,△:poor,×:no good)

	tree	trie	hash	array	list	BF
space	△	△	△	△	△	⊙
search	○	○	⊙	⊙	△	⊙
add	○	○	⊙	△	⊙	⊙
delete	○	○	⊙	△	⊙	×

For space efficiency, the evaluation focuses on the need to retain data elements. A Bloom filter is the best choice for this purpose because one can search a Bloom filter or add an element to it even if does not hold data. The evaluation focused on the amount of calculations entailed in performing search, add, and delete operations on sets. The Bloom filter is superior because it can search for or add elements in a fixed time. It can not however be used to remove an element.

The above considerations led us to choose the Bloom filter for a situation in which computational resources are limited.

3 Proposed Method

3.1 Scope of Method in Forensics

The network forensics process begins when an incident is discovered during operations. After the incident, evidence is acquired, saved, and analyzed [5][6]. In this study, we shall focus on the analysis.

3.2 Evidence Generation

The data packet and evidence packet are shown in Figure 2.

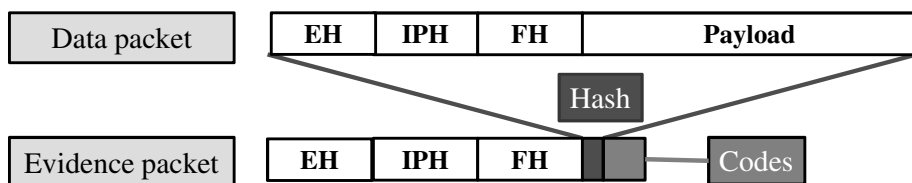


Fig. 2. Content of data packets and evidence packets

The process of evidence collection is as follows. First, the evidence collection node calculates a hash value for the data packet and make the size of the data uniform first by storing it in the Hash field. Accordingly, the content is less likely to be read illegally. Furthermore, the node takes a digital signature for this hash value and stores it in the code field and realizes that falsification of it is impossible. Next, it stores the IP address of the node generating the evidence in the FH of the data packet that it received. Finally, it stores the evidence generation time and evidence generation position information in the optional field of the FH of the data packet.

The content of the Evidence packet includes the Ethernet header to use when it sends back an evidence packet (EH), the IP header (IPH), Forensics header (FH), and the hash of the payload's IP address (Hash), and the device that performed a digital signature for the Hash (Codes).

3.3 Evidence Reception

The evidence collection node that receives an evidence packet distinguishes whether it came from upstream or downstream, and it stores this information in its evidence information table.

In this study, we shall deal with four important kinds of evidence:

1. Content of data packet (in Hash field),
2. Evidence generation time (in optional field of FH),
3. Node that generated evidence (Witness node IP in address part of FH),
4. The location of the generated evidence (in optional field of FH).

3.4 Evidence Analysis

The analysis is done in three steps: classification and organization of the evidence, search and extraction, and presentation of the results. Figure 3 shows the flow of the evidence analysis process.

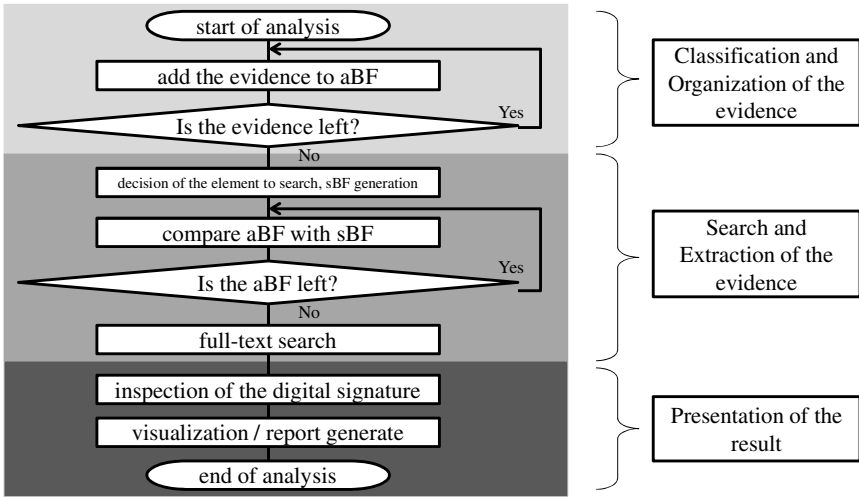


Fig. 3. Flow of evidence analysis

3.4.1 Classification and Organization of Evidence

The post-processing for the Bloom filter includes preparing m bit strings, setting all bits to 0, and building k hash functions. When the evidence is collected, the four pieces of information described in section 3.3 are extracted from the evidence packet. Next, the hash value of each piece of information that fit into m bit is calculated, and change the bit of the bit string matching to 1. n pieces of evidence are extracted from all the evidence and used to generate n Bloom filters. These filters are then logically added to form an assortment Bloom filter (aBF). The use of the aBF instead of the individual n Bloom filters helps to limit the search range and reduce the number of calculations. Figure 4 illustrates how the evidence is classified.

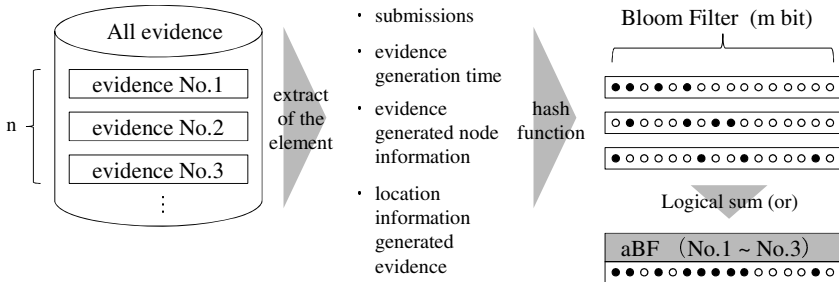


Fig. 4. Classification /organization of the evidence

3.4.2 Search and Extraction of Evidence

The Bloom filter (aBF) is used to find and extract certain elements from the evidence (four pieces of information described in Section 3.3). Suppose that we want to see if element d is in the evidence. First, the node sets all bits of the m bit string to 0 and creates k hash functions. Next, it calculates element d with the k hash functions and changes the bits of the bit string matching the hash value which it calculated to 1. The result is the Bloom filter to be searched (search Bloom filter, i.e., sBF). All aBFs are then compared with sBF. When an aBF encompassing sBF (i.e., all the 1 bits of sBF are included in some aBF) is discovered, the node conducts a full text search for the evidence in the range corresponding to aBF and extracts evidence including the objective element. As a result, all evidence including the certain element (a certain position or a certain time) is extracted. Figure 5 shows the process of searching for and extracting evidence.

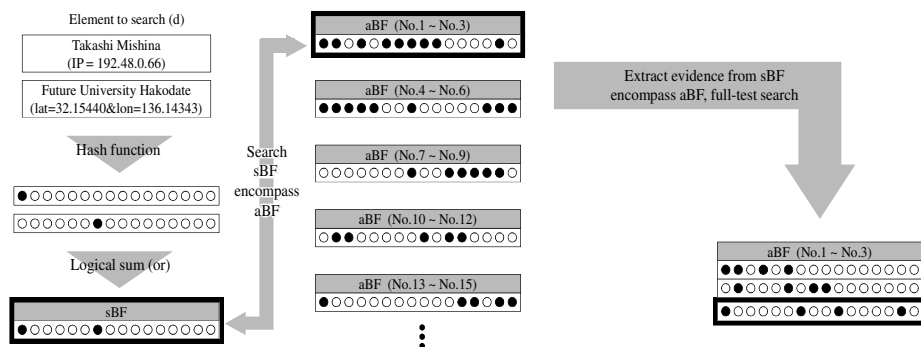


Fig. 5. Search / extraction of evidence

3.4.3 Presentation of the Result

After extraction with certain elements of evidence, performing the verification of digital signatures and abstraction of information to need technical knowledge. The digital signatures are verified and the log information is abstracted. The digital signature verification shows whether or not the evidence has been tampered with and thus increases the reliability of the evidence.

In addition, visualization has been reported to be effective in revealing network security risks [7][8]. IDS and firewall logs have a huge amount of output and are still not suitable for analysis. However, we can perform a moderate amount of abstraction to visualize the input. In the same way, it is important to report the output in an easily comprehensible way. It is necessary to compile appropriate information so that the output discloses information while preserving trust.

Although these aspects of security and visualization are important, we will have to forego discussion of them until we have finished examining methods to visualize important information in MANETs.

4 Evaluation

The evaluation used a full text search as the baseline for comparison. Table 2 shows the simulation environment.

Table 2. Simulation environment

OS	CentOS 5
CPU	AMD Athlon 64 X2 6000+ 3GHz
Memory	1024 MB
Number of evidence	10000
Number of bits of aBF (m)	8000
Number of evidence added to one aBF (n)	1 10 100 1000
Number of hash functions (k)	1400 140 14 1

The analyzed number of pieces of evidence was set at 10000, and the number of bits in the aBF was set at 8000. The number of pieces of evidence added to the aBF (n) was varied from 1 to 1000. From 1 to 1400 hash functions were used according to the value requested by the algorithm.

The experiment used the following indexes of evaluation.

(1) Runtime

This was the total of the time needed to make aBF, the time needed to compare sBF with all aBFs and to search for an aBF including sBF, and the time needed to extract evidence including a target element from all evidence within the range corresponding to the aBF including sBF.

(2) Memory usage

The memory area needed to generate aBF by using the proposal method.

(3) Probability of false positives

The probability of false positives generated when aBF including sBF is retrieved.

4.1 Results and Considerations

(1) Runtime

Figure 6 shows the runtime. The proposed method decreased the runtime compared with that of the baseline for $n=100$ and $n=1000$. Runtime decreased the most for $n=100$. This number yielded the shortest time probably because the number of times to compare sBF with aBF and the number of times to search the evidence in the range corresponding to aBF were fewer than in the other cases.

However, the amount of time taken to discover evidence for $n=1$ and $n=10$ was long. The cause is probably that the number of aBFs increases when the value of n is small and the hash values that have to be calculated increase, too. In particular, the runtime was very long for the case of $n=1$ i.e., one piece of evidence. It required a lot of aBFs and hash functions.

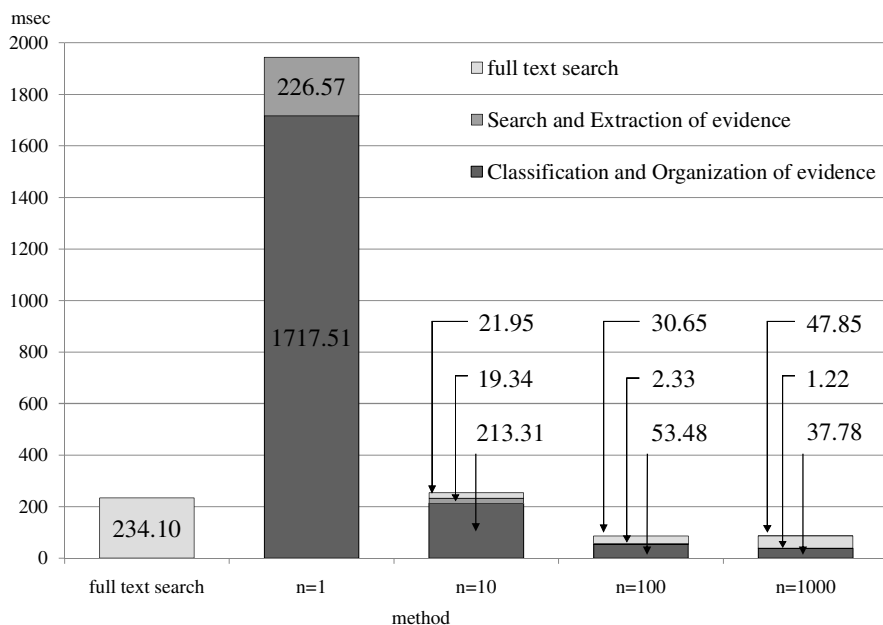


Fig. 6. Runtime result

(2) Memory usage

Table 3 shows the memory usage of aBF with the proposed method and the probability of false positives. Memory usage is computed with the following formula.

$$m \times \frac{\text{NumberOf Evidence}}{n} \quad (1)$$

The case of $n=1$ used the most memory (10 Mbytes), and the case of $n=1000$ used the least (10 kbytes). In other words, memory usage decreased when there was a lot of evidence for one aBF.

(3) Probability of false positives

The probability of false positives can be computed from the characteristics of the Bloom filter [4]. The hash values tended to collide when a lot of evidence was added to one aBF, and incorrect elements may be extracted. That is, the probability of the false positives decreased as n decreased.

Table 3. Memory usage and probability of false positives

	Number of generated aBFs	Memory usage (byte)	Probability of false positive (%)
$n = 1$	10000	10 M	0
$n = 10$	1000	1 M	1.87E-37
$n = 100$	100	100 k	0.006717766
$n = 1000$	10	10 k	39.34882957

4.2 Summary of Results and Considerations

The evaluation used 10000 pieces of evidence and an 8000-bit aBF, and it varied the amount of evidence added to aBF and the number of hash functions. It was found that the runtime, memory usage, and probability of false positive varied greatly depending on the number of evidence added to aBF, and these values were in trade-offs with each other. For evidence analysis in MANET forensics, we found that $n=100$ is the most suitable number to add, considering the limited calculation resources of nodes and the need for accuracy in extracting evidence.

5 Conclusion

We evaluated and implemented a forensics analysis method to prove whether content is relayed correctly or not in MANETs. The judgment at each node is done by using a Bloom filter in consideration of the probability of false positives. The experimental simulation showed that the proposed method is efficient: it has a short runtime and low retrieval frequency. Analysis of memory usage and the probability of the false positives confirmed that the method could extract the right evidence without using too many calculation resources if the value of n (evidence added to the filter) was adequately chosen.

In the future, we will ascertain the effect of varying the number of hash functions, amount of evidence, and number of bits in the aBF. Moreover, we will determine an appropriate set of values for evidence analysis in MANET forensics.

References

1. Jones, R.: The Internet forensics – Gathering and analysis of electronic evidence to solve crime. O'Reilly, Japan (2006)
2. Otaka, A., Takahashi, O., Takagi, T.: Reliable Method for Collecting and Evaluating Transmission Records for MANET Forensics. In: Proc. of International Workshop on Informatics (IWIN 2008), pp. 24–27 (2008)
3. Otaka, A., Takagi, T., Takahashi, O.: Network Forensics on Mobile Ad-hoc Networks. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 175–182. Springer, Heidelberg (2008)
4. Bloom, B.H.: Space/time trade-offs in hash coding with allowable errors. Communications of the ACM 13(7), 422–426 (1970)
5. Tsujii, S. (editorial supervisor): Digital forensics dictionary, Digital forensics society (2006)
6. Itou, T.N.M.: Trend of forensics technology that uses network information. NTT Journal, 36–40 (2004)
7. Abdullah, K., Lee, C., Conti, G.J., Copeland, J.A., Stasko, J.T.: IDS RainStorm: Visualizing IDS Alarms. In: Proc. IEEE Workshop on Visualization for Computer Security (VizSEC 2005), pp. 1–10. IEEE Computer Society, Los Alamitos (2005)
8. Koike, H., Ohno, K., Koizumi, K.: Visualizing Cyber Attacks using IP Matrix. In: Proc. IEEE Workshop on Visualization for Computer Security (VizSEC 2005), pp. 91–98. IEEE Computer Society, Los Alamitos (2005)

Diminished Reality for Landscape Video Sequences with Homographies

Kosuke Takeda and Ryuuki Sakamoto

Faculty of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama 640-8510, Japan
s111021@sys.wakayama-u.ac.jp

Abstract. This paper describes a Diminished Reality application to eliminate occluders on videos capturing distant landscapes. Although many tourist spots have lookouts to see the distant landscape, the video capturing from there usually includes obstructive occluders such as pillar and muntin. The application vanishes the occluders in the video in semi-interactively with homographies of distant landscape among frames. For this purpose, the regions of occluders in every frame are segmented by the GrabCut algorithm, and then the each region is recovered by the appropriate texture from another frame which is not occluded in same region.

Keywords: Diminished Reality, Video Inpainting, Distant Landscapes.

1 Introduction

As the camcorder has been the major equipment for consumers, many tourists have brought that to tourist spots. These many spots provide lookouts to see the beautiful distant landscape. Video sequences captured from such lookouts, however, tend to include obstructive objects such as pillar and muntin (Fig. 1).

The objective of Diminished Reality is to remove obstructive objects from video sequences. For realizing Diminished Reality, some techniques of computer vision are utilized for each frame of the video sequences. Image Inpainting [1] [2] [3] [4] [5] [6] [7] is a genre of computer vision techniques to recover lost portions in an image and recently applied to video sequences to remove obstructive objects, and thus the techniques can be used for Diminished Reality [8]. However, that is a quite time consuming process.

In this paper, we describe a handy Diminished Reality application to remove occluders from video sequences capturing a distant landscape from a lookout. In videos of distant landscapes, all objects on the ground such as buildings, trees, hills and etc. can be approximated as micro objects on one large plane, because the distance between the capturing spot and the landscape is quite long. In this case, the Homography can be used as the transformation matrix mapping from a pixel on one image to the pixel on another image. In the proposed application, we utilize the Homographies on the segmentation and compensation process to removing occluders.

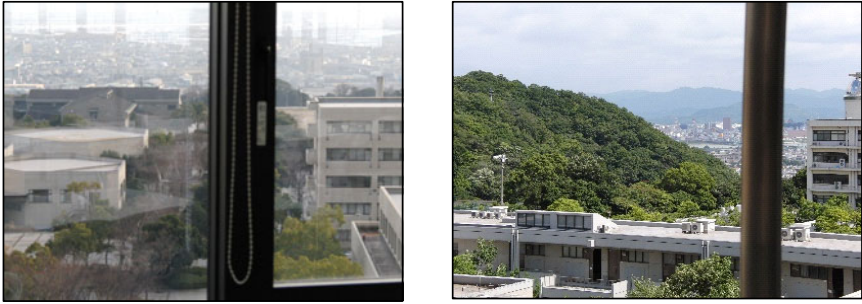


Fig. 1. Instances of occluders on distance landscapes

2 Estimating Homographies

To remove objects from video sequences, our application utilizes the subpixel-wise correspondence among the frames as transformation matrices. Hence the area which is far from camera can be approximated as an large flat plane, the matrix of the homography can be used as each transformation matrix.

The homographies corresponding frame by frame are utilized extracting occluder and recovering occluded pixels process. The extracting occluder process means segmenting between the distant landscape and occluders on each frame by using the GrabCut technique with the homographies. The recovering occluded pixels process maps the landscape textures obscured by occluder from the textures of another frame which is not occluded on the same position of the landscape (Fig. 2).

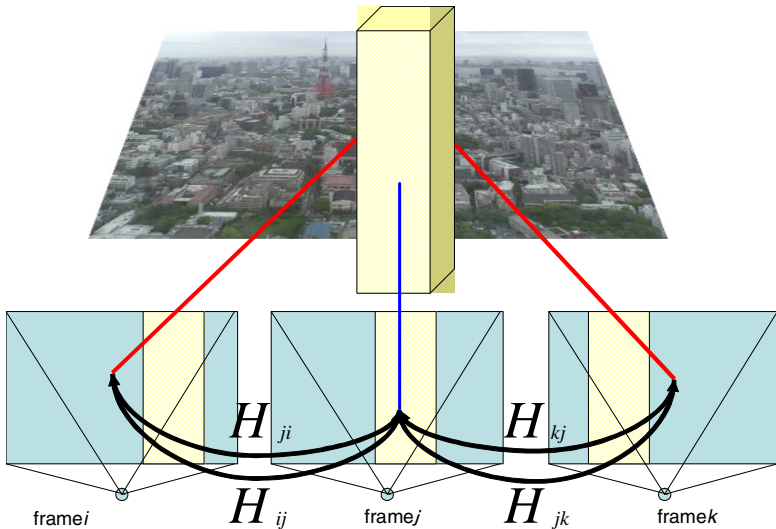


Fig. 2. Homographies for a distant landscape

To estimate the homography, more than four corresponding points are required. The corresponding points is a pair of points which point the same 3D scene in different images. Although many techniques to find corresponding points have been proposed, the approach finding feature points and then matching them is a most major technique. In our application, SIFT (Scale-Invariant Feature Transform) [9] is used to find feature points and ANN (Approximate Nearest Neighbor) is to match them. SIFT is an algorithm finding feature points, which have a 128 dimensional vector on each point, in sub-pixel order. After applying SIFT on every frame, many feature points are detected on each frame. All feature points in each frame are stored in the kd-tree of ANN, and then each feature point corresponds to the most appropriate feature point in another frame.

However, these pairs of feature points often include outlier. One reason is that ANN do not refer geometric relationships but 128 dimensional descriptor when matching feature points. In this case, the outliers can be detected by checking geometric consistency with the epipolar geometry.

In a pair of correspondent points between two camera images, the fundamental matrix (F) is carried by the following formula.

$$m'Fm = 0. \quad (1)$$

The m and m' means the position of a point and corresponding point of that, respectively. By using this formula for checking consistency with RANSAC (RANDOM Sample Consensus), inliers and outliers can be detected.

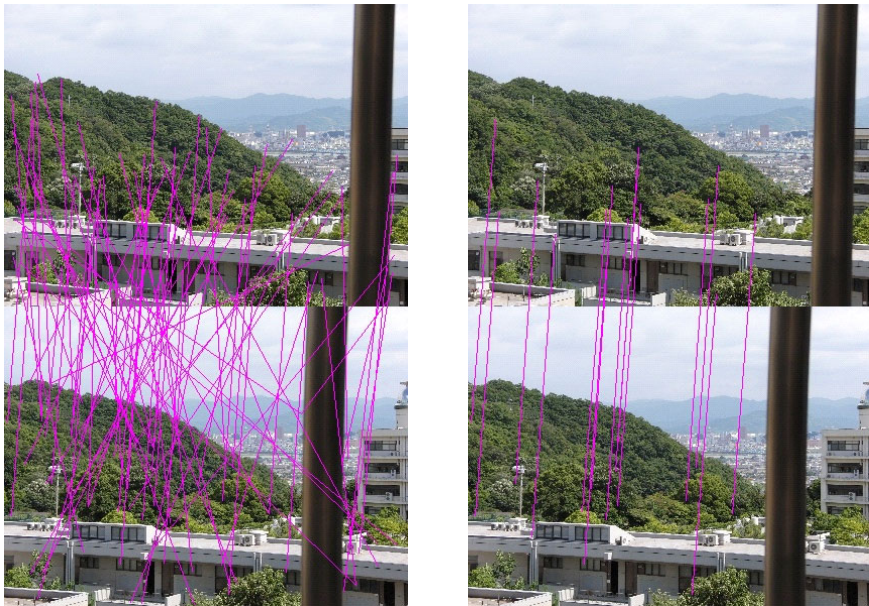


Fig. 3. Removing outliers

Another reason of including outlier is that SIFT may detect some feature points from occluder area. In this case, the corresponding point are correct as a pair in the epipolar geometry, and thus RANSAC cannot be applied for detecting outliers. However, there are some differences between corresponding points on the distant landscape and occluders. One of those is distance of translation from a frame to another frame. Because of perspective projection, the distance tends to be short if the point is on the distant landscape, and the distance is longer than that if on the occluders. Such difference can be detected by the following measure with the Mahalanobis' generalized distance. Fig. 3 depicts an input and result image of removing outliers.

$$\Delta d - \frac{\overline{\Delta d} - \Delta d}{\sigma} < Threshold. \quad (2)$$

3 Segmentation

The GraphCuts [10] [11] [12] had been a most important algorithm for image segmentations. In a typical application of the segmentation with GraphCuts, users roughly select two areas, which indicates foreground and background areas, with mouse operations. The application cuts out the foreground regions according to the result of the energy minimization with GraphCuts. The graph of GraphCuts means the two sorts of conceptual weighted links. One is the t-links which set between the terminal nodes and every pixel. Another is the n-links set among every neighbor pixel. In the GraphCuts algorithm, these links are cut to achieve a minimal cut of the graph by the max-flow min-cut theorem.

For applying the GraphCuts algorithm to videos, n-links must be set between not only neighbor pixels in an image but also same pixels on the back and forth frame. In distant landscape videos, movements may be occurred between the back and forth frame, and thus the two pixels connected by a n-link on back and forth frame should be decided as according to the movement. The movement can be estimated by the homographies mentioned in the Section 2.

In the video segmentation with the GraphCuts, it is preferable that the areas indicating background and foreground by users are selected for all frames. However, it involves an immense amount of time and effort. To avoid this effort, the selected areas for the background in the first frame can be propagated by using homography because background means distant landscape. On the contrary, the areas for the foreground meaning occluders cannot use the homographies. Therefore the movements such areas in another frames are estimated with the template matching method using SSD (Sum of Square Difference).

To improve the accuracy of the GraphCuts, Rother et al. proposed iterative algorithm called GrabCut [13] which applies a result of GraphCuts segmentation to the next input of that while the result is going divergent iteratively. Since the iterative approach of the algorithm can be applied to video sequences, our application has employed GrabCut to segment the occluders and distant landscape portion.

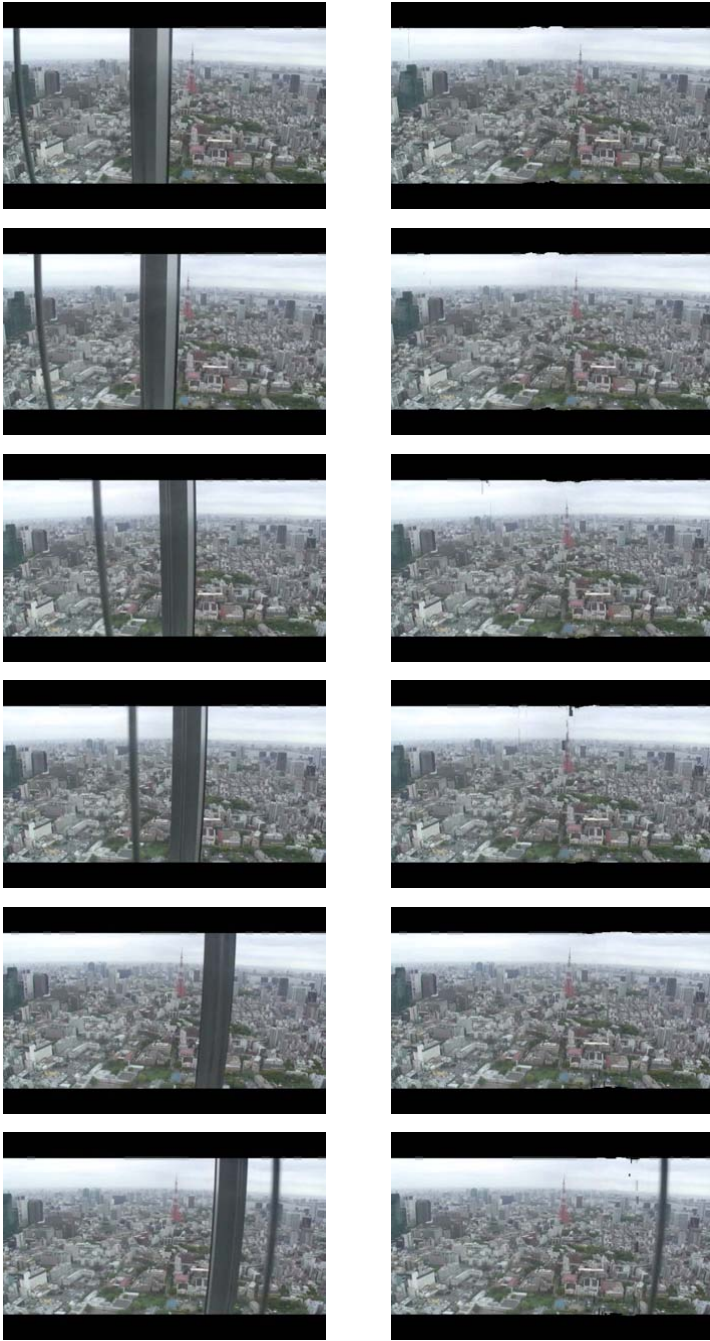


Fig. 4. Left: Input frames capturing a distant landscape from the observation room of a skyscraper. Right: Result of attempting to remove the muntin.



Fig. 5. Left: Input video containing small mountain in the near space. Right: Removed the pillar.

4 Removing Occluders

The basic idea to recover occluder areas by no-occluded views is assigning segmented pixels as the distant landscape at the another frame to detected pixels as occluders by segmentation. For every occluded pixel in the n -th frame, a the distant landscape pixel in the temporal nearest frame except the n -th frame are searched by using homographies. The pixel in the n -th frame are overwritten by the searched pixel detected as the distant landscape if such pixel is found.

Although such searching for the other frames can searched for every pixel individually, neighbor pixels in the same frame should be recovered by another same frame. For this purpose, the searching puts high priority on the frame which has large amount of pixels to be overriders.

5 Results

The Fig. 4 shows a result of the Diminished Reality application input video shown Fig. 4. Two pillars are almost vanished despite a high tower in the landscape region. However, the smaller pillar are slightly left in the later frame. It is caused by segmentation error deriving propagation errors of the user indication for occluders.

Fig. 5 is another video captured from a building. The video contains a small mountain and small buildings in near field. Fig. 5 shows a result applying the application and the pillar is almost vanished. The result, which recovering on the mountain and building with no unconformity of the appearance, suggests that using homographies for matching among frames in distant landscape region has propriety for this kind of Diminished Reality applications.

6 Conclusion

This paper has presented a handy Diminished Reality application for removing occluders from video sequences capturing a distant landscape from a lookout. Occluders was removed by landscape textures which behind them. The Landscape textures was estimated another frame in video sequences. The homography could utilize to know transformation on a image to another pixel on another image. In the proposed application, the Homographies was utilized on segmentation and compensation process to removing occluders.

References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual inproceedings on Computer graphics and interactive techniques, pp. 417–424. ACM Press/Addison-Wesley Publishing Co., New York (2000)
2. Chan, T., Shen, J.: Nontexture inpainting by curvature-driven diffusions. *Journal of Visual Communication and Image Representation* 12(4), 436–449 (2001)

3. Wei, L.Y., Levoy, M.: Fast texture synthesis using tree-structured vector quantization. In: Proceedings of the 27th annual inproceedings on Computer graphics and interactive techniques, pp. 479–488. ACM Press/Addison-Wesley Publishing Co, New York (2000)
4. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing* 13(9), 1200–1212 (2004)
5. Criminisi, A., Perez, P., Toyama, K.: Object removal by exemplar-based inpainting. In: *IEEE Computer Society inproceedings on Computer Vision and Pattern Recognition*, vol. 2, Citeseer (2003)
6. Wexler, Y., Shechtman, E., Irani, M.: Space-time video completion. In: *IEEE Computer Society inproceedings on Computer Vision and Pattern Recognition*, vol. 1. IEEE Computer Society, Los Alamitos (2004)
7. Jia, J., Wu, T.P., Tai, Y.W., Tang, C.K.: Video repairing: Inference of foreground and background under severe occlusion. In: *IEEE Computer Society inproceedings on Computer Vision and Pattern Recognition*, vol. 1. IEEE Computer Society, Los Alamitos (2004)
8. Lepetit, V., Berger, M.O.: A semi-automatic method for resolving occlusion in augmented reality. In: *IEEE Computer Society in Proceedings on Computer Vision and Pattern Recognition*, vol. 2 (2000)
9. Lowe, D.: Object recognition from local scale-invariant features. In: *International Conference on Computer Vision*, vol. 2, pp. 1150–1157 (1999)
10. Kwatra, V., Schodl, A., Essa, I., Turk, G., Bobick, A.: Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics* 22(3), 277–286 (2003)
11. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut / max -flow algorithms for energy minimization in vision. *PAMI* 26(9), 1124–1137 (2004)
12. Bugeau, A., Perez, P.: Joint tracking and segmentation of objects using graph cuts. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) *ACIVS 2007*. LNCS, vol. 4678, pp. 628–639. Springer, Heidelberg (2007)
13. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics* 23(3), 309–314 (2004)

Prediction of Combinatorial Protein-Protein Interaction Networks from Expression Data Using Statistics on Conditional Probability

Takatoshi Fujiki, Etsuko Inoue, Takuya Yoshihiro, and Masaru Nakagawa

Faculty of Systems Engineering, Wakayama University,
930, Sakaedani, Wakayama, 640-8510, Japan
{s101044,etsuko,tac,nakagawa}@sys.wakayama-u.ac.jp

Abstract. In this paper we propose a method to retrieve combinatorial protein-protein interaction to predict the interaction networks from protein expression data based on statistics on conditional probability. Our method retrieves the combinations of three proteins A, B and C which include combinatorial effects among them. The combinatorial effect considered in this paper does not include the “sole effect” between two proteins A-C or B-C, so that we can retrieve the combinatorial effect which appears only when proteins A, B and C get together. We evaluate our method with a real protein expression data set and obtain several combinations of three proteins in which protein-protein interactions are predicted.

Keywords: Protein-Protein Interaction, Expression Data, Data Mining, Interaction Network.

1 Introduction

After the whole DNA sequences have been in public, many post-genome researches began to investigate the system of living creatures. Creatures consist of vast sort of proteins and their bodies are maintained by complex interactions of genes and proteins. One of the major interests in this area is that how the characteristics of each creature appear and what kind of proteins, genes and their interactions are related with them.

Many approaches are tried to clarify the interactions of proteins (or genes) which build up the mechanisms of living creatures [1]. Among them, one major approach is to derive some patterns or rules from expression data which imply some interactions of genes or proteins [2][3][4][5][6][7]. In fact, much work from this approach has been done with gene expressions of microarrays. Microarrays are so useful because we can treat expression of thousands of genes simultaneously. Several statistical or informatical techniques are applied for expression data to find interactions among genes and further to infer their interaction networks. Bayesian networks [2] and Boolean networks [3] are the typical strategy among them and they have been contributed to predict functions of genes. Especially, bayesian networks are useful at computational speed so that it handles

vast combinations of genes to test whether each of the combinations of genes has any interaction with one another. However, the Bayesian network has a shortcoming that a group of genes (or proteins) in which every pair of genes (or proteins) has correlation with each other tends to be retrieved. It means that interactions within two genes (or proteins) may prevent from understanding interactions among more than three genes (or proteins).

In this paper, we concentrate on predicting protein-protein interaction from protein expression data and propose a method to predict combinatorial protein-protein interactions based on conditional probability and statistics. Our method retrieves from expression data the interactions among three proteins without considering the effect between two proteins. Namely, our method is able to retrieve the interaction of two proteins effecting on a protein, which appears only when the three proteins get together. We evaluate our method by applying it to a real protein expression data set and found several combinations of three proteins in which combinatorial interaction is predicted.

The rest of this paper is organized as follows. In Section 2, we describe the model of protein-protein interaction considered in our method, then present the method to retrieve the combinations of three proteins which is predicted to have interaction using statistics on conditional probability. In Section 3, we evaluate our method by applying to real protein-protein expression data, and finally in Section 4 we conclude the work.

2 The Method to Retrieve Combinatorial Effects

2.1 Expression Data Used In Our Method

In this section, we explain the operational process of the experiment to obtain protein expression data. Protein expression data represent the expression level of each protein i in sample j . Typically, the number of proteins in the data are about several hundreds to thousands while the number of samples is usually several tens and at most hundreds. The process of obtaining protein expression data in this work is somehow complicated compared to microarray, which measures gene expression levels. See Figure 1. First, we prepare target samples and obtain 2-dimensional electrophoresis images from each target sample through biological experimental processes. Second, we identify spots of separated proteins using an image processing software and measure the expression level of each spot. Third, we match the spots of the same protein among the images in the experiment. Finally, we normalize the values of expression levels using one of the normalization methods as a preprocess of the data mining processes. As a result, we have a set of protein expression values as shown in Figure 2, where you see expression level values of each protein in each sample.

2.2 The Combinatorial Protein-Protein Interaction Model

The protein-protein interaction model we try to predict in this paper is shown in Figure 3. Three proteins A, B, and C, are related to this model, where each

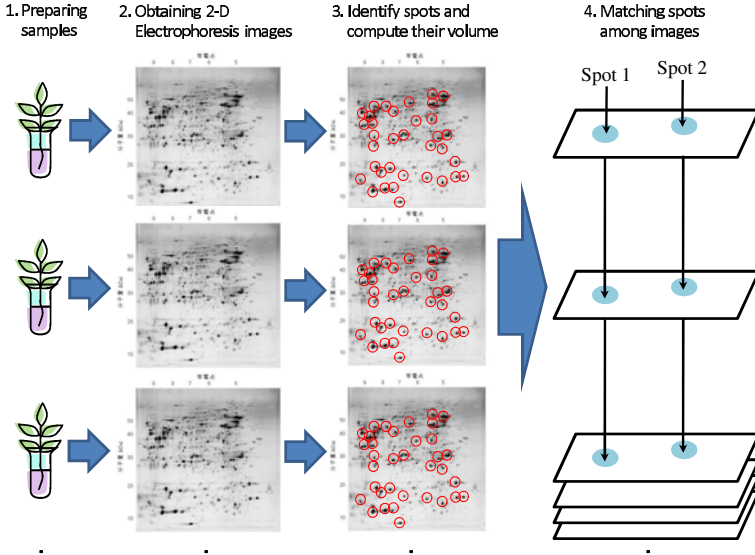


Fig. 1. The process of obtaining Proteome Expression Data

Sample ID	Protein ID				
	A	B	C	D	...
1	0.000582	0.000107	0.000338	0.000451	...
2	0.000563		0.000475	0.000458	...
3	0.000495	0.000126	0.000433	0.000565	...
4	0.000553	0.000153	0.000382	0.000486	...
5	0.000536	0.000134	0.000536	0.000471	...
6	0.000601	0.000185	0.000457	0.000513	...
:	:	:	:	:	:

Fig. 2. The Data Format for Our Data Mining Process

of A and B solely effects on the expression level of C, but if both A and B are expressed together, they have far larger effect on the expression level of C. We call the sum of two sole effects and the combinatorial effect, *the total effect*. What we want to retrieve from expression data is the combinatorial effect of A and B on C, which is not seen if A or B solely expressed in a sample. To measure this combinatorial effect, we first have to estimate total effect of A and B on C, and from the total effect level we have to subtract the sole effect of each of A-C and B-C to obtain the combinatorial effect level.

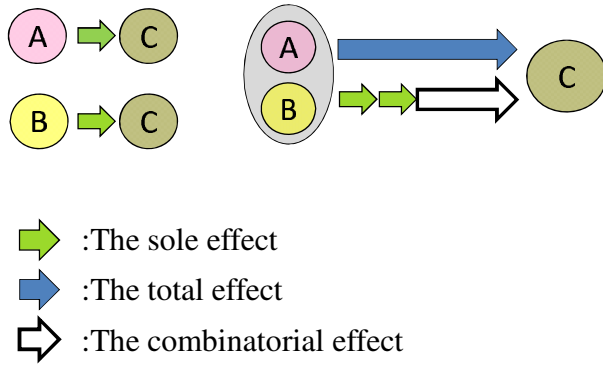


Fig. 3. The Interaction Model to Predict

2.3 Estimating Sole and Total Interaction Levels Based on Conditional Probability

Our idea to retrieve this interaction from expression data is to use conditional probability. The probability of the sole interaction of A-C and B-C is measured by conditional probability as shown in Figure 4. Namely, the sole interaction effect level of A on C is measured as the ratio of the number of samples where the expression levels of both A and C are sufficiently high out of the number of samples where the expression level of A is sufficiently high. The total interaction effect of A and B on C is also measured in a similar fashion, i.e., the ratio of the number of samples where the expression level of A, B and C are all sufficiently high out of the number of samples where the expression levels of both A and B are sufficiently high.

Now we give the definitions and formulation of our problems. We handle proteins $i(1 \leq i \leq I)$ and samples $j(1 \leq j \leq J)$, both of which are included in the input expression data. We sometimes call some of the proteins A, B, C, ..., and so on. As a parameter, we define r be the threshold on ratio to judge the expression, i.e., if the expression level of sample j for protein i is within top $r\%$ among all the expression levels of protein i , we call the protein i is “expressed” in sample j . Let $num(A)$ be the number of samples at which protein A is expressed, and similarly, let $num(A \cap B)$ be the number of samples at which both protein A and B is expressed. Then, we define $E_A^C = \frac{num(A \cap C)}{num(A)}$ as the sole effect level of A on C. Similarly, the sole effect level of B on C is defined as $E_B^C = \frac{num(B \cap C)}{num(B)}$, the total effect level of A and B on C is defined as $E_{A,B}^C = \frac{num(A \cap B \cap C)}{num(A \cap B)}$.

2.4 Retrieving Combinatorial Effect

What we want to estimate is the amount of the combinatorial interaction effect level, which can be estimated from the total interaction level (presented in the

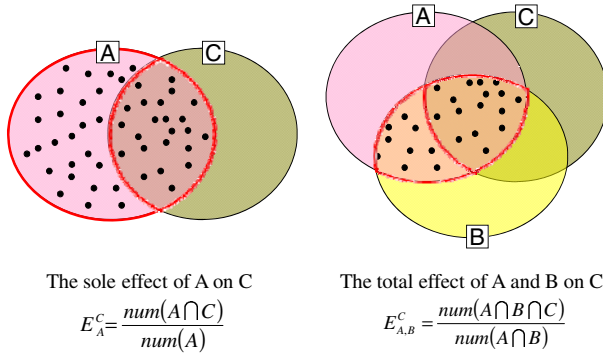


Fig. 4. How to Measure Sole and Total Effect Level of Protein A and B on C

previous section) and the sole effect level of A-C and B-C. See Figure 5. To estimate the combinatorial effect level for the combination of three proteins A, B, and C, we split the total interaction effect into two parts, i.e., two sole interaction effects and the combinatorial effect. Then, the difference between them is regarded as the combinatorial effect level that we object to compute. To obtain the combinatorial effect level, we compute the statistical distribution of the total effect level $E'_{A,B}{}^C$, which is computed through simulation under the assumption that no combinatorial effect exists over A, B and C. From the distribution of $E'_{A,B}{}^C$ and the total effect score $E_{A,B}{}^C$ (which is the total effect presented in the previous subsection), we can estimate the combinatorial effect level.

Computer simulation to compute the distribution of $E'_{A,B}{}^C$ is done as follows. For the corresponding value of α and β , which are the sole effect values for the combination A-C and B-C, we first create distributions of A, B and C randomly

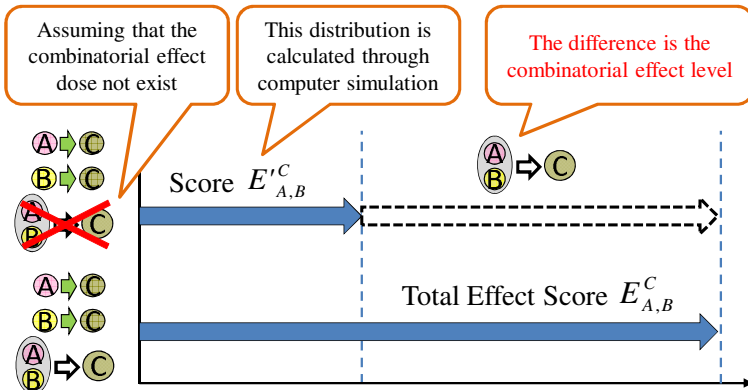


Fig. 5. Dividing Total Effect into Sole and Combinatorial Effect

such that the sole effect levels of A-C and B-C are α and β , respectively. Since those distributions are created randomly, it is possible to assume that they do not include any combinatorial effect. Then we compute the total effect score of the combination A, B and C. After the sufficient number of repetition of this process, we obtain the distribution of $E'_{A,B}{}^C$ as a pile of the total effect scores. Note that we do not care what kind of distribution that A, B and C follow in our method, since we judge whether the protein is expressed or not with the threshold r of the ranking in expression levels.

From this total effect distribution $E'_{A,B}{}^C$, we compute the combinatorial effect as a z-score in the distribution of $E'_{A,B}{}^C$. The z-score $z_{A,B}^C$ is defined as $z_{A,B}^C = \frac{(E_{A,B}^C - \mu)}{\sigma}$, where $E_{A,B}^C$ is the total effect level of A, B and C obtained from the real data, μ and σ are the average and the standard deviation of the distribution of $E'_{A,B}{}^C$ obtained from the computer simulation. Namely, z-score is the difference between the average μ of $E'_{A,B}{}^C$ distribution and the real total effect level obtained from the real data, which is measured as the number of the unit value σ . Intuitively, z-score indicates the probability of occurring the value $E_{A,B}^C$ assuming that the combinatorial effect does not exist, which implies the level of the combinatorial effect.

To compute the distribution of total effect levels through the simulation, however, much running time is required so that it is desirable to pre-compute the distribution. Thus, we prepared a distribution-table which shows the average and the standard deviation of the distribution for each values of α and β , as shown in Figure 6. Note that when we computed Figure 6, we prepare the data of A, B and C with 100 samples and we did 1,500,000 trials for each pair of α and β . Since we computed the table for 20 values of α and β between 0 and 1, when we want to obtain the corresponding values of μ and σ we use the value in the table which is the closest to the α and β of A, B and C.

Now we summarize the proposed method. First, we enumerate every combination of three proteins A, B and C from the input data set. For each of the combinations, we compute the total effect level $E_{A,B}^C$ of A, B and C. By referring the pre-computed distribution table, we find the distribution of $E'_{A,B}{}^C$ corresponding to the value α and β of A, B and C. From the distribution of $E'_{A,B}{}^C$ and the total effect level $E_{A,B}^C$, we obtain the combinatorial effect level of A, B and C as the corresponding z-score. Finally we create a ranking of all the combinations of three proteins by ordering them by z-score.

3 Evaluation

We evaluate the proposed method by applying it to the real protein expression data obtained by a 2-D electrophoresis-based experiment [8]. The data include 195 samples and 879 proteins, and have been processed by global normalization [9].

Although our method does not assume that protein expression levels follow any specific distributions, we first confirmed that the expression data we use

Average table

		The sole effect of B on C																				
		0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
The sole effect of A on C	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.05	0.0000	0.0064	0.0135	0.0212	0.0298	0.0393	0.0500	0.0620	0.0757	0.0913	0.1094	0.1305	0.1556	0.1858	0.2228	0.2694	0.3296	0.4107	0.5255	0.7005	1.0000
	0.10	0.0000	0.0135	0.0280	0.0437	0.0609	0.0795	0.1000	0.1225	0.1473	0.1750	0.2059	0.2406	0.2800	0.3251	0.3770	0.4377	0.5093	0.5953	0.7003	0.8315	1.0000
	0.15	0.0000	0.0212	0.0437	0.0678	0.0933	0.1207	0.1500	0.1815	0.2154	0.2520	0.2917	0.3348	0.3819	0.4334	0.4901	0.5528	0.6224	0.7002	0.7877	0.8868	1.0000
	0.20	0.0000	0.0298	0.0609	0.0933	0.1273	0.1628	0.2000	0.2390	0.2800	0.3231	0.3685	0.4163	0.4668	0.5201	0.5766	0.6365	0.7001	0.7679	0.8401	0.9173	1.0000
	0.25	0.0000	0.0393	0.0795	0.1207	0.1628	0.2059	0.2500	0.2952	0.3415	0.3889	0.4376	0.4874	0.5385	0.5910	0.6449	0.7001	0.7569	0.8152	0.8751	0.9367	1.0000
	0.30	0.0000	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5001	0.5501	0.6001	0.6501	0.7001	0.7501	0.8001	0.8501	0.9001	0.9500	1.0000
	0.35	0.0000	0.0620	0.1225	0.1815	0.2390	0.2952	0.3500	0.4035	0.4559	0.5069	0.5569	0.6057	0.6534	0.7001	0.7457	0.7904	0.8341	0.8769	0.9188	0.9598	1.0000
	0.40	0.0000	0.0757	0.1473	0.2154	0.2800	0.3415	0.4000	0.4559	0.5091	0.5601	0.6088	0.6554	0.7001	0.7429	0.7841	0.8236	0.8616	0.8982	0.9334	0.9673	1.0000
	0.45	0.0000	0.0913	0.1750	0.2520	0.3231	0.3889	0.4500	0.5069	0.5601	0.6097	0.6563	0.7001	0.7412	0.7801	0.8167	0.8514	0.8843	0.9154	0.9450	0.9732	1.0000
	0.50	0.0000	0.1094	0.2059	0.2917	0.3685	0.4376	0.5001	0.5569	0.6088	0.6563	0.7000	0.7405	0.7779	0.8126	0.8449	0.8750	0.9033	0.9297	0.9546	0.9780	1.0000
	0.55	0.0000	0.1305	0.2406	0.3348	0.4163	0.4874	0.5501	0.6057	0.6554	0.7001	0.7405	0.7779	0.8106	0.8412	0.8694	0.8954	0.9194	0.9417	0.9623	0.9819	1.0000
	0.60	0.0000	0.1556	0.2800	0.3819	0.4668	0.5385	0.6001	0.6534	0.7001	0.7412	0.7779	0.8106	0.8399	0.8667	0.8909	0.9131	0.9334	0.9520	0.9692	0.9852	1.0000
	0.65	0.0000	0.1858	0.3251	0.4334	0.5201	0.5910	0.6501	0.7001	0.7429	0.7801	0.8126	0.8412	0.8667	0.8894	0.9100	0.9286	0.9455	0.9609	0.9750	0.9880	1.0000
	0.70	0.0000	0.2228	0.3770	0.4901	0.5766	0.6449	0.7001	0.7457	0.7841	0.8167	0.8449	0.8694	0.8909	0.9100	0.9269	0.9423	0.9561	0.9686	0.9800	0.9904	1.0000
	0.75	0.0000	0.2694	0.4377	0.5528	0.6365	0.7001	0.7501	0.7904	0.8236	0.8514	0.8750	0.8954	0.9131	0.9286	0.9423	0.9545	0.9655	0.9754	0.9844	0.9925	1.0000
	0.80	0.0000	0.3296	0.5093	0.6224	0.7001	0.7569	0.8001	0.8341	0.8616	0.8843	0.9033	0.9194	0.9334	0.9455	0.9561	0.9655	0.9738	0.9815	0.9882	0.9944	1.0000
0.85	0.0000	0.4107	0.5953	0.7002	0.7679	0.8152	0.8501	0.8769	0.8982	0.9154	0.9297	0.9417	0.9520	0.9609	0.9686	0.9754	0.9815	0.9868	0.9917	0.9960	1.0000	
0.90	0.0000	0.5255	0.7003	0.7877	0.8401	0.8751	0.9001	0.9188	0.9334	0.9450	0.9546	0.9623	0.9692	0.9750	0.9800	0.9844	0.9882	0.9917	0.9947	0.9975	1.0000	
0.95	0.0000	0.7005	0.8315	0.8868	0.9173	0.9367	0.9500	0.9598	0.9673	0.9732	0.9780	0.9819	0.9852	0.9880	0.9904	0.9925	0.9944	0.9960	0.9975	0.9988	1.0000	
1.00	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	

Standard deviation table

		The sole effect of B on C																						
		0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00		
The sole effect of A on C	0.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		
	0.05	0.0000	0.0024	0.0032	0.0040	0.0046	0.0053	0.0059	0.0065	0.0070	0.0076	0.0082	0.0088	0.0095	0.0103	0.0111	0.0122	0.0137	0.0156	0.0183	0.0209	0.0000		
	0.10	0.0000	0.0032	0.0048	0.0055	0.0064	0.0071	0.0078	0.0084	0.0090	0.0096	0.0101	0.0107	0.0112	0.0117	0.0125	0.0133	0.0141	0.0150	0.0155	0.0144	0.0000		
	0.15	0.0000	0.0040	0.0055	0.0069	0.0076	0.0084	0.0090	0.0096	0.0101	0.0105	0.0109	0.0114	0.0117	0.0121	0.0125	0.0129	0.0132	0.0132	0.0127	0.0108	0.0000		
	0.20	0.0000	0.0046	0.0064	0.0076	0.0087	0.0092	0.0098	0.0103	0.0107	0.0110	0.0113	0.0115	0.0117	0.0119	0.0120	0.0121	0.0120	0.0116	0.0106	0.0085	0.0000		
	0.25	0.0000	0.0053	0.0071	0.0084	0.0092	0.0099	0.0103	0.0107	0.0110	0.0112	0.0113	0.0114	0.0115	0.0115	0.0114	0.0112	0.0109	0.0102	0.0091	0.0070	0.0000		
	0.30	0.0000	0.0059	0.0078	0.0090	0.0098	0.0103	0.0107	0.0109	0.0111	0.0111	0.0112	0.0112	0.0112	0.0111	0.0109	0.0107	0.0103	0.0098	0.0090	0.0078	0.0059	0.0000	
	0.35	0.0000	0.0065	0.0084	0.0096	0.0103	0.0107	0.0109	0.0111	0.0111	0.0111	0.0110	0.0110	0.0108	0.0106	0.0104	0.0100	0.0095	0.0089	0.0080	0.0068	0.0050	0.0000	
	0.40	0.0000	0.0070	0.0090	0.0101	0.0107	0.0110	0.0111	0.0111	0.0111	0.0111	0.0111	0.0110	0.0108	0.0104	0.0101	0.0098	0.0093	0.0087	0.0080	0.0071	0.0060	0.0044	0.0000
	0.45	0.0000	0.0076	0.0096	0.0105	0.0110	0.0112	0.0112	0.0111	0.0109	0.0112	0.0103	0.0100	0.0096	0.0092	0.0086	0.0080	0.0073	0.0064	0.0053	0.0038	0.0000	0.0000	
	0.50	0.0000	0.0082	0.0101	0.0109	0.0113	0.0113	0.0112	0.0110	0.0107	0.0103	0.0109	0.0095	0.0091	0.0085	0.0080	0.0073	0.0066	0.0057	0.0047	0.0033	0.0000	0.0000	
	0.55	0.0000	0.0088	0.0107	0.0114	0.0115	0.0114	0.0112	0.0108	0.0104	0.0100	0.0095	0.0104	0.0088	0.0079	0.0073	0.0067	0.0059	0.0051	0.0042	0.0029	0.0000	0.0000	
	0.60	0.0000	0.0095	0.0112	0.0117	0.0117	0.0115	0.0111	0.0106	0.0101	0.0096	0.0091	0.0085	0.0079	0.0073	0.0067	0.0061	0.0054	0.0046	0.0037	0.0026	0.0000	0.0000	
	0.65	0.0000	0.0103	0.0118	0.0121	0.0119	0.0115	0.0109	0.0104	0.0098	0.0092	0.0085	0.0079	0.0073	0.0068	0.0061	0.0055	0.0048	0.0041	0.0033	0.0023	0.0000	0.0000	
	0.70	0.0000	0.0111	0.0125	0.0125	0.0120	0.0114	0.0107	0.0100	0.0093	0.0086	0.0080	0.0073	0.0067	0.0061	0.0055	0.0049	0.0043	0.0036	0.0029	0.0020	0.0000	0.0000	
	0.75	0.0000	0.0122	0.0133	0.0129	0.0121	0.0112	0.0103	0.0095	0.0087	0.0080	0.0073	0.0067	0.0061	0.0055	0.0049	0.0043	0.0038	0.0032	0.0025	0.0017	0.0000	0.0000	
	0.80	0.0000	0.0137	0.0141	0.0132	0.0120	0.0109	0.0098	0.0089	0.0080	0.0073	0.0066	0.0059	0.0054	0.0048	0.0043	0.0038	0.0030	0.0022	0.0021	0.0015	0.0000	0.0000	
0.85	0.0000	0.0156	0.0150	0.0132	0.0116	0.0102	0.0090	0.0080	0.0071	0.0064	0.0057	0.0051	0.0046	0.0041	0.0036	0.0032	0.0027	0.0039	0.0018	0.0012	0.0000	0.0000		
0.90	0.0000	0.0183	0.0155	0.0127	0.0106	0.0091	0.0078	0.0068	0.0060	0.0053	0.0047	0.0042	0.0037	0.0033	0.0029	0.0025	0.0021	0.0018	0.0025	0.0010	0.0000	0.0000		
0.95	0.0000	0.0209	0.0144	0.0108	0.0085	0.0070	0.0059	0.0050	0.0044	0.0038	0.0033	0.0029	0.0026	0.0023	0.0020	0.0017	0.0015	0.0012	0.0010	0.0013	0.0000	0.0000		
1.00	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		

Fig. 6. The Distribution Table of $E_{A,B}^C$ Created through Simulation

follows normal distribution or not to comprehend the property of input data. We omitted expression levels which apart from the average by more than three-times of the standard deviation as outlier, and applied Jarque-Bera test to judge whether the expression levels of each protein follow normal distribution or not. The result is that 454 out of 879 proteins follow normal distribution with the level of significance of 5%.

We implemented the proposed method in C++ language. As computational parameters, the threshold ratio used is $r = 30\%$. To ensure the statistical significance, we omitted the combinations of three proteins if $num(A \cap B)$, which is the denominator in the total effect level, is less than 20. From the same reason we also omitted the combinations if $num(A \cap B \cap C)$ is less than 10. Note that the data include null expression values. We only used the expression values where all the expression values for three proteins are not null.

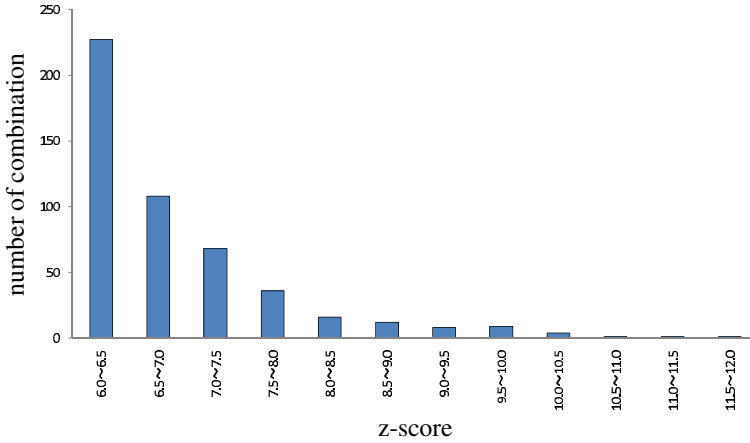


Fig. 7. The ranking of z-score

rank	A (spot No.)	B (spot No.)	C (spot No.)	z-score	E_A^C	E_B^C	$E_{A,B}^C$	num(A∩B)	num(A∩B∩C)
1	5957	6235	1500	11.9854	0.2857	0.3846	0.5000	20	10
2	2490	2755	5587	11.1771	0.2500	0.3704	0.4348	23	10
3	1531	5366	1526	10.6472	0.8800	0.5128	1.0000	20	20
4	6142	6236	1520	10.1442	0.3030	0.3846	0.5000	20	10
5	5052	1471	1476	10.0998	0.6176	0.8519	1.0000	20	20
6	6043	6233	5420	10.0876	0.3030	0.3226	0.4348	23	10
7	2490	2755	237	10.0750	0.2703	0.4000	0.4762	21	10
8	4710	5311	1520	9.9173	0.3125	0.3333	0.4545	22	10
9	3626	5366	1520	9.9035	0.2778	0.4167	0.5000	20	10
10	4710	6236	1500	9.8779	0.3235	0.3226	0.4545	22	10
11	6142	6236	1529	9.6033	0.3125	0.3571	0.4762	21	10
12	6142	6236	1531	9.6033	0.3125	0.3571	0.4762	21	10
13	1443	5620	6164	9.5513	0.8214	0.7059	1.0000	20	20
14	5957	6233	1500	9.5098	0.2941	0.3571	0.4545	22	10
15	5285	5957	1526	9.5064	0.3704	0.3226	0.5000	20	10
16	6142	6236	5420	9.5064	0.3226	0.3704	0.5000	20	10
17	1531	1476	1526	9.4073	0.8750	0.6061	1.0000	20	20
18	1762	4189	6036	9.1893	0.7333	0.8148	1.0000	20	20
19	1476	3880	5418	9.1210	0.8800	0.6111	1.0000	22	22
20	4292	6021	1526	9.0895	0.4348	0.2703	0.5000	20	10
21	2490	2755	2480	9.0784	0.2750	0.4074	0.4783	23	11
22	5060	1452	1466	9.0543	0.7222	0.8276	1.0000	24	24
23	5622	5984	1520	9.0393	0.3235	0.4231	0.5500	20	11
24	5266	4795	5207	9.0166	0.8276	0.7241	1.0000	20	20
25	1531	1429	1529	8.9543	0.8889	0.5946	1.0000	22	22
26	5675	6021	1500	8.9151	0.3226	0.3333	0.4545	22	10
27	5620	1443	1476	8.9116	0.6765	0.8571	1.0000	20	20
28	6236	4710	5420	8.8985	0.3548	0.3235	0.4783	23	11

Fig. 8. The histogram of z-score

The histogram of the z-scores is shown in Figure 7. In this figure the numbers of the combinations included in each section of z-score are shown out of ${}_{879}C_3$ combinations of three proteins. There are several very large z-scores which implies the strong possibility of existing combinatorial effects.

Figure 8 shows the ranking of top 28 combinations of proteins in the result. In this table several values for each combination is shown including $E_A^C, E_B^C, E_{A,B}^C, num(A \cap B), num(A \cap B \cap C)$. Note that in many cases $num(A \cap B)$ takes 20, which is the threshold value to omit the combination. Also in many cases

$num(A \cap B \cap C)$ takes the lowest value (i.e., the threshold value 10) or the highest value (i.e., the value of $num(A \cap B)$). Although the reason is not known, but this may indicate some property of this method. Note that the result includes both high and low values of total effect levels $E_{A,B}^C$. Our method succeeded to retrieve the combination of lower total effect levels (, but including high combinatorial effect level), which is not able to be retrieved by Bayesian network based methods.

4 Conclusion

In this paper we proposed a method to retrieve combinatorial protein-protein interactions based on statistics on conditional probability. Our method tries to retrieve the combinatorial effect level instead of retrieving the total effect level. This is the major difference from the other methods proposed ever in the literature. We evaluated our method using the real protein expression data obtained by a 2-D electrophoresis based experiment. We found that the result includes strange tendency that the high-score combinations are classified into two types, i.e., $num(A \cap B \cap C)$ takes the maximum value or the minimum value. Although it is desirable to clarify the reason of the tendency, we retrieved many combinations in which existence of the combinatorial effect is strongly predicted.

As future work, we would like to perform more experiments to clarify the whole tendency on the results of the proposed method. And, find the known interactions in our result to verify the availability of this data-mining method.

Acknowledgment

This work was partly supported by the Programme for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry, Japan.

References

1. Enright, A.J., Skrabanek, L., Bader, G.: Computational Prediction of Protein-Protein Interactions. In: *The Proteomics Protocols Handbook*, pp. 629–652. Humana Press (2005)
2. Wang, L., Chu, F., Xie, W.: Accurate Cancer Classification Using Expressions of Very Few Genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4(1), 40–53 (2007)
3. Imoto, S., Goto, T., Miyano, S.: Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. In: *Pacific Symposium on Biocomputing*, vol. 7, pp. 175–186 (2002)
4. Akutsu, T., Kuhara, S., Maruyama, O., Miyano, S.: Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model. *Theoretical Computer Science* 298, 235–251 (2003)

5. Yoshihiro, T., Inoue, E., Nakagawa, M.: Mining Combinatorial Effects on Quantitative Traits from Protein Expression Data. In: 8th Joint Conference on Knowledge-based Software Engineering 2008 (JCKBSE 2008), pp. 359–367 (2008)
6. Lu, Y., Liu, F., Sanchez, M., Wang, Y.: Interactive Semisupervised Learning for Microarray Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4(2), 190–203 (2007)
7. DeRisi, J.L., Lyer, V.R., Brown, P.O.: Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* 278, 680–686 (1997)
8. Nagai, K., Yoshihiro, T., Inoue, E., Ikegami, H., Sono, Y., Kawaji, H., Kobayashi, N., Matsuhashi, T., Ohtani, T., Morimoto, K., Nakagawa, M., Iritani, A., Matsumoto, K.: Developing an Integrated Database System for the Large-scale Proteomic Analysis of Japanese Black Cattle. *Animal Science Journal* 79(4), 467–481 (2008)
9. Lu, C.: Improving the Scaling Normalization for High-density Oligonucleotide GeneChip Expression Microarrays. *BMC Bioinformatics* 5(103) (2004)

Development and Evaluation of a Historical Tour Support System Using 3D Graphics and Mobile Terminal

Satoru Fujii¹, Takahiro Shima¹, Megumi Takahashi¹, and Koji Yoshida²

¹ Matsue National College of Technology, 14-4 Nishi-Ikuma Matsue, Shimane 690-8518, Japan
fujii@matsue-ct.ac.jp, {s0916, j0523}@stu.cc.matsue-ct.ac.jp

² Shonan Institute of Technology, 1-1-25, Tsujidonishikaigan, Fujisawa, Kanagawa
251-8511, Japan

yoshidak@info.shonan-it.ac.jp

Abstract. We have developed a historical tour support system based on Matsue Castle and the surrounding historic area. The system has pre tour mode using 3D graphics and tour mode using a mobile terminal. We used Virtools to modify our earlier system to enable mouse operation in 3D space in the pre-tour mode. Evaluation of the pre-tour and tour modes by users resulted in positive responses although there were areas for improvement. As a result, the tour mode was upgraded by using smart phone with touch pen draw and notation options. Other improvements and applications are proposed.

1 Introduction

In our daily life, we have many chances to study outdoors or go sightseeing. Information is usually accessed from the Internet or from guidebooks. There are still comparatively few systems that can provide individuals with information on-site and in real time. Equally, there are few systems to support sketching or notes^{1), 2)}.

We are developing a historical tour support system using Matsue Castle and the surrounding famous historical spots as a model. This system supports graphics and web pages to be used pre-tour, and a mobile terminal to be used during sightseeing. We evaluated the tour support system and found there were areas that needed improvement³⁾. We therefore developed a mouse operated option for pre-tour study and planning, a 3D perspective that adjusts for height of viewpoint, and an appended guide with voice and quiz options. We developed the tour support function using a smart phone. As a result, users will have many functions available and will be able to operate it easily. For example the mobile terminal in their pocket will allow access to voice information and the smart phone touch pen will enable users to draw sketches or take notes. Additionally, we are now developing a group learning support function. When completed, it will enable student group leader to track the whereabouts of students and offer advice or answer questions. This will enhance the learning experience.

This system would be useful, not only for historical tours, but also for social study activities such as visiting a zoo or school excursions.

2 System Structure

Figure 1 shows the structure of the historical tour support system. During preparation for the pre tour, a 3D graphic processor displays a 3D object data base (3D object DB). At the same time the historic site manager displays a historical information database referenced to Web pages. During the actual tour, the position data manager gives position data at the time of access. The position data DB uses GPS data. In addition, it stores any data such as sketches or notes entered on the mobile terminal screen. This data is stored in a data registration DB. Post tour, the quiz data handler has the capacity to create quizzes drawn from the “understanding check DB”. Users can use this option to check their understanding.

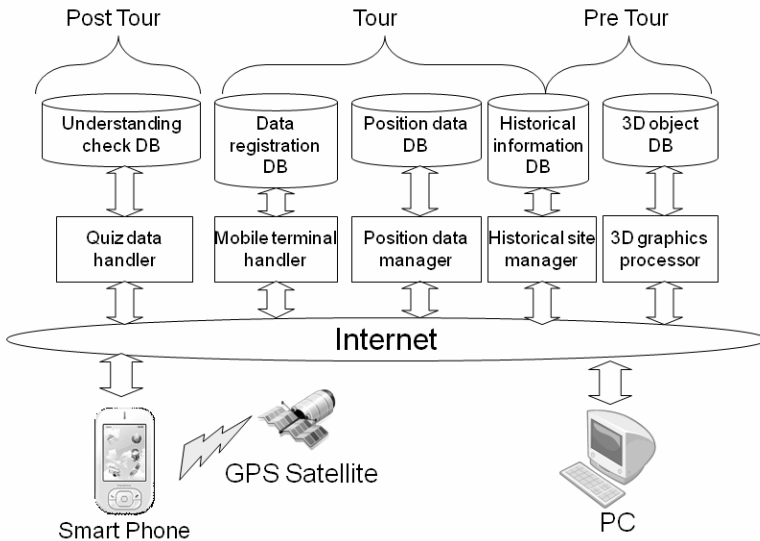


Fig. 1. System Structure

3 Summary of System with Proposed Enhanced Functions

At pre-tour, users can access locations and images of historical sites by taking a virtual tour based on 3D graphics and collect detailed information using the Web pages.

While on tour, the guidance function based on GPS gives position data and route guidance information can be accessed in real time.

By selecting ‘record view’ on the initial screen, users use the record function to note down their impressions and findings and can review their records after the tour.

They can browse information sent by the guide function and the learning history function enables information used pre tour to be accessed again.

The position data function displays a map of the area around the historical spot. This helps users to identify their own location and the position of the historic spot. Therefore, it is useful for users who stray off the route.

4 Pre Tour and Tour Function

By using a mouse, users can both travel in the 3D virtual space and study on Web pages using the pre tour function. In addition, having absorbed information from Web pages such as that shown in Figure 2, they can verify their understanding using the quiz function.

In our previous system, this function was activated by button click. We have modified this function to operate by mouse click and drag. In addition, function buttons to change the speed of movement of the virtual tour and the viewpoint are now included.

Operation is simple. Advance and go back are controlled by the left controller of the mouse, and change viewpoint by dragging the mouse up and down.



Fig. 2. Sample screen from Pre Tour

Figure 3 shows the two cameras in different positions that follow the avatar. Users can see the whole area by switching cameras.

An additional modification is that when the avatar approaches a historic site, voice guidance is activated automatically, providing concurrent explanation.

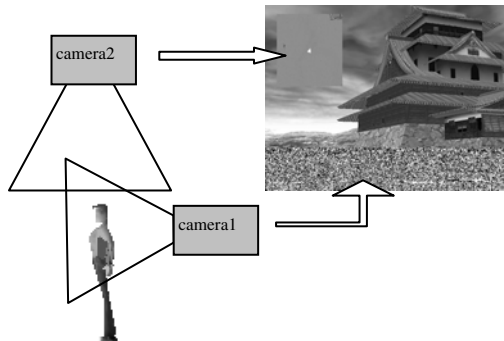


Fig. 3. Camera Layouts

Another new function identifies the avatar's position (x, y coordinates) and so, as the avatar moves, the scene changes to the next area automatically. Then users can "walk" through to another area. In addition, we are further developing the quizzes on Web pages to confirm understanding.



(a) Record about impressed memorandum

(b) Record of photo and note

Fig. 4. Screens of Tour Functions

We originally developed the tour function using PDA and a mobile phone connected with Bluetooth. However, it is inconvenient to carry both PDA and a mobile phone. At the same time, it is difficult for users to connect PDA and mobile phone with Bluetooth. That is the reason that we have sought to develop this improved system using a smart phone. Users can take advantage of this system more easily and have access to more function. The smart phone needs to have GPS, a camera and a touch screen. We are using PRO series T-01A of NTT Docomo that has a 4.1-inch screen. Figure 4 shows screens from the tour function using smart phone.

We are also developing a group learning function for students and their teachers. This function will be useful for student groups on study tours both from the group leader's perspective and from the enhanced learning experience for the students. Figure 5 shows the structure of the group learning function.

A mobile terminal will send address of partner, message and position data to the server. The server will identify the student location and transmit this to the partner (group member, leader or teacher). The receiver terminal will display the ID, message and position data. Thus, the teacher or group leader will be able to track the whereabouts of all group members.

We develop the 3D virtual space using Virtools instead of Java 3D since this enables a more advanced 3D space construction. We can develop this using Building Block (BB) of Virtools as shown in the flow chart such as Figure 6[4]. We can make programs having complicated movement in 3D virtual space using BB of Virtools.

When the user clicks the button on the screen, 'Push button BB' operates the program written by JavaScript that displays the Web page.

We implemented the application of mobile terminal using C# language and C#.NET mobile Framework.

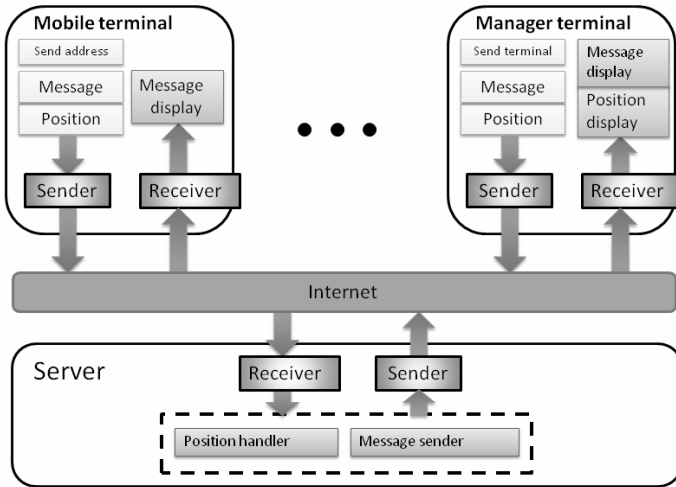


Fig. 5. Structure of group learning function

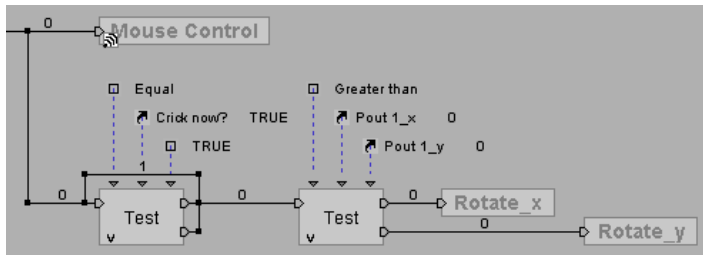


Fig. 6. Example of Building Block

5 Evaluation of the System

Eight students tried out the tour component and evaluated tour function by responding to a questionnaire. They used PDA and mobile phone connected with Bluetooth.

The questions were as listed below

- Q1. Was it easy to operate the system?
- Q2. Was it easy to have mobile terminal?
- Q3. Could you operate by touch pen smoothly?
- Q4. Was it easy to operate the touch pen?
- Q5. Could you input memorandum smoothly?
- Q6. Was the information on the historic site useful?
- Q7. Was the position data on the map useful?
- Q8. Was it easy to draw on the sketch screen?
- Q9. Do you hope to use this system on a tour?
- Q10. Do you think this system will be useful for historical tours?
- Q11. Do you have any desire to improve this system?
- Q12. If you have any comments about the system, please write them here.

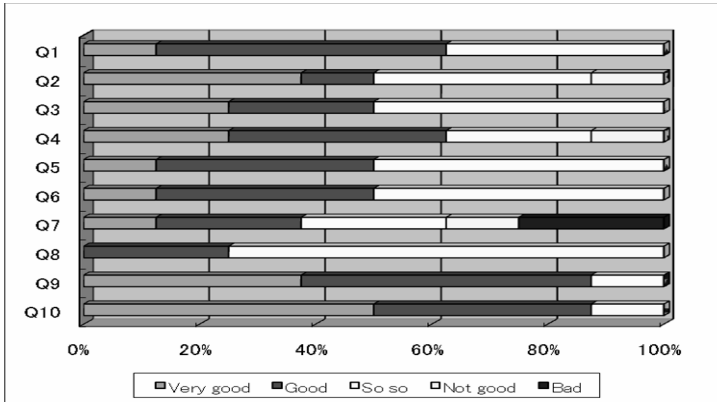


Fig. 7. Evaluation of tour function using PDA and mobile phone

Figure 7 displays the results of an evaluation of tour function. Answers to Q9 and Q10 indicated that PDA was found useful for a historical tour. Q7 of position display function received a low ranking because at the time of this trial, the map did not show position. Answers to Q12 showed that users wanted easier use of the guide function using voice and picture and user's interface. Especially at the time of this evaluation, the evaluators had to use both PDA and mobile phone, which they found inconvenient. These results led us to the smart phone as a preferable option.

In another evaluation twelve persons including our students, used the pre tour function and responded to a questionnaire. The questionnaire and results were as follows.

- Q1. Did you operate by mouse easily?
- Q2. Did you feel satisfied with the button functions?
- Q3. Did you transfer from one area to another smoothly?
- Q4. Was the quality of 3D objects good?
- Q5. Did you feel somewhat disoriented with background and other 3D graphics?
- Q6. Were the Web pages useful for pre tour?
- Q7. Was the voice guidance useful?
- Q8. Did you feel the quizzes were good?
- Q9. Did you feel the pre tour used 3D space effectively?

Q1 showed that three users felt operation was difficult. They felt that operation by clicking the mouse was awkward and the operation manual was not clear. Conversely, they felt it was easy to operate by keyboard. The evaluators found some bugs in button selection and area transfer (Q2, Q3). Answers to Q4 and Q6 were positive. Some users hoped that Web pages could be made more useful for pre tour study and wanted an on/off function for voice guide information. Some answers to Q5 confirmed disorientation with ground design. More quizzes were requested (Q8). As a result of Q9, almost all users evaluated the pre-tour function as useful as a learning tool that would increase enjoyment and understanding during the tour.

The next phase will be an evaluation of the total system combining pre-tour and tour functions.

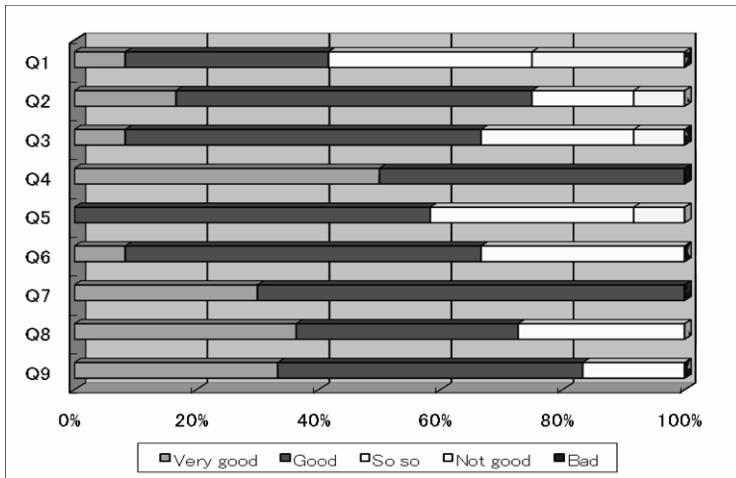


Fig. 8. Evaluation of pre tour functions

6 Conclusion

We developed a pre-tour component for the historical tour support system considering method of operation, 3D objects and a guide function. Twelve persons used this function and almost all evaluated this component as useful. All respondents liked each operation, so operation of the system should be flexible. We have improved the pre-tour component by including functions to change viewpoint and speed of walk through. We also modified the tour function by using smart phone instead of PDA and mobile phone. This has enhanced the active tour function by making it more convenient and able to offer more functions.

The total historical tour support system will be more useful if we are able to develop easier operation. The total historical tour support system must be evaluated by testing the pre tour and tour components in combination. The system has potential for a variety of outdoor learning and sightseeing applications. Our objective is to extend its usability beyond the example given in this paper. We must extend the outdoor learning support system to cover other types of activities such as a zoo tour or a botanical garden tour. Additionally, we will further develop the group learning support option as this offers potential for educationally oriented activities involving younger participants such as school students.

References

1. Tarumi, K., Tsurumi, Y., Yokoo, K., Nishimoto, S., Matsubara, T., et al.: Open Experiments of Sightseeing Support Systems with Shared Virtual Worlds for Mobile Phones. *Journal of Information Processing Society of Japan* 48(1), 110–124 (2007)

2. Hirashima, K., Watagoshi, K., Sakamoto, H., Yasukawa, N., Horiya, K., Sakai, T.: Proposal of Observation Class using Mobile Information Devices in Outdoor Social Education Facilities. *Transactions of Japanese Society for Information and Systems in Education* 24(4), 343–351 (2007)
3. Fujii, S., Takahashi, Y., Kageyama, H., Aoyama, H., Mizuno, T.: Development and evaluation of a ubiquitous historical tour support system. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) *KES 2007, Part III. LNCS (LNAI)*, vol. 4694, pp. 453–459. Springer, Heidelberg (2007)
4. 3dvia virttools (2010), <http://www.3ds-jp.com/virttools/>

Repetition of Dialogue Atmosphere Using Characters Based on Face-to-Face Dialogue

Junko Itou and Jun Munemori

Faculty of Systems Engineering, Wakayama University,
930, Sakaedani, Wakayama 640-8510, Japan
{itou,munemori}@sys.wakayama-u.ac.jp

Abstract. In this article, we analyze the relationships between nonverbal expressions and atmospheres during face-to-face dialogue by persons aiming to apply to an information character. Information characters provide information to viewers through their dialogue to get a better understanding. If the way which information is shown through an embodied characters' dialogue is adopted, the characters have the faces and bodies in order to be "embodied", so, the information provider must care about meanings of a state of nonverbal expressions displayed by the embodied characters. People also display and exchange nonverbal expressions including eye-gazes, noddings, and facial expressions in daily conversation. Nonverbal expressions convey various kinds of information that is essential to make our face-to-face communication successful. In the previous work on social psychology, it is known that there are interdependences among nonverbal expressions between those from different persons in conversation with each other. We investigate dialogue scenes of TV shows to convey information, with a goal of applying above knowledges to a dialogue between embodied characters, which provide information to users.

1 Introduction

Recent work on information character proposes to present information to the viewers in the form of dialogues between a pair of embodied characters [1-4], the viewers obtain information by just listening and watching to the dialogues between embodied characters instead of reading plain texts. This dialogue-based representation of information is also expected to serve easier for the viewers to acquire information than conventional text-based representations as a human interface. Whereas text-based representations such as a newspaper only play a role of the sender, dialogue-based representations employ the conversation partners as the sender and the receiver that ask questions to the sender. As the result, it becomes clear which part of the information the viewers are to be interested in.

In order to employ embodied characters for presenting information, we need to control not only their speech utterances but also their nonverbal expressions, which include eye gazes, noddings, facial expressions, gestures, and so on. As far

as characters have their own faces and bodies in order to be “embodied”, the viewers read various meanings in the nonverbal expressions displayed by the faces and the bodies of the characters even if the characters are not actually designed to send nonverbal expressions to their viewers but only to speak. Thus, for any embodied characters, we need to control their nonverbal expressions properly so that they convey appropriate meaning to the viewers.

Previous work on controlling nonverbal expressions of embodied characters mainly discusses the consistency of nonverbal expressions with the speech utterances or the goal of conversation for each character [3] [4]. However, when we consider dialogues between a pair of embodied characters, we need to consider interdependences between nonverbal expressions displayed by those two characters. As explained in more detail in the next section, it is reported in the field of social psychology that nonverbal expressions given by humans during their talks are not independent with each other, and nonverbal expressions with different modalities are not independent with each other either [5]-[10].

We need to apply this knowledge to the dialogues of embodied characters, and evaluate if any users feel the same atmosphere and the comparable reliability to the information. To achieve this goal, we initially investigate various dialogue scenes and we will evaluate if the presentation dialogue express the corresponding atmosphere to the expected atmosphere. In the remainder of this article, we discuss what kind of relationships among nonverbal expressions should apply to embodied characters to display proper atmosphere.

This paper is organized as follows: in section 2 we will describe the knowledge about the interdependences between nonverbal expressions in human communication reported in the previous work on social psychology. In section 3, we explain how to maintain the interdependences between nonverbal expression of embodied agents in dialogues. We will show the experimental result on analysis of dialogue scenes in section 4. Since this work is just at its very beginning, we will discuss our future steps towards our goal of realizing dialogue-based information presentation by embodied agents with appropriate nonverbal expressions in section 5.

2 Interdependence between Conversation Partners

In order to employ embodied characters as a communicator, we need to control their nonverbal expressions based on the relationships of nonverbal expressions existing in humans dialogue. For example, it would be strange if an character telling a funny story does not smile at all. As another example, when one character talks or smiles to the partner agent, if the partner character freeze with no response to the action, it looks also strange.

It has been investigated in social psychology what features are found in nonverbal expressions given by humans during their conversation. Through those investigations, it is known that nonverbal expressions given by humans in real conversation have some interdependences and synchronicity.

For example, it was reported that the test subjects maintained eye contact with their partners in lively animated conversation, whereas they avoided eye

contact when they are not interested in talking with their partners [5] [6]. In the experiments by Matarazzo, noddings by the listeners in conversation encouraged utterances of the speakers, and as the result, animated conversation between the speakers and the listeners is realized [7]. In another experiments by Dimberg, facial expressions of the test subjects were affected by those of their partners [8]. The subjects smiled when their partners gave smiles to them, whereas they gave expressions of tension when their partners had angry faces.

These results implies positive correlations or synchronicity between the non-verbal expressions given by the conversation partners for eye gazes, noddings and facial expressions. Table. 1 illustrates these positive correlations by *.

Table 1. Interdependences between nonverbal expressions by different persons during conversations

		person A			
		gaze	smile	nodding	speech
person B	gaze	*			*
	smile		*		*
	nodding				*
	speech	*	*	*	*

Furthermore, Watanabe [11] pointed we do not only exchange words, but also we share gestures and physical rhythm such as breath, so that we can feel an identification with the conversation partner. From this indication, kinds of synchronicity should be reflected to the relationships between an action and a reaction when the characters are caught up in conversation.

On the other hand, these interdependences are general tendencies in the usual conversation situation and can change by the second according with the conversation. Therefore, we can say that the strength of these interdependences are different by the atmosphere of the scene of talks.

3 Presentation Dialogue with Atmosphere

3.1 Goal

Our goal is the evaluation of our approach using the relationships among non-verbal expressions as mentioned in the previous section. It is clear that the atmosphere of the dialogue have to be different according to the mental states of speakers. On the other hand, even if whatever the embodied character has a face and an aspect, we have to serve a same atmosphere to users who obtain information through the characters' dialogue.

Therefore, a producer of an animated cartoon is required to control the movement of the embodied characters not to give an unexpected impression from cartoons, even if their faces, their aspects and the way of displaying nonverbal expressions are different. It is very troublesome for the producer to set manually

all actions and reactions at each message. By employing the knowledge described in the previous section, we aim to realize automatic actions and reactions of embodied characters in presentation dialogue.

In the remainder of this section, we will propose a system to realize this adjustment process especially for eye gazes, gestures, and facial expressions of the characters for the first step towards our goal.

3.2 Definition of Atmosphere

Let us consider a pair of embodied characters A and B to explain dialogue atmospheres. We denote a teacher agent by character A and a student agent by character B .

Character's actions and reactions can change by the information message. As discussed in section 2, the actions of character A are not only determined by the messages of user A but also the actions of character B . From these indication, kinds of synchronicity should be reflected to the relationships between an action and a reaction. For example, when character B laughs/doesn't laugh with the message of character B , character A should smile/shouldn't smile if the dialogue atmosphere is "joyful"/"serious". We define 3 kinds of dialogue atmosphere for providing information as follows: "joyful", "serious", and "thinking".

When an atmosphere is corresponding to "thinking", the partner shows the reaction "nodding" to stimulate their dialogue according to the research by Matarazzo.

3.3 Information Character and Animated Cartoons

The overview of the character is shown in Fig. 1.



(a) Tatsuya

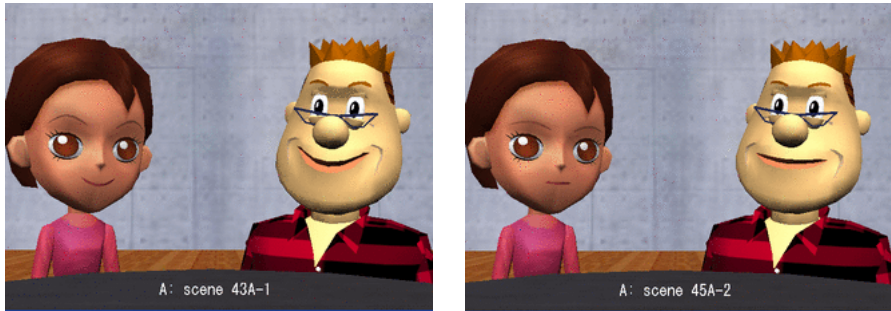


(b) Ai

Fig. 1. 3D Characters

Tatsuya and Ai are provided by TVML(TV program Markup Language), which is developed by NHK(Japanese Broadcasting Corporation), Hitachi Kokusai Electric, Hitachi, and Keio University.

An example of an animated cartoon is shown in Fig. 2. Each action and reaction takes one or two seconds. The next actions or the reactions are not displayed until the last action ends.



(b) Tatsuya & Ai

Fig. 2. Overview of animated cartoons

4 Experimental Results

We extracted the scenes from TV show constructed of face-to-face dialogue by two persons. The TV program shows scene of conversation between a host and a guest. The scenes which can be classified to three atmosphere have about 78% of all scenes to investigate, so we can say the 3 kinds of atmosphere are not sufficient one but necessary one for dialogue atmosphere.

We performed an evaluation experiment using these scenes to investigate what kind of atmosphere the viewers feel for the scenes. We prepare 50 dialogue scenes which include different 4 pairs. Each scene is 3 second movie that we can recognize the facial expressions of the guest and the host.

Experimental subjects were 12 college students. We instructed each subjects to watch the scenes 3 times and write down an answer in the form. There are 3 question in the form: 1. Which kind of atmosphere do you feel? 2. Do you feel another atmosphere? 3. What kind of the facial expression of the guest and the host? The participants selected one atmosphere of 3 atmospheres in question 1 and described freely their feelings in question 2. The result of a questionnaire is described in Table 2.

As shown in Table 2, there are 18 scenes more than 9 subjects chose a same atmosphere. In early 20 scenes of all 50 scenes the choices tend to vary widely but in the latter 30 scenes the degree of response tend to be higher because the subjects adjusted the classification work.

It was revealed that the subjects did not confused some atmospheres but felt the transition of atmospheres from their answers in question 2, in the scenes subjects' choice are divided into 3 atmospheres. They mainly wrote a degree of an atmosphere such as "somelike joyful", "a little serious", and the posture of

Table 2. Questionnaire result

The scenes more than 9 subjects chose a same atmosphere	18
joyful	9
serious	6
thinking	3
The scenes subjects' choice are divided into 3 atmospheres	8

the performers such as “speaking to viewers”, “explaining seriously”. We can classify these scenes into above 3 atmospheres and the state of transition.

As a next step, We performed a comparison experiment. We made two patterns animated cartoons to provide information through the dialogue between embodied characters whose nonverbal expressions were controlled by our approach.

The animated cartoons are made by using TVML for producing TV programs by embodied agents. Scripts of TVML can specify various factors for describing a TV program: positions of the camera and agents, speech, eye gaze, body movement and facial expression of each agent, and so on.

By setting the position of the agents and the cameras similar to the conversation between the original TV show, we adjusted nonverbal expressions using the procedure described in previous section while changing the amount of speech following to the video streams of the conversation.

Experimental subjects were 3 persons. We instructed each subject to watch the 6 animated cartoons and choose the atmosphere among three items, “joyful”, “serious”, and “thinking”. We instructed participants to answer which word was appropriate for each animated cartoons which we made and showed. Table. 3 shows the result.

Table 3. Average of recall and precision

Atmosphere	Recall	Precision
Joyful	83.3	100.0
Serious	83.3	62.5
Thinking	80.0	66.7

The ratio of the atmosphere “joyful” shows high rating. At the same time, the ratio of “serious” and “thinking” is low, compared to the former. It is because the difference between the former and latter is more clear than the differences between the two atmospheres of latter in the change of each nonverbal expressions.

Our goal is not precise reproduction of dialogists’ behaviors but reproduction of dialogue atmospheres. Therefore, even if each characters makes various impressions depending on individual viewers, it is important to give the atmosphere that the dialogue characters make. We can produce animated cartoons

reproducing dialogue atmospheres on some level as mentioned above using these 3D characters who have human-like faces and human-like bodies.

5 Conclusion

In this article, we discussed the dialogue atmosphere with nonverbal expressions displayed by a pair of two persons and the influence of an impression from animated cartoons for character-based information presentation. Considering the knowledge obtained by the previous work on social psychology, we applied the interdependences among nonverbal expressions from the agents to maintain the atmosphere of the dialogue appropriately.

In the experiments, we produced animated cartoons of a dialogue between a pair of embodied characters that display nonverbal expressions using TVML and the character from the chat system which we proposed. In the comparison test with two types of characters, we totally obtained fine rating for our method.

As one of the future steps, we plan to adopt the other kind of nonverbal expressions because nonverbal expressions don't only include gaze, smile and nodding but also gesture, physical distance, and so on.

As another future step, we also plan to conduct more comprehensive experiments for evaluation of the cartoons produced by our method. We will confirm the validity of our method based on the comparison and the evaluation by many subjects.

Acknowledgement

A part of this research was supported by Research for Promoting Technological Seeds of Japan Science and Technology Agency (JST) (11-213).

References

1. Andre, E., Rist, T.: Presenting Through Performing: On the Use of Multiple Lifelike Characters in Knowledge-Based Presentation Systems. In: Proc. of the Second International Conference on Intelligent User Interfaces (IUI 2000), pp. 1–8 (2000)
2. Kubota, H., Yamashita, K., Fukuhara, T., Nishida, T.: POC caster: Broadcasting Agent using Conversational Representation for Internet Community. *Journal of Japanese Society for Artificial Intelligence* 17(3), 313–321 (2002) [in Japanese, with English Abstract]
3. Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., Yan, H.: Embodiment in Conversational Interfaces: Rea. In: CHI 1999, pp. 520–527 (1999)
4. De Carolis, B., Pelachaud, C., Poggi, I., De Rosi, F.: Behavior Planning for a Reflexive Agent. In: Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2001), pp. 1059–1066 (2001)
5. Beattie, G.W.: Sequential patterns of speech and gaze in dialogue. *Semiotica* 23, 29–52 (1978)

6. Kendon, A.: Some functions of gaze direction in social interaction. *Acta Psychologica* 26, 22–63 (1967)
7. Matarazzo, J.D., Saslow, G., Wiens, A.N., Weitman, M., Allen, B.V.: Interviewer Head Nodding and Interviewee Speech Durations. *Psychotherapy: Theory, Research and Practice* 1, 54–63 (1964)
8. Dimberg, U.: Facial Reactions to Facial Expressions. *Psychophysiology* 19(6), 643–647 (1982)
9. Argyle, M.: *Gaze and Mutual Gaze*. Cambridge University Press (1976)
10. Argyle, M., Dean, J.: Eyecontact, distance and affiliation. *Sociometry* 28, 289–304 (1965)
11. Watanabe, T.: E-COSMIC: Embodied Communication System for Mind Connection. In: *Proc. of the 9th International Conference on Human-Computer Interaction (HCI International 2001)*, vol. 1, pp. 253–257 (2001)

CMOS-Based Radiation Movie and Still Image Pickup System with a Phototimer Using Smart Pattern Recognition

Osamu Yuuki¹, Hiroshi Mineno¹, Kunihiro Yamada², and Tadanori Mizuno¹

¹ Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu-shi,
Shizuoka 432-8011, Japan

² Tokai University, 2-2-12 takanawa, minato-ku, Tokyo-to 108-8619, Japan
yuki.osamu@canon.co.jp

Abstract. Radiation imaging systems are used in biology, medicine, industry, high-energy physics, and et ceteras application. However, it has been highly difficult to realize a radiation digital image pickup device that is adapted to show a moving image with a high resolution if it is made to show reduced dimensions at low cost. CMOSs can advantageously be used for the photoelectric converters. CMOSs are adapted to show a moving image with a high resolution and, at the same time, can be made to show reduced dimensions at low cost. Therefore, we propose a CMOS-based radiation movie and still image pickup device. Next we propose a phototimer for the device. The radiation image pickup system includes a phototimer using pattern recognition for X-ray emission. The system is able to recognize a pattern image of interest by allowing radiation from an X-ray source to pass through a object. The phototimer signal can be obtained from the image recognition unit by non-destructive reading in the image sensing operation. The radiation image pickup system can easily and accurately perform optimal exposure. Finally the system could obtain a high-quality image without any deterioration in S/N characteristics.

Keywords: CMOS, radiation, imager, Phototimer, pattern recognition.

1 Introduction

A film screen system realized by combining intensifying screens and an X-ray film is popularly used for X-ray photography for the purpose of medical diagnosis.

1.1 X-Ray Digital Image Pickup Devices

In recent years, X-ray digital image pickup devices have become a target of the study in the field of medical imaging. X-rays are converted into rays of visible light with intensities proportional to those of the original X-rays by means of a scintillator [1] and then obtained rays of visible light are converted into an electric signal by means of a photoelectric converter, which electric signal is then

transformed into a digital signal by means of an A/D converter. As one of studies, X-ray digital image pickup device had been proposed comprising an image pickup device formed by arranging elements on a glass substrate. A pixel of the device consisted of elements having an amorphous semiconductor sandwiched between a pair of electrodes [2]. And a scintillator was laid on the image pickup device in order to convert X-rays into rays of visible light. Another study of X-ray digital image pickup device had been proposed by two-dimensionally linking modules, each comprising a tapered optical fiber formed by heating and softening a bundle of optical fibers and drawing the softened bundle, a photoelectric converter such as a CCD [3] arranged at the tapered side of the optical fiber and a scintillator laid on the opposite side of the optical fiber [4]. More specifically, the device consists of “substrates carrying respective photoelectric converters”, “scintillators for converting X-rays into rays of visible light showing a wavelength that can be detected by the photoelectric converters”, a “base member”, “tapered optical fibers”, “protection glass plates” and “bonding wires”.

However, while X-ray digital image pickup devices comprising amorphous semiconductors typically made of silicon and arranged on a glass substrate are adapted to show a large sensor active area, they are accompanied by problems including that the size of pixels cannot be reduced because of the manufacturing process and the device characteristics and that the device sensitivity is limited. Therefore, devices of this type are not adapted to high speed operation particularly in terms of displaying moving images. On the other hand, X-ray digital image pickup devices comprising photoelectric converters such as CCDs realized by using a silicon substrate have a problem that they cannot show a large sensor active area mainly because of the restrictions in the manufacturing process and the high power consumption level that produces heat, although they are adapted to realize a small pixel size and pick up moving images because they are highly sensitive and can be driven at high speed. To solve such problems, there had been proposed a device comprising an increased number of elements, using optical fibers tapered in such a way that non-sensor areas of the photoelectric converters may not overlap in order to make it show an enlarged sensor active area. However, a tapered optical fiber is costly and the ratio of dimensional reduction is not stable because the tapering process involves dimensional dispersions. Furthermore, while several tapered optical fibers that are thick and heavy may be linked together, it is not realistic to link a large number of tapered optical fibers in order to produce a sensor having large active area. Additionally, tapered optical fibers show a poor light transmission factor to a great disadvantage of the device.

1.2 Phototimer for X-Ray Image Pickup System

Another aspect of the proposal relates to a controller for performing emission stop control on radiation such as X-rays. An exposure control method using a conventional phototimer had been proposed for an image pickup system. An X-ray image pickup system is shown in Fig. 1. This system included a gcontrolling method for emitting X-raysh and an gX-ray image pickup deviceh. The

X-ray image pickup device was comprised of a plurality of photoelectric conversion elements arranged two-dimensionally and a drive circuit of the elements. The X-ray image pickup device is driven by a panel drive circuit. A phototimer is arranged between the X-ray image pickup device and a human. The phototimer is a sensor for detecting X-rays transmitted through a reference part of the human during image sensing exposure. Outputs from the phototimer are integrated by an integrating circuit. The resultant value is output to a comparator. The comparator compares this integration output with a reference value V_{th} , and outputs the comparison result to an X-ray source drive circuit for driving the X-ray source. The X-ray source drive circuit is controlled by an output signal from the comparator. When the output from the integrating circuit exceeds the reference value V_{th} , the X-ray source drive circuit stops driving the X-ray source to stop X-ray emission. This system also includes a sensing start switch for designating the start of image pickup of the human, an A/D converter for A/D-converting a signal from the X-ray image pickup device, an image processing circuit for processing an image signal from the A/D converter, a monitor for displaying a sensed image, and a recording medium for recording sensed image data. When the sensing start switch is turned on, a signal for designating the start of image pickup operation is supplied from the sensing start switch to the panel drive circuit, integrating circuit, and X-ray source drive circuit. Upon reception of this signal from the sensing start switch, the X-ray source drive circuit starts driving the X-ray source to emit X-rays from the X-ray source. Upon reception of the signal from the sensing start switch, the panel drives circuit starts driving the X-ray image pickup device. The integrating circuit resets an output from the phototimer and starts integration. The X-rays emitted from the X-ray source are transmitted through the human. At this time, the X-rays transmitted through the human vary in transmission amount depending on the sizes and shapes of bones and internal organs, the presence/absence of a focus, and the like in the human, and include image information about them. The X-rays transmitted through the phototimer are converted into visible light by a phosphor. This light is incident on the X-ray image pickup device. In the X-ray image pickup device, sensed signals are accumulated by the photoelectric conversion elements arranged two-dimensionally. The integrating circuit integrates outputs from the phototimer. The output from the integrating circuit gradually increases. When the output from the integrating circuit exceeds the reference value V_{th} , the comparator outputs a signal for designating a driving stop to the X-ray source drive circuit, thereby stopping X-ray emission from the X-ray source. Thereafter, the image processing circuit reads the sensed signal through the A/D converter and performs predetermined image processing. The image processing circuit also displays the sensed image on the monitor or records the image data on the recording medium.

In the above conventional X-ray image pickup system, however, since interruption of X-rays is controlled by using the phototimer, the following problems arise. The phototimer is very expensive, and X-rays are slightly attenuated when they pass through the phototimer unit, resulting in deterioration in S/N

characteristics. When the phototimer is used, a doctor or examination technician selects one or two of switches SW1 to SW3 in accordance with an image sensing position before image sensing to select a sensor for detecting X-rays, as shown in FIG. 2. When, for example, the lung of the human is to be sensed, the two side sensors are selected, as shown in the lower left of FIG. 2. When the stomach is to be sensed, the central sensor is selected, as shown in the lower right of FIG. 2. In addition, the doctor or examination technician determines the reference value V_{th} for the integrating circuit in accordance with an image sensing position before image pickup. When, for example, the lung of a person is to be sensed, since a high S/N ratio is required, the reference value V_{th} is set to be high. When the stomach is to be sensed, since high contrast can be obtained owing to a contrast medium, the reference value V_{th} is set to be low. Selecting sensors and setting a reference value in accordance with an image sensing position lead to a deterioration in operability. In addition, if the image sensing position deviates from a proper position, the error between the X-ray amount at the actual image sensing position and the amount detected by the sensor increases. Optimal exposure cannot therefore be performed, resulting in a decrease in S/N ratio, an increase in X-ray dose, and deterioration in image quality.

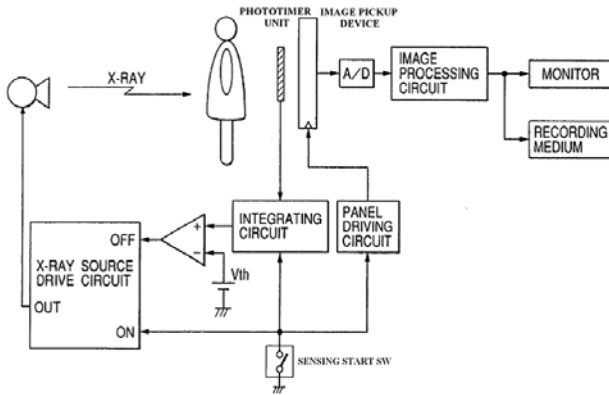


Fig. 1. The imaging system using a conventional phototimer

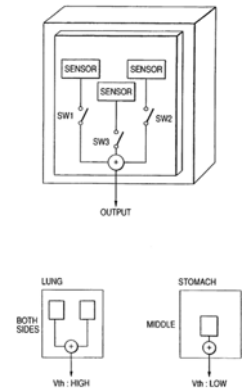


Fig. 2. The conventional phototimer

2 Proposed Imaging Pickup Device

In view of the above identified circumstances, it is the present proposal to provide a radiation movie and still image pickup device such as an X-ray image pickup device that is adapted to show a moving image with a high resolution and, at the same time, can be made to have a large-size sensor active area and show reduced dimensions at low cost.

2.1 Structure of Imaging Pickup Device

We describe the detailed structure of the prototype of radiation movie and still image pickup device. FIG. 3 is a schematic cross sectional lateral view of image pickup device. FIG. 4 is a schematic face view of the photoelectric converter substrates. Referring to FIGS. 3, the radiation irradiating a object produced information showing differences in the intensity thereof to reflect the state of the inside of the object. Then, the wavelength of the radiation is converted into visible light that can be detected by the photoelectric converters by the scintillator layer operating as wavelength converter. The visible light is then made to pass through the light guide section before being detected by the photoelectric converter substrates. The detect information is then led to the rear side of the photoelectric converter substrates. The photoelectric converter substrates are connected to flexible wiring terminals by way of bumps. The flexible wiring substrates are extended through gaps of the photoelectric converter substrates. The photoelectric converter substrates are arranged side by side to a common light guide section. The photoelectric converter substrates to transfer the electric charges detected from the photoelectric converters to a processing circuit so that the imaging active area can be increased. Referring to FIG. 4, the photoelectric converter substrates are arranged two-dimensionally in three rows and three columns and leads and flexible wiring substrates are extending from the photoelectric converter substrates to the rear side relative to the light guide section. CMOSs can be used as the photoelectric converters advantageously [5]. The light receiving pixels are arranged over the entire surface of the photoelectric converter substrate. Each of the “light receiving pixels along scan drive circuits” and the “light receiving pixels along the edges” have the light receiving quantity of pixels smaller than “the remaining light receiving pixels”. Because the receiving light is a lower rate, its output should be corrected to make it balanced with the output of other light receiving pixel. All most of pixels are arranged at a pitch rate of 100 %. However any two adjacently located photoelectric converter substrates are arranged with a gap by 80 % separating them due to taking the thickness of each flexible wiring substrate and the bonding accuracy into consideration. These pixels are arrayed by each pitch rate of 100 % - 80 % - 140 % - 80 % - 100 % . The defect of the irregularity is not visually remarkable. The quality of the produced image is not particularly bad even if compared with pixels of 100 % aperture. The adhesive transmits light very well and shows an excellent elasticity. Aligning all the photoelectric converter substrates in this way, the gaps separating the photoelectric converter substrates are sealed by an adhesive. Each of the flexible wiring substrates extending from the photoelectric converters are connected to related electronic parts including the signal processing circuit (such as A/D converters) on the corresponding base member as shown in FIG. 3. The light guide section is formed by using an optical fiber plate that is formed by cutting a large bundle of optical fibers to make it show a plate-like profile. An optical fiber plate can be prepared through a process that is by far simpler than the process for preparing a tapered bundle of optical fibers. If a light transmitting substrate such as a glass substrate is used, it cause

the scattering of light. An optical fiber plate can be used for the light guide section in order to guide light to the photoelectric converters without scattering it. Because the light guide member is made of a material containing lead, the X-rays that are not converted to rays of visible light by the scintillator can be effectively blocked by the lead contained in the light guide member. A member of light guide is transparent relative to visible light but opaque relative to radiation. It minimize the adverse effect of X-rays on the photoelectric converters and produce X-ray images with little noise. The scintillator is made of gadolinium sulfide (Gd₂O₂S:Tb) [6].

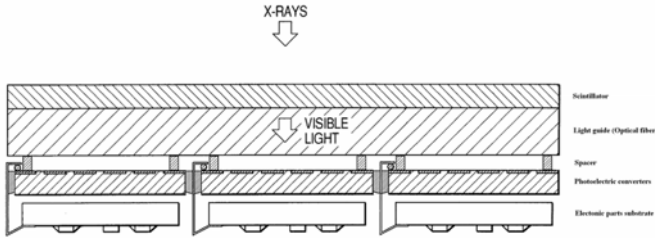


Fig. 3. Cross sectional lateral view of CMOS-based radiation movie and still image pickup device

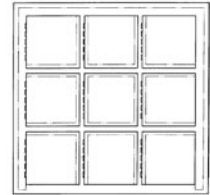


Fig. 4. Photoelectric converter substrate

2.2 Experiment of Movie Imaging Pickup

The experiment used a functional prototype of radiation movie and still image pickup device as shown in FIG. 5. The total active area of the device is 68.0mm × 45.4mm. Because the pixels of all borders have nearly 100 % aperture, it is possible to connect more many photoelectric converters seamlessly. Therefore, total active area of device can be increased more than it of the functional prototype device. A amount of pixels are about 28 million. The frame rate is 10 frame per second in case of movie mode. The frame rate was restricted due to reading the serial digital data from each individual A/D converters in the functional prototype device. However it can be increased to more than 30 frame per second by reading as parallel digital data. In movie mode, four integrated pixels were treated as a pixel. X-rays from an X-ray source are made to strike the radiation image pickup device arranged behind the object to be examined. Then, the wavelength of the incident radiation was changed to that of visible light by the scintillator and the obtained visible light was projected onto the light receiving pixel section of the CMOS-based image pickup elements by way of the FOP (Fiber Optical Plate). Then the visible light was converted into an electric signal by the photoelectric converter. After these analog signals were then converted to digital signals by A/D converter, the digital signals were transmitted to an image processor through the cable. These signals were processed to produce an image of the object by an image processor, which was then displayed on a monitor display or printed by a printer. The example of radiation image is shown

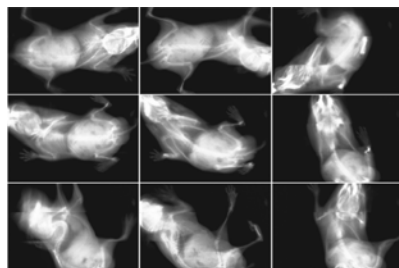
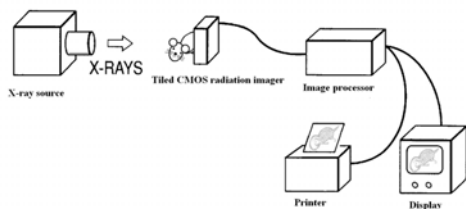


Fig. 5. Radiation movie and still image pickup system **Fig. 6.** Snap shots of the X-ray video system

in FIG. 6. The object was a mouse. Because we did not process the difference of scan timing between pixels, borders were seen between each of “photoelectric converter substrates” in case of the movie image.

3 Phototimer Using a Pattern Recognition

It is an object of the present proposal to provide a radiation image pickup system which can easily and accurately perform optimal exposure and obtain a high-quality image without any deterioration in S/N characteristics, and recognize an image from the radiation image pickup system.

3.1 Method of X-Ray Exposure Control by Phototimer

FIG. 7 is a block diagram showing the arrangement of a radiation image pickup system. In this FIG. 7, still images of the mouse’s lung, stomach, and the like are sensed, and X-ray emission stop control is performed by using a non-destructive reading output in place of a conventional phototimer unit. Referring to FIG. 7, a “radiation movie and still image pickup device” is an image pickup device capable of normal reading and non-destructive reading. The radiation image pickup device is comprised of a plurality of photoelectric conversion elements arranged two-dimensionally and a drive circuit of the elements. The circuit arrangement and operation of the radiation image pickup device will be described in detail later.

A mode switching circuit is a circuit for switching the reading mode of the radiation image pickup device to the normal reading mode or the non-destructive reading mode. The part of pattern recognizing circuit performs pattern recognition on the basis of the output values of all pixels from the A/D converter to identify the sensed image (e.g., the lung, stomach, leg, or the like). As a consequence, the position and size of the sensed image are known. The reference pattern optimizing circuit determines on the basis of the pattern recognition result which position of the image subjected to pattern recognition is to be

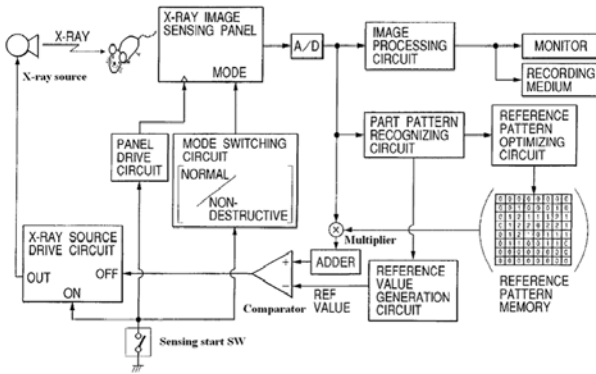


Fig. 7. Radiation movie and still image pickup system with a phototimer

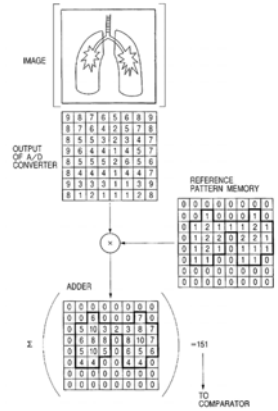


Fig. 8. Processing method of the pattern recognition

mainly seen, and then stores the numerical value determined for each pixel into a reference pattern memory. In this explanation, for the sake of descriptive convenience, as shown in FIG. 7, weighting is performed in three levels, namely 0, 1, and 2. In this case, g0h indicates a position other than a position to be examined (a position that need not be examined); g2h, a position that should be mainly examined; and g1h, a value that makes an image better as a whole. Obviously, however, an image with higher quality can be obtained by performing weighting in more levels. In the reference value generation circuit, REF values are determined in advance in accordance with images. For example, if a pattern recognition result indicates the lung, a large REF value is set because the lung demands a high S/N ratio. If the pattern recognition result indicates the stomach, a small REF value is set. Referring to FIG. 7, for the sake of illustrative convenience, the reference pattern memory has a storage area corresponding to 8 8 pixels, and stores numerical values each weighted with one of 0 to 2 for each pixel. In practice, however, the radiation image pickup device has much more pixels, and the reference pattern memory has a larger storage area accordingly. A multiplier is a circuit for multiplying the numerical value of each pixel in the reference pattern memory by the A/D conversion output value for each pixel of the radiation image pickup device. The multiplier multiplies the values of corresponding pixels and outputs the calculation result to an adder. An output value from the adder is the result of weighted addition of an A/D conversion output value and a corresponding value in the reference pattern memory. A comparator compares an output value from the adder with the REF value from the reference value generation circuit. If the output value from the adder becomes equal to or larger than the REF value, the comparator outputs to an X-ray drive circuit a signal for instructing an X-ray source to stop, thereby stopping X-ray emission.

FIG. 8 shows the processing method of the pattern recognition. In FIG. 8, the part of pattern recognition processes the lung image.

According to this X-ray exposure, optimal image sensing can be performed without requiring any cumbersome operation, and a high-quality image without a decrease in S/N ratio can be obtained. In addition, because no conventional phototimer is used, image sensing can be accurately performed regardless of deviations from appropriate image sensing positions and the like. This makes it possible to reduce deterioration in a S/N ratio and manufacture the product at a low cost. Although this experiment was demonstrated in case of X-rays, radiation such as α , β , or γ ray may be used.

3.2 Experiment of Still Imaging Pickup with Phototimer

At the results of experiment, the radiation still image with a proposed phototimer is shown in FIG. 9, 10 and 11. The object is mouse. First, we detected the head and tail positions of the mouse for finding the direction of body. We used a pattern recognition [7][8] to detect the head and tail position [9][10][11]. The rectangle of a white line is drawn around a recognized pattern. We show these photographs in Fig. 9 and 10. Next, the system performed the above processing. Finally, we could get the shown photograph in Fig. 11.

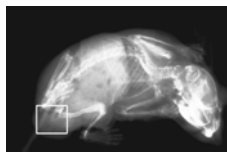


Fig. 9. Detecting the mouse tail

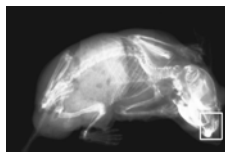


Fig. 10. Detecting the mouse head



Fig. 11. Radiation still image

4 Conclusion

We could propose a prototype of CMOS-based radiation movie and still image pickup system. And, we conducted an experiment of a phototimer using smart pattern recognition. Finally we could demonstrate radiation movie and still images by using the system.

As a future task, Because borders were seen between each of “photoelectric converter substrates” in case of the movie image, we will process the difference of scan timing between pixels. We will use the pattern recognition units for intelligent diagnosis too.

References

1. Nagarkar, V.V., Gupta, T.K., Miller, S.R., Klugerman, Y., Squillante, M.R., Entine, G.: Structured CsI(Tl) Scintillators for X-ray Imaging Applications. *IEEE Trans. Nuclear Science* 53(1), 49–53 (2006)
2. Fujieda, I., Nelson, S., Street, R.A., Weisfield, R.L.: Radiation imaging with 2D a-Si sensor arrays. *IEEE Trans. Nuclear Science* 39(4), 1056–1062 (1992)
3. Tipnis, S.V., Nagarkar, V.V., Gaysinskiy, V., O’Dougherty, P., Klugerman, Y., Miller, S., Entine, G.: Large area CCD based imaging system for mammography. *Nuclear Science Symposium 2*, 1043–1046 (1999)
4. Yamamoto, S.: Resolution improvement using a fiber optic plate for a small field-of-view NaI(Tl) gamma camera. *IEEE Trans. Nuclear Science* 53(1), 49–53 (2006)
5. Fossum, E.R.: CMOS image sensors: Electronic camera-on-a-chip. *IEEE Trans. Nuclear Science* 19(1), 81–86 (1972)
6. Fossum, E.R.: An Improved X-Ray Intensifying Screen. *IEEE Trans. Electron Devices* 44(10), 1689–1698 (1997)
7. Chen, C.H., Pau, L.F., Wang, P.S.: *Handbook of Pattern Recognition & Computer Vision*. World Scientific Publication Co. Pte. Ltd, Singapore (1993)
8. Brunelli, R.: *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley, Chichester
9. Viola, P., Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: *IEEE CVPR* (2001)
10. Lienhart, R., Maydt, J.: An Extended Set of Haar-like Features for Rapid Object Detection. In: *IEEE ICIP*, September 2002, pp. 900–903 (2002)
11. Yuuki, O., Yamada, K., Kubota, N.: Trajectory Tracking for a Pitching Robot based on Human-like Recognition. In: *IEEE CIRA 2009* (December 2009)

Optimal H_2 Integral Controller Design with Derivative State Constraints for Torsional Vibration Model

Noriyuki Komine¹ and Kunihiro Yamada²

¹ School of Information and Telecommunication Engineering, Tokai University
komine@keyaki.cc.u-tokai.ac.jp

² Professional Graduate School of Embedded Technology, Tokai University
yamadaku@keyaki.tokai.ac.jp

Abstract. An optimal H_2 integral controller designed with derivative state constraints for a torsional vibration model is presented. The proposed controller is designed based on the optimal H_2 control problem with a prescribed degree of stability in order to control the under damped response of an oscillatory system with a constant disturbance and to reduce the vibration of the system. It is shown in this paper that the derivative state constrained optimal H_2 integral problems can be reduced to the standard optimal H_2 control. The effectiveness of the controller in reducing the torsional vibration and the damping of two-inertia system are also shown by simulation and experimental results.

1 Introduction

A design method of the state feedback system is generally required to stabilize the closed-loop system. In practice, however, it is often desirable to be a controller that the closed-loop system dynamics are damped [1]. In the case of servo problem, it is also required that the output of a system has no steady-state error for a desired input even if the parameter variations or disturbances exist. In addition, the optimality in control was primarily concerned by R. E. Kalman to minimizing the quadratic performance index of state variables and inputs [2]. The optimal linear control with input derivative constraints is proposed to be advantage of the controller when the input energy is to be included in the performance index, by J. B. Moor and B.D.O. Anderson [3]. Beside, the design method of an optimal tracking system by introducing the integral action for the system using regulator theory was obtained and reported by T. Takeda and T. Kitamori [4]. However, it is difficult to select the proper values of the weighting matrices of performance index in the optimal servo problem to mitigate under damped responses of dynamic systems. The optimal H_2 servo problem is to find the optimal control such that the output tracks the desired trajectory, minimizing the tracking error cost and state excitation cost in the sense of an optimal H_2 control. Recently, the optimal H_2 control for oscillatory system minimizing a performance criterion involved time derivatives of state vector was formulated to control under damped responses of dynamic systems [6], [7]. On the other hand, Anderson and Moor introduced an optimal controller with a prescribed degree of stability affecting the locations of all closed-loop poles [5]. However, it

dose not necessarily ensure that the response of the closed-loop system are damped. In this paper, we derive a method which stabilizes a linear system such that the response of the closed-loop system is damped. The proposed controller obtained from derivative state constrained optimal H_2 integral servo theorem is applied to the two-inertia system. For the verification of the effectiveness of the proposed controller in mitigating an under damped responses of dynamic system, the proposed schemes can be applied to control the torsional vibration for the two-inertia system.

2 Derivative State Constrained H_2 Optimal Control Problem

Problem Statement:

In order to obtain the optimal H_2 integral servo controller, the following controlled plant equations are given as

$$\begin{aligned} \frac{d}{dt}x(t) &= Ax(t) + B_2u(t) + d, \quad x(0) = x_0 \\ y(t) &= C_2x(t) \end{aligned} \tag{1}$$

where $x(t)$, $u(t)$, d and $y(t)$ denote the state vector, the input vector, constant disturbance and the output vector, respectively. The integral $x_1(t)$ of the error vector $e(t)$ between the reference input $r(t)$ and controlled output $y_r(t)$ is defined as

$$\begin{aligned} \frac{d}{dt}x_1(t) &= e(t) = r(t) - y_r(t) \\ y_r(t) &= C_r x(t) \end{aligned} \tag{2}$$

where $y_r(t)$ denotes the controlled vector. Using Eq. (1) and Eq. (2), the augmented controlled plant is given by

$$\frac{d}{dt} \begin{bmatrix} x(t) \\ x_1(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ x_1(t) \end{bmatrix} + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} u(t) + \begin{bmatrix} d \\ r(t) \end{bmatrix} \tag{3}$$

In order to control the system (1) such that the controlled vector $y_r(t)$ tracks the reference $r(t)$, the derivative state constraint is introduced as in the following generalized equations:

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} &= \begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} \\ &+ \begin{bmatrix} \bar{B}_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \bar{B}_{11} & 0 & 0 & 0 & 0 \end{bmatrix} \dot{w}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \dot{u}(t) \end{aligned} \tag{4}$$

$$\dot{z}(t) = \begin{bmatrix} \bar{C}_1 & 0 \\ 0 & \bar{C}_{11} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{D}_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & \bar{D}_{111} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \dot{w}(t) + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \bar{D}_{12} \end{bmatrix} \dot{u}(t) \tag{5}$$

$$\begin{bmatrix} \dot{y}(t) \\ \dot{y}_1(t) \end{bmatrix} = \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{D}_{21} & 0 \\ 0 & \bar{D}_{211} \end{bmatrix} \dot{w}(t) \quad (6)$$

where

$$\dot{w}(t) = [\dot{w}_1(t) \ \dot{w}_2(t) \ \ddot{x}(t) \ \ddot{x}_1(t) \ \dot{w}_3(t) \ \dot{w}_4(t)]^T.$$

Disturbance $\dot{w}(t)$ in Eq. (4) to Eq. (6) is taken in the sense of a hyper function or generalized function.

It is seen that this system consists of a singular problem. The augmented generalized plant from Eq.(4) to Eq.(6) may be expressed in the following nonsingular generalized plant;

$$\frac{d}{dt} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} + \tilde{B}_1 \dot{w}(t) + \begin{bmatrix} B \\ 0 \end{bmatrix} \dot{u}(t) \quad (7)$$

$$\dot{z}(t) = \tilde{C}_1 \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} + \tilde{D}_{12} \dot{u}(t) \quad (8)$$

$$\begin{bmatrix} \dot{y}(t) \\ \dot{y}_1(t) \end{bmatrix} = \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix} + \tilde{D}_{21} \dot{w}(t)$$

where

$$\tilde{B}_1 = \begin{bmatrix} \bar{B}_{11} & 0 & 0 & A\bar{D}_{11} & 0 \\ 0 & \bar{B}_{11r} & 0 & -C_2\bar{D}_{11r} & 0 \end{bmatrix}, \tilde{C}_1 = \begin{bmatrix} \bar{C}_1 & 0 \\ 0 & \bar{C}_{1r} \\ \hline \bar{D}_{11}A & 0 \\ -\bar{D}_{11r}A & 0 \\ \hline 0 & 0 \end{bmatrix} \quad (9)$$

$$\tilde{D}_{12} = \begin{bmatrix} 0 \\ 0 \\ \hline \bar{D}_{12} \\ \bar{D}_{11}B_2 \\ 0 \end{bmatrix} \quad (10)$$

and

$$\tilde{D}_{21} = \begin{bmatrix} 0 & \bar{D}_{21} & 0 & \bar{C}_2\bar{B}_{11} & 0 \\ 0 & 0 & \bar{D}_{21r} & 0 & \bar{B}_{11r} \end{bmatrix} \quad (11)$$

In this paper, the integral servo controller is to be designed by deriving the H_2 controller to the above plant (7). The controller design problem is given by the following statement.

Statement of the problem:

The optimal H_2 servo problem is to obtain a method to find the admissible controller such that the reference output tracks the constant reference input while minimizing the H_2 norm of the transfer function from $\dot{w}(t)$ to $\dot{u}(t)$ of Eq. (7), and the controlled plant is stabilized with the prescribed degree α of stability.

Solution:

The solution to the derivative state constrained H_2 optimal control defined above is given by the loop shifting procedure. In order to consider the effect of the prescribed degree of stability to a controller, each vector variable is exponentially weighted as follows.

$$\begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_1(t) \end{bmatrix} = e^{\alpha t} \begin{bmatrix} \dot{x}(t) \\ \dot{x}_1(t) \end{bmatrix}, \quad \begin{matrix} \dot{\tilde{w}}(t) = e^{\alpha t} \dot{w}(t) & \dot{\tilde{z}}(t) = e^{\alpha t} \dot{z}(t) \\ \dot{\tilde{z}}(t) = e^{\alpha t} \dot{z}(t) & \dot{\tilde{u}}(t) = e^{\alpha t} \dot{u}(t) \end{matrix} \quad (12)$$

Hence, the generalized plant $\tilde{P}(s)_\alpha$ after applying the transformed vector variables of Eq. (12) is given by

$$\frac{d}{dt} \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_1(t) \end{bmatrix} = \left(\begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} + \alpha I \right) \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_1(t) \end{bmatrix} + \tilde{B}_1 \dot{\tilde{w}}(t) + \begin{bmatrix} B_2 \\ 0 \end{bmatrix} \dot{\tilde{u}}(t) \quad (13)$$

$$\dot{\tilde{z}}(t) = \tilde{C}_1 \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_1(t) \end{bmatrix} + \tilde{D}_{12} \dot{\tilde{u}}(t) \quad (14)$$

$$\begin{bmatrix} \dot{\tilde{y}}(t) \\ \dot{\tilde{y}}_1(t) \end{bmatrix} = \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}(t) \\ \dot{\tilde{x}}_1(t) \end{bmatrix} + \tilde{D} \dot{\tilde{w}}(t) \quad (15)$$

The solution to the derivative state constrained H_2 optimal control defined above is given by the following loop shifting procedure.

Singular value decomposition:

There always exist unitary matrices $U_i, V_i, i=1,2$ for the singular value decomposition of \tilde{D}_{12} and \tilde{D}_{21} ;

$$\tilde{D}_{12} = U_1 \begin{bmatrix} 0 \\ \Sigma_1 \end{bmatrix} V_1, \quad \Sigma_1 = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{1r} \end{bmatrix}, \quad r = \dim(u) \quad (16)$$

$$\tilde{D}_{21} = U_2 [0 \ \Sigma_2] V_2^T, \quad \Sigma_2 = \begin{bmatrix} \sigma_{11} & & \\ & \ddots & \\ & & \sigma_{ll} \end{bmatrix}, \quad l = \dim \begin{bmatrix} y(t) \\ y_1(t) \end{bmatrix} \quad (17)$$

where $\Sigma_i, i=1,2$ are the diagonal singular value matrices. Using the results obtained above, the input and output vectors as well as the generalized plant are transformed accordingly as follows.

Variable transformation:

The generalized plant can be obtained by using the following variable transformations defined by

$$\begin{aligned}\dot{\hat{w}}(t) &= V_2 \dot{\hat{w}}(t) \\ \dot{\hat{z}}(t) &= U_1^T \dot{\hat{z}}(t) \\ \dot{\hat{u}}(t) &= V_1 \Sigma_1^{-1} \dot{\hat{u}}(t)\end{aligned}\quad (18)$$

$$\begin{bmatrix} \dot{\hat{y}}(t) \\ \dot{\hat{y}}_i(t) \end{bmatrix} = \Sigma_2^{-1} U_2^T \begin{bmatrix} \dot{\hat{y}}(t) \\ \dot{\hat{y}}_i(t) \end{bmatrix}\quad (19)$$

Substituting Eq. (18) and Eq. (19) into the Eq. (13), Eq.(14) and Eq.(15) then the transformed generalized plant $\hat{P}(s)_\alpha$ which is reduced to the standard form of the H_2 control problem is then obtained as

$$\frac{d}{dt} \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_1(t) \end{bmatrix} = \left(\begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} + \alpha I \right) \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_1(t) \end{bmatrix} + \hat{B}_1 \dot{\hat{w}}(t) + \hat{B}_2 \dot{\hat{u}}(t)\quad (20)$$

$$\dot{\hat{z}}(t) = \hat{C}_1 \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_1(t) \end{bmatrix} + \hat{D}_{12} \dot{\hat{u}}(t)\quad (21)$$

$$\begin{bmatrix} \dot{\hat{y}}(t) \\ \dot{\hat{y}}_i(t) \end{bmatrix} = \hat{C}_2 \begin{bmatrix} \dot{\hat{x}}(t) \\ \dot{\hat{x}}_1(t) \end{bmatrix} + \hat{D}_{21} \dot{\hat{w}}(t)\quad (22)$$

where

$$\hat{B}_1 = \check{B}_1 V_2, \quad \hat{B}_2 = \begin{bmatrix} B_2 \\ 0 \end{bmatrix} V_1 \Sigma_1^{-1}, \quad \hat{C}_1 = U_1^T \check{C}_1, \quad \hat{C}_2 = \Sigma_2^{-1} U_2^T \begin{bmatrix} C_2 & 0 \\ 0 & I \end{bmatrix}\quad (23)$$

$$\hat{D}_{12} = U_1^T \check{D}_{12} V_1 \Sigma_1^{-1} = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad \hat{D}_{21} = \Sigma_2^{-1} U_2^T \check{D}_{21} V_2 = \begin{bmatrix} 0 & I \end{bmatrix},\quad (24)$$

The transformed generalized plant parameter matrices (23) and (24) are employed to obtain the main results as follows.

Hamiltonian matrices:

Under the above results of Eq. (20), Eq. (21) and Eq. (22), the optimal H_2 solution to the transformed generalized plant (20) is given as follows; A couple of Hamiltonian matrices are constituted as

$$H_2 = \begin{bmatrix} \left(\begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} + \alpha I \right) - \hat{B}_2 \hat{D}_{12}^T \hat{C}_1 & -\hat{B}_2 \hat{B}_2^T \\ -\hat{C}_1^T \hat{C}_1 + \hat{C}_1^T \hat{D}_{12} \hat{D}_{12}^T \hat{C} & -\left\{ \left(\begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} + \alpha I \right) - \hat{B}_2 \hat{D}_{12}^T \hat{C}_1 \right\}^T \end{bmatrix}\quad (25)$$

$$J_2 = \begin{bmatrix} \begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix}^T & -\hat{C}_2^T \hat{D}_{21} \hat{B}_1^T & -\hat{C}_2^T \hat{C}_2 \\ -\hat{B}_1 \hat{B}_1^T + \hat{B}_1 \hat{D}_{12} \hat{D}_{21}^T \hat{B}_1^T & -\left\{ \begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix}^T & -\hat{C}_2^T \hat{D}_{21} \hat{B}_1^T \right\}^T \end{bmatrix} \quad (26)$$

Then, it is guaranteed that the solutions exist, which make

$$\begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} + \alpha I + \hat{B}_2 \hat{F}_2 \quad \text{and} \quad \begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} + \alpha I + \hat{L}_2 \hat{C}_2 \quad \text{stable.}$$

From the couple of Riccati solutions,

$$\begin{aligned} X_2 &= \text{Ric}(H_2) \geq 0 \\ Y_2 &= \text{Ric}(J_2) \geq 0 \end{aligned} \quad (27)$$

It is able to construct the following optimal solution to the transformed generalized plant (20).

$$\hat{K}_{H_{2\alpha}}(s) = \left[\begin{array}{c|c} \left(\begin{bmatrix} A & 0 \\ -C_r & 0 \end{bmatrix} + \alpha I \right) + \hat{B}_2 \hat{F}_{2\alpha} + \hat{L}_{2\alpha} \hat{C}_2 & -\hat{L}_{2\alpha} \\ \hline \hat{F}_{2\alpha} & 0 \end{array} \right] \quad (28)$$

where

$$\begin{aligned} \hat{F}_{2\alpha} &= -(\hat{B}_2^T X_2 + \hat{D}_{12}^T \hat{C}_1) \\ \hat{L}_{2\alpha} &= -(Y_2 \hat{C}_2^T + \hat{B}_1 \hat{D}_{21}^T) \end{aligned}$$

A general control formulation with the derivative state constrained optimal H_2 integral servo controller is given by the general configuration shown in Fig.1.

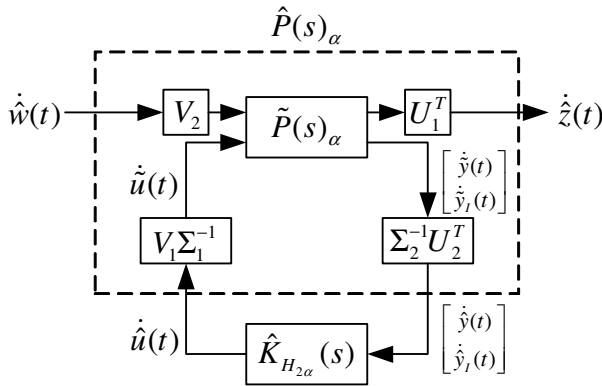


Fig. 1. Block diagram of the structure for controller $\hat{K}_{H_{2\alpha}}(s)$

Main results:

Theorem (Derivative State Constrained Optimal H_2 Integral Servo)

The derivative state constrained H_2 integral servo controller for the controlled plant (20) is given as

$$K_{F_{2\alpha}}(s) = \left[\begin{array}{c|c} \left[\begin{array}{cc} A & 0 \\ -C_r & 0 \end{array} \right] + \left[\begin{array}{c} B_2 \\ 0 \end{array} \right] F_{2\alpha} + L_{2\alpha} \left[\begin{array}{cc} C_2 & 0 \\ 0 & I \end{array} \right] & -L_{2\alpha} \\ \hline F_{2\alpha} & 0 \end{array} \right] \quad (29)$$

or its integral form

$$\frac{d}{dt} \begin{bmatrix} \hat{x}(t) \\ \hat{x}_1(t) \end{bmatrix} = \hat{A}_2 \begin{bmatrix} \hat{x}(t) \\ \hat{x}_1(t) \end{bmatrix} - L_{2\alpha} \begin{bmatrix} C_2 y(t) \\ \int_0^t e(t) dt \end{bmatrix}, \quad u(t) = F_{2\alpha} \begin{bmatrix} \hat{x}(t) \\ \hat{x}_1(t) \end{bmatrix} \quad (30)$$

where

$$\begin{aligned} \hat{A}_2 &= \left[\begin{array}{cc} A & 0 \\ -C_r & 0 \end{array} \right] + \left[\begin{array}{c} B_2 \\ 0 \end{array} \right] V_1 \Sigma_1^{-1} \hat{F}_{2\alpha} + \hat{L}_{2\alpha} \Sigma_2^{-1} U_2^T \left[\begin{array}{cc} C_2 & 0 \\ 0 & I \end{array} \right] \\ F_{2\alpha} &= V_1 \Sigma_1^{-1} \hat{F}_{2\alpha} = -V_1 \Sigma_1^{-2} V_1^T \left\{ \left[\begin{array}{cc} B_2^T & 0 \end{array} \right] X_2 + \check{D}_{12}^T \check{C}_1 \right\} \\ L_{2\alpha} &= \hat{L}_{2\alpha} \Sigma_2^{-1} U_2^T = - \left\{ Y_2 \left[\begin{array}{cc} C_2^T & 0 \\ 0 & I \end{array} \right] + \check{B}_1 \check{D}_{21}^T \right\} U_2 \Sigma_2^{-2} U_2^T \end{aligned}$$

Proof: The proof is neglected here.

3 Verification Experiment

A torsional vibration is occurred to the speed of motor by connecting flexible shaft. The vibration is an impediment to improve the characteristics of the two-inertia system. The experimental results of the speed control of the two-inertia system by proposed controller will be shown in this section. A structure of two-inertia system is shown in Fig.2.

The linear dynamic equation of the Two-inertia resonant system with constant disturbance T_L is represented by

$$\begin{aligned} J_m \frac{d\omega_m}{dt} + F_m \omega_m &= T_m + \tau_d \\ J_L \frac{d\omega_L}{dt} + F_L \omega_L &= \tau_d - T_L \\ \frac{d\tau_d}{dt} &= K_s (\omega_m - \omega_L) \end{aligned} \quad (31)$$

where J_m, J_L, F_m, F_L and K_s are the inertia of motor, the inertia of load, the friction of motor, friction of load and spring constant of the shaft, respectively. The integral $x_1(t)$ of the error vector $e(t)$ between the reference input $r(t)$ and controlled output $\omega_m(t)$ is defined as

$$\frac{d}{dt} x_1(t) = e(t) = r(t) - \omega_m(t). \tag{32}$$

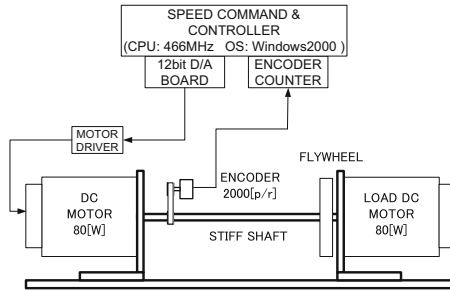


Fig. 2. Structure of the two-inertia System

The state equation with Eq. (32) is then given by

$$\begin{bmatrix} \dot{\omega}_m(t) \\ \dot{\omega}_L(t) \\ \dot{\tau}_d(t) \\ \dot{x}_1(t) \end{bmatrix} \dot{x}(t) = \begin{bmatrix} -\frac{F_m}{J_m} & 0 & \frac{1}{J_m} & 0 \\ 0 & -\frac{F_L}{J_L} & \frac{1}{J_L} & 0 \\ K_s & -K_s & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \omega_m(t) \\ \omega_L(t) \\ \tau_d(t) \\ x_1(t) \end{bmatrix} x(t) + \begin{bmatrix} \frac{1}{J_m} \\ 0 \\ 0 \\ 0 \end{bmatrix} T_m(t) + \begin{bmatrix} 0 \\ -T_L \\ 0 \\ r(t) \end{bmatrix} \tag{33}$$

where $\omega_m(t)$ denotes the speed of motor at time t , $\omega_L(t)$ denotes the speed of load at time t and $\tau_d(t)$ represents the torque of shaft. The numerical values of J_m, J_L and K_s are shown in Table 1. In the case of the numerical values, the friction of motor, friction of load and spring constant of the shaft are neglected, respectively.

Table 1. Numerical values of two-inertia system

J_m [kg · m ²]	J_L [kg · m ²]	K_s [N/m]
0.0866	0.0866	400

The designing parameters in the generalized plant $\bar{B}_1, \bar{D}_{12}, \bar{C}_1, \bar{D}_{11}, \bar{D}_{21}$ and \bar{D}_{211} are chosen as:

$$\bar{C}_1 = \bar{B}_1^T = \begin{bmatrix} \text{diag} \left[\sqrt{10^{q_1}} & \sqrt{10^{q_1}} & \sqrt{10^{q_1}} & \sqrt{20000} \right] \\ [0_{4 \times 4}] \\ [0_{1 \times 4}] \end{bmatrix} \tag{34}$$

$$\bar{D}_{11} = \text{diag} \left[\sqrt{e^{n_i}} \quad \sqrt{e^{n_i}} \quad \sqrt{e^{n_i}} \quad \sqrt{100} \right], \bar{D}_{12} = \sqrt{I} \text{ and } \begin{bmatrix} \bar{D}_{21} & 0 \\ 0 & \bar{D}_{211} \end{bmatrix} = \begin{bmatrix} \sqrt{0.01} & 0 \\ 0 & \sqrt{0.01} \end{bmatrix} \quad (35)$$

Simulation results:

The variation of closed-loop poles for varying $n_i = -10$ to $n_i = -3$ is shown in Fig.3. The original poles of the open-loop system locate on the imaginary axis. It verifies that the pair of poles with imaginary part approach to the real axis when the parameter n_i becomes large. The simulation results for step responses of the torque of the shaft with step disturbance shown in Fig.4 clearly explain the effectiveness of proposed controller. Significantly, the torsional resonance of two-inertia system is removed. This can be seen from designing parameter $n_i = -3$ is reduced torsional resonance compared with $n_i = -10$. Furthermore, it seen also from the Fig.4 that the proposed controller can reject the effect of the constant load disturbance.

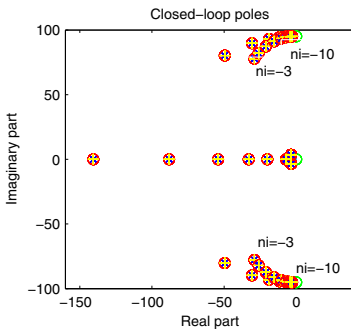


Fig. 3. Closed-loop poles location for values from $n_i = -10$ to $n_i = -3$

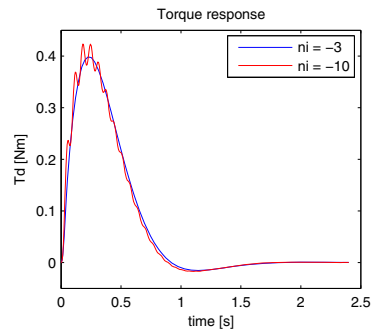


Fig. 4. Step responses with constant load disturbance

Experimental results:

In order to be recovered from the slow time response, the prescribed degree of stability $\alpha = 10$ is selected. Fig. 5 shows the variation of the closed-loop poles of the closed-loop system for the parameters varied from $n_i = -10$ to $n_i = -3$. The original poles of the open-loop system locate on the imaginary axis as shown in Fig.3. It verifies that the pair of poles with imaginary part approach to the real axis when the parameter n_i becomes large. The experimental results in controlling the speed of the two-inertia system at the speed of 2000 rpm and 1000 rpm are shown in Fig.5 and Fig.6, respectively. In order to verify the effectiveness of the prescribed degree of stability, the response of speed of two-inertia system at the reference speed 2000 rpm is shown in Fig.6. In the case of $\alpha = 20$, the speed of motor can reach the target speed at 2000 rpm rapidly than $\alpha = 0$. The step response for reference speed at 1000 rpm and the constant load disturbance entering at 0.6 sec is shown in Fig. 6. It seen also from the Fig.6 that the proposed controller can reject the effect of the load disturbance.

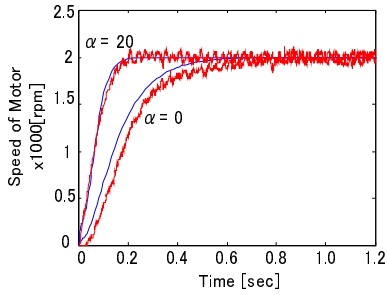


Fig. 5. Step responses of closed-loop system disturbance for $\alpha = 0$ and $\alpha = 20$

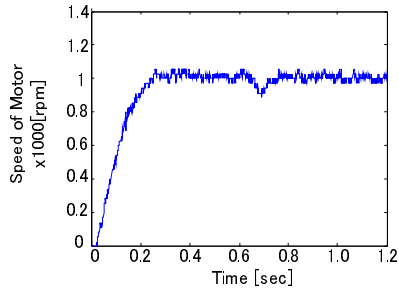


Fig. 6. Step response with constant load $\alpha = 10, n_i = -2, q_i = 2$

4 Conclusion

The optimal H_2 integral controller by using derivative state constraints has been proposed. The proposed controller is effective to control an under damped responses of the controlled system by H_2 control framework. It is recognized that the proposed controller can be applied to the systems whose reference inputs as well as disturbances are all given by step functions. The experimental results have verified that the proposed schemes can be applied to reduce the damping for the two-inertia system.

References

1. Hench, J.J., He, C., Kucera, V., Mehrmann, V.: Dampening Controllers via a Riccati Equation Approach. *IEEE Trans., on Automatic Control* 43(9), 1280–1284 (1998)
2. Kalman, R.E.: When Is a Linear Control System Optimal? *ASMEJ. Basic Engineering*, ser. D 86, 51–60 (1964)
3. Moor, J.B., Anderson, B.D.O.: Optimal Linear Control Systems with Input Derivative Constraints. *Proc. IEE* 114(12), 1987–1990 (1967)
4. Takeda, T., Kitamori, T.: A Design Method of Linear Multit-Output Optimal Tracking Systems. *Trans. SICE* 14(4), 13–18 (1978)
5. Anderson, B.D.O., Moore, J.B.: Linear System Optimization with Prescribed Degree of Stability. *Proc. IEE* 116(12), 2083–2089 (1969)
6. Trisuwanawat, T., Komine, N., Iida, M.: Optimal H_2 Control of Oscillations via Derivative State Constraints. In: *Proc. of 1999 American Control Conference (ACC)*, San Diego, California, USA, pp. 2305–2309 (1999)
7. Komine, N., Benjanarasuth, T., Ngamwiwit, J.: Derivative State Constrained Optimal H_2 Control for and its Application to Crane System. In: *ICCAS 2005, KINTEX, Korea*, June 2–5, pp. 2076–2080 (2005)

Utilization of Evolutionary Algorithms for Making COMMONS GAME Much More Exciting

Norio Baba¹, Hisashi Handa², Mariko Kusaka¹, Masaki Takeda¹,
Yuriko Yoshihara¹, and Keisuke Kogawa¹

¹ Information Science, Osaka Kyoiku University,

Asahiga-Oka, 4-698-1, Kashiwara City, Osaka Prefecture, 582-8582, Japan

² Information Science, Okayama University, Tsushima Naka, 3-1-1, Okayama City,
700-8530, Japan

Abstract. In this paper, we suggest that soft computing techniques such as NNs, MOEA-II, and FEP could be utilized for making the original COMMONS GAME much more exciting. Several game playing results by our students confirm the effectiveness of our approach.

Keywords: Neural Networks, Evolutionary Algorithms, Gaming, COMMONS GAME, Exciting Game.

1 Introduction

Historically speaking, gaming [1]-[3] has its origin in war games. However, after World War II, it has mainly been used for peaceful purposes. It has been used for various purposes such as education, training, decision-making, entertainment, and etc. [4]-[9]. Along with the appearance of various types of games, continuous effort has been done in order to let existing games be more exciting [8]-[13]. About 15 years ago, we suggested that NNs & GAs could be utilized for making the original COMMONS GAME, one of the most popular environmental games, become much more exciting [10]-[12]. Recently, we [14] also suggested that EAs & NNs could be utilized for constructing a more exciting version of the original COMMONS GAME.

In this paper, we shall compare these two games (Original COMMONS GAME, COMMONS GAME modified by EAs & NNs) by checking the data having been obtained after the several game playing by our university students.

2 COMMONS GAME

The COMMONS GAME [5] developed by Powers et al. is an educational game intended to help people learn about “COMMONS”. Since we live in a world having only finite natural resources, it is wise to consider their careful utilization. The COMMONS GAME may be quite helpful in stimulating discussion on this problem.

Fig. 1 illustrates the layout of the original COMMONS GAME.

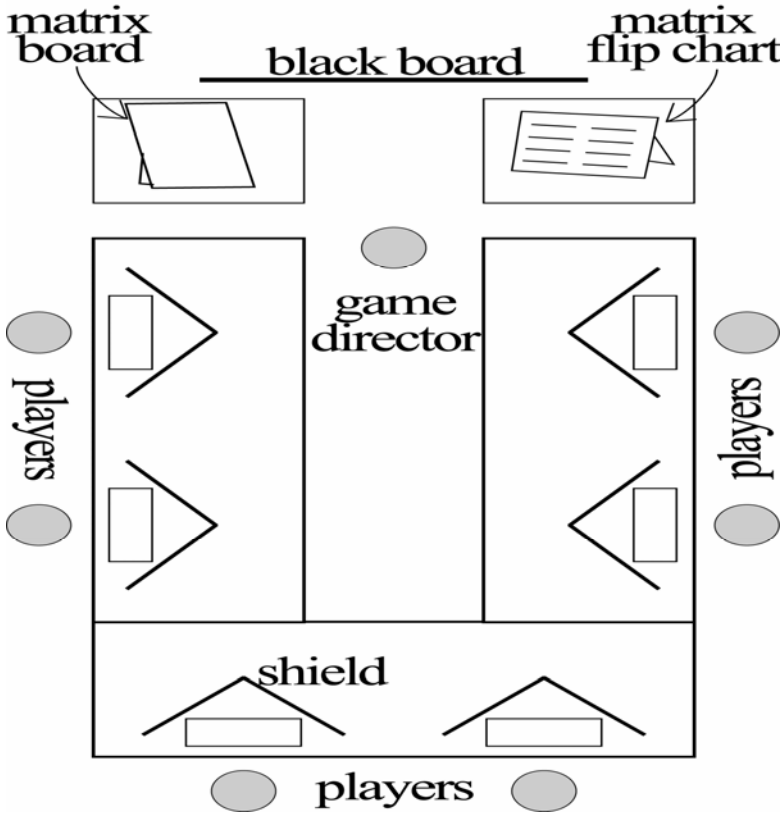


Fig. 1. Layout of the original COMMONS GAME

In the following, we give a brief introduction to this game:

First, six players are asked to sit around a table. Following a brief introduction concerning game playing, the game director tells the players that their objective is to maximize their gains by choosing one card among the five colored (Green, Red, Black, Orange, Yellow) cards in each round. In each round, players hide their cards behind a cardboard shield to ensure individual privacy. Each colored card has its own special meaning concerning the attitude toward the environmental protection, and has the following effect upon the total gain of each player:

- 1) Playing a green card represents high exploitation of the commons. Players who play a green card can get a maximum reward. However, they lose 20 points if one of the other players plays a black card in the same round.
- 2) A red card represents a careful utilization of the commons. Red card players can only get about forty percent as much as green in each round.
- 3) A black card has a punishing effect on the green card players. Players who have played a black card have to lose 6 points divided by the number of black card players, but are able to punish green card players by giving them - 20 points.

- 4) A yellow card represents a complete abstention from the utilization of the commons. Players who play this card get 6 points.
- 5) Orange cards give an encouraging effect to red card players. Players who have played this card have to lose 6 points divided by the number of the orange cards, but are able to add 10 to red players points.

Depending upon the players' strategies, the state of the commons changes. If players are too eager to exploit the commons, then they would face serious deterioration of the commons in a rather early stage of game playing. Although players are informed that there will be 60 rounds, each game ends after 50 rounds. After each 8th round, players have a three minute conference. They could discuss everything about the game and decide every possible way to play in future rounds.

We have so far explained a brief outline of the COMMONS GAME. Due to page, we don't go into further details. (Interested readers are referred to [5], [8], and [14]).

3 Modified COMMONS GAME Utilizing Evolutionary Algorithms (EAs)

We have so far enjoyed a large number of playing of the original COMMONS GAME. Those experiences have given us a valuable chance to consider seriously about the current situation of the commons. However, we did often find that some players lost their interest in game playing, even in the middle of the game.

We have tried to find the reason why some players lost their interest in game playing. We have come to the conclusion that the original COMMONS GAME is comparatively monotonous. Further, we have concluded that the following rule makes its game playing monotonous:

In the original COMMONS GAME, players who have chosen a green card receive a penalty, - 20 points, when some player chooses a black card in the same round. On the other hand, black card players receive a minus point $[- 6 / (\text{numbers of players who have chosen a black card})]$ even though they have contributed a lot in giving punishing effect toward green card players and maintaining current state concerning the commons. Orange card players (who have played an important role in recommending other players use of the red card (not green card)) also receive a minus point $[- 6 / (\text{numbers of players who have chosen an orange card})]$.

Only red card players who have not played any positive role in maintaining current state of the commons always receive gains.

We have considered that some change in the points "-20" and "-6" mentioned above would make the original COMMONS GAME much more exciting. In order to find an appropriate point for each card, we tried to utilize EAs [14]-[16].

In the following, we shall briefly explain "How the EAs have been utilized in order to find a better point of each card?"

3.1 Modified COMMONS GAME Constructed by the Use of the Two Evolutionary Algorithms

In the original COMMONS GAME, point of each colored card is fixed and any environmental change is not taken into account for deciding it. Recently, we have

tried to consider a new rule for assigning a point to each colored card which takes environmental changes into consideration.

In the followings, we shall briefly explain our new trial.

First, we set up the following framework for deciding a point of each colored card:

- a) Penalty P_G for green players: We proposed an appropriate way for penalizing green players which takes the environmental changes into account:
 $P_G = -W_G \times (\text{Gain } G)$, where W_G means the numerical value that is determined by the Evolutionary Programming, and "Gain G " denotes the return that the green players can get if any other player does not choose "Black Card."
- b) Point AOB that black players lose: We proposed an appropriate way (for asking black players pay cost in trying to penalize green players) which takes the environmental changes into account: $AOB = OB / NOB$ ($OB = -A \times (\text{Gain } R)$), where A means the numerical value that is determined by the Evolutionary Programming, and NOB means the number of players who have chosen the black cards, and "Gain R " denotes the return that the red player can get.
- c) Point OR that orange players add to the red players: We proposed an appropriate way (for helping red players maintain the commons) which takes the environmental changes into account: $OR = W_0 \times (\text{Gain } R)$, where W_0 means the numerical value that is determined by the Evolutionary Programming.

Then, we tried to utilize the two Evolutionary Algorithms NSGA-II [15] and FEP [16] for the following objectives:

- 1) NSGA-II was utilized for generating various intelligent computer players.
- 2) FEP was utilized in order to find an appropriate combination of the values of the three parameters W_G , A , and W_0 .

Thanks to NSGA-II and FEP, we could succeed in finding the following two combinations of the parameter values:

$$W_g = 0.27, A = 0.04, W_0 = 0.12 ; \quad W_g = 0.29, A = 0.2, W_0 = 0.1$$

Due to space, we don't go into details. Interested readers are referred to [14].

4 Game Playing Experiments for an Appropriate Evaluation of the Three Games

In order to investigate whether the two modified COMMONS GAMES are appropriate or not, one of the authors asked his students to play three games (original COMMONS GAME, two modified COMMONS GAMES). In the followings, we show several data relating to game playing of the three games. Fig.2 illustrates the changes of the total points of the 6 players in the game run of the original COMMONS GAME. Fig.3 illustrates the changes of the total points of the six players in the game run of the modified COMMONS GAMES where the parameter values ($W_g=0.27$, $A=0.04$, and $W_0=0.12$) have been utilized.

Fig.4 illustrates the changes of the total points of the six players in the game run of the modified COMMONS GAME where the parameter values ($W_g=0.29$, $A=0.2$, and $W_0=0.1$) have been utilized.

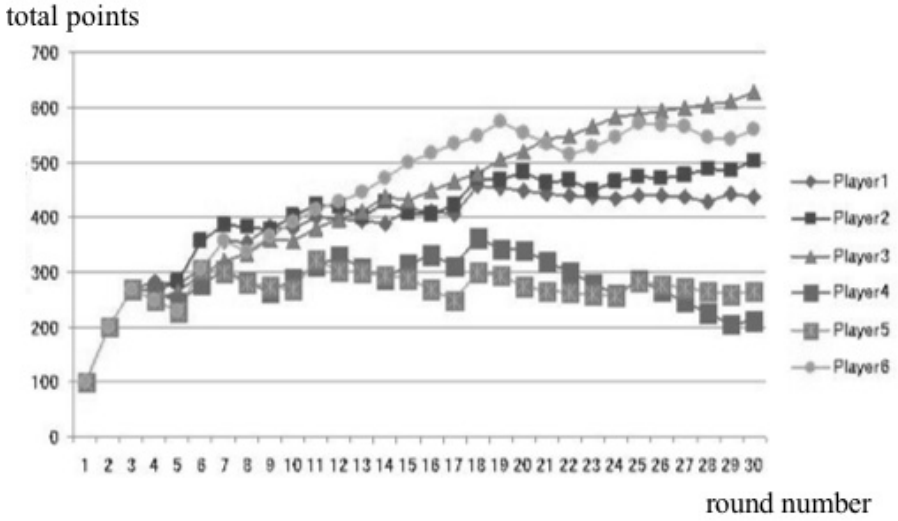


Fig. 2. Changes of the total points of the 6 players in the game run of the original COMMONS GAME

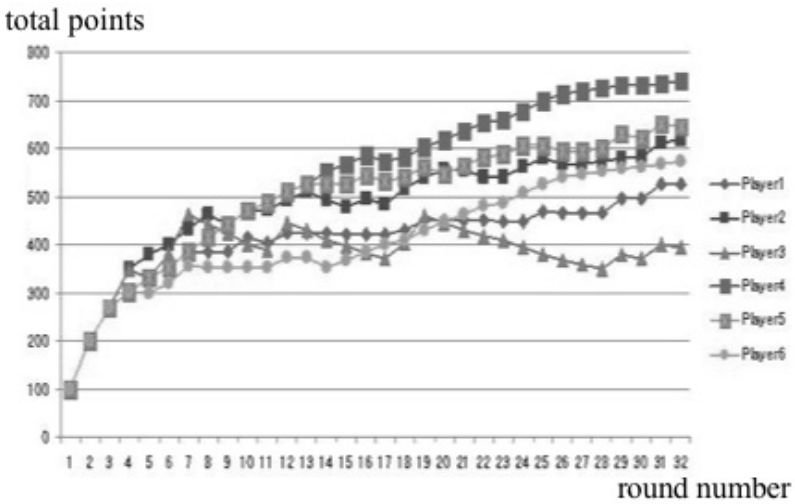


Fig. 3. Changes of the total points of the 6 players in the game run of the Modified COMMONS GAME ($W_g=0.27$, $A=0.04$, $W=0.12$)

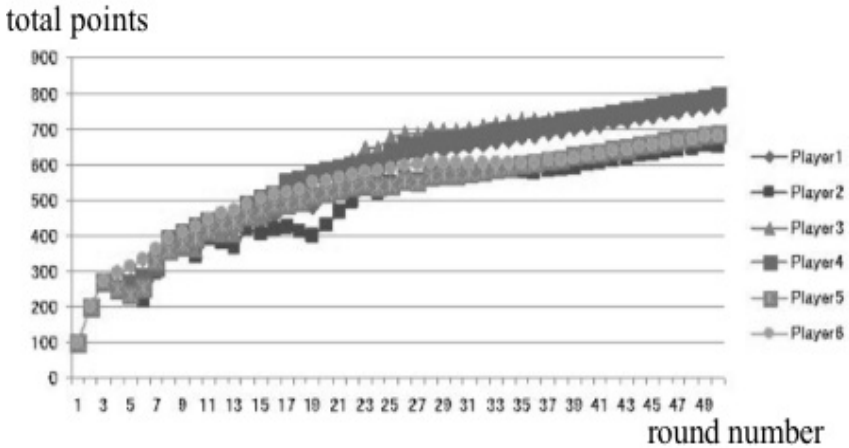


Fig. 4. Changes of the total points of the 6 players in the game run of the Modified COMMONS GAME ($W_g=0.29, A=0.2, W_0=0.1$)

From Fig.2, Fig.3, and Fig.4, we can easily observe:

- 1) Difference between the total point gained by the top player and those gained by the last player is almost always the smallest in the game playing of the modified COMMONS GAME (where the parameter values $W_g=0.29, A=0.2,$ and $W_0=0.1$ have been utilized) among those of the three games.
- 2) The number of the times of the changes of the rankings of the 6 players is also the highest in the game playing of the modified COMMONS GAME (where the parameter values $W_g=0.29, A=0.2,$ and $W_0=0.1$ have been utilized) among those of the three games.
- 3) From Fig.2 and Fig.3, we can also observe that the other modified COMMONS GAME (where the parameter values $W_g=0.27, A=0.04,$ and $W_0=0.12$ have been utilized) also provides a bit better gaming environment compared with that of the original COMMONS GAME.

There might be a number of indicators which can evaluate player’s involvement in the game playing. Among those, the authors believe that the two indicators having been utilized above (1)&2)) might be by far the two of the most important. After the game playing, one of the authors asked the players concerning their satisfaction about the game playing of the three games. Almost all of the players expressed their positive impression about the game playing of the modified COMMONS GAME (constructed by the use of the two Evolutionary Algorithms & NNs). He also asked the following question: “How do feel about the effectiveness of the game playing from the view point of “letting players consider seriously about the commons” ?” More than half of the players agreed with the effectiveness of the game playing of the modified COMMONS GAME (constructed by the use of the two Evolutionary Algorithms & NNs) toward this purpose.

5 Concluding Remarks

During the last several years, we have been involved to the study for finding the best game (from the point of view of “Which game is the best for letting people consider

seriously about the commons?”. In this paper, we have compared the modified COMMONS GAME utilizing EAs & NNs with the original COMMONS GAME. Several game playing having been done by our students confirm that the modified COMMONS GAME utilizing EAs & NNs can provide the better chance for letting players consider seriously about the commons. However, this has been confirmed only by several game playing by our students. Future research is needed to carry out lots of game playing by various people for the full confirmation of our research.

Further, continuous effort is also needed for finding a more advanced game for letting people have a chance to consider seriously about the commons.

Acknowledgements

The authors would like to express their heartfelt thanks to the Foundation for Fusion of Science & Technology (FOST) who has given them financial support.

References

1. Duke, R.: *Gaming: The Future's Language*. Sage Publications, Thousand Oaks (1974)
2. Shubik, M.: *Games for Society Business and War: Towards a Theory of Gaming*. Elsevier, Amsterdam (1975)
3. Hausrath, A.: *Venture Simulation in War, Business, and Politics*. McGraw-Hill, New York (1971)
4. Stahl, I. (ed.): *Operational Gaming: An International Approach*, IIASA (1983)
5. Powers, P., Duss, R., Norton, R.: *THE COMMONS GAME Manual*, IIASA (1980)
6. Ausubel, J.: *The Greenhouse Effect: An Educational Board Game*. Instruction Booklet, IIASA (1981)
7. Baba, N., Uchida, H., Sawaragi, Y.: A Gaming Approach to the Acid Rain Problem. *Simulation & Games* 15(3), 305–314 (1984)
8. Baba, N., et al.: Two Microcomputer-Based Games. IIASA Collaborative Paper, WP-86-79, 1–46 (1986)
9. Baba, N.: PC-9801 Personal Computer Gaming System. Nikkan Kogyo Publishing Company (1986) (in Japanese)
10. Baba, N.: An Application of Artificial Neural Network to Gaming. In: *Proceedings of SPIE*, vol. 2492, pp. 465–476 (1995) (Invited Paper)
11. Baba, N., Kita, T., Takagawara, Y., Erikawa, Y., Oda, K.: Computer Simulation Gaming Systems Utilizing Neural Networks & Genetic Algorithms. In: *Proceedings of SPIE*, vol. 2760, pp. 495–505 (1996) (Invited Paper)
12. Baba, N.: Application of Neural Networks to Computer Gaming. In: Tzafestas, S.G. (ed.) *Soft Computing in Systems and Control Technology*, ch.13, pp. 379–396. World Scientific, Singapore (1999)
13. Baba, N., Jain, L.C. (eds.): *Computational Intelligence in Games*. Springer, Heidelberg (2001)
14. Baba, N., Jain, L.C., Handa, H.: *Advanced Intelligent Paradigms in Computer Games*. Springer, Heidelberg (July 2007)
15. Deb, K., et al.: A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II. *IEEE Trans. EC* 6(2), 182–197 (2002)
16. Yao, X., Liu, Y., Lin, G.: Evolutionary Programming Made Faster. *IEEE Trans. EC* 3(2), 82–102 (1999)

Education of Embedded System by Using Electric Fan

Osamu Yuuki, Junji Namiki, and Kunihiro Yamada

Professional Graduate School Embedded Technology,
Tokai University,
2-2-12 takanawa, minato-ku, Tokyo-to 108-8619, Japan
yuki.osamu@canon.co.jp, yamadaku@tokai.ac.jp

Abstract. These days most electronic equipment uses system engineering incorporating computers. The increased number of semiconductor elements in computer-embedded systems causes both the hardware and the software to grow in size, making them increasingly complex. It is therefore an urgent task to develop compatible techniques and maintain quality. Considering this situation, this paper discusses a method of training engineers and managers for computer-embedded systems using fans.

Keywords: Education, Embedded System, Electric Fan.

1 Introduction

Since a computer-embedded system (referred to below as an embedded system) integrates hardware, software and mechanisms into one system, it requires engineers and managers to have a broad, comprehensive viewpoint. Such a viewpoint should include the technical boundaries of hardware, software and mechanisms and also extend to the business feasibility of the system. Technical departments have to do more than simply create element technologies and put them together. They should take the initiative which is the key to most successful business.

Progress and development in semiconductor technology increases the number of semiconductor elements employed, making hardware grow [1][2][3] in size. Software on the other hand has a great deal of freedom when dealing with hardware. Since this freedom is not always addressed properly, there are various types of software that can realize hardware processing. This in turn makes systems very complex, leading to delays in development and deterioration of quality.

Under these circumstances, this project aims at training human resources who understand the standard system development procedures, behave according to engineers' basic principles, recognize the scope of work and career [4], and carry out the duties and responsibilities of project managing engineers. From the technical viewpoint, the study is intended to train engineers who can develop embedded systems while appreciating embedded systems as products [5] and understanding the technological elements necessary to develop systems, hardware and software. Taking the familiar product of a fan as study material, this paper discusses the training of engineers who deal with hardware and software as well as project base engineers [6] who integrate [7] hardware and software.

2 Training Program

This section of the paper deals with the training program being undertaken at the graduate school of Tokai University for Toshiba Solutions Corporation and Renesas Technology Corporation. The following subsections describe the outline of the program. Figure 1 shows its framework diagrammatically.

The project at the graduate school of Tokai University, in particular, is designed so that it moves towards the goal while students with various unique talents from various faculties and people who already have jobs cooperate and supplement the others and share each specialty between them.

2.1 Outline of Classroom Sessions

The classroom sessions described below enable the trainees to learn the skills for the cooperative development of hardware, software, mechanisms and systems regarding the embedded systems.

Through group activities, the trainees learn the following from the study material of the product (fan).

- Business
- Product specifications
- Functions of system, mechanism, hardware and software to realize specifications
- Developing a new product (fan)
- Presentation and description of achievements of the project through group activities

(1) To physically simulate the investigation of embedded system specifications, their testing and quality through practical drills while following the standard development procedures of an embedded system so that the trainees learn the skills and techniques to develop and market an embedded system.

(2) The trainees survey in advance the production scales, producing countries, market and technical trends around the world and report their findings. Also without prior notice, they plan a new product development strategy and a business strategy.

(3) The trainees operate the fan to theorize the how it realizes its operating functions in terms of system, mechanism, hardware and software.

(4) The trainees verify what they theorized in Step (3). They disassemble the fan to verify their theory in terms of the mechanism and system. They then use an oscilloscope to observe the signals and electrical waveforms and interpret the electronic circuit of the circuit board and the electrical circuit of the motor and switches to draw up circuit diagrams to verify their theory of realizing the operating functions of the hardware and software.

(5) Based on the business strategy planned in Step (2), the trainees write out the product specifications to develop the product.

2.2 Classroom Session Details

As a drill for the embedded system development project, a comprehensive program for embedded technology training is implemented using a fan.

(1) Drill for software development:

To physically simulate the investigation of embedded system specifications, their testing and quality through practical drills while following the standard development procedures of embedded systems so that the trainees learn the skills and techniques to develop and market the embedded system.

(2) Planning business strategy

The trainees survey in advance the production scale, producing countries, market and technical trends around the world and report their findings. Also, without prior notice, they plan a new fan development strategy and a business strategy.

(3) Estimating product functions

The trainees operate the fan to theorize how it realizes its operating functions in terms of system, mechanism, hardware and software.

(4) Analyzing actual product

The trainees verify what they have theorized in Step (3). They disassemble the fan to verify their theory in terms of the mechanism and system. They then use an oscilloscope to observe the signals and electrical waveforms and interpret the electronic circuit of the circuit board and the electrical circuit of the motor and switches to draw up the circuit diagrams to verify their theory of realizing the operating functions of the hardware and software.

(5) Developing product

Based on the business strategy planned in Step (2), the trainees write out the product specifications to develop the product.

(6) Presentation

The trainees make a presentation of the business feasibility with the developed fan.

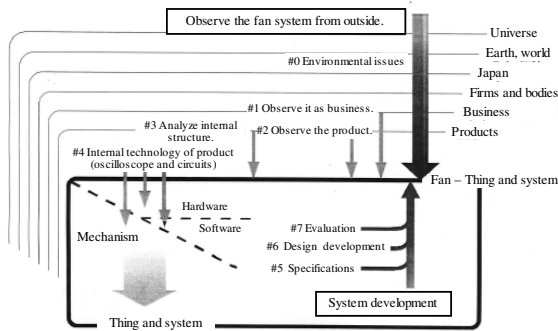


Fig. 1. Diagrammatic representation of embedded system development program (System - Observe from outside, Develop it.)

While the program deals with product development, the program development only takes place while the specifications describe what the mechanism and hardware are like. Fans to neutralize a typhoon or a tornado and endure rainfall are also included. The following section describes a case of the simplest form of fan which the project members worked out.

3 Classroom Sessions

The following is part of a case example that the project members worked out in the classroom sessions.

3.1 Interpreting Fan Functions

(1) Specifications

- Air flow volume
The air flow volume changes in three steps; Low, Medium, and High.
- Rhythm
The air flow volume changes after a certain interval.
- Timer
The fan turns itself off in one, two, four and six hours.
- Swing
The fan starts to swing when pressing the switch on its back.

Even when the fan is switched off, it memorizes the operating parameters and settings immediately before it was turned off as long as electric power is supplied from a power outlet. When the fan is turned on for the first time, it operates in the default Low setting mode. A parameter transition table is created to describe the relationships such as those between the on/off pushbutton switch and the timer button and between the elapsed time and input.

Table 1. Parameter Transitions between On/Off Pushbutton Switch and Timer Button

Status/input	On/off	Timer	Elapsed time
Off	None	Off	Off
None	Off	1 hour	None
1 hour	Off	2 hours	Off
2 hours	Off	4 hours	1 hour
4 hours	Off	6 hours	2 hours
6 hours	Off	None	4 hours

Table 2. Parameter Transitions between Air Flow Volume Button and Rhythm Button

Status/input	Air flow volume	Rhythm
Low	Medium	Low (rhythm)
Medium	High	Medium (rhythm)
High	Low	High (rhythm)
Low (rhythm)	Medium (rhythm)	Low
Medium (rhythm)	High (rhythm)	Medium
High (rhythm)	Low (rhythm)	High

Table 1 shows the parameter transitions between the on/off pushbutton switch and the timer button. Table 2 shows the parameter transitions between the air flow volume button and the rhythm button. Figure 2 depicts the parameter transitions of the entire fan.

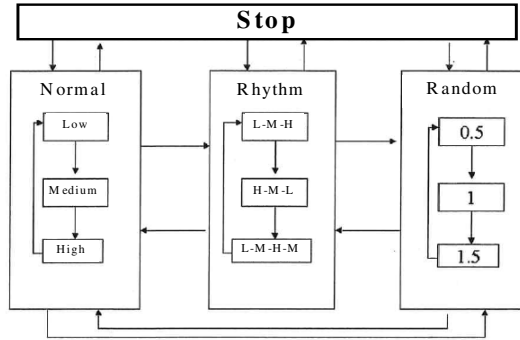


Fig. 2. Parameter transitions of entire fan

3.2 Analyzing Internal Mechanism

From this part on, the trainees investigate the actual hardware structure of the fan so that they acquire the techniques of using a microcomputer to control the mechanism.

(1) Internal structure

1) Function switches

The fan has four function switches, on/off, air flow volume, rhythm and timer. Pressing these buttons changes the fan's operating conditions. Figure 2 is a photograph of the pushbutton switches of the fan.

2) Motor

Driven by the power IC, the coil and the permanent magnet run the motor. Figure 3 is a photograph of the fan motor.

3) LEDs

These LEDs are turned on and off by the switches in the microcomputer to indicate the control status of the fan.

Figure 4 is a photograph of the LEDs

(2) Electric system (circuit)

1) Main circuit board

The main circuit board is responsible for controlling the elements of the mechanism to generate wind as one of its basic functions and control the air flow volume as a value-added operation. Figure 5 is a photograph of the main circuit board. Note that the red lines show the connections read out from the main circuit board by tracing its metal wiring. Figure 6 is the main board circuit diagram that shows the electronic device layout and the wiring connections which the trainees read out.



Fig. 3. Pushbutton

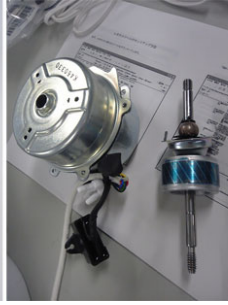


Fig. 4. Fan motor



Fig. 5. LEDs

2) LED circuit board

Corresponding to the LEDs shown in Figure 4, the LEDs on the circuit board indicate the operating status of the fan.

Figure 7 shows the LED circuit board and connectors.

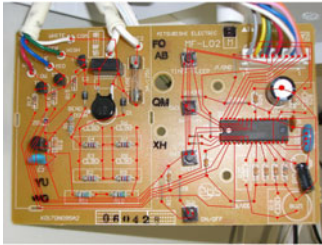


Fig. 6. Main circuit board

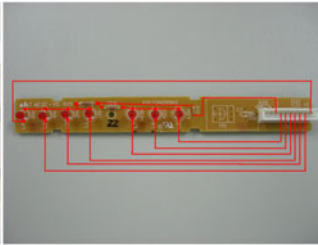


Fig. 7. LED circuit board

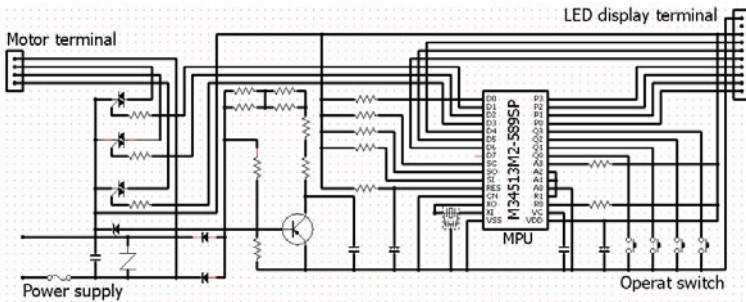


Fig. 8. Circuit diagram

3) Operation

- Turning on or off three TRIACs controls the motor speed.
- A Zener diode, two diodes and a PNP transistor convert the 100 V AC supply to 5 V DC.
- The ceramic oscillator operates at 1.8 MHz.
- A bypass capacitor is installed immediately next to the IC.

- A measure against chattering is considered to be provided by the software that checks the on status and other parameters.
- The microcomputer controls the TRIACs at its output port. The rhythm duration is controlled by turning on and off at a certain interval.
- The low voltage detection of the microcomputer is connected to the 5 V DC power source. The microcomputer therefore appears to have a safety function to prevent it from running away.
- The microcomputer turns the indicator LEDs on or off.
- The microcomputer is considered to control the timer.

The trainees investigate the datasheet to see the MPU pin layout. Figure 8 shows the pin layout of the MPU (M34513E4SP) used on this fan.



Fig. 9. MPU (M34513E4SP)

M34513 M2 589SP

Descriptions of terminals

- 01 D0_ 5V (with resister)
- 02 D1_T3 (TRIAC_Gate_Line GREEN_HIGH-Control of wind power)
- 03 D2_T2 (TRIAC_Gate_Line BLUE_MIDDLE- Control of wind power)
- 04 D3_T1 (TRIAC_Gate_Line RED_LOW- Control of wind power)
- 05 D4_LED_L5 (Purple-LOW-LED Drive)
- 06 D5_LED_L6 (Red-MIDDLE-LED Drive)
- 07 D6_LED_L7 (Black-HIGH-LEDDrive)
- 08 D7 (N.C.)
- 09 P20_Input (with resister)
- 10 P21_Input (with resister)
- 11 P22_Input (with resister)
- 12 RESET Input (Line is set LOW→RESET)
- 13 CNVss (GROUND)
- 14 Xout (Ceramic resonator)
- 15 Xin (Ceramic resonator)

16 V_{ss} (GROUND)
 17 V_{dd} (5V power source)
 18 VDCE (The voltage reduction detection of enable_V_{dd}-5V_Line)
 19 P30 (LED Power: 5V)
 20 P31 (GROUND)
 21 Ain0 (GROUND)
 22 Ain1 (GROUND)
 23 Ain2 (GROUND)
 24 Ain3 (Input)
 25 P00 (SWITCH1-ON/OFF input)
 26 P01 (SWITCH2-SPEED)
 27 P02 (SWITCH3-RHYTHM)
 28 P03 (SWITCH4-TIMER/SLEEP)
 29 P10_LED_L4_6HR (Orange-HIGH-LED Drive)
 30 P10_LED_L3_4HR (Pink-HIGH-LED Drive)
 31 P10_LED_L0_2HR (Blue-HIGH-LED Drive)
 32 P10_LED_L0_1HR (White-HIGH-LED Drive)

The trainees then use an oscilloscope to check the relationships between the operating status from the main circuit board and the voltage changes across the terminals on the circuit board.

Figure 9 shows the voltage changes across the terminals when the fan is operating in the medium (rhythm) mode. It shows that a voltage may not be applied to D3, D2 or D1 on some occasions in this condition and, when turned off, a voltage is applied tentatively to all the diodes. The fan repeats the cycle of Low (1s) to Medium (3s) to High (1s) to Medium (4s) and to Off (4s).

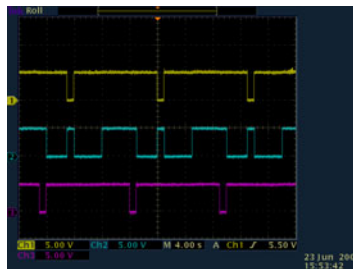


Fig. 10. Voltage in Medium (rhythm) mode

(3) Developing a fan with new specifications

While the program deals with product development, the program development only takes place while the specifications describe what the mechanism and hardware are like. Fans to neutralize a typhoon or a tornado and endure rainfall are also included.

Here the trainees propose the specifications of a fan with new functions so that the project can respond to market needs. The trainees create a new program or modify an existing one and download it onto the microcomputer so that they physically experience the actual programming. Figure 10 shows part of the originally created program of a fan with new functions.

```

                                  扇風自作制御プログラム.txt
/*****
/* 扇風機制御処理教材サンプルプログラム */
/* ファイル名: ElectricFan.c */
/* 内容: ボード上のCPUを介し扇風機を制御 */
/* 日付: 2007.7.24 */
/* コンパイラ: NC30WA (Ver. 5.00) */
/* 作成者: 鈴木 文彦 */
*****/

/* インクルードファイル */
#include "sfr26a.h" /* OAKSmini用定義ファイル */
#include "ComTypeDef.h" /* Typedef & Inline マクロ*/
#include "stdlib.h"
#include <time.h>

/* プロトタイプ宣言 */
void main(void); /* メイン関数 */
static INT16 GetSw( void ); /* スイッチ確認関数*/
static void Initializing( void ); /* M16C26Aの初期化と 内部処理領域初期化*/
static void ControlStop( void ); /* 扇風機停止制御*/
static void ControlLow( void ); /* 扇風機低速制御*/
static void ControlMid( void ); /* 扇風機中速制御*/
static void ControlMid2( void ); /* 扇風機中速制御*/
static void ControlHig( void ); /* 扇風機高速制御*/
static void ControlHig2( void ); /* 扇風機高速制御*/
static void ControlTime( void ); /* タイマー監視起動*/

void ta0int0(); /* 割込み関数 */
#pragma INTERRUPT ta0int

```

Fig. 11. Original program of fan with new functions

Fifty-one graduate school students of Tokai University (total number of students from 2007 to 2009), 48 persons from Toshiba Solutions Corporation and 16 persons from Renesas Technology Corporation participated in this program of training embedded system engineers. The classroom sessions of 45 hours, 105 hours and 30 hours were allocated to the graduate school of Tokai University, Toshiba Solutions Corporation and Renesas Technology Corporation, respectively [8].

4 Conclusion

Through a series of disassemblies and analyses as well as the development of fans with new functions, this embedded system engineer training program is effective, though at an initial level, for the trainees to acquire the techniques and skills of measurement with an oscilloscope, circuit interpretation, investigation of transistor functions, disassembling and reassembling mechanical components, soldering, program development using C programming language, proposing business plans, recognizing safe products and work, planning product specifications and quality, and diverse aspects of product development.

In addition, for the trainees to acquire these techniques and skills in a short period of time, it is effective to have the opportunity to learn together through (i) presentations and reporting to identify clear targets, (ii) project learning from several instructors, and (iii) analysis and development through group activities.

Acknowledgments. The authors would like to thank Prof. S. Oohara, F. Suzuki and all the participants for their support in the development of the project.

References

1. Yamada, K.: New System Structuring Method that Adapts to Technological Progress of Semiconductors. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based and Intelligent Information and Engineering Systems. LNCS, vol. 5712, pp. 189–196. Springer, Heidelberg (2009)
2. Chen., I.J., Su., H.-M., Liu, J.-H.: A curriculum design on embedded system education for first-year graduate students. In: 2007 International Conference on Parallel and Distributed Systems (2004)
3. Limin, C.: Thought on embedded system education for application-oriented undergraduates of electronics major. In: 4th International Conference on Computer Science & Education, ICCSE 2009, pp. 1476–1478 (2009)
4. Bobbie, P.O., Uboh, J., Davis, B.: A project in embedded systems design and development: a partnership with area high school scholars. In: 34th Annual Frontiers in Education, FIE 2004, vol. 2, pp. F4D -14 -17 (2004)
5. Hall, T.S., Bruckner, J., Halterman, R.L.: A Novel Approach to an Embedded Systems Curriculum. In: 36th Annual Frontiers in Education, pp. 15–20 (2006)
6. Hsu., R.C., Liu, W.-C.: 3rd International Conference on Information Technology: Research and Education, ITRE 2005, pp. 362–366 (2005)
7. Mitsui, H., Kambe, H., Endo, S., Koizumi, H.: A Student Experiment Method for Learning the Basics of Embedded Software Development Including HW/SW Co-design. In: 22nd International Conference on Advanced Information Networking and Applications - Workshops, AINAW 2008, pp. 31–37 (2008)
8. Shi., X., Zhang., J., Ju, Y.: Research and Practice in Undergraduate Embedded System Course. In: The 9th International Conference for Young Computer Scientists, ICYCS 2008, pp. 2659–2663 (2008)

Development and Evaluation of a Routing Simulator for a Mutually Complementary Network Incorporating Wired and Wireless Components

Hiroki Morita¹, Naoki Yusa¹, Noriyuki Komine², Kouji Yoshida³, Masanori Kojima⁴,
Tadanori Mizuno⁵, and Kunihiko Yamada¹

¹ Professional Graduate School of Embedded Technology, Tokai University 2-2-12, Takanawa,
Minatoku, Tokyo, 108-0074, Japan

² School of Information and Telecommunication Engineering, Tokai University
³ Shonan Institute of Technology

⁴ Osaka Institute of Technology

⁵ Shizuoka University

9aej015@mail.tokai-u.ac.jp, yamadaku@tokai.ac.jp

Abstract. Networks, whether wired or wireless, used at homes are now 100% mutually complementary in terms of communication performance, but that performance can be as low as 4.1% if used in a school building, for example. The objective of the study described herein is to improve communication performance using additional routing capability. A simulator was developed to ascertain and evaluate the characteristics of such a mutually complementary network with an added routing capability.

Assuming that we have up to 15 s available for practical use of communication, such a judgment was made that up to 11 nodes for every five-story building was about reasonable to use for practical uses of communications. With that fact in mind, a mutually complementary network was shown to be available for use in a school building by allocating every domain unit of 11 nodes to the entire building.

Keywords: simulation, network, PLC, ZigBee, building.

1 Mutually Complementary Network Using Wired and Wireless Communications

A mutually complementary network supports communication using two or more different communication methods concurrently to improve the characteristic of the communication performance [1]. Figure 1 presents a diagram showing the concept of that communication method. In the study described herein, Power Line Communication (PLC) [2] was used for wired communication, with Zigbee [3] for a wireless communication. For a three-story and 200 m² ferroconcrete house or condominium, it has been recognized that the use of PLC alone achieves 70% of communication performance, and that the use of Zigbee alone gives 82%. The mutual complementary network using those two communication methods concurrently is

expected to increase to 94.6% of theoretical communication performance from a mathematical perspective, but it actually attains 100% [1]. As described above, a mutually complementary network, if used for a private homes, is understood to have sufficient communication performance that can be available for practical use. However, its performance for such buildings as schools is only 4.1% in practice, which is far from satisfactory [4]. To confront this problem, a routing method is expected to improve the situation.

A study has already been made of using a routing method to improve communication performance by allowing every node to recognize its location [5]. One problem is the fact that home electronic appliances or office machines that are placed and used at home or in the office are of lower functionality than a PC, for example. To have a network incorporating such equipment, we must therefore use a less-burdened routing method. One idea was to have a simply processed routing capability with no node location information given to each device at all, thus reducing the burden. That routing takes such a method that it tries to communicate with one node; if it finds that it failed, it moves to try every other node one by one in an arbitrary order. For one-time communication, it bypasses the node with which the communication fails. No re-execution would be made in such a case. Because this method is provided with no mutual location information, it is called a location-undefined communication method. Incidentally, the reported idea of using both wired and wireless communications used in that study resembles the mode used for the study described herein [6]. However, an important difference is that it uses a method by which either of the wired or wireless communications is selected for every communication space concerned, which means that it does not always use both communications concurrently. The mutually complementary network described herein has a characteristic of using both wired and wireless communications concurrently.

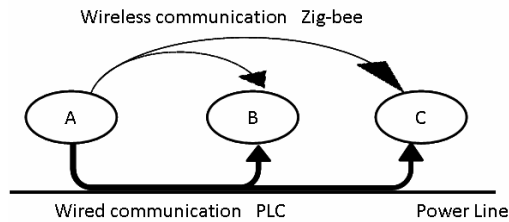


Fig. 1. Concept of a mutually complementary network

2 Necessity for Simulator Use

To have a mutually complementary network for a school building, for example, a routing method is taken to improve its communication performance. Here, a location-undefined communication method is used as described above. To accomplish a mutually complementary network with it, the routing performance must be verified first. In other words, we must check to determine whether or not communication can be accomplished within a practically allowable period of time. For that, we must know the averaged communication time period that is taken technically for the

shortest and longest routes as well as that for the entire route. However, we understand that vastly numerous combinations of routing exist when we examine the entire route, which indicates to us that we must determine the existing node disposition before anything else. Because it is practically difficult to do that manually, we must proceed to check with the help of a simulator. For the node distribution shown in Fig. 1 with nine nodes for a three-story building, the route combinations can be as many as 950 in all, for which we need a simulator for correct modeling.

Each number appearing in Fig. 1 represents a serial number given to a node. The electric power supply to Japanese homes is 100 V of single-phase three-wire type, separated into phase A and phase B. Communication, if tried using between phase A and the phase B, requires passage through a transformer, which lengthens the communication path distance, thereby increasing the load to its signal transmission, and further tends to pick up more noise. Overall, the communication performance becomes hampered significantly. Furthermore, with the Zigbee wireless communication used, we know that its communication performance is degraded by obstacles in its straight communication path line. That might happen often when trying to communicate with the node on another floor of a building, for example.

During the simulation we performed, communication performance lower than the threshold value of 50% is taken as 0%, and that higher than the threshold value as 100%. For this reason, communication using PLC is accomplished within the same phase only, in which sense only communication between the node number groups of 1, 3, 5, 7, and 9 and another node groups of 2, 4, 6, and 8 is available. With Zigbee used, it is available only between the node number group of 1, 2, 3, and 4 and another group of 5, 6, 7, and 8.

The simulator that is used shows the necessary routing trials made to accomplish communication between the associated nodes. It computes and outputs the minimum number of trials made, the maximum number of trials made, and the averaged value of those trials. The output is selectable either in a CSV format, a text format or an HTML format. The simulator specifications are presented in Table 1.

Table 1. Simulator specification

Specification	
Input data	Number of nodes
	Comm. availability table
Output data	Path for all nodes
	No. of comm. trials for each node path
	Max and min numbers for all comm. trials
	Averaged value of all comm. trials
Output method	CSV format
	Text format
	HTML format

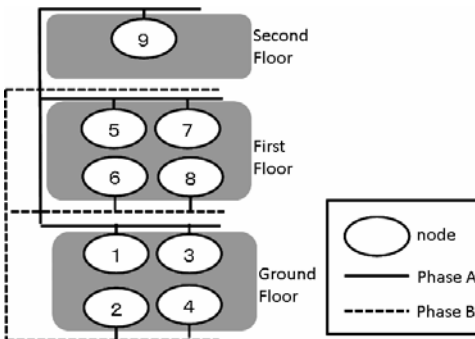


Fig. 2. Example of nine-node distribution in a three-story building

3 Development of a Simulator

The simulator was developed with C language using a currently available standard PC. The total lines of code of the program were 1065; its executable file size was 48 KB. Figure 3 presents the order in which the simulator actions take place. The simulator first detects the state of each evaluated node—where it is located on what floor of the building—and creates a communication availability table based on the phase relation. It then executes the simulator itself with the nodes given. Table 2 shows the communication availability table for nine nodes spread all on the three-story building floors, as shown in Fig. 2.

As described previously, because the communication performance level is determined at the threshold value of 50% with respect to whether the communication would be possible or not, the communication availability table for the node distribution of Fig. 2 is given as that shown in Table 2. In the table, element P represents that the communication is available using PLC, and Z indicates that it is available using Zigbee. The element value x indicates that the communication is not available in that position. Although the communication between nodes 1 and 3 of Table. 2 is available using either PLC or Zigbee, only the value of P is shown there for reference purposes.

Table 2. Transfer enable table for nine nodes

-	1	2	3	4	5	6	7	8	9
1	-	Z	P	Z	P	x	P	x	P
2	Z	-	Z	P	x	P	x	P	X
3	P	Z	-	Z	P	x	P	x	P
4	Z	P	Z	-	x	P	x	P	X
5	P	X	P	X	-	Z	P	Z	P
6	x	P	x	P	Z	-	Z	P	X
7	P	X	P	X	P	Z	-	Z	P
8	x	P	x	P	Z	P	Z	-	X
9	P	X	P	x	P	x	P	x	-

Figure 4 shows the simulator process sequence. As the simulator process operates, if the communication between two node points is found to be unavailable, then it is bypassed. That process progresses as depicted in Fig. 6 as a bypass process of simulator in the way to add a new table for it. Upon receipt of the information of the nodes and the communication availability table, the simulator creates the Combination List 1 (CL1), which is the list of the nodes telling whether the communication for each pair of them is available or not. First, whether the communication between each pair is directly available or not is checked based on the information of the communication availability table. If any node combination for which no direct communication is available is found, then the simulator halts checking with CL1, creates a new combination list (CL2) to use for a bypass process, and performs a routing process to check whether or not the communication would be available with the route that is newly found. Completing the check on the current combination list, it reverts to the previous list and starts checking it similarly. While repeating such process, it checks all possible combinations in that manner.

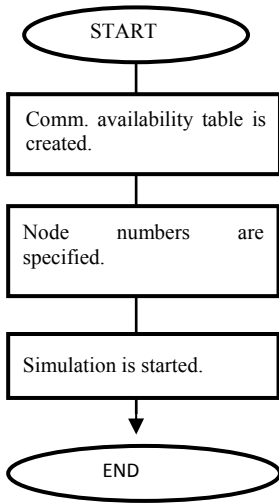


Fig. 3. Simulator action sequence

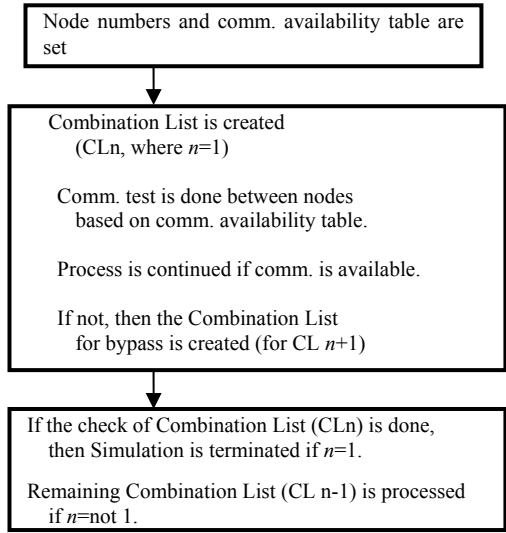


Fig. 4. Simulator process sequence

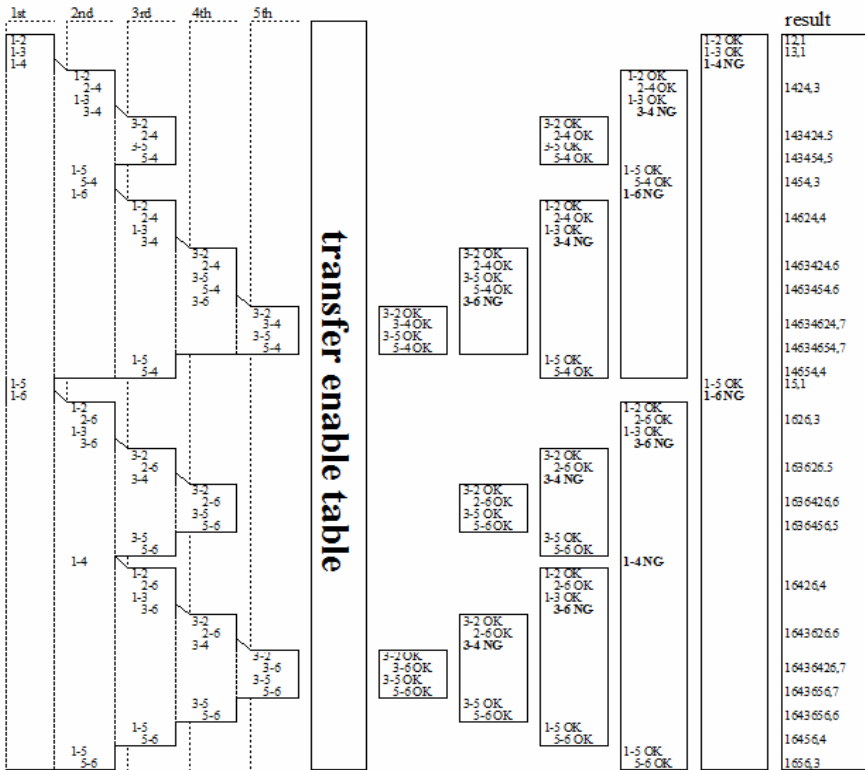


Fig. 5. Working flow of simulator

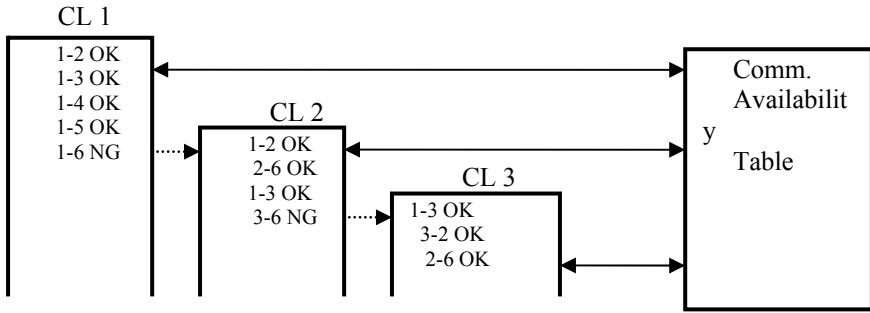


Fig. 6. The bypass process of simulator: CL2 is created because the 1–6 communication of CL1 fails so that it is bypassed. Similarly, it tries to connect via node 3. Because it is still not possible, it further bypasses to move to create CL3.

Figure 5 shows working of the simulator. At first, the simulator makes the 1st combination list. Then simulator checks transfer enable table. If can't communication, simulator makes new list name 2nd and make new combination use easy routing algorithm. Then the simulator change it form 1st to 2nd. Then the simulator checks transfer enable table. If the simulator reaches the end of it, change the list above of current list. The simulator continues for finish checking all combination lists.

4 Simulation Results

Table 3 presents simulation results obtained for eight nodes spread throughout the two-story building. In Table 3, “SN” stands for serial number, each node parenthesized in the path line that is given under “Route” signifies that no data were received there, “adr” stands for the address to, “Tn” stands for the serial number that falls in the same address, “R” for the number of route combinations to come to the same address, “Try” for how many times the communication availability table was referenced, and “TryAv” for the averaged number of times the communication availability table was referenced in the same address. Figure 7 portrays the relation between the number of nodes and the associated number of routings made based on the simulation results available. Figure 8 shows the total number of routes. Up from nine nodes, node distribution of two kinds was taken in the simulation. The solid line portion represents the case in which nodes are added on the two-story building, and the dotted line portion represents the case in which every new node is gradually added to one floor up. There were 379 route combinations and 7 times of maximum trial for 10 nodes in a two-story building, with 1,697,456 route combinations and 23 maximum trials for 10 nodes in a four-story building. For 11 nodes in a five-story building, there were over 400 million route combinations. The simulation duration lasted over 70 hr. Examining the nodes and the number of routings made in Fig. 6 reveals that up to 29 times routing was made for 11 nodes in a five-story building. Whether the communication is available or not was determined by sending a packet to

the node; if no response was returned, then it is judged that the communication was unavailable there. Presuming that it takes 0.5 s every time to check whether the communication is available or not, then this example of 29 times of routing takes 14.5 s in all, meaning that up to a 14.5 s delay would occur to start communication. Furthermore, for the case of 10 nodes in a four-story building, 11.5 s are caused to delay, and for nine nodes in a three-story building, 6 s delay. For the 12-node distribution in the two-story building, the maximum number of trials was 13 times, and the time necessary for the routing process was, consequently, 6.5 s.

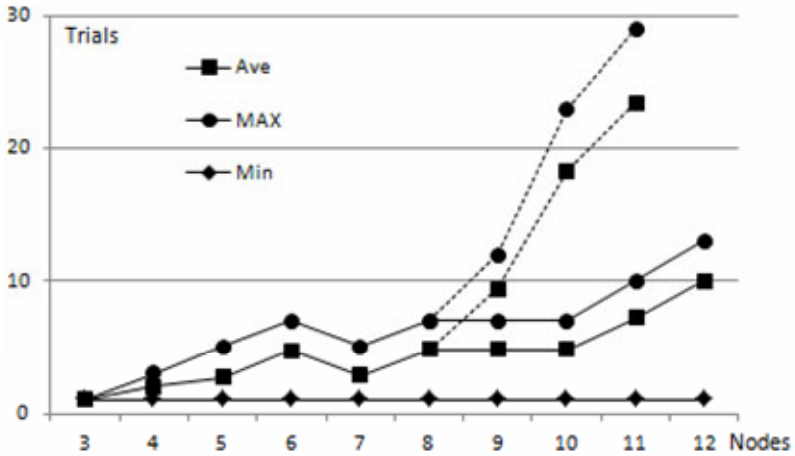


Fig. 7 Node and routing trial relations: ● maximum value, ■ averaged value, and □ minimum value

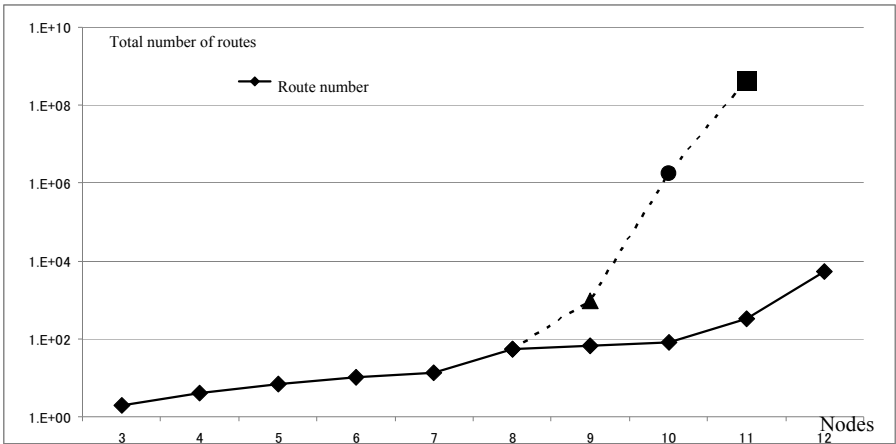


Fig. 8. Total number of route combinations

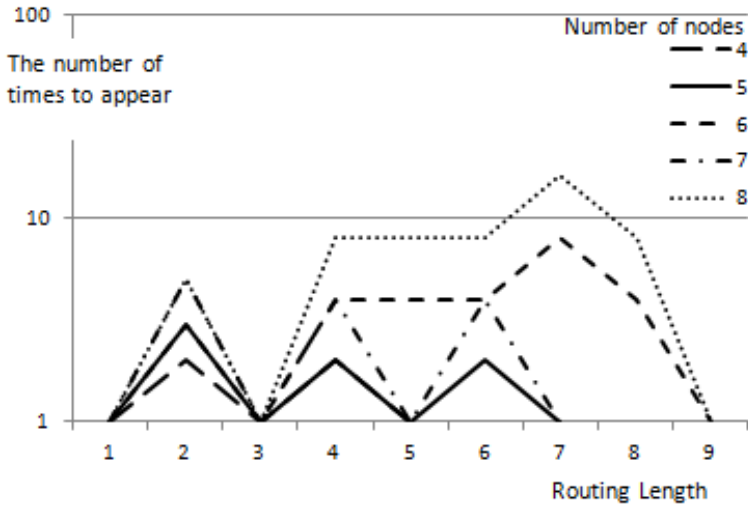


Fig. 9. Histogram of simulation result: This histogram contains 4 nodes for 8 nodes. This histogram show less difference with number of nodes. So few nodes and less spread nodes can make mutually complementary network using wired and wireless communications.

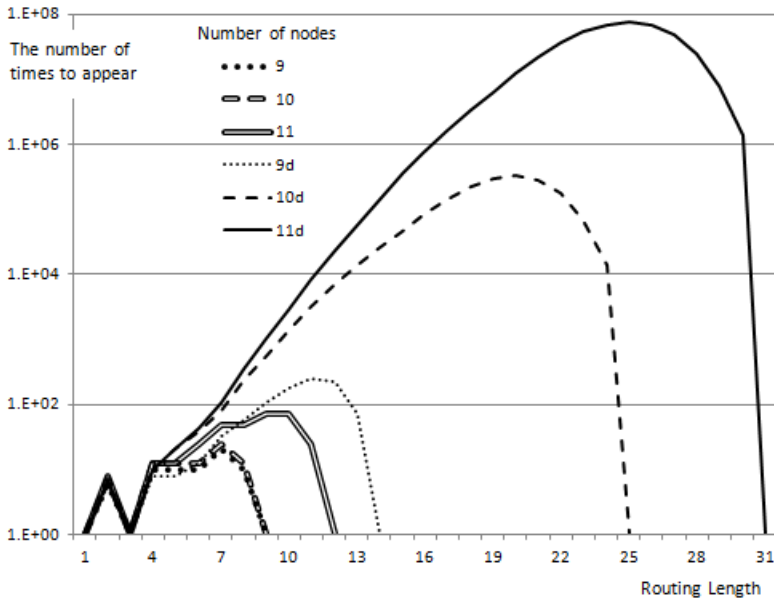


Fig. 10. Histogram of simulation result: This histogram contains 9 nodes for 11 nodes. This result means a lot of nodes and wide spread nodes to need some roles to make mutually complementary network using wired and wireless communications.

Table 3. Results of routing simulation: eight nodes in a two-story building

SN	Route	adr	Tn	R	Try	TryAv
1	[1→2]	1→2	1	1	1	1
2	[1→3]	1→3	1	1	1	1
3	[1→4]	1→4	1	1	1	1
4	[1→5]	1→5	1	1	1	1
5	[1→(6)→2→6]	1→6	1		3	
6	[1→(6)→3→(6)→2→6]	1→6	2		5	
7	[1→(6)→3→(6)→4→6]	1→6	3		5	
8	[1→(6)→3→(6)→5→6]	1→6	4		5	
9	[1→(6)→3→(6)→7→6]	1→6	5		5	
10	[1→(6)→3→(6)→(8)→2→6]	1→6	6		6	
11	[1→(6)→3→(6)→(8)→4→6]	1→6	7		6	
12	[1→(6)→3→(6)→(8)→5→6]	1→6	8		6	
13	[1→(6)→3→(6)→(8)→7→6]	1→6	9		6	
14	[1→(6)→4→6]	1→6	10		3	
15	[1→(6)→5→6]	1→6	11		3	
16	[1→(6)→7→6]	1→6	12		3	
17	[1→(6)→(8)→2→6]	1→6	13		4	
18	[1→(6)→(8)→3→(6)→2→6]	1→6	14		6	
19	[1→(6)→(8)→3→(6)→4→6]	1→6	15		6	
20	[1→(6)→(8)→3→(6)→5→6]	1→6	16		6	
21	[1→(6)→(8)→3→(6)→7→6]	1→6	17		6	
22	[1→(6)→(8)→3→(6)→(8)→2→6]	1→6	18		7	
23	[1→(6)→(8)→3→(6)→(8)→4→6]	1→6	19		7	
24	[1→(6)→(8)→3→(6)→(8)→5→6]	1→6	20		7	
25	[1→(6)→(8)→3→(6)→(8)→7→6]	1→6	21		7	
26	[1→(6)→(8)→4→6]	1→6	22		4	
27	[1→(6)→(8)→5→6]	1→6	23		4	
28	[1→(6)→(8)→7→6]	1→6	24		4	5.167
29	[1→7]	1→7	1	1	1	1
30	[1→(8)→2→8]	1→8	1		3	
31	[1→(8)→3→(8)→2→8]	1→8	2		5	
32	[1→(8)→3→(8)→4→8]	1→8	3		5	
33	[1→(8)→3→(8)→5→8]	1→8	4		5	
34	[1→(8)→3→(8)→(6)→2→8]	1→8	5		6	
35	[1→(8)→3→(8)→(6)→4→8]	1→8	6		6	
36	[1→(8)→3→(8)→(6)→5→8]	1→8	7		6	
37	[1→(8)→3→(8)→(6)→7→8]	1→8	8		6	
38	[1→(8)→3→(8)→7→8]	1→8	9		5	
39	[1→(8)→4→8]	1→8	10		3	
40	[1→(8)→5→8]	1→8	11		3	
41	[1→(8)→(6)→2→8]	1→8	12		4	
42	[1→(8)→(6)→3→(8)→2→8]	1→8	13		6	
43	[1→(8)→(6)→3→(8)→4→8]	1→8	14		6	
44	[1→(8)→(6)→3→(8)→5→8]	1→8	15		6	
45	[1→(8)→(6)→3→(8)→(6)→2→8]	1→8	16		7	
46	[1→(8)→(6)→3→(8)→(6)→4→8]	1→8	17		7	
47	[1→(8)→(6)→3→(8)→(6)→5→8]	1→8	18		7	
48	[1→(8)→(6)→3→(8)→(6)→7→8]	1→8	19		7	
49	[1→(8)→(6)→3→(8)→7→8]	1→8	20		6	
50	[1→(8)→(6)→4→8]	1→8	21		4	
51	[1→(8)→(6)→5→8]	1→8	22		4	
52	[1→(8)→(6)→7→8]	1→8	23		4	
53	[1→(8)→7→8]	1→8	24	24	3	5.167

Figure 9 and figure 10 is histogram of simulation result. Horizontal line means number of routing. And vertical line means number of appareling the number of routing. Figure 9 contains 4 nodes for 8 nodes. And figure 10 contains 9 nodes for 11 nodes. 9d for 11d means difference node spreading pattern. The additional nodes set new room without other nodes. So the new nodes can communicate only use PLC. Then the simulator try a lot of pattern to check can communicating nodes. Figure 9 shows less changes because the nodes spread pattern is similar each other. So the result is similar too. However, figure 10 shows big changes. In these cases, node

spreads widely than the case of 4 nodes for 8 nodes. So the histogram shows big differences. This result means the lay out of nodes is an important thing to make mutually complementary network using wired and wireless communications.

5 Summary

For the study described herein, a mutually complementary network of wired and wireless communications was applied to a building such as a school; a simulator was developed to determine how effective a routing capability would be. Then its results were evaluated. Assuming that we can have only 10 s available for communication knowing it based on the number of nodes considered, results show that up to nine nodes for a three-story building are just about the limit available for a practical use with a location-undefined communication method used. If 15 s or less are assumed to be available, then it is up to 11 nodes for five-story building. Based on this evaluation to the case to apply the mutually complementary network for a school building, for example, we find it possible by allocating a domain unit of 11 nodes to the entire building.

Future studies to be made should first use a building with as few stories as possible and have a greater number of nodes under the associated communication duration time conditions given. Second, they should include that we have a network capability in which we have a domain unit of multiple nodes to allocate to the entire building. With a simulator used for the simulation to the setting of 11 nodes for a five-story building, the number of routing combinations that existed with our routing capability method exceeded 400 million, and the simulation duration conducted exceeded 70 hr in all. Even under those circumstances, the maximum number of routings performed was 29 times for up to 14.5 s spent, which indicates that the routing method introduced herein can be available for practical use. It would be necessary in any case to reduce the simulation hours greatly.

Regarding mutually complementary networks incorporating wired and wireless technologies, it is important to have the right communication duration time and communication performance to make it available for a practical use, and to show a typical example of its use as well, thereby contributing to the security and the energy control for such a wide area that covers private homes and other buildings for which no proper communication method is available [7].

References

- [1] Yamada, K., et al.: Dual Communication System Using Wired and Wireless Correspondence in Small Space. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 898–904. Springer, Heidelberg (2004)
- [2] M16C/6S, <http://www.renesas.com>
- [3] <http://www.zigbee.org>
- [4] Watabe, D., Yamada, K.: Route simulation in mutually complementary network between wired and wireless. In: Information Processing Society of Japan, pp. 3313-3314 (2010)

- [5] Yamada, K., Furumura, T., Naoe, Y., Kitazawa, K., Shimizu, T., Yoshida, K., Kojima, M., Mizuno, T.: Home-Network of a Mutually Complementary Communication System by Wired and Wireless. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 189–196. Springer, Heidelberg (2006)
- [6] Incorporating the Best of Both Worlds for Improved Functionality Industrial Environments, Hybrid Wired/Wireless Network for Real-Time Communications. IEEE Industrial Electronics Magazine 2(1), 8–20 (March 2008)
- [7] Yamada, K.: The home network by cooperation communication technology - mutually complementary network incorporating. In: Intellectual property strategy headquarters. Industry-university cooperation technolog Tokai University, Collection of the research starting points, p. 29. Tokai University

On the Impact of the Metrics Choice in SOM Learning: Some Empirical Results from Financial Data

Marina Resta

DIEM, sez. Matematica Finanziaria, University of Genova,
via Vivaldi 2, 16126 Genova, Italy

mresta@unige.it

http://www.diem.unige.it/Marina_RESTA.html

Abstract. This paper studies the impact of the metrics choice on the learning procedure of Self Organizing Maps (SOM). In particular, we modified the learning procedure of SOM, by replacing the standard Euclidean norm, usually employed to evaluate the similarity between input patterns and nodes of the map, with the more general Minkowski norms:

$$\|X\|_p = \left(\sum_i |X_i|^p \right)^{\frac{1}{p}}, \text{ for } p \in \mathbb{R}_+.$$

We have then analyzed how the clustering capabilities of SOM are modified when both prenorms ($0 < p < 1$), and ultrametrics ($p \gg 1$) are considered. This was done using financial data on the Foreign Exchange Market (FOREX), observed at different time scales (from 1 minute to 1 month). The motivation inside the use of this data domain (financial data) is the relevance of the addressed question, since SOM are often employed to support the decision process of traders. It could be then of interest to know if and how the results of SOM can be driven by changes in the distance metric according to which proximities are evaluated. Our main result is that concentration seems not to be the unique factor affecting the effectiveness of the norms (and hence of the clustering procedure); in the case of financial data, the time scale of observations counts as well.

Keywords: Self-Organizing Maps, Metrics choice, Financial data.

1 Introduction

In the past decade various contributions focused, mostly from the theoretical point of view, on the observation that standard metrics seem to be inappropriate, when dealing with data embedded into high-dimensional spaces [4], [6]. In particular, [6] has shown that for random vectors with independent and identically distributed components (i.i.d), the mean of their Euclidean norm increases as the square root of the dimension of the space, while the variance maintains stable.

In practice, this means that in multivariate datasets all pairwise distances could seem equal or more similar than effectively they are: this might lead to

regrettable inconvenients, especially in cases where the distance among various patterns is the fundament for more complex content retrieval tasks.

As a consequence, the relevance of the Euclidean norm has been questioned, when this is used to measure the similarity on high-dimensional datasets [1], [2]; obviously, the adequateness of all soft computing techniques based on it could be misdoubted in turn. Towards such direction, [4] proved that in a nearest neighbour search framework, all points tend to converge to approximately the same distance from the query point.

Being aware of such curse of dimensionality, [3] focused on the analysis of the concentration in the alternatives to the standard Euclidean norm, and stressed the attention on the family of generalised Minkowsky norms:

$$\|X\|_p = \left(\sum_i |X_i|^p \right)^{\frac{1}{p}}. \quad (1)$$

Where p is a strictly positive real value (fractional norms). Using the family defined by [1], [2] observed that nearest neighbour search is meaningless in high-dimensional spaces for integer p values equal or greater that 2 (the so called *ultrametrics*). In [3] such results were generalized to cases in which p is not restricted to assume positive integer values, but it also can take real values. In addition, [8] evidenced that the optimal distance to be used can also depend on the type of noise on the data: fractional norms should be preferable in the case of colored noise, while in the case of Gaussian noise, the Euclidean metrics should be more robust than fractional ones. This led [9] to suggest the use in kernel methods of the so called p -Gaussian kernels:

$$K(x, y) = \exp(-d(x, y)^p / \sigma^p). \quad (2)$$

Where p and σ are two parameters allowing to adjust both the slope of the function, and the shift to larger distances.

More recently, [14] gave also proof that, in contrast to what expected, prenorms ($0 < p < 1$) are not always less concentrated than higher order norms, and suggests to use a kind of rule of thumb to find the best metric which fits better to the requirements of the data domain.

This paper starts from that point to observe the impact of the choice of metrics in the case of Kohonen's Self Organizing Maps (SOM), and examines it experimentally on financial data observed at different time scales (from 1 minute to 1 month). We are interested to test if the performances of SOM may take advantage from changes in the adopted similarity measures: to the best of our knowledge this aspect was only examined in [12], in the case of 5 minutes data of the S&PMIB index. The motivation inside the use of this data domain (financial data) is the relevance of the addressed question, since SOM are often employed to support the decision process of traders. It could be then of interest to know if and how the results of SOM can be driven by changes in the distance metric according to which proximities are evaluated.

With respect to previous contributions on the concentration of norms (and in particular to [13]), our work introduces at least two novel elements: (i) we focus

on the algorithm of Kohonen, evaluating the impact of the choice of different distance metrics on its clustering capabilities; (ii) we run our experiments on financial data observed at very different time scales, in order to verify whether in addition to high dimensionality, the time scale can influence the choice of optimal distance metric or it does not count.

This paper is organized as follows: Section 2 briefly illustrates the changes we have made to Kohonen’s algorithm, in order to use it with different metrics derived from (1), Section 3 illustrates the details of our simulations, while Section 4 discusses the results obtained. Finally, Section 5 concludes.

2 Modified Metrics in the Kohonen’s Algorithm

The Self Organizing Map (SOM) (10) is a projection method based on the principle of space representation through dimension reduction: a finite set of input patterns is represented by means of a smaller number of nodes (neurons), sharing with inputs the same format, and arranged into a mono or bi-dimensional grid; in order to avoid hedges effects, wraparound versions can be also implemented (11), (15).

A generic step of the procedure may be then summarized as follows: we will refer to the case of a mono-dimensional SOM, but the layout presented can be easily generalized to higher dimensional grids.

If $\mathbf{x}(t) = \{x_j(t)\}_{j=1,\dots,n} \in \mathbb{R}^n$ is the input item presented to a map M with q nodes with weights $\mathbf{m}_i(t) = \{m_{i,j}(t)\}_{j=1,\dots,n} \in \mathbb{R}^n$, ($i = 1, \dots, q$), i_t^* will be claimed the winner neuron at step t iff:

$$i_t^* = \operatorname{argmin}_{i \in M} \left(\sum_{i \in M} \sum_{j=1}^n |x_j(t) - m_{i,j}(t)|^p \right)^{1/p}, \quad p \in \mathbb{N}. \tag{3}$$

Where p is the distance parameter. More common choices for p include $p = 1$ (Manhattan or city block distance), and $p = 2$ (Euclidean distance).

Once the leader has been identified according to (3), the correction of nodes in the map takes place; if $N_{i^*}(t)$ is the set of neurons in the map belonging to the neighbourhood of i^* (in a topological sense), then:

$$\mathbf{m}_i(t + 1) = \mathbf{m}_i(t) + h_{i^*,i}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)]. \tag{4}$$

Where $h_{i^*,i}(\cdot)$ is an interaction function, governing the way the nodes adjust in relation to the winning neuron on the grid.

After iterating such procedure over a number of epochs, the map should tend to a steady organized state, and neighbouring neurons should represent similar inputs. The degree of organization reached by the map can be checked by means of convergence indexes, such as those described in (7) or in (5); in this way, the learning procedure is arrested once a proper convergence threshold level is reached.

To the purpose of this paper it is relevant the way the Kohonen’s algorithm couples each input pattern to the more similar unit in the neural space.

In accordance to the studies presented in [3] and [14], in this work we have considered various extensions of the standard Minkowski metrics appearing as argument in (3), and we trained SOM accordingly. To such purpose, we have modified (3), replacing $p \in \mathbb{N}$ by $p \in \mathbb{R}^+$, to include both ultrametrics ($p \gg 1$), and prenorms ($0 < p < 1$).

3 Simulation

3.1 Description of Data

We run our simulation using FOREX data. FOREX is the acronym for Foreign Exchange Market. The relevance of this market is manifold:

- FOREX is the largest, most liquid and most transparent financial market in the world. Daily average turnover has now exceeded 2 trillion USD.
- The market operates 24 hours a day, each day of the week: this means that one could trade everytime he wants.
- It is possible to exploit leverage effects, depending on the broker one decides to operate with.
- Tradings are not affected by transaction fees.
- The market provides facilities to potential traders that are incited to test with real data (but with virtual money) the success of their trading strategies.

The above mentioned features make FOREX very appealing to investors. In addition, FOREX makes possible to trade a variety of assets, including currencies, stock exchange indexes and commodities, towards derivatives. Table 1 provides an overview on the more relevant and typical assets negotiated on this market, and on related symbols used to refer to them.

The level of each traded asset is expressed in terms of points or *pip*, whose value varies according to the asset itself. For example, on a EUR/USD position, 1 pip=0.0001, therefore if EUR moved from 1.2760 to 1.2820, the difference between the two positions amounts to $1.2820 - 1.2760 = 0.0060$, and on a 100 000 EUR/USD position, 1 pip will worth 10 US Dollars.

Table 1. Main assets traded on FOREX

Symbol	Descr.	Symbol	Descr.
EUR	Euro/US Dollar	GBP	British Pound/US Dollar
CHF	US Dollar/ Swiss Franc	JPY	US Dollar/Japanese Yen
EURGBP	Euro / British Pound	EURCHF	Euro /Swiss Franc
EURJPY	Euro / Japanese Yen	GBPCHF	British Pound /Swiss Franc
GBPCHF	British Pound/ Japanese Yen	CHFJPY	Swiss Franc/ Japanese Yen
CAD	US Dollar/Canadian Dollar	AUDJPY	Australian Dollar/Japanese Yen
AUD	Australian Dollar/US Dollar	NZD	New Zealand Dollar/US Dollar
NZDJPY	New Zealand Dollar/Japanese Yen	XAU	Gold Spot/ US Dollar
XAG	Silver Spot/US Dollar		

Table 2. EUR data used in this study. ED is the abbreviation for Embedding Dimension. For intraday data, the columns Starting and Final $r_{\Delta t}(t)$ include also the first (last) block of minutes it was considered.

Δ_t	Descr.	Starting $r_{\Delta t}(t)$	Final $r_{\Delta t}(t)$	Length ED
Δ_1	One minute data	11/23/2009 00.1	04/09/2010 22.00	139 503 10
Δ_5	Five minutes data	02/16/2009 00.5	04/09/2010 22.00	85 018 10
Δ_{10}	Ten minutes data	07/06/2009 00.10	04/09/2010 22.00	28 231 10
Δ_{30}	Thirty minutes data	07/06/2009 00.30	04/09/2010 22.00	9 412 10
Δ_{60}	Sixty minutes data	02/16/2009 1.00	04/09/2010 22.00	7 086 14
Δ_{120}	Two-hours data	07/06/2009 2.00	04/09/2010 22.00	2 373 6
Δ_{240}	Four-hours data	01/03/2001 4.00	04/09/2010 22.00	9 646 5
Δ_{1d}	Daily data	01/03/2001	04/09/2010	1 594 5
Δ_{1w}	Weekly data	01/05/2001	04/09/2010	321 7
Δ_{1m}	Monthly data	01/31/2001	04/09/2010	75 8

Table 3. Simulations features. The SOM set associated to each input data matrix was trained for various values of p : $p = 0.2; 0.5; 1$, and hence: $p = 1 + \kappa \cdot 0.5, k = 1, \dots, 4$.

Δ_t	Input Matrix Dim.	SOM sets ID	Average SOM sets Dim.	Average Nr. of Epochs
Δ_1	$139\,494 \times 10$	SOM_{Δ_1}	27×22	10
Δ_5	$85\,009 \times 10$	SOM_{Δ_5}	26×18	8
Δ_{10}	$28\,222 \times 10$	$SOM_{\Delta_{10}}$	18×15	3
Δ_{30}	$9\,403 \times 10$	$SOM_{\Delta_{30}}$	14×11	3
Δ_{60}	$7\,073 \times 14$	$SOM_{\Delta_{60}}$	12×9	3
Δ_{120}	$2\,368 \times 6$	$SOM_{\Delta_{120}}$	11×9	2
Δ_{240}	$9\,642 \times 5$	$SOM_{\Delta_{240}}$	16×14	3
Δ_{1d}	$1\,590 \times 5$	$SOM_{\Delta_{1d}}$	10×9	2
Δ_{1w}	315×7	$SOM_{\Delta_{1w}}$	6×5	1
Δ_{1m}	68×8	$SOM_{\Delta_{1m}}$	5×3	1

Although we run experiments on all the assets described in Table 1, in this paper we will report and discuss the results obtained on EUR, i.e. on the positions on Euro vs US Dollar, that we have studied by examining historical data over different time scales. For each series of pips we have considered the log returns obtained as follows:

$$r_{\Delta t}(t) = \log \frac{L_{\Delta t}(t)}{L_{\Delta t}(t-1)} . \tag{5}$$

where $L_{\Delta t}(\cdot)$ is the position expressed in pip, referred to the proper time unit Δt . Table 2 shows main features of EUR data, over the different time scales.

3.2 SOM Training

Our simulation procedure may be summarized into a number of steps. For each dataset:

1. we estimated its Embedding Dimension (ED) i.e. the dimension of the space into which the trajectories of the observed variable is merged. In this way, we were able to assign a (fixed) length to patterns to be presented to SOM: the original array of data, at this point, was replaced by a matrix of dimension: $(Length_{tot} - ED + 1) \times ED$, where $Length_{tot}$ is the original length of the data (presented in Table 2, 5th column), and ED is as described in the above rows (see Table 2, 6th column).
2. For all the matrices given at point 1., we run the training on a group of 100 SOM for each value of p , as given in Table 3: the results we will discuss need then to be assumed as average performances. In order to guarantee simulations' replicability we have settled the random seed generator to: 123456.

4 Discussion of the Results

The steps described in the previous section led us to obtain for each dataset, an input space representation potentially varying in subordination to the metric in use, i.e. to the value of p : Figure 1 illustrates a typical outcome of our simulations, referred to the case of 1 minute data.

From Figure 1, it is clear that the choice of p affects the final representation of the input space. In the case of 1 minute EUR data, for instance, in contrast to common knowledge, it seems that higher p values produce more *clusterized* maps.

We have also examined the Quantization Error (QE) of each SOM set varying p , as it can be seen from Figure 2 and Table 4. Note that due to changes in the way SOM couple the best matching unit to input patterns, we have had to modify the procedure that computes QE accordingly. In this way, we have kept QE results conformal with respect to the way distance is evaluated.

It can be noticed that for some of the examined input datasets the dynamics of QE seems to be monotonic (not decreasing or not increasing): in this case, the choice of p appear to be relatively simple. When not monotonic, however the behaviour of QE exhibits minima, hence making the choice of final p value, if possible, even easier than in the previous case.

Table 4. Quantization error for different values of p . Values need to be multiplied for 10^{-3}

Δ_t	p values										
	p=0.25	p=0.5	p=1	p=1.5	p=2	p=2.5	p=3	p=3.5	p=4	p=4.5	p=5
Δ_1	0.233	0.245	0.088	0.112	0.131	0.175	0.223	0.177	0.276	0.3	0.318
Δ_5	0.496	0.353	0.127	0.113	0.162	0.175	0.206	0.224	0.264	0.271	0.276
Δ_{10}	0.442	0.354	0.103	0.081	0.113	0.141	0.179	0.16	0.198	0.23	0.249
Δ_{30}	0.349	0.193	0.091	0.064	0.078	0.097	0.131	0.122	0.157	0.176	0.192
Δ_{60}	0.366	0.18	0.12	0.048	0.048	0.004	0.004	0.004	0.004	0.004	0.004
Δ_{120}	0.434	0.18	0.114	0.06	0.106	0.083	0.111	0.089	0.117	0.163	0.206
Δ_{240}	0.264	0.154	0.094	0.084	0.094	0.002	0.002	0.002	0.002	0.002	0.002
Δ_{1d}	0.27	0.097	0.06	0.047	0.007	0.007	0.006	0.006	0.006	0.006	0.006
Δ_{1w}	0.35	0.175	0.025	0.05	0.024	0.022	0.022	0.022	0.021	0.02	0.025
Δ_{1m}	0.1	0.055	0.05	0.043	0.038	0.035	0.033	0.033	0.031	0.03	0.03

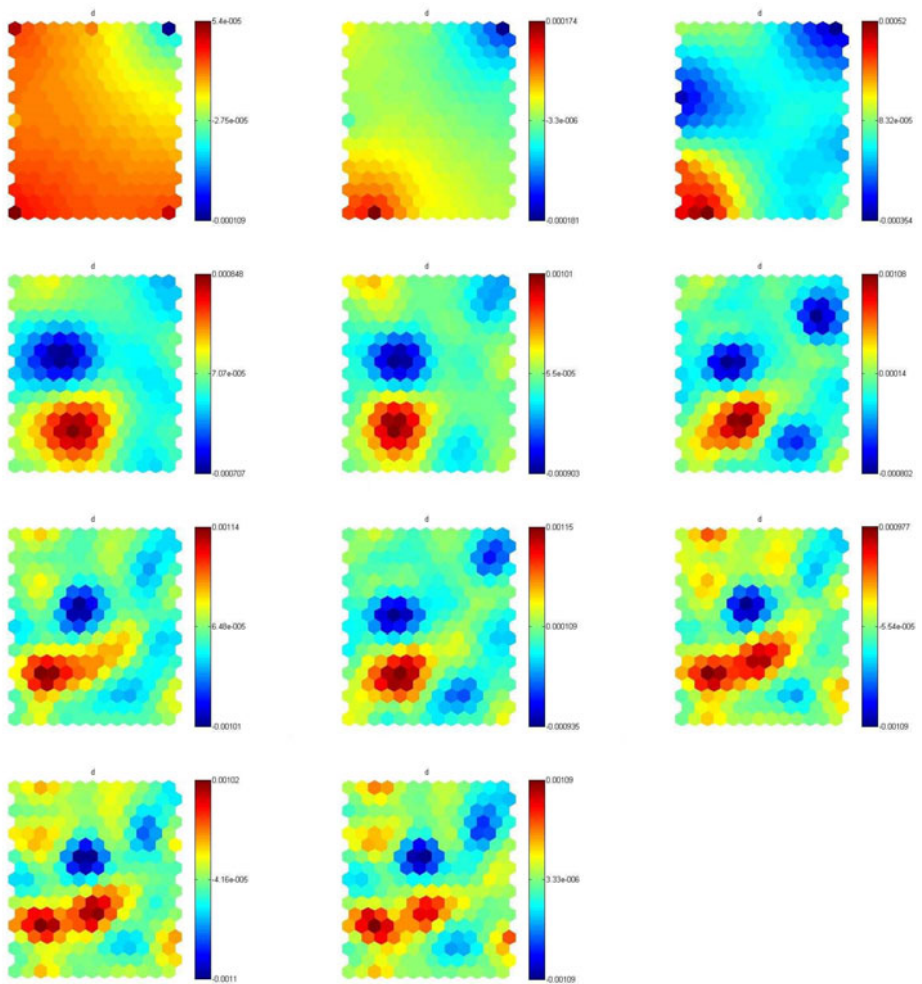


Fig. 1. Evolution of SOM for Δ_1 dataset for different p values. Values of p need to be read in ascending order from left to right.

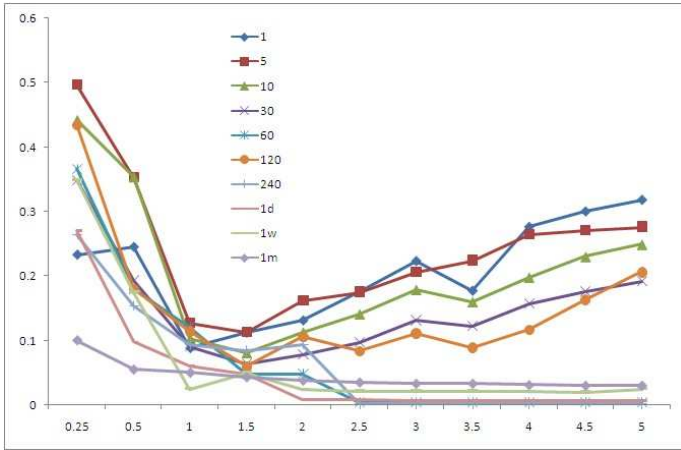


Fig. 2. Quantization Error (QE) for SOM sets working on input data. On x-axis we reported the values of p , while on y-axis we put QE values.

5 Conclusion

In this paper we have empirically tested whether the choice of metrics can affect the overall performance of Kohonen’s Self Organizing Maps (SOM), or not. To do that, we have considered data from the FOREX market at very different time scales. We have then run a modified version of Kohonen’s algorithm, where the distance between input and nodes in the map was evaluated with a general-

ized version of Minkowski metrics: $\|X\|_p = \left(\sum_i |X_i|^p \right)^{\frac{1}{p}}$, $p \in \mathbb{R}$, allowing the

computation of both prenorms ($0 < p < 1$), and ultrametrics ($p \gg 1$). The results we have obtained confirm that the same data can have very different SOM representation, depending on the chosen value for p . We also found out that there is not any universal p value which can guarantee to minimize the quantization error independently of the input dataset in use, i.e. of the time scale under observation; however, inside each time frame, it should be possible to associate to each time scale a p value which is definitely better than others. In addition, we have observed that, in some cases, p behaves monotonically, i.e. the higher/lower p is, the most the quantization error maintains at lower levels. This, in our opinion, is quite relevant for every task involving the use of SOM to support trading decisions, because the procedure could be driven to fit better to the data, if the best suitable p is chosen in accordance with the time scale (1 minute rather than 10 minutes or 1 day, and so on) at which traders set their positions in the market. Such remarks could give room to further investigations, in order to verify, for instance, if either the features discussed in previous rows are typical of financial data at different time frequencies, independently of the market the corresponding assets are traded into, or rather they are a typical of

FOREX, descending as consequences of the peculiarity of the market itself (for instance: its extreme liquidity, or the 24 hours operativity).

References

1. Aggarwal, C.C., Yu, P.S.: The IGrid Index: Reversing the Dimensionality Curse For Similarity Indexing in High Dimensional Space. In: Proc. of KDD, pp. 119–129 (2000)
2. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: What Is the Nearest Neighbor in High Dimensional Spaces? In: Abbadì, A., Brodie, M.L., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G., Whang, K.-Y. (eds.) Proc. of VLDB 2000, 26th Intl. Conference on Very Large Data Bases, Cairo, Egypt, September 10-14, pp. 506–515. Morgan Kaufmann, San Francisco (2000)
3. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, p. 420. Springer, Heidelberg (2000)
4. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When Is Nearest Neighbor Meaningful. In: Beerì, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 217–235. Springer, Heidelberg (1998)
5. Cattaneo Adorno, M., Resta, M.: Reliability and convergence on Kohonen maps: an empirical study. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3213, pp. 426–433. Springer, Heidelberg (2004)
6. Demartines, P.: Analyse de Donnes par Rseaux de Neurones Auto-Organiss. PhD dissertation, Institut Nat'l Polytechnique de Grenoble, Grenoble, France (1994)
7. De Bodt, E., Cottrell, M., Verleysen, M.: Statistical tools to assess the reliability of Self-Organizing Maps. *Neural Networks* 15, 967–978 (2002)
8. Francois, D., Wertz, V., Verleysen, M.: Non-euclidean metrics for similarity search in noisy datasets. In: Proc. of ESANN 2005, European Symposium on Artificial Neural Networks (2005)
9. Francois, D., Wertz, V., Verleysen, M.: On the locality of kernels in high-dimensional spaces. In: Proc. of ASMDA 2005, Applied Stochastic Models and Data Analysis, Brest, France (2005)
10. Kohonen, T.: *Self-Organizing Maps*. Springer, Berlin (1982)
11. Liou, C.Y., Tai, W.P.: Conformal self-organization for continuity on a feature map. *Neural Networks* 12, 893–905 (1999)
12. Resta, M.: Seize the (intra)day: Features selection and rules extraction for tradings on high-frequency data. *Neurocomputing* 72(16-18), 3413–3427 (2009)
13. Verleysen, M., Francois, D.: The Curse of Dimensionality in Data Mining and Time Series Prediction. In: Cabestany, J., Prieto, A.G., Sandoval, F. (eds.) IWANN 2005. LNCS, vol. 3512, pp. 758–770. Springer, Heidelberg (2005)
14. Verleysen, M., Francois, D.: The Concentration of Fractional Distances. *IEEE Trans. on Knowledge and Data Engineering* 19(7), 873–886 (2007)
15. Wu, Y., Takatsuka, M.: Spherical Self-Organizing Map using efficient indexed geodesic data structure. *Neural Networks* 19(6-7), 900–910 (2006)

Reinforcement Learning Scheme for Grouping and Characterization of Multi-agent Network

Koichiro Morihiro¹, Nobuyuki Matsui²,
Teijiro Isokawa², and Haruhiko Nishimura³

¹ Hyogo University of Teacher Education, Hyogo 673-1494, Japan
mori@hyogo-u.ac.jp

² Graduate School of Engineering, University of Hyogo, Hyogo 671-2201, Japan
matsui@eng.u-hyogo.ac.jp, isokawa@eng.u-hyogo.ac.jp

³ Graduate School of Applied Informatics, University of Hyogo,
Hyogo 650-0044, Japan
haru@ai.u-hyogo.ac.jp

Abstract. Several models have been proposed for describing grouping behavior such as bird flocking, terrestrial animal herding, and fish schooling. In these models, a fixed rule has been imposed on each individual a priori for its interactions in a reductive and rigid manner. We have proposed a new framework for self-organized grouping of agents by reinforcement learning. It is important to introduce a learning scheme for developing collective behavior in artificial autonomous distributed systems. This scheme can be expanded to cases in which predators are present. We integrated grouping and anti-predator behaviors into our proposed scheme. The behavior of agents is demonstrated and evaluated in detail through computer simulations, and their grouping and anti-predator behaviors developed as a result of learning are shown to be diverse and robust by changing some parameters of the scheme. In this study, we investigate the network structure of agents in the process of learning these behaviors. From the view point of the complex network, the average shortest path length and clustering coefficient are evaluated through computer simulations.

1 Introduction

The collective behavior of creatures can often be observed in nature. Bird flocking, terrestrial animal herding, and fish schooling are the typical well-known cases. Several observations suggest that there are no leaders in such groups who control the behavior of the group. Collective behavior develops from the local interactions among agents in groups [1-3]. Several models have been proposed to describe grouping behavior. In these models, a fixed rule has been imposed on each agent a priori for its interactions [4-11]. This reductive and rigid approach is suitable for modeling groups of biological organisms since they appear to inherit the ability to form groups. However, it is important to introduce a learning scheme that develops collective behavior in artificial autonomous distributed systems.

The characteristic feature of reinforcement learning [12, 13] is unsupervised learning by trial and error, i.e., by exploration, in order to maximize rewards obtained from the environment. Introducing appropriate relations between an agent's behavior (action) and its reward, we could design a new scheme for the development of grouping behavior by reinforcement learning. We have proposed a new framework for self-organized grouping of agents by reinforcement learning [14]. This scheme can be expanded to cases in which predators are present [15]. We integrated grouping and anti-predator behaviors into our proposed scheme [16]. The behavior of agents is demonstrated and evaluated in detail through computer simulations, and their grouping and anti-predator behaviors developed as a result of learning are shown to be diverse and robust by changing some parameters of the scheme.

In the current framework, agents interact as a result of reinforcement learning and produce grouping and anti-predator behaviors. Interaction as part of learning can also be conceived of as a network for information exchange independently constructed by agents. In order to assess and analyze the structure of such a network with reinforcement learning agents as nodes and to compare its characteristics with those of a complex network [17-19] as would be found in the real world, this study examined the relationship between network characteristics and the process of achieving grouping and anti-predator behaviors.

2 Reinforcement Learning

Reinforcement learning originated from the experimental studies on learning in the field of psychology. Almost all reinforcement learning algorithms are based on the estimation of value functions. Computer systems receive only an evaluative scalar feedback for a value function from their environment and not an instructive feedback as in supervised learning. Q-learning [20] is known as the best-understood reinforcement learning technique. A value function in Q-learning consists of values determined from a state and an action, which is called Q-value. In Q-learning, the learning process consists of acquiring a state (s_t), deciding an action (a_t), receiving a reward (r) from an environment, and updating the Q-value ($Q(s_t, a_t)$). The Q-value is updated by the following equation:

$$Q(s_{t+1}, a_{t+1}) = Q(s_t, a_t) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s', a') - Q(s_t, a_t)] \quad , \quad (1)$$

where A denotes the set of actions; α , the learning rate ($0 < \alpha \leq 1$); and γ , the discount rate ($0 < \gamma \leq 1$). Q-learning is one of the reinforcement learning techniques used for maximizing the sum of the rewards received. It attempts to learn the optimal policy by compiling a table of Q-values $Q(s, a)$ according to Eq. (1). $Q(s, a)$ provides the estimated value of the expected response action a for state s . Once these Q-values are learned, the optimal action for a state is the action with the highest Q-value. In the original Q-learning algorithm, a greedy policy with pure exploitation has been adopted. However, it is generally difficult to obtain satisfactory results by employing this policy. Therefore, in the present study, a policy that allows the adoption of a nonoptimal action is introduced.

In reinforcement learning, many types of exploration policies have been proposed for learning by trial and error, such as ϵ -greedy, softmax, and weighted roulette action selection. In the present study, we adopt the softmax action selection method, and the rule is given as $p(a|s) = \frac{\exp\{Q(s,a)/T\}}{\sum_{a_i \in A} \exp\{Q(s,a_i)/T\}}$, where T is a positive parameter called temperature. High temperatures cause all the actions to be (nearly) equiprobable.

3 Model and Method

3.1 Internal Perceptual Space of Each Agent

We employ a configuration where N agents that can move to any direction are placed in a two-dimensional field. The agents act in discrete time, and at each time-step an agent (agent i) finds other agent (agent j) among $N - 1$ agents. In the perceptual internal space, the state s_t of $Q(s_t, a_t)$ for the agent i is defined as $[R]$ by Gauss' notation, the maximum integer not surpassing the Euclidean distance from agent i to agent j , R . For the action a_t of $Q(s_t, a_t)$, four kinds of action patterns (a_1, a_2, a_3, a_4) are taken as follows, illustrated in Fig 1.

- a_1 : Attraction to agent j
- a_2 : Parallel positive orientation to agent j ($\mathbf{m}_a \cdot (\mathbf{m}_i + \mathbf{m}_j) \geq 0$)
- a_3 : Parallel negative orientation to agent j ($\mathbf{m}_a \cdot (\mathbf{m}_i + \mathbf{m}_j) < 0$)
- a_4 : Repulsion to agent j

Here, \mathbf{m}_a is the directional vector of a_t , and \mathbf{m}_i and \mathbf{m}_j are the velocity vectors of agents i and j , respectively. Agent i moves in accordance with \mathbf{m}_i in each time step, and \mathbf{m}_i is updated by the expression

$$\mathbf{m}_i \leftarrow (1 - \kappa)\mathbf{m}_i + \kappa\mathbf{m}_a, \tag{2}$$

where κ is a positive parameter ($0 \leq \kappa \leq 1$) called the inertia parameter.

In this study, as we consider same types of agents and the perceived object as a predator, two types of corresponding Q-values should be introduced.

3.2 Learning Modes against Agents of the Same Type and against Predators

In our proposed model, we offer the reward r for $Q(s_t, a_t)$ to each agent according to the distance R from the perceived agent of the same type. The learning of the agents proceeds according to the positive or negative reward, as shown in Table 1, in which $R_1 < R_2 < R_3$. In the case of $0 < [R] \leq R_3$, agent i can perceive another agent of the same type with a probability in proportion to $R^{-\beta}$, where β is a positive parameter. This implies that the smaller the value of R is, the easier the selection of the agent at the position is. When $0 < [R] \leq R_1$,

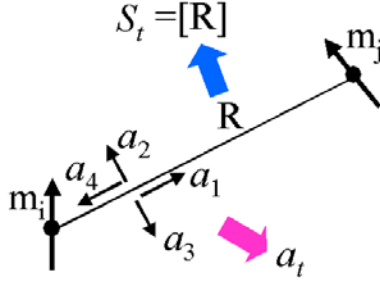


Fig. 1. Constitution of internal perceptual space of each agent

the agent receives a positive reward (+1) if it assumes a repulsive action against the perceived agent (a_4); otherwise, it receives the penalty (-1). In the case of $R_1 < [R] \leq R_2$ and $R_2 < [R] \leq R_3$, the agent also receives the reward or penalty defined in Table 1 depending on the actions. In the case of $[R] > R_3$, agent i cannot perceive agent j ; hence, it receives no reward and chooses an action from the four action patterns (a_1, a_2, a_3 , and a_4) randomly.

When there is a predator within R_3 , agent i perceives the predator with a probability of 1, and the learning mode against agents of the same type is switched to the learning mode against predator. In this case, agent i gets the positive reward (+1) if it takes a repulsive action to evade the predator (a_4); otherwise, it gets the penalty (-1), as defined in Table 1.

Table 1. Reward r for selected action a_t in state $s_t = [R]$

	Learning mode against agents of the same type				Learning mode against predators	
	$0 < [R] \leq R_1$	$R_1 < [R] \leq R_2$	$R_2 < [R] \leq R_3$	$R_3 < [R]$	$0 < [R] \leq R_3$	$R_3 < [R]$
a_t	$a_4; a_1, a_2, a_3$	$a_2; a_1, a_3, a_4$	$a_1; a_2, a_3, a_4$	a_1, a_2, a_3, a_4	$a_4; a_1, a_2, a_3$	a_1, a_2, a_3, a_4
r	1; -1	1; -1	1; -1	0	1; -1	0

4 Simulations and Results

In the computer simulations, we have assumed the following experimental conditions: $\alpha = 0.1$, $\gamma = 0.7$ in Eq. (1), $T = 0.5$ (under learning) for the softmax action selection method, $\kappa = 0.4$ in Eq. (2), $\beta = 0.5$ for the distance dependence of $R^{-\beta}$, $d_i = 1$, and $o_i = 0$. The initial velocities of the same type of agents are set to one body length (1 BL). The velocity $|\mathbf{m}_a|$ which is the directional vector of a_t is also set to one body length (1 BL). The velocity of the predator is set to two body lengths (2 BL). We have simulated our model for the number of agents $N = 100$ and $R_1 = 4$ (BL), $R_2 = 20$ (BL), and $R_3 = 50$ (BL).

4.1 Evaluation in No Predator Case and in the Case Predator Appears

In order to quantitatively evaluate how the agents develop grouping behavior, we introduce the measure $|\mathbf{M}|$ of the uniformity in direction and the measure E of the spread of agents.

$$|\mathbf{M}| = \frac{1}{N} \left| \sum_{i=1}^N \mathbf{m}_i \right|, \tag{3}$$

$$E = \frac{1}{N} \sum_{i=1}^N \sqrt{(x_i - x_G)^2 + (y_i - y_G)^2}, \tag{4}$$

where (x_i, y_i) and (x_G, y_G) are the two-dimensional coordinate of agent i and the barycentric coordinate among the agents, respectively. The value of $|\mathbf{M}|$ becomes closer to 1 when the directions of agents increase their correspondence. The agents come close when the value of E becomes small. In the evaluation, we take 100 events of simulation with various random series in exploration in both no predator case and the case predator appears.

Figure 2 shows the time step dependences of averaged $|\mathbf{M}|$ and E for no predator case. The transition of $\langle |\mathbf{M}| \rangle$ evolves good in every time step. The value of $\langle E \rangle$ takes a large value at the early stage of learning, after which it decreases to a value around 8 as learning proceeds.

In the case predator appears, the predator approaches the agents from behind and passes straight through the center of the group of agents. The predator appears in every 500th time step up to 5000 time steps. Figure 3 shows the average of non-splitting 94 events in 100 events. When the predator appears, the learning mode is changed. Hence, $\langle E \rangle$ takes a large value and $\langle |\mathbf{M}| \rangle$ decreases to around 0.2. This implies that the agents do not exhibit grouping behavior. When the predator disappears, the learning mode is reverted to the original mode. $\langle E \rangle$ takes a small value and $\langle |\mathbf{M}| \rangle$ increases again to around 0.9 because of the grouping behavior exhibited by the agents.

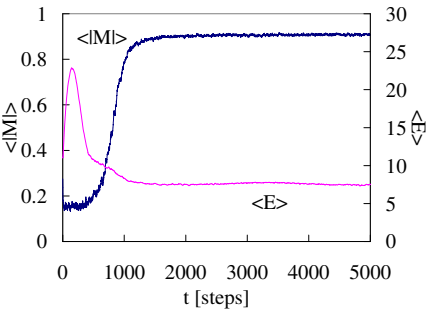


Fig. 2. Time step dependence of averaged $|\mathbf{M}|$ and E in 100 events for no predator case

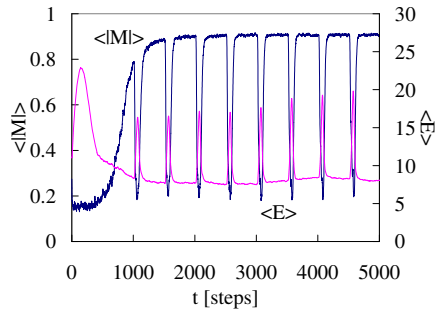


Fig. 3. Time step dependence of averaged $|\mathbf{M}|$ and E in non-splitting 94 events for the case predator appears

4.2 Trajectories of Agents and Predator

Figure 4 shows the trajectories of the agents in 1700 steps against the predator after learning. In this case, each agent uses fixed Q-value at $t=5000$ under learning and by setting $T \rightarrow 0$ in the softmax action selection method as the greedy behavioral policy. Through the learning stages, they have learned grouping and anti-predator behaviors. The magnification of 100 steps in Fig. 4 is shown in Fig. 5. On spotting the predator, the agents form a shape resembling a (polarized) fountain to escape from it. This suggests that the adaptive behaviors of agents, including escaping from the predator, is developed as a result of the two learning modes. Many kinds of anti-predator strategy are observed and recorded from a field study on predator-prey interactions [3, 9]. In our simulation, such anti-predator behaviors of agents like vacuole and herd are also observed.

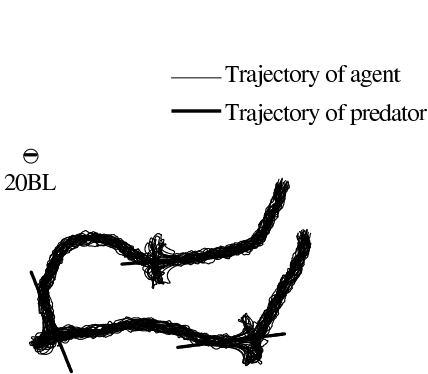


Fig. 4. Trajectories of agents in 1700 steps after learning

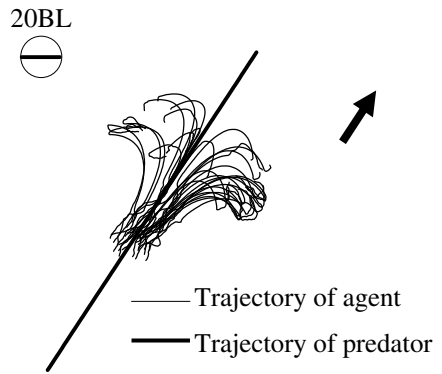


Fig. 5. Magnification of Fig. 4 near appearance of predator

5 Features of Behavioral Structure in Multi-agent Network

In the current framework, agents interact as a result of reinforcement learning and produce grouping and anti-predator behaviors. Interaction as part of learning can also be conceived of as a network for information exchange independently constructed by agents. In order to assess and analyze the structure of such a network with reinforcement learning agents as nodes and to compare its characteristics with those of a complex network [17–19] as would be found in the real world, this study examined the relationship between network characteristics and the process of achieving grouping and anti-predator behaviors. This paper focused on 2 characteristics, namely 1) the average shortest path length L , indicating the average distance between nodes(agents), i.e. the minimum number of edges meeting at 2 nodes(agents), and 2) the clustering coefficient C ,

indicating the rate of cluster formation in the network. L serves as a measure of communications effectiveness while C serves as a measure of fault tolerance in the network. Here, there are 100 nodes(agents), as indicated in Fig. 3

$$L = \frac{1}{N(N-1)} \sum_{i,j(i \neq j)} d_{ij}, \quad (5)$$

where a distance d_{ij} between nodes(agents) i and j in a network is the length of the shortest path between them through the multi-agent network. The clustering coefficient is given by the average of the local clustering coefficient C_i ,

$$\begin{aligned} C &= \frac{1}{N} \sum_i C_i, \\ &= \frac{1}{N} \sum_i \frac{2e_i}{k_i(k_i - 1)} \end{aligned} \quad (6)$$

when node(agent) i has k_i nearest neighbours with e_i connections between them.

5.1 Changes in Network Characteristics during Learning

In the current framework, agent status was determined based on the distance between a selected agent and other agents during learning. Here, agents that behaved commensurate with an appropriate level of grouping behavior ($R_1(= 4) < [R] \leq R_2(= 20)$) while learning were presumed to be connected by the network. Changes in the average shortest path length L and the clustering coefficient C in such instances are shown in Fig. 6. With the achievement of grouping behavior, the average shortest path length L decreased to about 1.2 and the clustering coefficient C exceeded 0.8 and remained somewhat constant. When a predator appeared, in accordance with avoidance behavior by agents the average shortest path length L increased to about 1.6 and the clustering coefficient C decreased to about 0.7. After the predator disappeared, the average shortest path length L again decreased and the clustering coefficient C increased. The theoretical values for both characteristics in a complex network (with small-world properties) were given by $L = \log(n)$ for the average shortest path length and $C = 0.6 \sim 0.7$ for the clustering coefficient. Comparing these values indicated that anti-predator behavior and subsequent group re-formation resulted in values closer to those of a complex network (with small-world properties) than did individual agents with consistent grouping behavior.

The location of agents with consistent grouping behavior prior to a predator appearing (pink circles) and with anti-predator behavior once a predator has appeared (grey circles) is shown in Fig. 7. All of the circles have a diameter of 20 (BL) with the individual agent positioned at the center. When each of the above behaviors was evident, agents represented by intersecting circles were connected

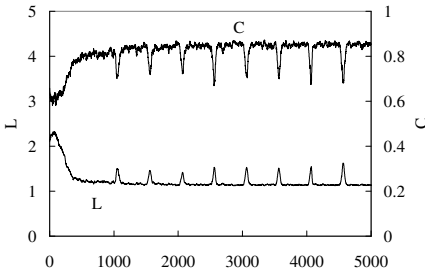


Fig. 6. Average shortest path length L and clustering coefficient C in the case of Fig. 3

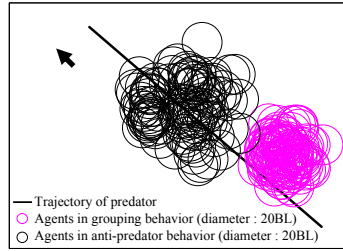


Fig. 7. Distribution of agents in grouping behavior (at $t=2457$) and anti-predator behavior (at $t=2557$)

as a network, as the figure indicates. With consistent grouping behavior, the distance between agents closed, and numerous network connections were evident. In addition, with anti-predator behavior the distance between agents grew and a decrease in network connections was evident.

5.2 Network Characteristics in After Learning

The average shortest path length L and the clustering coefficient C during grouping and anti-predator behaviors after learning ($Q(t = 5000)$) are shown in Fig. 8. Since effective grouping and anti-predator behaviors had already been learned, there were no gradual changes in those values immediately after the start. Moreover, there were no fluctuations due to exploration, so there were substantial changes in both characteristics in comparison to conditions during learning, which are shown in Fig. 6. Similar fluctuations are evident, but the values for the average shortest path length L and the clustering coefficient C were closer to those of a complex network during anti-predator behavior and subsequent group re-formation than during learning.

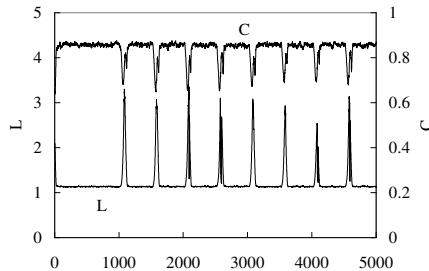


Fig. 8. Average shortest path length L and clustering coefficient C in after learning

6 Conclusion

We have demonstrated a scheme for forming autonomous groups of agents by reinforcement Q-learning. In addition to the grouping behavior of agents, the anti-predator behavior exhibited while escaping from predators can be developed by learning. This indicates the adaptive flexibility of our proposed scheme.

Moreover, this study examined the relationship between network characteristics for reinforcement learning agents, in the form of the average shortest path length and the clustering coefficient, and the process of achieving grouping and anti-predator behaviors in order to compare those characteristics with the characteristics of a complex network as is found in the real world.

Several topics for the future are the need for further study of the network structure for the various avoidance behaviors and the presence of heterogeneous agents. In addition, the relationship between reward provision and behavioral structure in multi-agent network should be studied.

References

1. Shaw, E.: Schooling Fishes. *American Scientist* 66, 166–175 (1978)
2. Partridge, B.L.: The structure and function of fish schools. *Scientific American* 246, 90–99 (1982)
3. Pitcher, T.J., Wyche, C.J.: Predator avoidance behaviour of sand-eel schools: why schools seldom split. In: Noakes, D.L.G., Lindquist, B.G., Helfman, G.S., Ward, J.A. (eds.) *Predators and Prey in Fishes*, pp. 193–204. Junk, The Hague (1983)
4. Aoki, I.: A Simulation Study on the Schooling Mechanism in Fish. *Bulletin of the Japanese Society of Scientific Fisheries* 48(8), 1081–1088 (1982)
5. Reynolds, C.W.: Flocks, Herds, and Schools: A Distributed Behavioral Model. *Computer Graphics* 21(4), 25–34 (1987)
6. Huth, A., Wissel, C.: The Simulation of the Movement of Fish Schools. *Journal of Theoretical Biology* 156, 365–385 (1992)
7. Niwa, H.-S.: Self-organizing dynamic model of fish schooling. *Journal of theoretical Biology* 171, 123–136 (1994)
8. Shimoyama, N., Sugawara, K., Mizuguchi, T., Hayakawa, Y., Sano, M.: Collective Motion in a System of Motile Elements. *Physical Review Letters* 76, 3870–3873 (1996)
9. Vabo, R., Nottestad, L.: An individual based model of fish school reactions: predicting antipredator behaviour as observed in nature. *Fisheries Oceanography* 6, 155–171 (1997)
10. Inada, Y., Kawachi, K.: Order and Flexibility in the Motion of Fish Schools. *Journal of theoretical Biology* 214, 371–387 (2002)
11. Oboshi, T., Kato, S., Mutoh, A., Itoh, H.: A Simulation Study on the Form of Fish Schooling for Escape from Predator. *Forma* 18, 119–131 (2003)
12. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research* 4, 237–285 (1996)
13. Sutton, R.S., Barto, A.G.: *Reinforcement Learning*. MIT Press, Cambridge (1982)
14. Morihiro, K., Isokawa, T., Nishimura, H., Tomimasu, M., Kamiura, N., Matsui, N.: Reinforcement Learning Scheme for Flocking Behavior Emergence. *Journal of Advanced Computational Intelligence and Intelligent Informatics(JACIII)* 11(2), 155–161 (2007)

15. Morihiro, K., Nishimura, H., Isokawa, T., Matsui, N.: Reinforcement Learning Scheme for Grouping and Anti-predator Behavior. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS (LNAI), vol. 4694, pp. 115–122. Springer, Heidelberg (2007)
16. Morihiro, K., Nishimura, H., Isokawa, T., Matsui, N.: Learning Grouping and Anti-predator Behaviors for Multi-agent Systems. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 426–433. Springer, Heidelberg (2008)
17. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–442 (1998)
18. Barabási, A.-L., Albert, R.: Emergence of Scaling in Random Networks. *Science* 286(5439), 509–512 (1999)
19. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: Structure and dynamics. *Physics Reports* 424, 175–308 (2006)
20. Watkins, C.J.C.H., Dayan, P.: Q-learning. *Machine Learning* 8, 279–292 (1992)

Extracting Principal Components from Pseudo-random Data by Using Random Matrix Theory

Mieko Tanaka-Yamawaki

Department of Information and Knowledge Engineering
Graduate School of Engineering
Tottori University, Tottori, 680-8552 Japan
Mieko@ike.tottori-u.ac.jp

Abstract. We develop a methodology to grasp temporal trend in a stock market that changes year to year, or sometimes within a year depending on numerous factors. For this purpose, we employ a new algorithm to extract significant principal components in a large dimensional space of stock time series. The key point of this method lies in the randomness and complexity of the stock time series. Here we extract significant principal components by picking a few distinctly large eigenvalues of cross correlation matrix of stock pairs in comparison to the known spectrum of corresponding random matrix derived in the random matrix theory (RMT). The criterion to separate signal from noise is the maximum value of the theoretical spectrum of W . We test the method using 1 hour data extracted from NYSE-TAQ database of tickwise stock prices, as well as daily close price and show that the result correctly reflect the actual trend of the market.

Keywords: Stock Market, Trend, Principal Component, RMT, Correlation, Eigenvalues.

1 Introduction

In a stock market, numerous stock prices move under a high level of randomness and some regularity. Some stocks exhibit strong correlation to other stocks. A strong correlation among eminent stocks should result in a visible global pattern. However, the networks of such correlation are unstable and the patterns are only temporal. In such a condition, a detailed description of the network may not be very useful, since the situation quickly changes and the past knowledge is no longer valid under the new environment. If, however, we have a methodology to extract, in a very short time, major components that characterize the motion of the market, it should give us a powerful tool to describe temporal characteristics of the market and help us to set up a time varying model to predict the future move of such market.

Recently, there have been wide interest on a possible candidate for such a methodology using the eigenvalue spectrum of the equal-time correlation matrix between pairs of price time series of different stocks, in comparison to the corresponding matrix computed by means of random time series [1-4]. Plerau, et. al. [1] applied this technique on the daily close prices of stocks in NYSE and S&P500.

We carry on the same line of study used in Ref. [1] for the intra-day price correlations on American stocks to extract principal components. In this process we clarify the process in an explicit manner to set up our algorithm of RMT_PCM to be applied on intra-day price correlations. Based on this approach, we show how we track the trend change based on the results from year by year analysis.

2 Cross Correlation of Price Time Series

It is of significant importance to extract sets of correlated stock prices from a huge complicated network of hundreds and thousands of stocks in a market. In addition to the correlation between stocks of the same business sectors, there are correlations or anti correlations between different business sectors.

For the sake of comparison between price time series of different magnitudes, we often use the profit instead of the prices [1-5]. The profit is defines as the ration of the increment ΔS , the difference between the price at t and $t + \Delta t$, divided by the stock price $S(t)$ itself at time t .

$$\frac{S(t + \Delta t) - S(t)}{S(t)} = \frac{\Delta S(t)}{S(t)} \tag{1}$$

This quantity does not depend on the unit, or the size, of the prices which enable us to deal with many time series of different magnitude. More convenient quantity, however, is the log-profit defined by the difference between log-prices.

$$r(t) = \log(S(t + \Delta t)) - \log(S(t)) \tag{2}$$

Since it can also be written as

$$r(t) = \log\left(\frac{S(t + \Delta t)}{S(t)}\right) \tag{3}$$

and the numerator in the log can be written as $S(t) + \Delta S(t)$,

$$r(t) = \log\left(1 + \frac{\Delta S(t)}{S(t)}\right) \cong \frac{\Delta S(t)}{S(t)} \tag{4}$$

It is essentially the same as the profit $r(t)$ defined on Eq. (1). The definition in Eq.(2) has an advantage.

The correlation $C_{i,j}$ between two stocks, i and j , can be written as the inner product of the two log-profit time series, $r_i(t)$ and $r_j(t)$,

$$C_{i,j} = \frac{1}{T} \sum_{t=1}^T r_i(t)r_j(t) \tag{5}$$

We normalize each time series in order to have the zero average and the unit variances as follows.

$$x_i(t) = \frac{r_i(t) - \langle r_i \rangle}{\sigma_i} \quad (i=1, \dots, N) \tag{6}$$

Here the suffix i indicates the time series on the i -th member of the total N stocks.

The correlations defined in Eq. (5) makes a square matrix whose elements are in general smaller than one.

$$|C_{i,j}| \leq 1 \quad (i=1, \dots, N; j=1, \dots, N) \tag{7}$$

and its diagonal elements are all equal to one due to normalization.

$$C_{i,i} = 1 \quad (i=1, \dots, N) \tag{8}$$

Moreover, it is symmetric

$$C_{ij} = C_{ji} \quad (i=1, \dots, N; j=1, \dots, N) \tag{9}$$

As is well known, a real symmetric matrix C can be diagonalized by a similarity transformation $V^{-1} C V$ by an orthogonal matrix V satisfying $V^t = V^{-1}$, each column of which consists of the eigenvectors of C .

$$v_k = \begin{pmatrix} v_{k,1} \\ v_{k,2} \\ \cdot \\ v_{k,N} \end{pmatrix} \tag{10}$$

Such that

$$C v_k = \lambda_k v_k \quad (k=1, \dots, N) \tag{11}$$

where the coefficient λ_k is the k -th eigenvalue.

Eq.(11) can also be written explicitly by using the components as follows.

$$\sum_{j=1}^N C_{i,j} v_{k,j} = \lambda_k v_{k,i} \tag{12}$$

The eigenvectors in Eq.(10) form an ortho-normal set. Namely, each eigenvector v_k is normalized to the unit length

$$v_k \cdot v_k = \sum_{n=1}^N (v_{k,n})^2 = 1 \tag{13}$$

and the vectors of different suffices k and l are orthogonal to each other.

$$v_k \cdot v_l = \sum_{n=1}^N v_{k,n} v_{l,n} = 0 \tag{14}$$

Equivalently, it can also be written as follows by using Kronecker's delta.

$$v_k \cdot v_l = \delta_{k,l} \tag{15}$$

The right hand side of Eq.(15) is zero(one) for $k \neq l(k = l)$. The numerical solution of the eigenvalue problem of a real symmetric matrix can easily be obtained by repeating Jacobi rotations until all the off-diagonal elements become close enough to zero.

3 RMT-Oriented Principal Component Method

The diagonalization process of the correlation matrix C by repeating the Jacobi rotation is equivalent to convert the set of the normalized set of time series in Eq. (6) into the set of eigenvectors.

$$y(t) = (V) \begin{pmatrix} x_1(t) \\ \cdot \\ \cdot \\ \cdot \\ x_N(t) \end{pmatrix} = Vx(t) \tag{16}$$

It can be written explicitly using the components as follows.

$$y_i(t) = \sum_{j=1}^N v_{i,j} x_j(t) \tag{17}$$

The eigenvalues can be interpreted as the variance of the new variable discovered by means of rotation toward components having large variances among N independent variables. Namely,

$$\begin{aligned} \sigma^2 &= \frac{1}{T} \sum_{t=1}^T (y_i(t))^2 \\ &= \frac{1}{T} \sum_{t=1}^T \sum_{l=1}^N v_{i,l} x_l(t) \sum_{m=1}^N v_{i,m} x_m(t) \\ &= \sum_{l=1}^N \sum_{m=1}^N v_{i,l} v_{i,m} C_{l,m} \\ &= \lambda_i \end{aligned} \tag{18}$$

Since the average $\langle y_i \rangle$ of y_i over t is always zero based on Eq.(6) and Eq.(17). For the sake of simplicity, we name the eigenvalues in descending order, $\lambda_1 > \lambda_2 > \dots > \lambda_N$.

The theoretical base underlying the principal component analysis is the expectation of distinguished magnitudes of the principal components compared to the other components in the N dimensional space. We illustrate in Fig.1 the case of 2 dimensional data (x,y) rotated to a new axis $z=ax+by$ and w perpendicular to z , in which z being the principal component and this set of data can be described as 1 dimensional information along this principal axis.

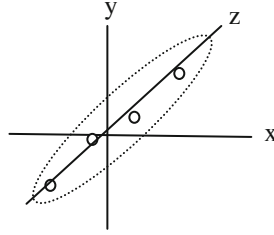


Fig. 1. A set of four 2-dimensional data points are characterized as a set of 1-dimensional data along z axis

If the magnitude of the largest eigenvalue λ_1 of C is significantly large compared to the second largest eigenvalue, then the data are scattered mainly along this principal axis, corresponding to the direction of the eigenvector v_1 of the largest eigenvalue. This is the first principal component. Likewise, the second principal component can be identified to the eigenvector v_2 of the second largest eigenvalue perpendicular to v_1 . Accordingly, the 3rd and the 4th principal components can be identified as long as the components toward these directions have significant magnitude. The question is how many principal components are to be identified out of N possible axes.

One criterion is to pick up the eigenvalues larger than 1. The reason behind this scenario is the conservation of trace, the sum of the diagonal elements of the matrix, under the similarity transformation. Due to Eq. (8), we obtain

$$\sum_{k=1}^N \lambda_k = N \tag{19}$$

which means there exists m such that $\lambda_k > 1$ for $k < k_m$, and $\lambda_k < 1$ for $k > k_m$. As was shown in Re1.[1], this criterion is too loose to use for the case of the stock market having $N > 400$. There are several hundred eigenvalues that are larger than 1, and many of the corresponding eigenvector components are literally random and do not carry useful information.

Another criterion is to rely on the accumulated contribution. It is recommended by some references to regard the top 80% of the accumulated contribution are to be regarded as the meaningful principal components. This criterion is too loose for the stock market of $N > 400$, for m easily exceeds a few hundred.

A new criterion proposed in Ref. [1-4] and examined recently in many real stock data is to compare the result to the formula derived in the random matrix theory [6].

According to the random matrix theory (RMT, hereafter), the eigenvalue distribution spectrum of C made of random time series is given by the following formula[7]

$$P_{RMT}(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \tag{20}$$

in the limit of $N \rightarrow \infty, T \rightarrow \infty, Q = T/N = const.$

where T is the length of the time series and N is the total number of independent time series (i.e. the number of stocks considered). This means that the eigenvalues of correlation matrix C between N normalized time series of length T distribute in the following range.

$$\lambda_- < \lambda < \lambda_+ \tag{21}$$

Following the formula Eq. (19), between the upper bound and the lower bound given by the following formula.

$$\lambda_{\pm} = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \tag{22}$$

The proposed criterion in our RMT_PCM is to use the components whose eigenvalues, or the variance, are larger than the upper bound λ_+ given by RMT.

$$\lambda \gg \lambda_+ \tag{23}$$

4 Cross Correlation of Intra-day Stock Prices

In this chapter we report the result of applying the method of RMT_PCM on intra-day stock prices. The data sets we used are the tick-wise trade data (NYSE-TAQ) for the years of 1994-2002. We used price data for each year to be one set. In this paper we mention our result on 1994, 1998 and 2002.

One problem in tickdata is the lack of regularity in the traded times. We have extracted N stocks out of all the tick prices of American stocks each year that have at least one transaction at every hour of the days between 10 am to 3 pm. This provides us a set of price data of N symbols of stocks with length T , for each year. For 1994, 1998 and 2002, the number of stock symbols N as 419, 490, and 569, respectively. The length of data T was 1512, six (per day) times 252, the number of working days of the stock market in the above three years.

The stock prices thus obtained becomes a rectangular matrix of $S_{i,k}$ where $i=1, \dots, N$ represents the stock symbol and $k=1, \dots, T$ represents the executed time of the stock.

The i -th row of this price matrix corresponds to the price time series of the i -th stock symbol, and the k -th column corresponds to the prices of N stocks at the time k .

We summarize the algorithm that we used for extracting significant principal components from 1 hour price matrix in Table 1.

Following the procedure described so far, we obtain the distribution of eigenvalues shown in Fig. 2 for the 1-hour stock prices for $N = 419$ and $T = 1512$ in 1994.

The histogram shows the eigenvalues (except the largest $\lambda_1 = 46.3$), $\lambda_2 = 5.3$, $\lambda_3 = 5.1$, $\lambda_4 = 3.9$, $\lambda_5 = 3.5$, $\lambda_6 = 3.4$, $\lambda_7 = 3.1$, $\lambda_8 = 2.9$, $\lambda_9 = 2.8$, $\lambda_{10} = 2.7$, $\lambda_{11} = 2.6$, $\lambda_{12} = 2.6$, $\lambda_{13} = 2.6$, $\lambda_{14} = 2.5$, $\lambda_{15} = 2.4$, $\lambda_{16} = 2.4$, $\lambda_{17} = 2.4$ and the bulk distribution of eigenvalues under the theoretical maximum, $\lambda_+ = 2.3$. These are compared with the RMT curve of Eq. (20) for $Q = 1512 / 419 = 3.6$.

Table 1. The algorithm to extract the significant principal components (RMT_PCM)

<p>Algorithm of RMT_PCM:</p> <ol style="list-style-type: none"> 1. Select N stock symbols for which the traded price exist for all $t=1, \dots, T$. (6 times a day, at every hour from 10 am to 3 pm, on every working day of the year). 2. Compute log-return $r(t)$ for all the stocks. Normalize the time series to have mean=0, variance=0, for each stock symbol, $i=1, \dots, N$. 3. Compute the cross correlation matrix C and obtain eigenvalues and eigenvectors 4. Select eigenvalues larger than λ_+ in Eq.(22), the upper limit of the RMT spectrum, Eq. (20).
--

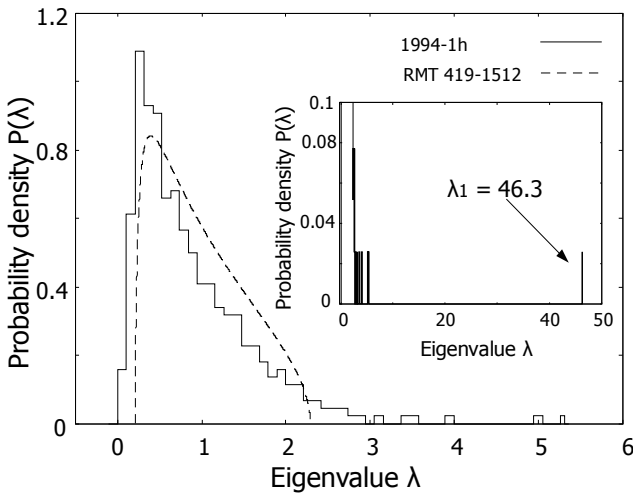


Fig. 2. Distribution of eigenvalues of correlation matrix of N=419 stocks for T=1512 data in 1994 compared to the corresponding RMT in Eq. (20) for Q= T/N =3.6

Corresponding result of 1998 data gives, for N=490, T=1512, there are 24 eigenvalues: $\lambda_1=81.12, \lambda_2=10.4, \lambda_3= 6.9, \lambda_4= 5.7, \lambda_5= 4.8, \lambda_6= 3.9, \lambda_7= 3.5, \lambda_8= 3.5, \lambda_9= 3.4, \lambda_{10}= 3.2, \lambda_{11}= 3.1, \lambda_{12}= 3.1, \lambda_{13}= 3.0, \lambda_{14}= 2.9, \lambda_{15}= 2.9, \lambda_{16}= 2.8, \lambda_{17}= 2.8, \lambda_{18}= 2.8, \lambda_{19}= 2.7, \lambda_{20}= 2.7, \lambda_{21}= 2.6, \lambda_{22}= 2.6, \lambda_{23}= 2.5, \lambda_{24}= 2.5$ and the bulk distribution of eigenvalues under the theoretical maximum, $\lambda_+ = 2.46$. These are compared with the RMT curve of Eq. (20) for $Q = 1512 / 490 = 3.09$.

Similarly, we obtain for 2002 data, for N=569, T=1512, there are 19 eigenvalues, $\lambda_1= 166.6, \lambda_2= 20.6, \lambda_3= 11.3, \lambda_4= 8.6, \lambda_5= 7.7, \lambda_6= 6.5, \lambda_7= 5.8, \lambda_8= 5.3, \lambda_9= 4.1, \lambda_{10}=4.0, \lambda_{11}= 3.8, \lambda_{12}= 3.5, \lambda_{13}= 3.4, \lambda_{14}= 3.3, \lambda_{15}= 3.0, \lambda_{16}= 3.0, \lambda_{17}= 2.9, \lambda_{18}= 2.8= 3.0, \lambda_{19}= 2.6,$ and the bulk distribution under the theoretical maximum, $\lambda_+ = 2.61$. These are compared with the RMT curve of Eq. (12) for $Q = 1512 / 569 = 2.66$.

However, a detailed analysis of the eigenvector components tells us that the random components are not necessarily reside below the upper limit of RMT, λ_+ , but percolates beyond the RMT limit if the sequence is not perfectly random. Thus it is more reasonable to assume that the border between the signal and the noise is somewhat larger than λ_+ . This interpretation also explains the fact that the eigenvalue spectra always spreads beyond λ_+ . It seems there is no more mathematical reason to decide the border between signal and noise. We return to data analysis in order to obtain further insight for extracting principal components of stock correlation.

5 Eigenvectors as the Principal Components

The eigenvector v_1 corresponding to the largest eigenvalue is the 1st principal component. For 1-hour data of 1994 where we have $N=419$ and $T=1512$, the major components of U_1 are giant companies such as GM, Chrysler, JP Morgan, Merrill Lynch, and DOW Chemical. The 2nd principal component v_2 consists of mining companies, while the 3rd principal component v_3 consists of semiconductor manufacturers, including Intel. The 4th principal component v_4 consists of computer and semiconductor manufacturers, including IBM, and the 5th component v_5 consists of oil companies. The 6th and later components do not have distinct features compared to the first 5 components and can be regarded as random.

For 1-hour data of 1998 where we have $N=490$ and $T=1512$, the major components of v_1 are made of banks and financial services. The 2nd principal component v_2 consists of 10 electric companies, while v_3 consists of banks and financial services, and U_4 consists of semiconductor manufacturers. The 6th and later components do not have distinct features compared to the first 5 components and regarded as random.

For 1-hour data of 2002 where we have $N=569$ and $T=1512$, the major components of v_1 are strongly dominated by banks and financial services, while v_2 are strongly dominated by electric power supplying companies, which were not particularly visible in 1994 and 1998.

The above observation summarized in Table 2 indicates that Appliances/Car and IT dominated the industrial sector in 1994, which have moved toward the dominance of Finance, Food, and Electric Power Supply in 2002.

Table 2. Business sectors of top 10 components of 5 principal components in 1994, 1998 and 2002

v_k	1994	1998	2002
v_1	Finance(4), IT(2), Appliances/Car(3)	Finance(8)	Finance(9)
v_2	Mining(7), Finance(2)	Electric(10)	Food(6)
v_3	IT (10)	Finance(3)	Electric(10)
v_4	IT(7),Drug(2)	IT (10)	Food(4), Finance(2),Electric (4)
v_5	Oil(9)	Mining(6)	Electric (9)

6 Separation of Signal from Noise

Although this method works quite well for $v_1 - v_5$, the maximum eigenvalue λ_+ seems too loose to be used for a criterion to separate signal from the noise. There are many eigenvalues near λ_+ which are practically random. In Fig.2, for example, only the largest five eigenvalues exhibit distinct signals and the rest can be regarded more or less random components. In this respect, we examine the validity of RMT for finite values of N and T . Is there any range of Q under which the RMT formula breaks down?

First of all, we examine how small N and T can be. We need to know whether $N=419-569$ and $T=1512$ in our study in this paper are in any adequate range. To do this, we use two kinds of computer-generated random numbers, the random numbers of normal distribution generated by Box-Muller formula, and the random numbers of uniform distribution generated by the `rand()` function. However, if we shuffle the generated numbers to increase randomness, the eigenvalue spectra perfectly match the RMT formula.

The above lesson tells us that the machine-generated random numbers, independent of their statistical distributions, such as uniform or Gaussian, become a set of good random numbers only after shuffling. Without shuffling, the random series are not completely random according to the sequence, while the evenness of generated numbers is guaranteed.

Taking this in mind, we test how the formula in Eq.(20) works for various values of parameter N and Q . Our preliminary result shows that the errors are negligible in the entire range of $N > 50$ for $T > 50$ ($Q > 1$), after shuffling.

7 Summary

In this paper, we propose a new algorithm RMT-oriented PCM and examined its validity and effectiveness by using the real stock data of 1-hour price time series extracted from the tick-wise stock data of NYSE-TAQ database of 1994, 1998, and 2002. We have shown that this method provides us a handy tool to compute the principal components $v_1 - v_5$ in a reasonably simple procedure.

We have also tested the method by using two different machine-generated random numbers and have shown that those random numbers work well for a wide range of parameters, N and Q , only if we shuffle to randomize the machine-generated random series.

References

1. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E.: Random matrix approach to cross correlation in financial data. *Physical Review E*, American Institute of Physics 65, 66126 (2002)
2. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E.: *Physical Review Letters*. American Institute of Physics 83, 1471–1474 (1999)
3. Laloux, L., Cizeaux, P., Bouchaud, J.-P., Potters, M.: *American Institute of Physics*, vol. 83, pp. 1467–1470 (1999)

4. Bouchaud, J.-P., Potters, M.: Theory of Financial Risks. Cambridge University Press, Cambridge (2000)
5. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge (2000)
6. Mehta, M.L.: Random Matrices, 3rd edn. Academic Press, London (2004)
7. Sengupta, A.M., Mitra, P.P.: Distribution of singular values for some random matrices. Physical Review E 60, 3389 (1999)

Music Impression Detection Method for User Independent Music Retrieval System

Masato Miyoshi¹, Satoru Tsuge², Hillary Kipsang Choge¹, Tadahiro Oyama³, Momoyo Ito¹, and Minoru Fukumi¹

¹ The University of Tokushima, 2-1, Minami-josanjima, Tokushima, Tokushima, 770-8506, Japan

{m.miyoshi,choge,momoito,fukumi}@is.tokushima-u.ac.jp

² Daido University, 10-3, Takiharu-cho, Minami-ku, Nagoya, Aichi, 457-8530, Japan
tsuge@daido-it.ac.jp

³ Kobe City College of Technology, 8-3, Gakuen-higashimachi, Nishi-ku, Kobe, Hyogo, 651-2194, Japan
oyama@kobe-kosen.ac.jp

Abstract. In previous work, we have proposed the automatic sensitive word score detection system for a user dependent music retrieval system. However, the user dependent method causes a lot of burdens to the user because the system requires a lot of data for adapting it to each user. Hence, in this paper, we propose an automatic sensitive word score detection method for a user independent music retrieval system and evaluate the proposed method using 225 music data. Experimental results show that 87.5% of music patterns succeeded in detection of sensitive word score in the case that the difference between estimated and evaluated score is 1 (Error 1 rate). Moreover, we conduct subjective evaluation experiments to evaluate the proposed method as a utility method. From this experiment, it is observed that the user satisfaction level of the proposed method is higher than random selection impression detection.

Keywords: Sensitive word, Feature extraction, Neural network, Subjective evaluation.

1 Introduction

In recent years, it has become possible to access a lot of music data because we can get music from distributions on the Internet and save these on large capacity portable digital audio players and personal computers. As a result, it is difficult to retrieve and classify these music data by hand. Hence, efficient music retrieval and classification methods have been developed. Most traditional retrieval systems required music information, such as an artist name, musical genre, and so on, to users. These systems are not convenient for users because users do not necessarily know the information of all retrieval/classification music. If the music retrieval/classification systems using the sensitive words are developed, the

users can use the sensitive words, such as Brightness, Cheerful, and so on, without music information for retrieval/classification. In order to retrieve the music by sensitive words, it is necessary to set suitable words to each song and/or set a score of each sensitive word to each song.

Therefore, recently, the retrieval systems using sensitive words [1] [2] [3] and impression estimation method of music [4] [5] have been proposed. We also have proposed an automatic sensitive word score detection method for music retrieval system using sensitive words [6]. Most of these methods are designed for a user dependent system and have not been investigated user satisfaction. In case of the user dependent system, each user needs to prepare the data for training the system. This places a big burden on the user and causes a degradation of the system's convenience for the user. Moreover, we cannot know the performance of the system without a subjective evaluation of the system even if some methods improve the accuracy. Therefore, in this paper, we propose a feature extraction method from musical audio signals and a sensitive word score detection method for a user independent music retrieval system. For investigating the user satisfaction of the proposed method, we conduct subjective evaluation experiments on the proposed method.

First, in Sect. 2, we describe the details of the proposed method. In Sect. 3, we conduct the sensitive word score detection experiment and show experimental results. In Sect. 4, we conduct subjective evaluation experiments to investigate the user satisfaction level of the proposed method. Finally, in Sect. 5, we present conclusions of this paper and future works.

2 Automatic Sensitive Word Score Detection Method

In this section, we describe the proposed automatic sensitive word score detection method. An overview of the proposed method is shown in Fig. 1. This method consists of two parts; feature extraction and sensitive word score detection. We describe the details of these methods in the following sections.

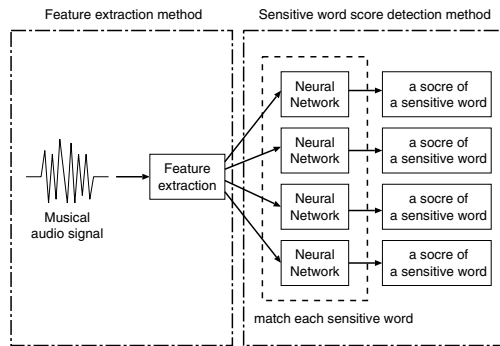


Fig. 1. Overview of the proposed method

2.1 Feature Extraction Method

The feature parameters extracted from music data by this feature extraction method are intensity, timbre, rhythm and harmony. We consider that these parameters are sufficient for the sensitive word score detection. We describe each feature as follows:

Intensity feature. The following 3 features are used as intensity features:

- **Frame Energy**[\[4\]](#)
Power of audio signals is used as an intensity feature.
- **Log Spectrum Sum and Δ Log Spectrum Sum**[\[6\]](#)
Log Spectrum Sum is the summation of log power spectrum and Δ Log Spectrum Sum is the linear regression coefficients of the Log Spectrum Sum. These features represent cheerful characteristics and variation of music.
- **Low Energy Frame**[\[6\]](#)
Low Energy Frame is the number of frames that Frame Energy is less than the threshold value. We consider that the music with many low frame energy frames expresses quietness.

Timbre feature. The following 6 features are used as timbre features:

- **Centroid**[\[4\]](#)
This represents the centroid frequency of the spectrum. We use this feature as a brightness feature of music.
- **Bandwidth**[\[4\]](#)
If the number of instruments in music is high, we consider that this music is cheerful. We use the power distribution –Bandwidth– as the cheerful feature of music.
- **Roll-off**[\[7\]](#)
We consider that the music in which the low-frequency band is emphasised expresses darkness. On the contrary, the music in which the high-frequency band is emphasised expresses brightness. For representing the frequency band emphasis, we use the Roll-off feature.
- **Variation of timbre**
We use Flux[\[7\]](#) and Cosine Similarity[\[6\]](#) as the feature for the variation of timbre in music.
- **Zero-Crossing**[\[6\]](#)
For the pitch feature, we use the Zero-Crossing which is related to fundamental frequency in music.

Rhythm feature. Power Spectrum Peak[\[6\]](#) is used as a rhythm feature. The percussive sounds, such as bass drum, high-hat, cymbal, and so on, are impulse signals. These frequency elements exist up to the high-frequency band. Hence, we detect power spectrum peaks and use these as the rhythm feature of music.

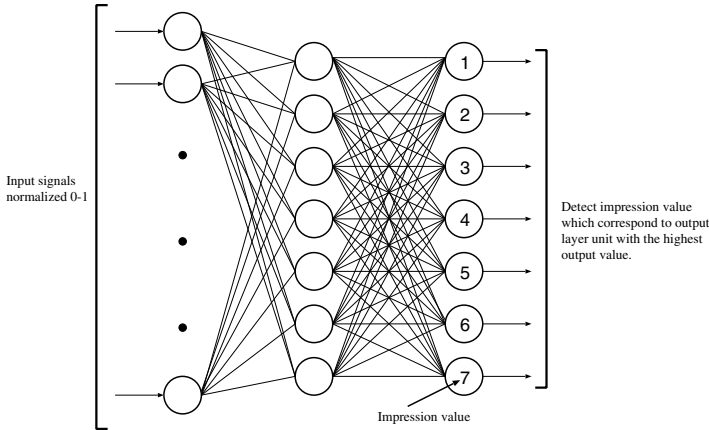


Fig. 2. Impression detection method using neural networks

Harmony feature. We calculate chroma vectors^[8] to extract the harmony features of music. The following 2 harmony features are calculated using chroma vectors.

- **Chroma Vector Flux**^[6]
To extract variation of pitch, Chroma Vector Flux is calculated.
- **Major and Minor Key Components**^[6]
The key of music is one of the mood features of music. For example, a major key music is felt as brightness, minor key music is felt as darkness. Hence, we calculate major and minor key components and use them as harmony features.

2.2 Sensitive Word Score Detection Method

We assume that human sensitive spaces are non-linear spaces^[4]. Therefore, Multi-Layer Neural Networks (MLNNs) are used for detection of sensitive word score because they have a high discriminant ability for non-linear problems. MLNNs are trained using the Back-Propagation algorithm (BP). These MLNNs are trained to use input audio features to detect each sensitive word score for every music pattern. In the proposed method, each output layer unit corresponds to 7 levels of impression scores. These are shown in Fig. 2. Estimated sensitive word score is given as an impression score which corresponds to the output layer unit with highest output value. Also, input values are normalized in the 0 to 1 range on each feature axis. MLNNs are constructed for each sensitive word score estimation.

3 Automatic Sensitive Word Score Detection Experiment in a User Independent System

In this section, in order to evaluate whether the proposed method is suitable for a user independent system, we conduct automatic sensitive word score detection

experiments. In Sect. 3.1 feature extraction conditions are described. Next, we describe the experimental conditions in Sect. 3.2 and report the experimental results in Sect. 3.3.

3.1 Feature Extraction Conditions

In this experiment, we used stereo audio signals which are sampled at 44.1kHz, quantized in 16 bits. Feature extraction conditions are shown in Table 1. Means and standard deviations of Frame Energy, Log Spectrum Sum, Δ Log Spectrum Sum, Centroid, Roll-off, Flux, Cosine Similarity, and Zero-Crossing are calculated every 5 seconds. Means of Bandwidth and Power Spectrum Peak are also calculated every 5 seconds. Means of Chroma Vector Flux are calculated every 1 second. The dimension of feature vector which is used for music impression score estimation is 72. The Blackmann window is used as the windowing function.

3.2 Experimental Conditions

The sensitive word score detection experiment is conducted to investigate the effectiveness of the proposed method. In this experiment, we use the RWC Music Database [9]. From this database, we use 401 music patterns which are selected from 225 music data [1]. Impression of 401 music patterns are measured against each sensitive word, which are Brightness, Cheerful, Up-tempo, and Airiness, by 6 test subjects. The impression score of each sensitive word is represented by 7 levels. In order to deal with a user independent system, medians of impression scores which are detected by 6 test subjects are used as impression score of each music pattern in each sensitive word. This experiment is conducted by Cross-validation method. Music patterns selected from a music datum are used

Table 1. Feature extraction conditions

Feature	Frame length	Frame shift length
Intensity	23.2ms	11.6ms
Timbre		
Rhythm		
Harmony	185.8ms	80.0ms

for evaluation data. For training data, we use other music patterns selected from music data in which the evaluation music patterns are not included. This experiment is run over only the number of music data. The number of input layer units, hidden layer units, and output layer units are 72, 7, and 7, respectively. The number of iterations for training is 50,000. The training coefficient and the momentum term are set to 0.1 and 0.7, respectively. The MLNNs are used as the detectors of the sensitive word score. The proposed method accuracy is evaluated by the difference between estimated and evaluated score. “Error”

¹ Some music patterns have been selected from a music datum.

indicates the average error rate between estimated and evaluated scores, “Error 0 Rate” indicates the rate of the music patterns for which the difference between estimated and evaluated score is 0, “Error 1 Rate” indicates the rate of the music patterns for which the difference between estimated and evaluated score is 1, respectively.

3.3 Experimental Results

The experimental results of each sensitive word are shown in Table 2. From the experimental results, we can see that “Error” is less than 1.0, “Error 0 Rate” is 43.0%, and “Error 1 Rate” is 87.5% as average accuracies. Especially, we can see that the “Error 1 Rate” of all sensitive words are higher than 80%. For investigating the details of this result, we show the distributions of impression score of each sensitive word in Fig. 3. From this figure, we can see that impression score of each music is concentrated on the range from 2 to 6. In this experiment, we use the median of the 6 test subjects’ impression score as the user independent score. As a result, the distribution of the impression score is narrowed, making the impression score estimation easier. From this investigation, we conclude that “Error 1 Rate” of the proposed method is high.

4 Subjective Evaluation Experiment

In the previous section, we described the accuracies of sensitive word score detection by the proposed method. However, the relationship between the accuracy and the satisfaction of the user is not clear. Hence, in this section, we conduct the subjective evaluation experiment for investigating the effectiveness of the proposed method as a utility method. The details of the subjective evaluation experiment are described in Sect. 4.1 and the experimental results are shown in Sect. 4.2.

4.1 Experimental Conditions

In this experiment, it is assumed that the user retrieves music data using the music retrieval system shown in Fig. 4. In this system, first, the user selects impression score of sensitive words of music which the user requires. Then, the

Table 2. Experimental results

Sensitive word	Error	Error 0 Rate	Error 1 Rate
Brightness	0.681	44.1%	88.8%
Cheerful	0.611	51.6%	89.0%
Up-tempo	0.753	39.2%	87.8%
Airiness	0.808	37.2%	84.5%
Average	0.713	43.0%	87.5%

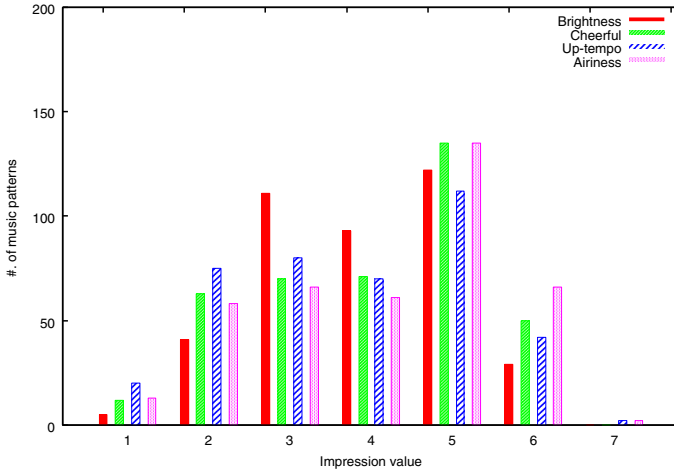


Fig. 3. Distribution of impression score of each sensitive word in the user independent system

system retrieves the music by using these scores. In this experiment, we use the words shown in Table 3 as sensitive words.

The number of test subjects is 5. The number of retrieval patterns is 12. These retrieval patterns are shown in Table 3. In this table, the digit indicates the impression score of each sensitive word and the “×” indicates that this sensitive word is not used for retrieval. We show the 3 music patterns against each retrieval pattern to test subjects. These music patterns are;

1. Detected by the proposed method (named “Proposed”)
2. Detected by the listening experiment (“Ideal”)
3. Detected by random selection (“Random”).

The test subjects evaluate a total of 36 music patterns because the number of retrieval patterns is 12 and 3 music patterns are shown to test subjects by each

Input impression								
Sensitive word	Weak ← → Strong							Cancel
	1	2	3	4	5	6	7	
Brightness	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Cheerful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Up-tempo	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Airiness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fig. 4. The retrieval system

Table 3. Retrieval patterns in subjective evaluation experiment

Retrieval pattern No.	Brightness	Cheerful	Up-tempo	Airiness
1	6	×	×	×
2	2	×	×	×
3	×	6	×	×
4	×	2	×	×
5	×	×	6	×
6	×	×	2	×
7	×	×	×	6
8	×	×	×	2
9	2	2	2	2
10	5	5	5	5
11	3	4	2	2
12	4	5	6	6

Table 4. User satisfaction levels

Test subject	Proposed	Random	Ideal
A	3.58	2.67	4.42
B	4.00	3.08	4.42
C	3.75	2.17	3.83
D	4.08	2.83	4.50
E	3.75	2.50	4.42
Average	3.83	2.65	4.32

retrieval pattern. The retrieval music patterns are shown in random order. The test subjects listen to these and evaluate user satisfaction levels of each music pattern. In this experiment, the user satisfaction levels are expressed in 5 levels: 5: Satisfied, 4: Moderately satisfied, 3: Neutral, 2: Moderately dissatisfied, 1: Dissatisfied.

4.2 Experimental Results

Table 4 shows average user satisfaction levels as experimental results. This table shows that the user satisfaction level of the proposed method, shown as “Proposed” in this table, are higher than that of “Random” under the condition of all test subjects. From this result, we consider that the proposed method is able to present the music data required by the user. However, we can see that the user satisfaction level of the proposed method does not match that of “Ideal” from this table. We need to investigate the factors which affect the user satisfaction and improve the user satisfaction level of the proposed method.

Next, Table 5 shows the number of the music patterns of each user satisfaction level as the details of the experimental results. From Table 5, we can see that the rates of music patterns for which the user satisfaction level is more than 4 are 71.7% for “Proposed” ((28 (level 5) + 15 (level 4))/60 (in total)), 81.7%

Table 5. User satisfaction level distribution against each impression detection method

Satisfaction level	Proposed	Random	Ideal
5	28	6	37
4	15	14	12
3	2	8	5
2	9	16	5
1	6	16	1

$((37+12)/60)$ for “Ideal”, and 33.3% $((6+14)/60)$ for “Random”, respectively. Hence, we consider that the proposed method is useful for the music retrieval system because the users are satisfied with the proposed method in more than 70% of the music patterns.

This table shows that the number of music patterns in the proposed method for which user satisfaction level is 5 is smaller than that of “Ideal”. However, this table also shows that the number of music patterns of the proposed method for which the user satisfaction level is 4 is more than that of “Ideal”. For this reason, we consider that “Error 0 rate” and “Error 1 rate”, which are shown in Table 2, affect these subjective evaluation results. In the proposed method, “Error 0 rate” is low although “Error 1 rate” is high. In other words, the proposed method can estimate the impression score of sensitive word but not detect it correctly. Therefore, we conclude that the proposed method can present the music pattern similar to that which user requires. In the future, we plan to investigate the relationships between the user satisfaction and the accuracy of the proposed method.

5 Conclusions

In this paper, we proposed an automatic sensitive word score detection method for a user independent music retrieval system. In the proposed method, audio features related to intensity, timbre, rhythm and harmony are extracted to estimate impression score in each sensitive word score. These features are input into MLNNs and sensitive word scores are detected for each sensitive word.

We conducted the automatic sensitive word score detection experiment for investigating the effectiveness of the proposed method. Experimental results showed that the proposed method could achieve 87.5% of “Error 1 Rate” and 43.0% in “Error 0 Rate”. From these results, we concluded that the proposed method could present the music data similar to that which a user requires.

In addition, we conducted a subjective evaluation experiment for investigating user satisfaction. This experiment showed the user satisfaction of the proposed method was higher than that of random selection. However, the user satisfaction of the proposed method was lower than that of the listening detection. From this experiment, we could realize that Error 0 rate affected the user satisfaction level 5.

In the future, we shall investigate methods to improve the sensitive word detection accuracy of the proposed method. Therefore, we plan to investigate the feature parameters which affect the user satisfaction level.

Acknowledgements. This research has been partially supported by the Japan Society for the Promotion of Science, Grant-in-Aid for Young Scientists(B), 19700172.

References

1. Sugihara, T., Morimoto, K., Kurokawa, T.: m-RIK: Music Retrieval System Specialized in Individual Kansei Characteristic. *The Journal of Information Processing Society* 46(7), 1560–1570 (2005) (in Japanese)
2. Ikezoe, T., Kajikawa, Y., Nomura, Y.: Music Database Retrieval System with Sensitivity Words Using Music Sensitivity Space. *The Journal of Information Processing Society* 42(12), 3201–3212 (2001) (in Japanese)
3. Otsuka, L., Kajikawa, Y., Nomura, Y.: A Study on a Music Retrieval System for PCM Data by Sensitivity Words. In: *Proc. of DEWS 2003*, 8-p-05 (2003) (in Japanese)
4. Hirae, R., Nishi, T.: Mood Classification of Music Audio Signals. *The Journal of the Acoustical Society of Japan* 64(10), 607–615 (2008) (in Japanese)
5. Tanoue, N., Nishi, T.: Sensibility Classification of Music Audio Signals Using Time-Frequency Image Pattern. In: *Proc. of ASJ 2009* (2009) (in Japanese)
6. Miyoshi, M., Oyama, T., Choge, H.K., Karungaru, S.G., Tsuge, S., Fukumi, M.: Music Classification using Sensitive Words. In: *Proc. of NCSP 2010*, pp. 211–214 (2010)
7. Tzanetakis, G., Essl, G., Cook, P.: Automatic Genre Classification Of Audio Signals. In: *Proc. of ISMIR 2001* (2001)
8. Goto, M.: A Real-time Music Scene Description System: A Chorus-Section Detecting Method. In: *MUS 2002*, vol. 2002(100), pp. 27–34 (2002) (in Japanese)
9. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: RWC Music Database: Database of Copyright-cleared Musical Pieces and Instrument Sounds for Research Purposes. *The Journal of Information Processing Society* 45(3), 728–738 (2004) (in Japanese)

Applying Fuzzy Sets to Composite Algorithm for Remote Sensing Data

Kenneth J. Mackin¹, Takashi Yamaguchi¹, Jong Geol Park², Eiji Nunohiro¹,
Kotaro Matsushita², Yukio Yanagisawa³, and Masao Igarashi³

¹ Department of Information Systems, Tokyo University of Information Sciences
1200-2 Yatoh-cho, Wakaba-ku, Chiba, Japan
mackin@rsch.tuis.ac.jp

² Department of Environmental Information, Tokyo University of Information Sciences
1200-2 Yatoh-cho, Wakaba-ku, Chiba, Japan

³ College of Bioresource Sciences, Nihon University
1866 Kameino, Fujisawa, Kanagawa, Japan

Abstract. Remote sensing of the earth surface using satellite mounted sensor data is a major method for global environmental monitoring today. However, when using satellite sensor data, clouds in the atmosphere can interfere with the readings, and specific land points may not be correctly monitored on a given day. In order to overcome this problem, multiple day composite data is frequently used. Multiple day composite data uses several consecutive days' remote sensing data, and picks the most accurate data within the temporal dataset for the same land point. This allows creating a more complete dataset by patching together data not interfered by clouds during a specified time period, to create a clearer, more usable dataset. In this paper, we propose applying fuzzy set logic in order to select the clearest data in the temporal interval for the composite data. Moderate resolution remote sensing data of areas in Japan were used for evaluation.

Keywords: Fuzzy Sets, remote sensing, MODIS, composite algorithm.

1 Introduction

Remote sensing of the earth surface using satellite mounted sensor data is one of the most important methods for global environmental monitoring today. For example, changes in land surface use can be monitored using remote sensing data. By monitoring the changes of vegetation coverage, the amount of carbon held in vegetation and how much is released into the atmosphere can be calculated.

However, when using satellite sensor data, clouds in the atmosphere can interfere with the remote sensing, and specific land points may not be correctly monitored on a given day (Fig.1). In order to overcome this problem, a common workaround is to use multiple day composite data. Multiple day composite data uses several consecutive days' remote sensing data, and picks the most accurate data within the temporal dataset for the same land point (Fig.2). This allows creating a more complete dataset by patching together data uninterfered with clouds during a specified time period, to create a clearer, more usable dataset.

There have been many data composite algorithms proposed, and different methods have their merits and demerits, but all methods require tuning of parameters to achieve best accuracy for a specific region. In this paper, we propose applying a soft computing approach, namely fuzzy logic[3], in order to facilitate tuning of the data composite algorithm in order to achieve the best accuracy for a specific region. Moderate resolution remote sensing data of areas in Japan were used for evaluation, and results were compared with previous composite methods.

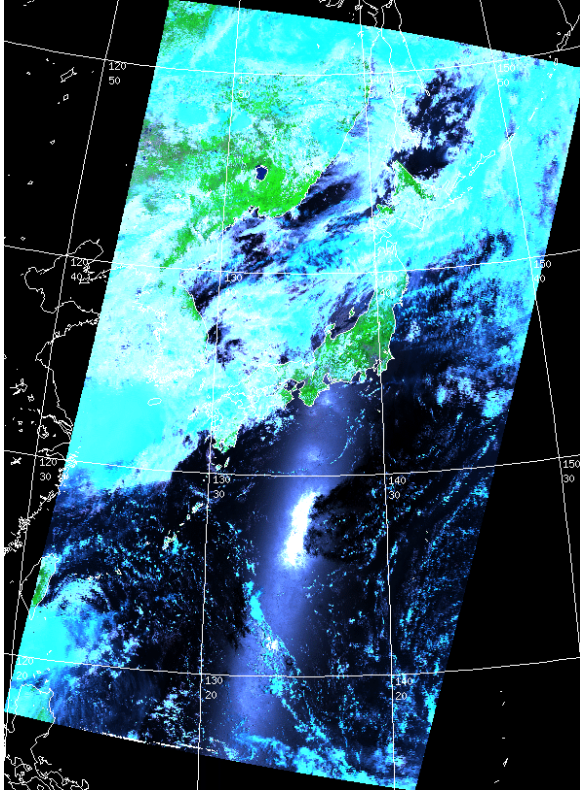


Fig. 1. 1 pass of MODIS data with cloud interference

2 Satellite Data

With the increased interest in monitoring the global ecological changes, the demand for satellite remote sensing has increased. NASA-centered international Earth Observing System project has launched many satellites to monitor the earth for scientific purposes, including Terra and Aqua. (Fig.3)

A key instrument aboard the Terra and Aqua satellites is MODIS (Moderate Resolution Imaging Spectroradiometer). Terra's orbit around the Earth is timed so that it passes from north to south across the equator in the morning, while Aqua passes

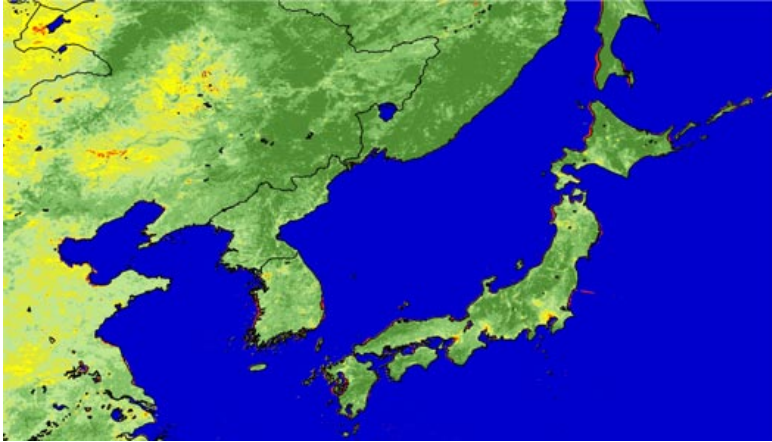


Fig. 2. 30-day composite data with cloud interference removed

south to north over the equator in the afternoon. Terra MODIS and Aqua MODIS enable the viewing of the entire Earth's surface every 1 to 2 days. MODIS captures data in 36 spectral bands, or groups of wavelengths. Moderate resolution remote sensing allows the quantifying of land surface type and extent, which can be used to monitor changes in land cover and land use for extended periods of time. This data is used to monitor and understand global dynamics and processes occurring on the land, oceans, and lower atmosphere.

For this paper, we used MODIS data collected at Tokyo University of Information Sciences (TUIS), Japan. TUIS receives satellite MODIS data over eastern Asia, and provides this data for open research use.



Fig. 3. Terra and Aqua satellites ©NASA

3 Composite Algorithm

There have been many different satellite data composite algorithms proposed [1][2], and each method have different merits depending on the target usage. Different

composite methods include 1) MVC (maximum value composition) using maximum NDVI (normalized difference vegetation index), 2) MVC using maximum temperature reflectance, 3) minimum scan angle with high NDVI, 4) minimum scan angle with high NDVI and temperature reflectance, 5) minimum blue reflectance with high NDVI.

In this research, we use MODIS sensors satellite data for evaluation of the composite algorithms.

The MODIS data is a raster format image file where each pixel of the image is the reflectance value of a specific bandwidth for the location. MODIS sensor data contains 36 different bands, including visible red (band 1), near infra-red (band 2). MODIS has sensors in 3 different spatial resolutions, 250m² resolution (band 1-2), 500m² resolution (band 3-7), 1km² resolution (band 8-36), so each pixel data will be the average reflectance value for the location in the specified spatial resolution.

NDVI is a standard vegetation index defined by the following equation,

$$NDVI = \frac{NIR - R}{NIR + R} \quad (1)$$

where R is band 1 (visible red : bandwidth 620 - 670 nm) reflectance value, and NIR is band 2 (near infra-red : bandwidth 841 - 876 nm) reflectance value for the given location.

NDVI reflects the growth of vegetation, so for a given temporal interval the value is not expected to change drastically. It is also known that NDVI values become lower when clouds in the atmosphere interfere with the reflection. MVC with NDVI takes into account these 2 assumptions, and selects the highest NDVI value within a set temporal interval for the same location, to pick the best cloud-free data for the location within the temporal interval.

To create a 10-day composite image, assuming there is 1 MODIS reading every day for a total of 10 files, the maximum NDVI value is selected from the 10 different MODIS readings for the exact same location. The maximum NDVI marks the best cloud-free data for the temporal interval for the location, and the reading is copied to the composite image file for the same location. The reflective values of each of the different bands for the same location and day are also copied to the respective composite image files for each band. This is repeated for every location (pixel) in the MODIS NDVI image file to create a composite image for the specified temporal interval.

4 Proposed Method

The composition method using minimum blue criterion from among high NDVI values is also widely used. This method is based on the observation that visible blue wavelengths are highly sensible to atmospheric interference, and the blue reflectance values increase with rise in atmospheric interference. Therefore, the clearest days should have the lowest blue reflectance readings. This method first selects a set of candidate readings with high NDVI values, and then selects the reading with the lowest blue value from among the candidates. The candidates are selected for

example by selecting NDVI values within 80% range of the maximum NDVI for the temporal interval.

In this research we propose applying soft computing approach, namely fuzzy logic[3], in order to create a composition method which can be more easily tuned for specific regions.

In this paper we take the minimum blue criterion as the initial model. In the standard rule-based minimum blue criterion, there is a crisp criterion of high NDVI candidates (e.g. 80% of maximum NDVI) and crisp criterion of lowest blue value from among the candidate. In this standard crisp rule-based method, a very low blue value with 79% NDVI value will be discarded. Similarly, even if the maximum NDVI value has second lowest blue value, the maximum NDVI value will be discarded.

We extend the blue criterion composition algorithm by defining a fuzzy set of high NDVI values, and a fuzzy set of low blue values, and selecting the best value from the combined fuzzy set of high NDVI and low blue using fuzzy logic.

We set the fuzzy set N as the set of readings with high NDVI values and fuzzy set B as the set of readings with low blue values, defined by the membership function m_N and m_B on total space X and Y of NDVI and blue readings, respectively.

$$N = \int_X m_N / x \quad x \in X \tag{2}$$

$$B = \int_Y m_B / y \quad y \in Y \tag{3}$$

$$m_N : X \rightarrow [0,1] \tag{4}$$

$$m_B : Y \rightarrow [0,1] \tag{5}$$

The membership value or grade $m_N(x)$, $m_B(y)$ for reading x and y is similarly defined as below.

$$m_N(x) \in [0,1] \tag{6}$$

$$m_B(y) \in [0,1] \tag{7}$$

From fuzzy logic, the fuzzy set of readings with high NDVI values and low blue values are defined as the fuzzy intersection of set N and B , and given as below.

$$\begin{aligned} N \cap B &\Leftrightarrow m_{N \cap B}(x, y) \\ &= m_N(x) \wedge m_B(y) \\ &= \min\{m_N(x), m_B(y)\} \end{aligned} \tag{8}$$

By defining an appropriate membership function for fuzzy set N and B , the reading with best value can be calculated as the reading with the highest grade in $N \cap B$.

For this experiment, the membership grade for $m_N(x)$ and $m_B(x)$ was defined as below.

$$m_N(x) = \begin{cases} \frac{(x - T_N)}{(x_{MAXN} - T_N)}, & x \geq T_N \\ 0, & x < T_N \end{cases} \quad (9)$$

$$m_B(y) = \begin{cases} \frac{(T_B - y)}{(T_B - y_{MINB})}, & y \leq T_B \\ 0, & y > T_B \end{cases} \quad (10)$$

x_{MAXN} is the maximum NDVI value, y_{MINB} is the minimum blue value, T_N is the threshold value for large NDVI candidates, T_B is the threshold value for low blue candidates.

5 Conclusion

The proposed method was applied to create a monthly composite data for July 2001 over regions in Japan. This was compared with composite data for the same temporal interval and region using MVC with maximum NDVI. The proposed method was able to remove clouds successfully, and the final composite result was very close to that of standard MVC using maximum NDVI.

In this paper, an extension to minimum-blue criterion composition method for remote sensing data was proposed. The proposed method applies fuzzy set theory and fuzzy logic to facilitate parameter tuning, as well as produce a composite method allowing a more flexible selection of data for the composite image.

For future works, we need to closely examine the differences of output between the proposed method and standard method, and evaluate the effectiveness and accuracy of the proposed method.

Acknowledgments. This research was supported by the Research project of Tokyo University of Information Sciences for the sustainable development of economic and social structure dependent on the environment in eastern Asia.

References

1. Takeuchi, W., Yasuoka, Y.: Comparison of composite algorithm for South East Asia using MODIS data. In: Proceeding of Annual Conference of Japan Society of Photogrammetry and Remote Sensing, CD-ROM, Tokyo (2003) (in Japanese)
2. van Leeuwen, W.J.D., Huete, A.R., Laing, T.W.: MODIS Vegetation Index Compositing Approach: A Prototype with AVHRR Data. *Remote Sens. Environ.* 69, 264–280 (1999)
3. Zadeh, L.A.: Fuzzy sets. *Informat. Control* 8(3), 338–353 (1965)
4. National Aeronautics and Space Administration (NASA), MODIS Web, <http://modis.gsfc.nasa.gov/>

Evaluations of Immunity-Based Diagnosis for a Motherboard

Haruki Shida¹, Takeshi Okamoto¹, and Yoshiteru Ishida²

¹ Dept. of Information Network and Communication, Kanagawa Institute of Technology
1030, Shimo-ogino, Atsugi, Kanagawa 243-0292, Japan

haruki.shida@gmail.com
take4@nw.kanagawa-it.ac.jp

² Dept. of Knowledge-Based Information Engineering, Toyohashi University of Technology,
1-1, Tempaku, Toyohashi, Aichi 441-8580, Japan

ishida@tutkie.tut.ac.jp

Abstract. We have utilized immunity-based diagnosis to detect abnormal behavior of components on a motherboard. The immunity-based diagnostic model monitors voltages of some components, CPU temperatures, and Fan speeds. We simulated abnormal behaviors of some components on the motherboard, and we utilized the immunity-based diagnostic model to evaluate motherboard sensors in two experiments. These experiments showed that the immunity-based diagnostic model was an effective method for detecting abnormal behavior of components on the motherboard.

Keywords: immunity-based system, fault diagnosis, sensor, motherboard, immune network.

1 Introduction

The technology of cloud computing has become prevalent, and the demand for data centers that provide such cloud computing has increased. Each server in the data center must be highly available for data processing and data transmission. To maintain system availability, it is important to detect abnormalities during their early stages, before system failure.

The simplest way of diagnosing abnormalities consists of evaluating each component individually by comparing the output value of its sensor with a predetermined threshold value. However, it is difficult to identify the abnormal component using this method [1].

Another method of diagnosis uses an immunity-based diagnostic model [2-5], which is derived primarily from the concept of an immune network [6]. In this diagnostic model, mutual tests are performed among nodes and the dynamic propagation of active states. This diagnostic model has been applied to node fault diagnosis in processing plants [7], to self-monitoring/self-repairing in distributed intrusion detection systems [3], and to sensor-based diagnostics for automobile engines [4].

This paper reports on the use of an immunity-based diagnostic model for detecting the abnormal behavior of components on a motherboard, including CPUs, memories, chipsets and Fans.

2 Embedded Sensors on the Motherboard

Since a motherboard has multiple sensors, including voltage, temperature, and fan speed sensors, abnormalities on the motherboard can be detected by monitoring these sensors. We therefore used sensor output values for diagnosis of the motherboard.

We collected sensor output values on a server from July 27th to September 18th. The specification of the server is shown in Table 1. The average air temperature during that period was 25.3 °C, ranging from 20.1°C to 32.8°C. Data were collected using `lm_sensors`, a hardware health monitoring package for Linux that allows information to be obtained from temperature, voltage, and fan speed sensors.

Table 1. Server specification

Motherboard	Supermicro® X7DVL-I
OS	Debian GUN/Linux 5.0
Kernel	2.6.26-2-amd64
Module	lm-sensors version 3.0.2 with libesensors version 3.0.2
CPU	Intel® Xeon E5410 2.33GHz×2
Power supply	Thermaltake Toughpower 700w
Fan	XFan model: RDM8025B×2 Gantle Typhoon D0925C12B2AP×2 ADDA CFX-120S

We collected the output values from all 29 sensors on the motherboard, from which we calculated the correlation coefficients of all sensors. We observed correlations between 5 sensors (Table 2), and we therefore, used these 5 sensors for evaluation.

Table 2. Sensors used for evaluation and the range of sensor output values

Sensor	Component	Range	Mean	Standard deviation
CPU1	CPU temperature	11.00-48.00(°C)	18.68	4.550
Core2	Core2 temperature	35.00-72.00(°C)	42.79	4.450
VcoreA	CoreA voltage	1.11-1.19(V)	1.121	0.007
Vbat	Internal battery voltage	3.23-3.26(V)	3.237	0.009
Fan5	Fan speed	1012-1044(RPM)	1034	5.021

3 Immunity-Based Diagnostic Model

The immunity-based diagnostic model has the feature of a dynamic network [7], in which diagnoses are performed by mutually testing nodes, i.e., sensors and by dynamically

propagating their active states. In this paper, the targets of the immunity-based diagnosis are components with a sensor embedded on a motherboard. Each sensor can test linked sensors and can be tested by linked sensors. Each sensor is assigned a state variable R_i indicating its credibility.

The initial value of credibility $R_i(0)$ is 1. The aim of the diagnosis is to decrease the credibility of all the abnormal sensors. If the credibility of a sensor is less than a threshold value, the sensor is considered abnormal in this model.

When the value of credibility R_i is between 0 and 1, the model is called a *gray model*, reflecting the ambiguous nature of credibility. The *gray model* is formulized by the equation:

$$\frac{dr_i(t)}{dt} = \sum_j T_{ji}^+ R_j(t) - r_i(t), \quad (1)$$

where

$$R_i = \frac{1}{1 + \exp(-r_i(t))}, \quad (2)$$

$$T_{ij}^+ = \begin{cases} T_{ij} + T_{ji} - 1, & \text{if one of evaluation from } i \text{ to } j \text{ or } j \text{ to } i \text{ exists,} \\ 0, & \text{if neither evaluation from } i \text{ to } j \text{ nor } j \text{ to } i \text{ exists,} \end{cases} \quad (3)$$

$$T_{ij} = \begin{cases} 1, & \text{if a balance formula between sensors } i \text{ and } j \text{ is satisfied,} \\ -1, & \text{if a balance formula between sensors } i \text{ and } j \text{ is not satisfied,} \\ 0, & \text{if there is no balance formula between sensors } i \text{ and } j. \end{cases} \quad (4)$$

In the right-hand side of Equation (1), the first term is the sum of evaluations from other nodes for node i . The second term is an inhibition term that maintains ambiguous states of credibility. In this model, equilibrium points satisfy the equation $r_i(t) = \sum_j T_{ji}^+ R_j(t)$. Thus R_i monotonically reflects the value of $\sum_j T_{ji}^+ R_j(t)$. If $\sum_j T_{ji}^+ R_j(t)$ is close to 0, then R_i is close to 0.5. The balance formulas are shown in Table 3. We determined the balance formulas by calculating the relationships of the output value of the sensors.

4 Evaluations of Immunity-Based Diagnosis of the Motherboard

We evaluated the immunity-based diagnostic model for motherboard sensors by two experiments.

In the first experiment, we compared two diagnostic models: a stand alone diagnostic model and a mutual diagnostic model, i.e., an immunity-based diagnostic model. In the second experiment, we compared two networks in the immunity-based diagnostic model: a fully-connected network and a correlation-based network. Each evaluation was based on the four test cases shown in Table 4.

The test cases in (a) and (b) assumed that the speed of Fan5 was largely out of the range shown in Table 2. A significant decrease in Fan speed would therefore cause the CPU temperature to rise, with the overheated CPU causing the server to crash.

Table 3. Balance formulas between sensors

Sensor	Balance formula
CPU1-Core2	$ CPU1-Core2 \leq 26$
CPU1-VCoreA	$ CPU1-VCoreA \times 25 \leq 20$
CPU1-Vbat	$ CPU1-Vbat \times 9 \leq 18$
CPU1-Fan5	$ CPU1-Fan5/34 \leq 18$
Core2-VCoreA	$ Core2-VcoreA \times 45.5 \leq 28$
Core2-Vbat	$ Core2-Vbat \times 16 \leq 20$
Core2-Fan5	$ Core2-Fan5/19 \leq 21$
VCoreA-Vbat	$ VCoreA-Vbat/2.8 \leq 0.05$
VCoreA-Fan5	$ VCoreA-Fan5/893 \leq 0.07$
Vbat-Fan5	$ Vbat-Fan5/316 \leq 0.07$

Conversely, a significant increase in Fan speed would waste power and decrease the life span of the Fan. Therefore, we determined that the test cases of (a) and (b) are abnormal.

The test cases of (c) and (d) assumed that the output values of the sensors were slightly out of the range shown in Table 2. The test case of (c) assumed that the speed of Fan5 was slightly higher than that of Table 2, but that Fan5 was not abnormal. The test case of (d) assumed that the temperature of CPU1 was slightly higher than that of Table 2, but that CPU1 was not abnormal. Temperatures outside the range are not always abnormal, because these temperatures depend on room temperature. Therefore, we determined that the test cases of (c) and (d) are normal.

Table 4. Test cases

Case	Sensor output value					State
	CPU1	Core2	VcoreA	Vbat	Fan5	
(a) : Fan speed is very low.	70	65	1.12	3.23	200	Abnormal
(b) : Fan speed is very high.	9	35	1.12	3.23	2000	Abnormal
(c) : Fan speed is slightly high.	14	35	1.12	3.23	1050	Normal
(d) : CPU temperature is slightly high.	50	60	1.12	3.23	1020	Normal

4.1 Stand Alone vs. Mutual Diagnosis

We evaluated a stand alone diagnosis and a mutual diagnosis. According to the stand alone diagnosis, a component is considered abnormal if the sensor output value is outside the range shown in Table 2. In contrast, mutual diagnosis uses the immunity-based diagnostic model.

Tables 5 and 6 show the results of the stand alone and mutual diagnoses, respectively. In Table 5, a credibility of 0 indicates that the output value was not within

range, i.e., it was abnormal, whereas a credibility of 1 indicates that the output value was within range, i.e., it was normal. In Table 6, the credibility corresponds to R_i of Equation (2), i.e., it expresses the probability that component i is normal. We assumed that a component on the motherboard was abnormal if its credibility was less than 0.1. A diagnosis of “X” indicates an abnormality, whereas a diagnosis of “O” indicates an absence of abnormality. An accuracy of “O” indicates a correct decision, an accuracy of “X” indicates an incorrect decision, and an accuracy of “P” indicates that the diagnostic model could not identify the abnormal component, although it detected multiple abnormalities.

Table 5. Results of the stand alone diagnosis

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0	1	1	1	0	X	P
(b)	0	1	1	1	0	X	P
(c)	1	1	1	1	0	X	X
(d)	0	1	1	1	1	X	X

Table 6. Results of the mutual diagnosis

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0.00	0.83	0.78	0.78	0.00	X	P
(b)	0.73	0.99	0.99	0.73	0.00	X	O
(c)	0.99	0.91	0.99	0.99	0.99	O	O
(d)	0.00	1.00	1.00	1.00	1.00	X	X

The stand alone diagnostic model detected abnormalities in all test cases, because all test cases have values out of the range. In test cases (a) and (b), the stand alone diagnostic model failed to identify the abnormal component. This model also misdiagnosed test cases (c) and (d), judging them abnormal since the output values were slightly out of the range. In contrast, the mutual diagnosis model identified the abnormal Fan in test case (b) since only the credibility of Fan5 was 0.00. In test case (c), the mutual diagnosis made a correct decision. Consequently, the mutual diagnosis model is more accurate than the stand alone diagnosis model.

4.2 Fully-Connected Network vs. Correlation-Based Network

The immunity-based diagnostic model contains a network for mutually testing the credibility of nodes. In the above section, the network of the immunity-based diagnostic model was fully-connected, with each sensor connected to all other sensors, and each sensor mutually tested by all other sensors. A fully-connected network can include some connections between sensors with weakly correlated output values. These connections may be unreliable for mutually testing the credibility of their sensors. Therefore, we removed such connections from a fully-connected network, forming a correlation-based network.

We used the immunity-based diagnostic model to evaluate two network models, a fully-connected network and a correlation-based network. Figure 1 shows the correlation coefficients among the 5 sensors in Table 2. Any pair of sensors with a correlation greater than a threshold value was defined as connected. In this experiment, we built correlation-based networks for all the thresholds, using the correlation coefficients shown in Figure 1. Typical correlation-based networks are shown in Figure 2. All test cases were the same as those in Table 4. Table 7 shows the results of correlation-based networks. A network with a threshold less than 0.01 was identical to a fully-connected network, whereas a network with a threshold greater than 0.90 had no connection between any pair of sensors, i.e., a diagnostic model with a threshold greater than 0.90 was identical to a stand alone diagnostic model. These diagnostic models were evaluated in the previous section.

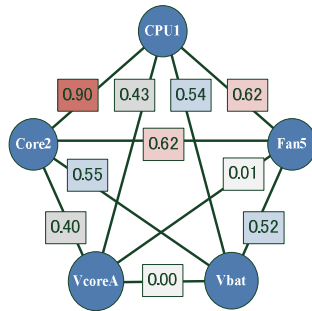


Fig. 1. Correlation coefficients among 5 sensors

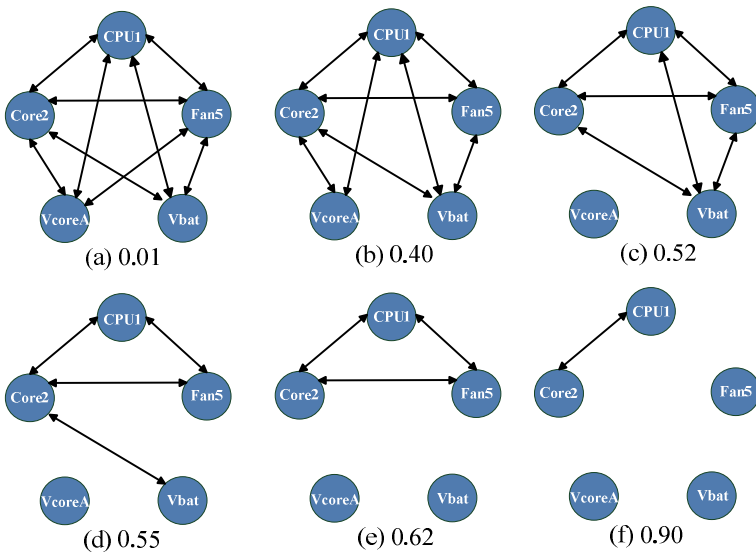


Fig. 2. Correlation-based networks for thresholds of (a) 0.01, (b) 0.40, (c) 0.52, (d) 0.55, (e) 0.62, and (f) 0.90

Table 7(A). Results of a correlation-based network with a threshold of 0.01

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0.00	0.97	0.87	0.87	0.00	X	P
(b)	0.96	0.98	0.98	0.12	0.00	X	O
(c)	0.99	0.99	0.98	0.51	0.98	O	O
(d)	0.00	0.98	0.99	0.98	0.99	X	X

Table 7(B). Results of a correlation-based network with a threshold of 0.40

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0.00	0.97	0.87	0.87	0.00	X	P
(b)	0.00	0.98	0.98	0.12	0.00	X	P
(c)	0.99	0.99	0.98	0.73	0.73	O	O
(d)	0.00	0.99	0.88	0.98	0.98	X	X

Table 7(C). Results of a correlation-based network with a threshold of 0.52

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0.34	0.67	0.50	0.34	0.01	X	O
(b)	0.81	0.61	0.50	0.81	0.00	X	O
(c)	0.99	0.99	0.50	0.73	0.73	O	O
(d)	0.00	0.98	0.50	0.98	0.98	X	X

Table 7(D). Results of a correlation-based network with a threshold of 0.55

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0.87	0.97	0.50	0.87	0.00	X	O
(b)	0.87	0.97	0.50	0.87	0.00	X	O
(c)	0.98	0.99	0.50	0.88	0.98	O	O
(d)	0.67	0.95	0.50	0.87	0.67	O	O

Table 7(E). Results of a correlation-based network with a threshold of 0.62

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0.84	0.84	0.50	0.50	0.00	X	O
(b)	0.84	0.84	0.50	0.50	0.00	X	O
(c)	0.61	0.81	0.50	0.50	0.81	O	O
(d)	0.81	0.61	0.50	0.50	0.81	O	O

Table 7(F). Results of a correlation-based network with a threshold of 0.90

Test case	Credibility					Decision	Accuracy
	CPU1	Core2	VcoreA	Vbat	Fan5		
(a)	0.84	0.84	0.50	0.50	0.50	O	X
(b)	0.84	0.84	0.50	0.50	0.50	O	X
(c)	0.84	0.84	0.50	0.50	0.50	O	O
(d)	0.84	0.84	0.50	0.50	0.50	O	O

In Table 7(A) the diagnostic models with thresholds of 0.01 misidentified the normal CPU1 in test cases (a) and (d). In Table 7(B), the diagnostic models with thresholds of 0.40 misidentified the normal CPU1 in test cases (a), (b) and (d). In Table 7(C), the diagnostic model with a threshold of 0.52 identified the abnormal Fan in test cases (a) and (b), and did not falsely identify an abnormality in test case (c), but misidentified the abnormal CPU1 in test case (d) as normal. In Tables 7 (D) and (E), the diagnostic models with thresholds of 0.55 and 0.62 correctly identified the abnormal Fan in test cases (a) and (b) and did not falsely identify abnormalities in test cases (c) and (d). In Table 7(F), the diagnostic model with a threshold of 0.90 identified only test case (c), because the abnormal sensor of Fan5 was isolated from the correlation-based network. This diagnostic model could not diagnose the isolated sensors, because the credibility of each was always 0.50.

Even networks with the best thresholds, of 0.55 and 0.62, have isolated sensors of VcoreA and Vbat. The sensor output values of VcoreA and Vbat were approximately constant over time, i.e., their standard deviations were very small (Table 2), such that the stand alone diagnostic model would correctly detect their abnormalities. Therefore, we applied stand alone diagnosis only to these isolated sensors (Figure 3). In other words, we use a hybrid diagnosis model, using both stand alone and immunity-based diagnosis. Sensors on the correlation network were diagnosed by the immunity-based diagnostic model, and isolated sensors were diagnosed by the stand alone diagnostic model.

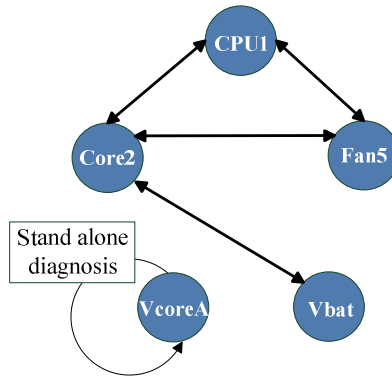


Fig. 3. Example of a hybrid diagnostic model with a threshold of 0.55

5 Conclusions

We applied immunity-based diagnosis to the detection of abnormal behaviors of components on a motherboard. We simulated the abnormal behaviors of some components on the motherboard, and we evaluated the ability of this model to diagnose abnormalities of components of motherboard sensors by two experiments. In the first experiment, which compared an immunity-based with a stand alone diagnostic model, we found that the immunity-based diagnostic model outperformed the standalone diagnostic model. In the second experiment, which compared a fully-connected network with a correlation-based network for mutually testing the credibility of sensors, we found that the correlation-based network improved the diagnosis accuracy in all test cases.

In addition, we utilized a hybrid model, consisting of the standalone and immunity-based diagnostic models, to diagnose nodes connected to the network, as well as nodes isolated from the network. The accuracy of hybrid diagnosis, however, depends on the stand alone diagnosis for the isolated nodes. In future, we will attempt to improve the accuracy of diagnosis of isolated nodes.

References

1. Tanaka, T., Kawazu, T., Kanda, S.: Computer-assisted Diagnostic System Applied with ANFIS. *Biomedical Fuzzy System Association* 5(1), 49–54 (2003)
2. Ishida, Y.: An immune network approach to sensor-based diagnosis by self-organization, vol. 10, pp. 73–90. *Complex Systems Publication* (1996)
3. Watanabe, Y., Ishida, Y.: Immunity-based Approaches for Self-monitoring in Distributed Intrusion Detection System. In: Palade, V., Howlett, R.J., Jain, L. (eds.) *KES 2003, Part II. LNCS*, vol. 2774, pp. 503–510. Springer, Heidelberg (2003)
4. Ishida, Y.: Designing an Immunity-Based Sensor Network for Sensor-based diagnosis of Automobile Engines. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) *KES 2006. LNCS (LNAI)*, vol. 4252, pp. 146–153. Springer, Heidelberg (2006)
5. Watanabe, Y.: Mutual tests among agents in distributed intrusion detection systems using immunity-based diagnosis. In: *Proc. of AROB 8th 2003*, vol. 2, pp. 682–685 (2003)
6. Jerne, N.K.: The immune system. *Scientific American* 229(1), 52–60 (1973)
7. Ishida, Y.: *Immunity-Based Systems: A Design Perspective*. Springer, Heidelberg (2004)

A Note on Dynamical Behaviors of a Spatial Game Operated on Intercrossed Rules

Kouji Harada and Yoshiteru Ishida

Toyohasi University of Technology,
1-1, Tenpaku, Toyohashi-shi, Aichi 441-8585, Japan
{harada,ishida}@tutkie.tut.ac.jp

Abstract. The present study discusses a twist in hierarchical rules embedded in a spatial game system. In the typical spatial games, each player has its own strategy, and determines its action on the strategy; its action does not determine its strategy. On the dominance relationship between rules, the strategy is an upper-level rule against the action. This means the game has a hierarchy on rules. The present study in order to discuss a twist between rules, introduced a rule determining a player's strategy from its immediate neighbor's actions. This introduction results in not to decide which of a strategy and an action is placed in an upper-level. The present paper reports the results about dynamical behaviors observed in a spatial game system with a twist in its hierarchical rules.

1 Introduction

In the game theory, a player has its own strategy and decides its action on the strategy. This means a strategy is an upper-level rule against an action (because an action does not decide a strategy). What would occur in a game system if a rule that an action determines an strategy is introduced into the system. Then an action is placed in an upper-level against a strategy. As the result, a strategy is placed on the upper-level as well as the lower-level against an action. Namely a twist of logical levels on rules occurs in the game system.

We take some examples of systems involved with an intercrossed logical levels. The first example is the immune systems. It is well-known that the antibody (Ab) recognizes the antigen (Ag): Ab is a “recognizing” object; Ag is a “recognized” one. The recognizing object can decide how it recognizes the recognized object; the reverse case does not succeed. Thus the recognizing object against the recognized one is in a dominant position on recognition relationship. In 1974, Jerne proposed an idiotypic network hypothesis [1] in which he pointed out that Ab can recognize the other Ab' . It implies antibody Ab is recognising itself as well as is being recognized by itself. In this situation, we can not say which Ab is in a dominant level on the recognition relationship. This mentions a twist on the recognition levels. The second example is from the art. M.C.Escher left a piece of work, “Drawing hands” in which the left and right hands draw each other. In other words, a hand draws itself through the other hand. This work poses an issue on a twist in the relationship of “drawing” and “being drawn”. The last

example is the “self modifying protocol game” [2]. The crotchety of this game lies in the character that its rules bind player’s action; these rules are varied depending on a spatial pattern of a game board determined by player’s action. This game focus on an issue of a twist in the relationship of a game’s rule and a player’s action.

This present study aims at examining dynamical features of a game system with a twist between rules. For the aim, we propose a spatial game with a twist on determination rules of player’s action and player’s strategy. The twist means a player’s strategy determines a player’s action as well as player’s actions determines a player’s strategy. On the other hand, the general and common spatial game models [3] usually have a loop structure that a player’s strategy determines a player’s action, the player’s action determines a player’s score and the player’s score determines a player’s strategy for the next round. In these models, a scoring process makes the alternate determination process between player’s strategy and its action indirect; in our model the process is direct. Thus to discuss the twist between rules our proposed model is more appropriate than the general ones.

The next section explains details of the spatial game model. The third section shows results from exhaustive investigation on dynamical features such as the periodicity of dynamical attractors and a structure of their basins. The last section discusses some interesting points in results.

2 Spatial Game Model

We consider a two-dimensional $L \times L$ square lattice, and place one player on each lattice-site. To designate each lattice-site, we prepare a horizontal axis with a suffix j and a vertical one with a suffix i ($i, j = 0 \sim L - 1$). Each player has two kind of actions: 0 or 1.

Here we introduce some symbols:

- $P_{i,j}$: the player at the lattice-site (i, j) ,
- $A_{i,j}^r$ (1 and 0): the player $P_{i,j}$ ’s action at the round r ,
- $S_{i,j}^r$: the player $P_{i,j}$ ’s strategy at the round r .

One round of the game composes of the following two steps:

Step1: Each player determines its action based on its own strategy,

Step2: Each player updates its own strategy.

Next, we mention details of the each step.

Step1: Each player determines its action for the next round based on its own strategy. Our strategy is so-called “spatial strategy” [4], and it means that each player’s action is determined by the total amount (k) of players taking “1” in its eight neighborhood players called the *Moore* neighbor (M). The quantity k takes a value from 0 to 8. We show an example of a strategy in table 1. This strategy means when k is 0, 1, 2, 3, 4, 5, 6, 7 and 8, the next action A become 0, 1, 0, 1, 1, 0, 1, 0 and 1, respectively.

Table 1. Spatial strategy. The quantity k is the total amount of players taking the particular action: 1 in the Moore neighbors. This strategy means when k is 0, 1, 2, 3, 4, 5, 6, 7 and 8, the next player’s action, A become 0, 1, 0, 1, 1, 0, 1, 0 and 1, respectively.

k	0	1	2	3	4	5	6	7	8
A	0	1	0	1	1	0	1	0	1

As the quantity k varies depending on each player or each round, we designate k of the player $P_{i,j}$ as $k_{i,j}^r$; $k_{i,j}^r$ is formulated as

$$k_{i,j}^r = \sum_{(i',j') \in M \text{ of } (i,j)} A_{i',j'}^r. \tag{1}$$

Step2: Each player’s strategy at the next round is determined by a spatial configuration of actions of its Moore neighbors. In concrete, the player $P_{i,j}$ ’s strategy $S_{i,j}^{r+1}$ is constituted on the following rule:

- When $k_{i,j}^r$ is 0, then $A_{i,j}^{r+1}$ is equal to $A_{i-1,j-1}^r$: the player’s action in the upper right against the site (i, j) .
- When $k_{i,j}^r$ is 1, then $A_{i,j}^{r+1}$ is equal to $A_{i-1,j}^r$: the player’s action in the right above against the site (i, j) .
- When $k_{i,j}^r$ is 2, then $A_{i,j}^{r+1}$ is equal to $A_{i-1,j+1}^r$: the player’s action in the upper left against the site (i, j) .
- When $k_{i,j}^r$ is 3, then $A_{i,j}^{r+1}$ is equal to $A_{i,j+1}^r$: the player’s action in the right against the site (i, j) .
- When $k_{i,j}^r$ is 4, then $A_{i,j}^{r+1}$ is equal to $A_{i+1,j+1}^r$: the player’s action in the lower right against the site (i, j) .
- When $k_{i,j}^r$ is 5, the $A_{i,j}^{r+1}$ is equal to $A_{i+1,j}^r$: the player’s action in the right below against the site (i, j) .
- When $k_{i,j}^r$ is 6, then $A_{i,j}^{r+1}$ is equal to $A_{i+1,j-1}^r$: the player’s action at the left lower against the site (i, j) .
- When $k_{i,j}^r$ is 7, then $A_{i,j}^{r+1}$ is equal to $A_{i,j-1}^r$: the player’s action at the left against the site (i, j) .
- When $k_{i,j}^r$ is 8, then $A_{i,j}^{r+1}$ is equal to $A_{i,j}^r$: the player, $P_{i,j}$ ’s action.

Namely, player $P_{i,j}$ ’s strategy, $S_{i,j}^{r+1}$ at the round $r + 1$ is represented as a table (see. Table 2).

Figure 1 displays that on the present coding rule, a concrete process to construct the strategy, $S_{1,1}^{r+1}$ from player’s actions.

Table 2. Player $P_{i,j}$ ’s strategy, $S_{i,j}^{r+1}$ at the round $r + 1$

$k_{i,j}^r$	0	1	2	3	4	5	6	7	8
$A_{i,j}^{r+1}$	$A_{i-1,j-1}^r$	$A_{i-1,j}^r$	$A_{i-1,j+1}^r$	$A_{i,j+1}^r$	$A_{i+1,j+1}^r$	$A_{i+1,j}^r$	$A_{i+1,j-1}^r$	$A_{i,j-1}^r$	$A_{i,j}^r$

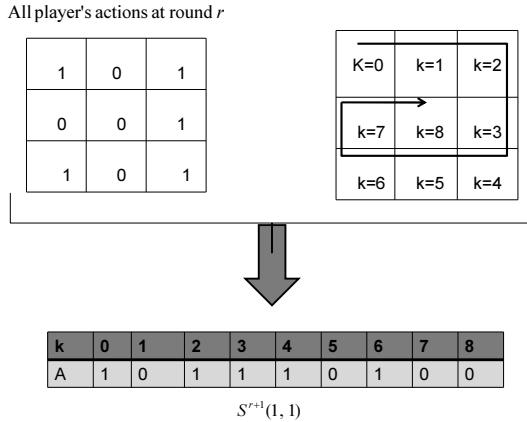


Fig. 1. A constituent process of the strategy, $S_{1,1}^{r+1}$ of $P_{1,1}$. $S_{1,1}^{r+1}$ is constituted by a combination of info. on a spatial configuration of Moore neighbor's actions of $P_{1,1}$ at the round r and the k -rule that k is spirally-arranged in a clockwise direction.

By the way, we note that the common spatial games [3] have one more step between the step 1 and 2. In the extra step, each player receives a score determined by a combination of its action and an action of each player in its Moore neighborhood. In these games, the score info. is very important because it is utilized to update a player's strategy. On the other hand, in our proposed game, the extra step is omitted because the game does not use the score information to update the strategy.

3 Dynamical Features

In order to examine complete structures of all dynamical attractors, choosing the lattice-size L to be its minimum value: three, we prepare a two dimensional 3×3 square lattice. Thus there are nine players in the game. A boundary condition of the square lattice space is periodic so that its shape becomes a torus.

As identifying a spatial configuration on all player's actions as a spatial pattern of 0's and 1's, we can consider one-round game as a process that a spatial bit pattern transits into some spatial bit pattern. The transition process is completely deterministic.

The total number of possible spatial bit patterns is $512 (= 2^9)$. We examined which spatial bit pattern transits into which dynamical attractors.

3.1 Periodicity of Dynamical Attractors

Through exhaustive analyses, we figured out the total number of dynamical attractors is 46: the proposed game system has multi-attractors; these periodicity are classified as four types: fixed point, period-3, period-6 and period-9.

We here defines the following symbols:

- $A, A1, A1', A1''$: each of them represents a different spatial bit pattern,
- $A \rightarrow B$: a spatial bit pattern A transits a pattern B after one-round game.

Also we define “basin size” of a dynamical attractor as an amount of spatial bit patterns which transit to the attractor.

Fixed Point Attractor. Two fixed-point attractors (sinks) exist, the one is a case of all player’s actions are 1 and another one is a case of all player’s are 0. Figure 2(a) displays a structure of the fixed-point attractor. Its basin size is three, and it is very narrow if we take it into account that 512 spatial bit patterns exist in the system.

We think the existence of the fixed-point attractor represented by all-1 spatial pattern is very interesting because if we consider the action 1 and 0 as “cooperation” and “defection” in the context of the Prisoner’s Dilemma game, its existence suggests that all players can cooperate; however, to achieve the cooperative state, there is a limit to choose the initial configuration on player’s actions from the three configurations included in the basin.

Period-3 Attractor. Figures 3 and 4 display diverse basin structures on period-3 attractors. The basin structures are quite different on those size: 0 (Fig. 3(a)), 3 (Fig. 3(b)), 6 (Fig. 3(c)), 9 (Fig. 3(d)) and 54 (Fig. 4); however, all of the observed basin structures hold symmetrical.

Here we especially focus on a specific attractor with extremely broad basin size: 54. Concerning this basin size, two attractors exist. Thus the total basin size of the two attractors is 108. This means that if we start a game with a randomly selected initial configuration on player’s actions, with a comparatively high probability: 0.22¹, which of those attractors appears. In other words, games represented by these attractors frequently occur. On the other hand, if we pay attention to “branches” in Fig. 4, each periodic point of the attractor has three paths to itself. Any path takes five rounds to reach the periodic point. If we call it “transient steps” that the number of round to reach the periodic point, this attractor’s transient steps are five, and is longer than that of the other period-3 attractors.

Period-6 Orbit. Only one orbit was observed (Fig. 2(b)), and its basin size is zero (therefore it is not an attractor). If we start a game from a randomly selected initial configuration, the probability of a game represented by this orbit starting is quite low; it is still open question why the number of period-6 orbits is very low.

Period-9 Attractor. Observed the longest period of the present model is nine. Six attractors exist, and these are classified into two types which differ in these basin sizes: 15 (the left in Fig. 5) and 21 (the right in Fig. 5). Each three attractors belongs in each type. The probability that one of period-9 attractors is observed when we starts a game from a random initial configuration of player’s actions is 0.32². Thus period-9 attractor as well as the period-3 attractor with the largest basin size:54 appears with a comparatively high probability.

¹ $((54 + 3) \times 2) / 512 \approx 0.22$

² $((15 + 9) \times 3 + (21 + 9) \times 3) / 512 \approx 0.32$

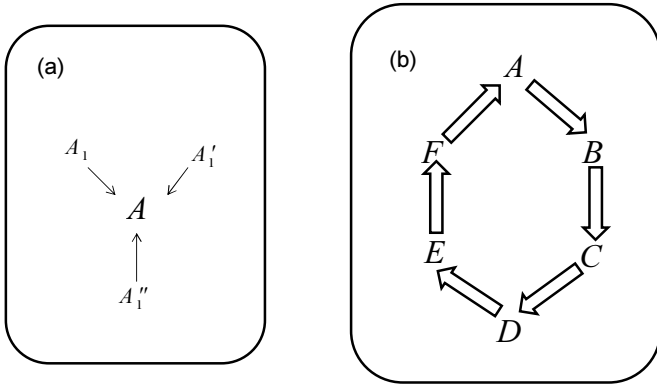


Fig. 2. Fixed-point attractor (a) Period-6 orbit (b)

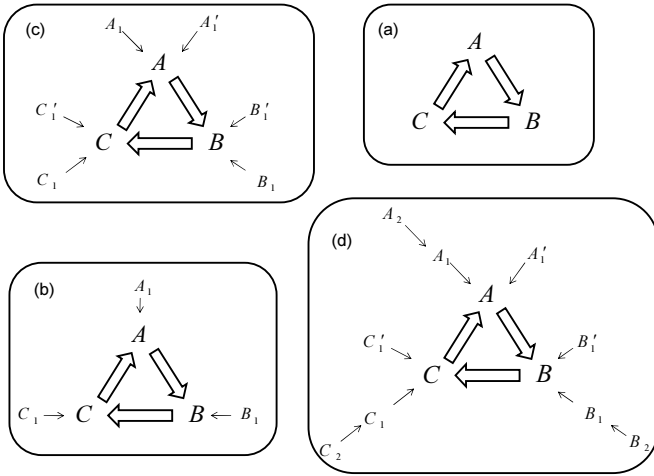


Fig. 3. Period-3 orbit (a), period-3 attractor with its basin size: 3 (b), period-3 attractor with its basin size: 6 (c), period-3 attractor with its basin: 9 (d)

3.2 Statistical Feature of Basin Size

Figure 6 shows a relationship of the number of attractors with a certain basin size versus the basin size. Interestingly the basin size is only in multiples of three. We figured out that the distribution in Fig. 6 represented by a double logarithmic plot was approximated in a straight line, thus it was a power law distribution (critical exponent: -0.67). This result means many attractors with small basin size exist; slight number of attractors with a quite large basin size exist.

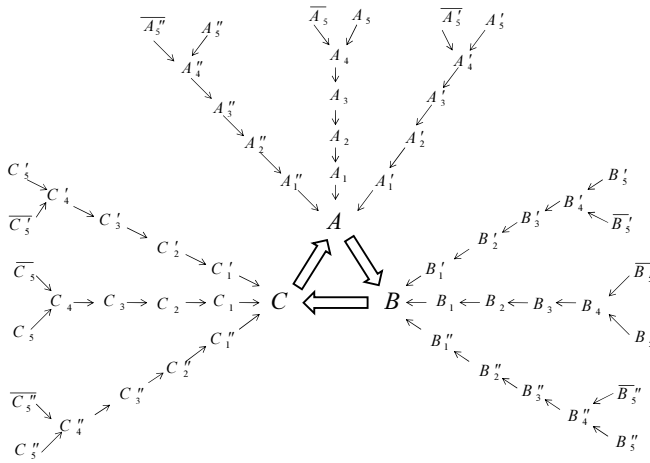


Fig. 4. Period-3 attractor with its basin size: 54

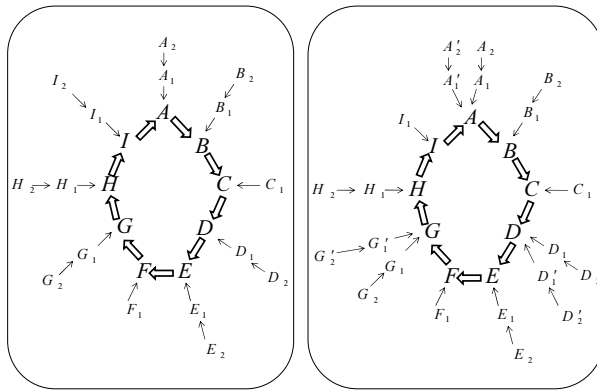


Fig. 5. Period-9 attractors

4 Discussions

We discuss to correspond the proposed game model with a model of spatially configured living cell population. One possible correspondence relation between these is followings:

- Each lattice-site corresponds with a cell.
- Player's action 0 (1) corresponds with an inactive (active) state of a cell respectively.

On this correspondence, we can consider the proposed model as a spatially configured living cell population model which describes each cell's activity dynamically changes depending on its neighbor s's activity state. It is interesting that a

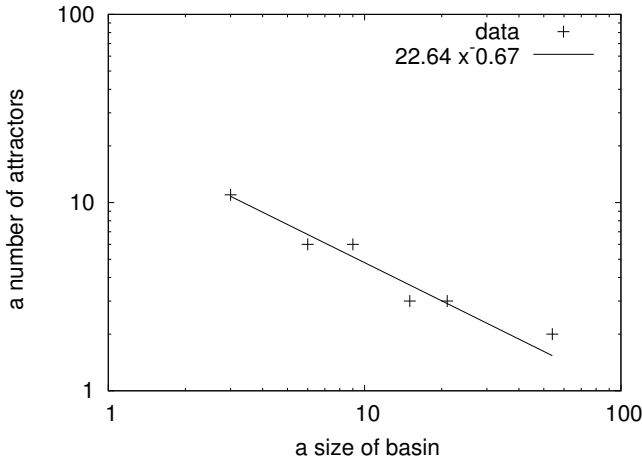


Fig. 6. A relationship of the number of attractors with a certain basin size versus the basin size

group of only nine cells has 46 dynamical attractors. Namely a number of the cell population's activity patterns amounts to 46; however an appearance ratio of an activity pattern is proportional to its basin size. We can imagine that a living cell population become desirable and acceptable for surviving of an individual organism if its activity patterns with a high appearance ratio represent its "normal" state; the other cell population's activity patterns with a low appearance ratio "abnormal" state.

In order to examine how a twist between rules in a system affects its dynamical behaviors, the present study dealt with the spatial game with the alternate determination process between a player's strategy and a player's action. However this study's results are not enough for its original purpose. To step this study ahead for the original purpose, we are necessary to review carefully the present result in comparison with that of when a spatial game does not include any twist in rules, in other words, a player's strategy determines a player's action but the inverse determination process is not prepared. In the future, from this point of view, we will advance our own research.

References

1. Jerne, N.K.: Towards a network theory of the immune system. *Annals d'Immunologie* 125C, 373–389 (1974)
2. Hofstadter, D.R.: *GÖDEL, ESCHER, BACH*. Basic Books, Inc., New York (1979)
3. Nowak, M.A., May, R.M.: Evolutionary games and spatial chaos. *Nature* 356, 826–829 (1992)
4. Ishida, Y., Katsumata, Y.: A Note on Space-Time Interplay through Generosity in a Membrane Formation with Spatial Prisoner's Dilemma. *LNCS*, vol. 5917, pp. 448–455. Springer, Heidelberg (2008)

Asymmetry in Repairing and Infection: The Case of a Self-repair Network

Yoshiteru Ishida and Kei-ichi Tanabe

Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology
Tempaku, Toyohashi 441-8580, Japan
<http://www.sys.cs.tut.ac.jp>

Abstract. A self-repair network is a model consisting of autonomous nodes capable of repairing connected nodes. Self-repairing by mutual repair involves the “double-edged sword” problem where repairing could cause adverse effects when done by abnormal nodes. Although self repair (as opposed to mutual repair) is not susceptible to this problem because an abnormal node always repairs itself, the possibility of normal nodes repairing abnormal nodes is lost. This note compares these two types of repair: mutual and self for a self-repair network with infections. With this extended model, abnormal nodes can spread not only by repair failures but also by infections.

Keywords: self-repairing network, distributed autonomy, anti-virus program, infection model.

1 Introduction

Recent progress in network technology has made possible not only huge computer networks but also several innovative systems on them such as cloud computing, grid computing [1] and parasitic computing [2]. However, large-scale information systems involve such risks as large-scale malfunctions and even outbreaks of infections. When systems become too large to be dealt with by only a central authority, distributed autonomy will be essential. Although distributed autonomy incurs the risk of malicious agents other than machine failures, it is useful not only for control and management purposes but also for robustness. Further, computer viruses exploit networked computers as infection paths, and hence anti-virus programs must be installed on the computers in a distributed autonomous fashion rather than a central management fashion.

Regarding distributed autonomy, we examined a network cleaning problem and proposed a self-repair network model [3, 4] that can be examined using knowledge of probabilistic cellular automata [5]. This note reports on the model involving infection in addition to repairing (whose effect seems asymmetric to the infection).

Section 2 revisits the self-repair network and extends it by involving infections as well as repair. Section 3 examines the model with a steady-state analysis and a computer simulation. Section 4 discusses the implications of the model and simulation considering the current situation of computer networks.

2 Basic Model

2.1 Definitions and Extensions

The self-repair network consists of autonomous nodes capable of repairing neighbor nodes (i.e. connected nodes) by copying their contents. Each node has a binary state: normal (0) and abnormal (1). Each node repairs the neighbor nodes with a probability P_r (called the repair rate). The repair will succeed with a probability P_m (called the repair success rate by normal nodes) when it is done by a normal node, but with a probability P_{ra} (called the repair success rate by abnormal nodes) when done by an abnormal node.

Repair may be divided into two types depending on the target of repair: *self repair* (Fig. 1 (a)) targets the repairing node itself; *mutual repair* (Fig. 1 (b)) targets the node connected to the repairing node. In the self repair by each node, there is no interaction among nodes and hence the transition is trivial (Table 1). In the state transition, the self-state is shown in parentheses. The self-state will be changed to the state indicated to the right of the arrow.

Table 1. Transition probability in each state transition (self repair)

State Transition	Transition Probability
(0)→1	$P_r(1 - P_m)$
(1)→1	$P_r(1 - P_{ra}) + (1 - P_r)$

Mutual repair can be further subdivided into two types: AND repair and OR repair. In AND repair, all the repairs must be successful when the repairs are done by multiple nodes simultaneously, while OR repair requires at least one repair to be successful out of multiple repairs done simultaneously. We consider only AND repair here, for OR repair can eradicate abnormal nodes when repair done by normal nodes always succeeds ($P_m = 1$).

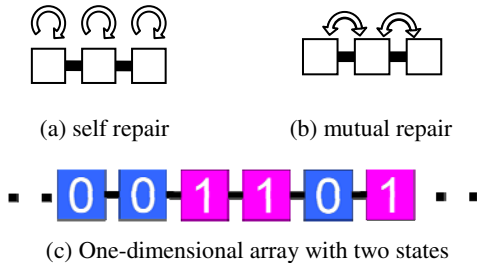


Fig. 1. (a) self repair; (b) mutual repair; (c) One-dimensional array with two states: normal (0) and abnormal (1). Each node is connected to two nodes, one on either side (neighbor nodes).

In a one-dimensional array with two adjacent neighbor nodes (Fig. 1 (c)), the probabilities for each state transition are listed in Table 2. In the state transition, the self-state is the center in parentheses and the two neighbor states are on the left and the

Table 2. Transition probability in each state transition

State Transition	Transition Probability (mutual AND repair)	Transition Probability (infection)
(000) →1	$P_r(1 - P_{rn})(2 - P_r + P_r P_{rn})$	0
(001) →1	$P_r^2(1 - P_{rn} P_{ra}) + P_r(1 - P_r)((1 - P_{rn}) + (1 - P_{ra}))$	P_i
(101) →1	$P_r(1 - P_{ra})(2 - P_r + P_r P_{ra})$	$P_i(2 - P_i)$
(010) →1	$1 - P_r P_{rn}(2(1 - P_r) + P_r P_{rn})$	1
(011) →1	$1 - P_r((P_{rn} + P_{ra})(1 - P_r) + P_r P_{ra} P_{rn})$	1
(111) →1	$1 - P_r P_{ra}(2(1 - P_r) + P_r P_{ra})$	1

Table 3. Transition probability in each state transition (infection before repair)

State Transition	Transition Probability (mutual AND repair)	Transition Probability (self repair)
(000) →1	$P_r(1 - P_{rn})(2 - P_r + P_r P_{rn})$	$P_r(1 - P_{rn})$
(001) →1	$P_r^2(1 - P_{rn} P_{ra}) + P_r(1 - P_r)((1 - P_{rn}) + (1 - P_{ra})) + P_i(1 - P_r)^2$	$(1 - P_i)P_r(1 - P_{rn}) + P_i(1 - P_r P_{ra})$
(101) →1	$P_r(1 - P_{ra})(2 - P_r + P_r P_{ra}) + P_i(2 - P_i)(1 - P_r)^2$	$(1 - P_i(2 - P_i))P_r(1 - P_{rn}) + P_i(2 - P_i)(1 - P_r P_{ra})$
(010) →1	$1 - P_r P_{rn}(2(1 - P_r) + P_r P_{rn})$	$1 - P_r P_{ra}$
(011) →1	$1 - P_r((P_{rn} + P_{ra})(1 - P_r) + P_r P_{ra} P_{rn})$	$1 - P_r P_{ra}$
(111) →1	$1 - P_r P_{ra}(2(1 - P_r) + P_r P_{ra})$	$1 - P_r P_{ra}$

right in parentheses. The self-state will be changed to the state indicated to the right of the arrow. (001)→1, for example, indicates that the normal node with right neighbor abnormal and left neighbor normal will change to abnormal with the probability stated in Table 2.

When infection is involved in a one-dimensional array with a probability P_i (called infection rate), the probability of each state transition is listed in the right-most column of Table 2. Infection occurs only when at least one infected node exists in the neighborhood.

2.2 Repair Coupled with Infection

When repair is coupled with infection, the order of the repair and infection counts, that is, whether infection is after repair or before repair leads to different results. This note focuses on the case in which infection occurs before repair (Table 3), for simulations indicated that the case of repair before infection makes little difference.

3 Analysis and Simulations

3.1 Steady-State Analysis

Under the approximation that the probability of the state of a node being abnormal is constant and equated with a density ρ_1 ($\rho_0 = 1 - \rho_1$) of abnormal nodes (mean field approximation and steady state), the following differential equation is obtained:

$$\frac{d\rho_1}{dt} = A\rho_1^3 + B\rho_1^2 + C\rho_1 + D$$

where the coefficients A , B , C and D are constants determined by the parameters of the self-repair network (Table 4).

Three roots are obtained in the third-order algebraic equation given by the steady state, that is the roots of $A\rho_1^3 + B\rho_1^2 + C\rho_1 + D = 0$. For simplicity, we limit ourselves to the case that the repair by normal nodes always succeeds ($P_m = 1$). These three roots are fixed points of the differential equation above, and we can obtain the density of abnormal nodes in the steady state determined by the stable point corresponding to the root. Figures 2 and 3 plot the density obtained by a numerical study of the root of the above algebraic equation.

Although mutual repair involves both a single repair (repair from one node) and a simultaneous repair (repair from two nodes, one on either side), self repair includes only the single repair (repair by itself). Under the mean field approximation, averaged repair success rates for the following two cases are:

- $\{(1 - \rho_1) + P_{ra}\rho_1\}^2$: simultaneous repair of a node by two nodes, one on either side;
- $(1 - \rho_1) + P_{ra}\rho_1$: single repair of a node by one node on one side.

In the mutual repair with AND repair scheme, both simultaneous repair and single repair occur (although in the self repair, only single repair occurs). Because the repair success rate of single repair is greater than that of simultaneous repair, single repair is

Table 4. Coefficients of the equation expressed by parameters of the self-repair network (infection before repair)

Con- stants	Constants Expressed by Parameters (mutual AND repair)	Constants Expressed by Parameters (self repair)
A	$P_i^2(1-P_r)^2$	$P_i^2\{(1-P_rP_{ra})-P_r(1-P_m)\}$
B	$-P_r^2(P_m-P_{ra})^2-P_i(2+P_i)(1-P_r)^2$	$-P_i(2+P_i)\{(1-P_rP_{ra})-P_r(1-P_m)\}$
C	$-2P_r(1-P_m)(P_r(P_m-P_{ra})+1)+P_r(P_r-2P_{ra})+2P_i(1-P_r)^2$	$-P_r(1+2P_i)(1-P_m+P_{ra})+2P_i$
D	$P_r(1-P_m)(2-P_r+P_rP_m)$	$P_r(1-P_m)$

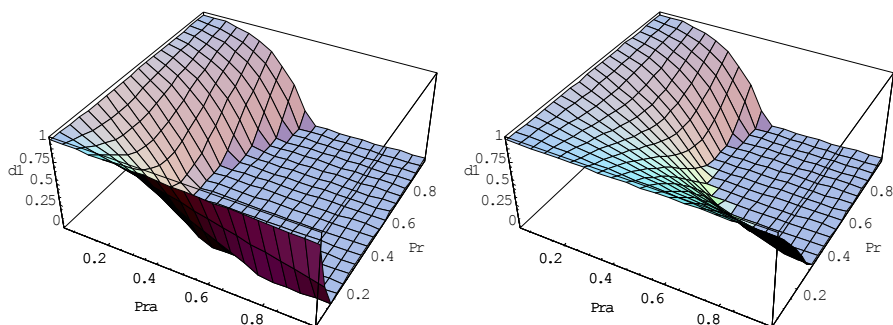


Fig. 2. Density of abnormal nodes (d_i) plotted for an infection rate of 0.1 (left) and 0.5 (right) as a function of the repair success rate by abnormal nodes (P_{ra}) and the repair rate (P_r) in mutual repair

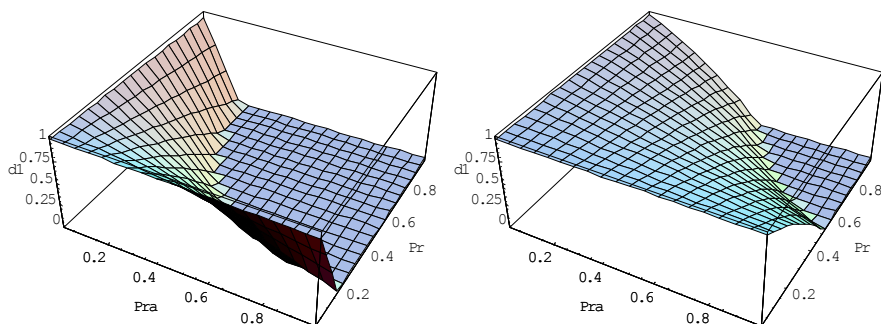


Fig. 3. Density of abnormal nodes (d_i) plotted for an infection rate of 0.1 (left) and 0.5 (right) as a function of the repair success rate by abnormal nodes (P_{ra}) and the repair rate (P_r) in self repair

avored. In mutual repair, since single repair occurs (with a rate $2P_r(1 - P_r)$) more often than simultaneous repair (with a rate P_r^2) when $P_r < 2/3$, mutual repair can utilize repairs more successfully if the same amount of repairs is allowed.

3.2 Simulation Results

Simulations are conducted with the following parameters (Table 5) under the conditions that the repair by normal nodes always succeeds ($P_m = 1$) and infections occur before repairing.

Table 5. Parameters for the simulations of the self-repair network with infection

Number of nodes	1000
Initial number of normal nodes	500
Number of steps	500
Repair rate P_r	0.00–1.00 (in 0.01 increments)
Repair success rate by normal nodes P_m	1.0
Repair success rate by abnormal nodes P_{ra}	0.00–1.00 (in 0.01 increments)
Infection rate P_i	0.0, 0.5, 1.0

Figure 2 shows a phase diagram plotted by the simulations when the infection rate is changed. Each plot indicates when all the abnormal nodes can be eradicated. The curves in each plot separate the area into two parameter regions: lower-left (including the origin) and upper-right (including upper-right $P_i = 1$ and $P_{ra} = 1$). The lower-left region is the *active phase* where some abnormal nodes remain in the network, and the upper-right region is the *frozen phase* where all the nodes become normal.

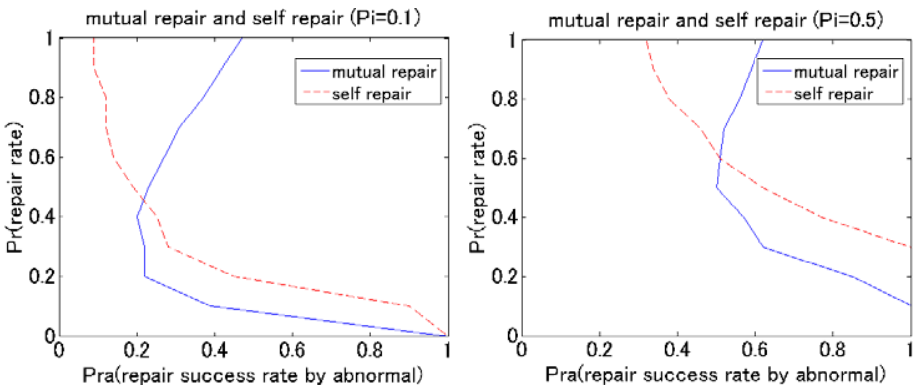


Fig. 4. Active phase (*left region* where some nodes remain abnormal) and frozen phase (*right region* where all the nodes are normal). The solid (dotted) line is the border separating two phases in mutual (self) repair.

As the infection rate increases from 0.0 to 1.0 (only 0.1 and 0.5 shown), these curves are dragged to the upper-right, for the frozen region shrinks due to the infection. These simulation results match those of numerical studies based on the mean field approximation (Figs. 2 and 3).

When self repair and mutual (AND) repair are compared, it is observed that self repair outperforms mutual repair when the repair rate is high, while mutual repair outperforms when the repair rate is low.

4 Discussion

It must be noted that the model with self repair and infection still does not reflect real-world situations. In real computer networks, the network topology is neither necessarily a ring nor even a regular structure; several parameters such as infection rate, repair rate, and repair success rate are not uniform; and so on; real-world situations are not that simple. However, a simple model has the merit of yielding some knowledge, which, although it cannot be directly applied to real situations, may be applied in many situations under several limiting conditions.

With these limitations in mind, the self repair with infection reflects the current situation where an anti-virus program is installed in personal computers. Although anti-virus programs are often installed in server-client systems, which may be close to mutual (but one direction) repair, we do not consider the case here.

Since there are so many computer viruses both active and dormant, the infection rate is certainly not zero ($P_i > 0$).

Focusing on self repair, the repair by a normal node will always succeed, hence the repair success rate by a normal node is 1 ($P_m = 1$). Repair by an abnormal node may not succeed, and hence the repair success rate by an abnormal (infected) node is smaller than 1 ($P_{ra} < 1$). However, this parameter will decrease unless signature files are regularly updated (which imposes a cost on each computer). If the repair success rate by abnormal nodes (P_{ra}) is high enough, then the simulation indicates that mutual repair is preferable because it can eradicate abnormal nodes even with a lower repair rate (P_r) and hence low cost.

One difficulty of mutual repair is that distributed autonomy encourages nodes to ignore other nodes' problems (self-governance). However, a game theoretic approach with a systemic payoff that evaluates the costs and benefits throughout the system over the long term indicates that even for selfish agents, they will be more likely to cooperate [4].

References

1. Foster, I., Kesselman, C.: *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers, San Francisco (1999)
2. Barabasi, A.-L., Freeh, J.V.W., Brockman, J.B.: Parasitic computing. *Nature* 412, 894–897 (2000)
3. Ishida, Y.: A Critical Phenomenon in a Self-Repair Network by Mutual Copying. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) *KES 2005. LNCS (LNAI)*, vol. 3682, pp. 86–92. Springer, Heidelberg (2005)
4. Ishida, Y.: Complex Systems Paradigms for Integrating Intelligent Systems. *Studies in Computational Intelligence (SCI)*, vol. 115, pp. 155–181 (2008)
5. Domany, E., Kinzel, W.: Equivalence of cellular automata to Ising models and directed percolation. *Phys. Rev. Lett.* 53, 311 (1984)

A Note on Symmetry in Logic of Self-repair: The Case of a Self-repair Network

Yoshiteru Ishida

Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology
Tempaku, Toyohashi 441-8580, Japan
<http://www.sys.cs.tut.ac.jp>

Abstract. A self-repair network consists of nodes capable of repairing other nodes where the repair success rate depends on the state (normal or abnormal) of the repairing node. This recursive structure leads to the “double-edged sword” of repairing, which could cause outbreaks in case the repairing causes adverse effects. The self-repair network can be equated to a probabilistic cellular automaton. Because of the distinction between repair by normal nodes and that by abnormal nodes, transition probabilities as a probabilistic cellular automaton exhibit symmetry.

Keywords: self-repair network, probabilistic cellular automaton, cost of autonomy, symmetry.

1 Introduction

As a candidate for the autonomous defense and maintenance of information systems, we consider self-repairing of a network by mutual copying [1]. In biological epidemics (e.g. [2]), spontaneous recoveries can occur and even the immune system could arise, because autonomous nodes repair other nodes which leads to increased reliability, availability or longer life time at the system level. However, distributed autonomy, which offers autonomous defense, could cause adverse effects and even complete system failure or outbreaks.

We consider the possibility of cleaning up the network by mutual copying. In information systems, repair by copying is a “double-edged sword” and so it is important to identify under what conditions the network can eradicate abnormal nodes from the system. We consider a probabilistic cellular automaton (p-CA) [3] to model the situation where computers in a LAN perform mutual repair by copying their contents. Since the problem could lead to a critical phenomenon, repairs must be decided by considering the eradication of abnormal nodes and the network environment [1].

Our model consists of nodes capable of repairing other connected nodes. We call the connected nodes “neighbor nodes” based on the terminology of CA. The repairing of a node may be done by overwriting its content to other nodes, since information systems depend on large-scale networks such as LAN-connected computer networks, and the grid system in electric power supply.

Once it became possible on the Internet for many selfish activities to proliferate, several utilities and protocols converged on the Nash equilibrium from which no players want to deviate [4]. Researchers focused on algorithms and computational complexity for obtaining equilibrium when selfish nodes (or agents when emphasizing autonomy) compete for resources. The cost for the Nash equilibrium relative to the optimized solution has also been discussed to measure the cost of “anarchy” [5].

This note focuses on the self maintenance task, and on self-repairs by mutual copying in particular. The double-edged sword of self-repair may be considered to be a cost of autonomy in attaining higher system reliability. However, without an autonomous distributed approach, high reliability comparable to that of biological systems will not be possible.

Section 2 revisits a self-repair network, and extends the model to include not only AND repair but also OR repair. The transition probabilities of these two types of repair are formulated in a symmetric form. Based on the transition probabilities, section 3 presents a mean field approximation to investigate the density of abnormal nodes in the steady state.

2 Basic Model

2.1 Definitions and Extensions

In a mathematical formulation, the model consists of three elements (U, T, R) where U is a set of nodes, T is a topology connecting the nodes, and R is a set of rules of the interaction among nodes. We assume that a set of nodes is a finite set with N nodes, and the topology is restricted to the one-dimensional array shown in Fig. 1 (which could be an n-dimensional array, complete graph, random graph, or even scale-free network) that could have S neighbors for each node. Also, we restrict the study to the case where each node has a binary state: normal (0) or abnormal (1).

A p-CA requires probabilistic rules for interactions. The model controls the repairing of all the nodes uniformly. That is, each node tries to repair the neighbor nodes in a synchronous fashion with a probability P_r (repair rate). The repairing will be successful with the probability P_m when it is done by a normal node, but with the probability P_{ra} when done by an abnormal node. The repaired nodes will be normal when all the repairing is successful (called *AND repair*). Thus, when repairing is done by the two neighbor nodes simultaneously (called *simultaneous repair*), both of these two repairs must be successful in order for the repaired node to be normal.



Fig. 1. One-dimensional array with two states: normal (0) and abnormal (1)

In OR repair, on the contrary, at least one repair must succeed for the simultaneous repair to succeed. Therefore, OR repair can repair more successfully than AND repair for the same number of repairs. The repair rate P_r controls not only the frequency of repairs but also the synchrony of repair actions: if P_r is close to 1, simultaneous repair

is more likely to occur. The transition probabilities as a probabilistic cellular automaton are listed in Table 1. In the state transition, the self-state is the center in parentheses and the two neighbor states are on the left and the right in parentheses. The self-state will be changed to the state indicated to the right of the arrow. (001)→1, for example, indicates that the normal node with right neighbor abnormal and left neighbor normal will change to abnormal with the probability stated in Table 1.

Table 1. Rules for state transition in the probabilistic cellular automaton

State Transition	Transition Probability (AND repair)	Transition Probability (OR repair)
(000)→1	$P_r(1 - P_m)(2 - P_r + P_r P_m)$	$P_r(1 - P_m)(2 - P_r - P_r P_m)$
(001)→1	$P_r^2(1 - P_m P_{ra}) + P_r(1 - P_r)((1 - P_m) + (1 - P_{ra}))$	$P_r(2 - P_r - P_m - P_{ra} + P_r P_m P_{ra})$
(101)→1	$P_r(1 - P_{ra})(2 - P_r + P_r P_{ra})$	$P_r(1 - P_{ra})(2 - P_r - P_r P_{ra})$
(010)→1	$1 - P_r P_m(2(1 - P_r) + P_r P_m)$	$(1 - P_r P_m)^2$
(011)→1	$1 - P_r((P_m + P_{ra})(1 - P_r) + P_r P_{ra} P_m)$	$1 - P_r(P_m + P_{ra} - P_r P_m P_{ra})$
(111)→1	$1 - P_r P_{ra}(2(1 - P_r) + P_r P_{ra})$	$(1 - P_r P_{ra})^2$

The transition probability satisfies several symmetries and asymmetries when they are viewed as a function of the repair parameters (P_r, P_m, P_{ra}) . Let $R^{\wedge}_{xyz}(P_r, P_m, P_{ra})$ ($R^{\vee}_{xyz}(P_r, P_m, P_{ra})$) denote the transition probability from the configuration (xyz) to (xIz) as a function of the parameters when the AND repair (OR repair) is applied where $x, y, z \in \{0,1\}$. For example, $R^{\wedge}_{000}(P_r, P_m, P_{ra}) = P_r(1 - P_m)(2 - P_r + P_r P_m)$. The bar as in \bar{p} indicates negation (flipping probability p to $1-p$). Then, the following properties can be observed.

Duality in P_m, P_{ra} exchange:

Dual: $R^{\wedge}_{1y1}(P_r, P_{ra}, P_m) = R^{\wedge}_{0y0}(P_r, P_m, P_{ra})$, $R^{\wedge}_{0y0}(P_r, P_{ra}, P_m) = R^{\wedge}_{1y1}(P_r, P_m, P_{ra})$

Self-dual: $R^{\wedge}_{0y1}(P_r, P_{ra}, P_m) = R^{\wedge}_{0y1}(P_r, P_m, P_{ra})$, $R^{\wedge}_{1y0}(P_r, P_{ra}, P_m) = R^{\wedge}_{1y0}(P_r, P_m, P_{ra})$

Non-sensitive: $R^{\wedge}_{0y0}(P_r, P_m, P_{ra}) = R^{\wedge}_{0y0}(P_r, P_m, -)$, $R^{\wedge}_{1y1}(P_r, P_m, P_{ra}) = R^{\wedge}_{1y1}(P_r, -, P_{ra})$

Degenerate: $R^{\wedge}_{0y1}(P_r, P_m, P_{ra}) = R^{\wedge}_{0y0}(P_r, P_m, P_{ra})$, $R^{\wedge}_{0y1}(P_r, P_{ra}, P_{ra}) = R^{\wedge}_{1y1}(P_r, P_m, P_{ra})$

Duality in P_m, P_{ra} flip:

$$R^{\wedge}_{xyz}(P_r, P_m, P_{ra}) = \overline{(R^{\vee}_{xyz}(P_r, \overline{P_m}, \overline{P_{ra}}))}$$
, $R_{xyz}((0), P_m, P_{ra}) = \overline{(R_{xyz}((0), \overline{P_m}, \overline{P_{ra}}))}$

When expressed in a form similar to *de Morgan's* theorem:

$$\overline{(R^{\wedge}_{xyz}(P_r, P_m, P_{ra}))} = (R^{\vee}_{xyz}(P_r, \overline{P_m}, \overline{P_{ra}}))$$
, $\overline{(R_{xyz}((0), P_m, P_{ra}))} = R_{xyz}((0), \overline{P_m}, \overline{P_{ra}})$

The involvement of probability in logic reveals a continuous structure of the logical operators AND and OR. Indeed, we can observe when the distinction between AND and OR disappears. It is also noted that the repair rate P_r can be an indicator of synchrony in repairing: the repair is synchronous when it is close to 1, and asynchronous when close to 0 (measured by the probability of the simultaneous repair).

Although it is tempting to formulate the self-repair network based on the formulation (starting from the algebra satisfying the above), lack of space precludes us from doing so here.

Since AND repair and OR repair match when the second-order terms of P_r are neglected (Table 2), we will use $R_{xyz}((0), P_m, P_{ra})$ for both of them:

$$R^{\vee}_{xyz}((0), P_m, P_{ra}) = R^{\wedge}_{xyz}((0), P_m, P_{ra}) \equiv R_{xyz}((0), P_m, P_{ra})$$

Using $R_{xyz}((0), P_m, P_{ra})$ which appears both in AND repair and OR repair, they can be formulated in a simple form (Table 3).

Table 2. Rules for state transition in the probabilistic cellular automaton (first order approximation)

State Transition	Transition Probability (AND OR repair common) $P_r \equiv 0$ (the second-order terms of P_r are neglected)	Transition Probability (AND repair) $P_r \equiv 1$ (the second-order terms of $\overline{P_r}$ are neglected)	Transition Probability (OR repair) $P_r \equiv 1$ (the second-order terms of $\overline{P_r}$ are neglected)
(000) → 1	$2P_r \overline{P_m}$	$R_{xyz}((0), P_m, P_{ra}) - (\overline{P_m})^2 (P_r - \overline{P_r})$	$R_{xyz}((0), P_m, P_{ra}) - (\overline{P_m})(1 + P_m)(P_r - \overline{P_r})$
(001) → 1	$P_r (\overline{P_m} + \overline{P_{ra}})$	$R_{xyz}((0), P_m, P_{ra}) - (\overline{P_m})(\overline{P_{ra}})(P_r - \overline{P_r})$	$R_{xyz}((0), P_m, P_{ra}) - (1 - P_m P_{ra})(P_r - \overline{P_r})$
(101) → 1	$2P_r \overline{P_{ra}}$	$R_{xyz}((0), P_m, P_{ra}) - (\overline{P_{ra}})^2 (P_r - \overline{P_r})$	$R_{xyz}((0), P_m, P_{ra}) - (\overline{P_{ra}})(1 + P_{ra})(P_r - \overline{P_r})$
(010) → 1	$1 - 2P_r P_m$	$R_{xyz}((0), P_m, P_{ra}) + P_m (1 + \overline{P_m})(P_r - \overline{P_r})$	$R_{xyz}((0), P_m, P_{ra}) + P_m^2 (P_r - \overline{P_r})$
(011) → 1	$1 - P_r (P_m + P_{ra})$	$R_{xyz}((0), P_m, P_{ra}) + (P_{ra} + P_m \overline{P_{ra}})(P_r - \overline{P_r})$	$R_{xyz}((0), P_m, P_{ra}) + P_m P_{ra} (P_r - \overline{P_r})$
(111) → 1	$1 - 2P_r P_{ra}$	$R_{xyz}((0), P_m, P_{ra}) + P_{ra} (1 + \overline{P_{ra}})(P_r - \overline{P_r})$	$R_{xyz}((0), P_m, P_{ra}) + P_{ra}^2 (P_r - \overline{P_r})$

It is first noted that $R_{xyz}((0), P_{ra}, P_m)$ has symmetry in flipping P_m and P_{ra} :

$$R_{xyz}((0), P_m, P_{ra}) = \overline{R_{xyz}((0), \overline{P_m}, \overline{P_{ra}})}$$

Further, the following inequalities hold:

$$R_{x0z}((0), P_m, P_{ra}) > R^{\wedge}_{x0z}((1), P_m, P_{ra}) > R^{\vee}_{x0z}((1), P_m, P_{ra})$$

$$R^{\wedge}_{x1z}((1), P_m, P_{ra}) > R^{\vee}_{x1z}((1), P_m, P_{ra}) > R_{x1z}((0), P_m, P_{ra})$$

Table 3. Rules for state transition in the probabilistic cellular automaton

State Transition	Transition Probability $R_{xyz}((0), P_r, P_m, P_{ra})$	Transition Probability (AND repair) $R^{\wedge}_{xyz}(P_r, P_m, P_{ra})$	Transition Probability (OR repair) $R^{\vee}_{xyz}(P_r, P_m, P_{ra})$
(000) → 1	$2P_r \overline{P_m}$	$R_{xyz}((0), P_r, P_{ra}) - P_r^2 (\overline{P_m})^2$	$R_{xyz}((0), P_r, P_{ra}) - P_r^2 (\overline{P_m})(1 + P_m)$
(001) → 1	$P_r (\overline{P_m} + \overline{P_{ra}})$	$R_{xyz}((0), P_r, P_{ra}) - P_r^2 (\overline{P_m})(\overline{P_{ra}})$	$R_{xyz}((0), P_r, P_{ra}) - P_r^2 (1 - P_m P_{ra})$
(101) → 1	$2P_r \overline{P_{ra}}$	$R_{xyz}((0), P_r, P_{ra}) - P_r^2 (\overline{P_{ra}})^2$	$R_{xyz}((0), P_r, P_{ra}) - P_r^2 (\overline{P_{ra}})(1 + P_{ra})$
(010) → 1	$1 - 2P_r P_m$	$R_{xyz}((0), P_r, P_{ra}) + P_r^2 P_m (1 + \overline{P_m})$	$R_{xyz}((0), P_r, P_{ra}) + P_r^2 P_m^2$
(011) → 1	$1 - P_r (P_{ra} + P_{ra})$	$R_{xyz}((0), P_r, P_{ra}) + P_r^2 (P_{ra} + P_m \overline{P_{ra}})$	$R_{xyz}((0), P_r, P_{ra}) + P_r^2 P_m P_{ra}$
(111) → 1	$1 - 2P_r P_{ra}$	$R_{xyz}((0), P_r, P_{ra}) + P_r^2 P_{ra} (1 + \overline{P_{ra}})$	$R_{xyz}((0), P_r, P_{ra}) + P_r^2 P_{ra}^2$

From this Table 3, the following inequalities are observed:

$$R_{x0z}((0), P_r, P_m, P_{ra}) > R^{\wedge}_{x0z}(P_r, P_m, P_{ra}) > R^{\vee}_{x0z}(P_r, P_m, P_{ra})$$

$$R^{\wedge}_{x1z}(P_r, P_m, P_{ra}) > R^{\vee}_{x1z}(P_r, P_m, P_{ra}) > R_{x1z}((0), P_r, P_{ra})$$

2.2 Related Models

The Domany-Kinzel (DK) model [2] is a one-dimensional two-state and totalistic p-CA in which the interaction timing is specific. The interaction is done in an alternated synchronous fashion: the origin cell with state 1 is numbered as 0. The numbering proceeds {1,2,... } to the right, and {-1,-2,... } to the left. At the N-th step the even numbered cells will act on the odd numbered cells and the odd numbered cells will act at the next step. The neighbors are the two cells adjacent to oneself without self-interaction. The interaction rule is as shown in Table 4.

Table 4. Rules for the DK model where p_1 and p_2 are two parameters for the DK model and symbol * is a wildcard

State Transition	Transition Probability
(0*0) → 0	1
(0*1) → 1	p_1
(1*1) → 1	p_2

The self-repair network with AND repair can be equated to the DK model [2] when $P_r = 1$ (i.e. nodes always repair) with the parameters $p_1 = 1 - P_{ra}$, $p_2 = 1 - P_{ra}^2$, i.e. the case of the directed bond percolation.

3 Steady-State Analysis

3.1 Mean Field Approximation

Let ρ_1 ($\rho_0 = 1 - \rho_1$) be a mean field approximation of the density of abnormal nodes (normal nodes). The dynamics of ρ_1 can be described by the following equation by letting a, b, c, d, e and f denote the transition probabilities $(000) \rightarrow 1, (001) \rightarrow 1, (101) \rightarrow 1, (010) \rightarrow 0, (011) \rightarrow 0,$ and $(111) \rightarrow 0,$ respectively.

$$\frac{d\rho_1}{dt} = a\rho_0^3 + 2b\rho_0^2\rho_1 + c\rho_0\rho_1^2 - d\rho_0^2\rho_1 - 2e\rho_0\rho_1^2 - f\rho_1^3$$

By eliminating ρ_0 , we obtain the following equation with only ρ_1 :

$$\frac{d\rho_1}{dt} = A\rho_1^3 + B\rho_1^2 + C\rho_1 + D$$

where A, B, C and D are constants determined by the three parameters of the self-repairing network as follows:

$$\begin{aligned} A &= -(a + c + d + f) + 2(b + e), \\ B &= (3a + c + 2d) - 2(2b + e), \\ C &= -(3a + d) + 2b, \\ D &= a. \end{aligned}$$

As in the transition probability (Table 5), a first-order approximation again makes no difference between AND repair and OR repair.

Table 5. Coefficients of the equation expressed by parameters of the self-repair network

Con- stants	Constants Expressed by Parameters (AND repair)	Constants Expressed by Parameters (OR repair)
A	0	0
B	$-P_r^2(P_m - P_{ra})^2$	$P_r^2(P_m - P_{ra})^2$
C	$2P_r(P_m - P_{ra} - 1) +$ $P_r^2\{-2(1 - P_m)(P_m - P_{ra}) + 1\}$	$2P_r(P_m - P_{ra} - 1) +$ $P_r^2\{-2P_m(P_m - P_{ra}) + 1\}$
D	$2P_r(1 - P_m) - P_r^2(1 - P_m)^2$	$2P_r(1 - P_m) - P_r^2(1 - P_m)^2$

The mean field approximation of the ratio of abnormal nodes ρ_1 may be obtained by solving a second-order algebraic equation; however, the root can be a complicated form in general. Nevertheless, some of them can be simple in specific cases.

First, when $P_{rn} = 1$ both AND repair and OR repair have a simple solution. For abnormal nodes eradicated in AND repair [6], the condition $\frac{P_{ra}}{P_r} \geq \frac{1}{2}$ must be satisfied, since the time derivative $\frac{d\rho_1}{dt}$ must be negative. Otherwise, the density of abnormal nodes converges on $\frac{P_r - 2P_{ra}}{P_r(1 - P_{ra})^2}$.

In OR repair, for example, the density of abnormal nodes converges on 0. This means that OR repair can eradicate all the abnormal nodes at $P_{rn} = 1$ regardless of the values of P_r and P_{ra} .

When $P_{rn} = P_{ra}$, the fixed point of the equation is $\frac{(1 - P_m)(2 - P_r + P_r P_m)}{2 - P_r}$ for AND repair, and $\frac{(1 - P_m)(2 - P_r - P_r P_m)}{2 - P_r}$ for OR repair. Since they are stable points, the density of abnormal nodes converges on the points.

When the first-order approximation of P_r is adopted, the fixed point of the equation is $\frac{1 - P_m}{1 - P_m + P_{ra}}$, which is stable for both AND repair and OR repair, hence the density of abnormal nodes converges on the point.

3.2 Synchrony or Asynchrony

In the self-repair network, the chances for repair are given equally and synchronously to all nodes. However, whether the action of repair takes place or not is a probabilistic event. (Further, even when the repair action takes place, whether the repair is successful or not is another probabilistic event.) Thus, although the synchrony of repair actions is not explicitly controlled, it is controlled through the repair rate: the closer P_r is to 1, the more synchronous repair actions are.

Because we have the following inequality, if there are many abnormal (1) nodes, P_r closer to 0 (but not 0), hence more asynchrony, is favored (to eradicate abnormal nodes). If there are more normal (0) nodes, P_r closer to 1, hence more synchrony, is favored. Further, if we can choose AND repair or OR repair, OR repair is favored.

$$R_{x0z}((0), P_m, P_{ra}) > R_{x0z}((1), P_m, P_{ra}) > R_{x0z}^\vee((1), P_m, P_{ra})$$

$$R_{x1z}((1), P_m, P_{ra}) > R_{x1z}^\vee((1), P_m, P_{ra}) > R_{x1z}((0), P_m, P_{ra})$$

Since we have the following inequalities, OR repair is always favored if we were to choose AND repair or OR repair. However, P_r closer to 0 is favored only when there are relatively many abnormal nodes.

$$R_{x0z}((0), P_m, P_{ra}) > R_{x0z}(P_r, P_m, P_{ra}) > R_{x0z}^\vee(P_r, P_m, P_{ra})$$

$$R_{x1z}(P_r, P_m, P_{ra}) > R_{x1z}^\vee(P_r, P_m, P_{ra}) > R_{x1z}((0), P_m, P_{ra})$$

4 Summary and Discussion

Our interest in the self-repair network is motivated by von Neumann’s probabilistic logic and we attempted to explore the conjecture that: “The ability of a natural organism

to survive in spite of a high incidence of error (which our artificial automata are incapable of) probably requires a very high flexibility and ability of the automaton to watch itself and reorganize itself. And this probably requires a very considerable autonomy of parts” [7].

Although we focused on a self-repair task as an example, the recursive structure should be stressed that the task depends upon the node that is doing the task. Indeed, for a case of repair, the repair success rate depends upon whether the repairing node is normal (0) or abnormal (1). In a conventional logic such as Boolean logic [8], true (0) or false (0) of a logical variable will not in general affect the truth value of other logical variables.

This recursive structure in a logic leads to a complicated form in state transition probability, which depends upon not only the repair rate but also the repair success rate (by abnormal nodes as well as normal ones). Also, this recursive structure creates a double-edged sword effect of the repairing actions.

Nevertheless, several symmetries and asymmetries are identified in terms of several operations such as an exchange of *twin* probabilities: P_{rn} and P_{ra} ; and logical operators AND, OR and negation (flipping probability p to $1-p$). And, these symmetries (and asymmetries) further result in duality when expressed in a form similar to logic. Although a duality similar to Boolean logic [8] holds, the attempt here shares a *self-reference* with Brown logic [9, 10].

References

1. Ishida, Y.: A Critical Phenomenon in a Self-Repair Network by Mutual Copying. In: Kholasla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3682, pp. 86–92. Springer, Heidelberg (2005)
2. Rhodes, C.J., Anderson, R.M.: Dynamics in a Lattice Epidemic Model. *Phys. Rev. Lett.* A 210, 183–188 (1996)
3. Domany, E., Kinzel, W.: Equivalence of cellular automata to Ising models and directed percolation. *Phys. Rev. Lett.* 53, 311 (1984)
4. Nash, J.: The bargaining problem. *Econometrica* 18, 155–162 (1950)
5. Koutsoupias, E., Papadimitriou, C.: Worst-case equilibria. In: 16th Annual Symposium on Theoretical Aspects of Computer Science, pp. 404–413 (1999)
6. Ishida, Y.: Complex Systems Paradigms for Integrating Intelligent Systems. *Studies in Computational Intelligence (SCI)*, vol. 115, pp. 155–181 (2008)
7. von Neumann, J.J.: Theory of Self-Reproducing Automata. In: Burks, A.W. (ed.). University of Illinois Press, Urbana (1966)
8. Boole, G.: An investigation of the laws of thought. Macmillan, London (1854)
9. Spencer-Brown, G.: Laws of Form. Allen & Unwin, London (1969)
10. Varela, F.: Principles of Biological Autonomy. North Holland, New York (1979)

An Immunity-Based Scheme for Statistical En-route Filtering in Wireless Sensor Networks

Yuji Watanabe

Graduate School of Natural Sciences, Nagoya City University,
Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya, Aichi 467-8501 Japan
yuji@nsc.nagoya-cu.ac.jp

Abstract. Statistical en-route filtering schemes do not deal with the identification of compromised nodes injecting bogus reports. In this paper, we propose an immunity-based scheme to identify compromised nodes combining with the statistical en-route filtering in wireless sensor networks. In the proposed scheme, each node has a list of neighborhood and assigns credibility to each neighbor node. Each node can not only update the credibility of neighbor node based on success or failure of filtering and communication but also use the updated credibility as the probability of next communication. Our scheme will be expected to inhibit neighbor nodes of compromised nodes from forwarding false reports.

Keywords: wireless sensor networks, statistical en-route filtering, immunity-based diagnosis.

1 Introduction

Wireless sensor networks have lately drawn considerable attention because of the popularization of sensor nodes that are smaller, cheaper, and intelligent [1]. These nodes are equipped with one or more sensors, a processor, memory, a power supply and wireless interfaces to communicate with each other. Wireless sensor networks may be deployed in potentially adverse or hostile environment, so that the issue of security and privacy must be addressed. Adversaries can capture or compromise sensor nodes to inject false data reports of non-existing or bogus events using the compromised nodes. Such an attack is called *false data injection attack* [2]. The attack may cause not only false alarms but also the depletion of the limited energy of the nodes forwarding these reports to base station. Node and message authentication mechanisms can prevent naive impersonation and false reports injection by outside attackers. However, they cannot block false injection of nodes compromised by the attackers that may know the basic approaches of the deployed security mechanisms. On the one hand, straightforward usage of symmetric keys for authentication mechanisms is not available because once a node is compromised, all the shared security information stored in the node can be used by an attacker. On the other hand, authentication mechanisms based on asymmetric cryptography or tamper-resistant hardware are also infeasible due to the computation and storage constraints of small sensor nodes. It is

necessary to detect and eliminate false data injection attacks in the presence of compromised sensor nodes simply and early.

Several research efforts [2,3,4,5] have proposed schemes to combat such attacks. The *statistical en-route filtering* (SEF) scheme [3] can probabilistically filter out false reports en-route in the dense deployment of large sensor networks. In SEF, assuming that the same event can be detected by multiple nodes, forwarding nodes along the way to base station can statistically detect false reports. SEF has achieved the early detection of false data reports with low computation and communication overhead. There are several revised en-route filtering scheme, for example, the dynamic en-route filtering [4] to deal with dynamic topology of sensor networks, and the multipath en-route filtering [5] to tackle false negative attacks such as blocking event reports and selective forwarding attacks. However, these schemes do not address the identification of compromised nodes injecting false reports. If the compromised nodes are successfully detected, then neighbor nodes of the compromised nodes can drop false reports at an earlier stage.

For the detection of fault nodes on networks, an *immunity-based diagnostic model* [6] has been proposed inspired by the *Jerne's idiotypic network hypothesis* [7]. The immunity-based diagnostic model has also been performed in wireless sensor network [8]. In the immunity-based diagnosis, each node has the capability of testing the neighbor nodes, and being tested by the adjacent others as well. Based on the test outcomes, each node calculates its credibility. However, compromised nodes can not only output bogus test outcomes but also calculate the credibility at random. It is necessary to modify the immunity-based diagnosis to identify compromised nodes combining with SEF.

In this paper, we propose an immunity-based scheme for SEF in wireless sensor networks. In the proposed scheme, each node has a list of neighborhood and assigns credibility to each neighbor node. Each node can not only update the credibility of neighbor node based on success or failure of filtering and communication but also use the updated credibility as the probability of next communication. Our scheme will be expected to inhibit neighbor nodes of compromised nodes from forwarding false reports.

The rest of the paper is organized as follows: In Section 2, we describe the statistical en-route filtering in detail. Section 3 presents our immunity-based scheme for SEF in wireless sensor networks. Section 4 concludes the paper.

2 Statistical En-route Filtering (SEF) [3]

SEF can probabilistically filter out false reports en-route. SEF exploits collective decision-making by multiple detecting nodes and collective false detection by multiple forwarding nodes in the dense deployment of large sensor networks.

SEF consists of three major components: 1) key assignment and report generation, 2) en-route filtering, and 3) base station verification. The process of key assignment and report generation is as follows:

- (1) The base station (BS) maintains a global key pool of N keys $\{K_i, 0 \leq i \leq N - 1\}$, divided into n non-overlapping partitions. Each partition has m keys.
- (2) Before each sensor node is deployed, it stores randomly chosen k ($k < m$) keys from a randomly selected partition in the key pool.
- (3) When an event (sensing target) appears, multiple surrounding nodes can detect the event. Specific event can be settled according to application. A cluster head (CH) is elected using gradient broadcast proposed in [9] to generate the event report. Note that SEF assume that the same event can be detected by multiple nodes.
- (4) Each of the nodes that detected the event generates a keyed message authentication code (MAC) M_i using the event report E (for example, the location, the time, and the type of event) and randomly selected K_i , one of its k stored keys. The node then sends $\{i, M_i\}$, the key index and the MAC, to the CH.
- (5) The CH collects all the $\{i, M_i\}$ s from the nodes that detected the event and attaches T MACs randomly chosen from distinct partitions to the report. This set of multiple MACs acts as the proof that the report is legitimate. Then the CH sends the final report with T key indices and T MACs like $\{E, i_1, M_{i_1}, i_2, M_{i_2}, \dots, i_T, M_{i_T}\}$ toward the BS.

Figure 1 illustrates the example of the key assignment and report generation in SEF. In this figure, the BS maintains a global key pool of $N = 12$ keys divided into $n = 4$ partitions, each of which has $m = 3$ keys. Each sensor node randomly picks $k = 2$ secret keys from one partition of a global key pool. After each node that detected the event endorses the event report by producing a keyed MAC using one of its stored keys, the CH collects all the MACs from the nodes that detected the event and attaches randomly selected $T = 3$ MACs, that is, M_2 , M_9 and M_{10} to the event report E .

In en-route filtering process, intermediate forwarding nodes verify the correctness of the MACs probabilistically and drop a report with forged MACs en-route. Since a legitimate report carries exactly T MACs produced by T keys of distinct partitions, a report with less than T MACs or more than one MACs in the same partition is dropped. Because of the randomized key assignment, each forwarding node has certain probability to possess one of the keys that are used to produce the T MACs. If forwarding node finds out that it has one of the T keys checking key indices in the report, it reproduces the MAC using its stored key and compares the result with the corresponding MAC attached in the report. If the attached MAC is different from the reproduced one, the report is dropped. When intermediate node does not have any of the T keys, the node forwards the report to the next hop. The key assignment ensures that each node can produce only *partial* proof for a report. A single compromised node has to forge MACs to assemble a seemingly complete proof in order to forward false reports. In Fig. 2, since a malicious node possess 2 keys from only partition 1, it

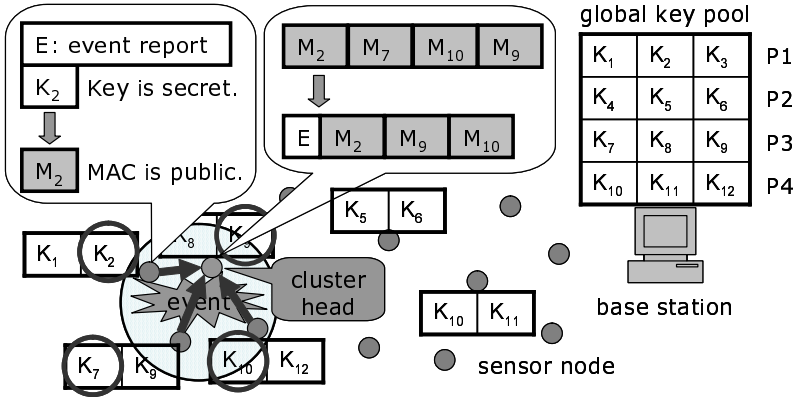


Fig. 1. Example of the key assignment and report generation in SEF with $N = 12$ keys, $n = 4$ partitions, $m = 3$ keys in each partition, $k = 2$ keys in each node, and $T = 3$ MACs attached to event report

needs to forge the other 2 MACs, M_9 and M_{10} . The report with forged MACs is dropped because the correctness of the MACs can be verified by the intermediate node with K_{10} .

Due to the statistical nature of the detection mechanism, a few bogus reports with invalid MACs may escape en-route filtering and reach the BS. In base station verification process, the BS further verifies the correctness of each MAC and eliminates false reports that elude en-route filtering.

3 Proposed Scheme

The filtering schemes do not deal with the identification of compromised nodes injecting false reports. If the compromised nodes are successfully detected, then neighbor nodes of the compromised nodes can drop false reports at an earlier stage. To detect fault nodes in networks, the *immunity-based diagnostic model* [6] is a promising approach. In the immunity-based diagnosis, each node has the capability of testing the neighbor nodes, and being tested by the adjacent others as well. Based on the test outcomes, each node calculates its own credibility. However, malicious nodes can not only output bogus test outcomes but also calculate the credibility at random.

Therefore, we propose an immunity-based scheme to identify compromised nodes combining with SEF in wireless sensor networks. In the proposed scheme, each node has a list of neighborhood and assigns a state variable $R \in [0, 1]$ indicating *credibility of neighbor* to each neighbor node. Note that each node does not have its own credibility. Node j updates the credibility R_{ji} of the previous neighbor node i sending the event report based on its filtering result and the reply from next neighbor node k as follows:

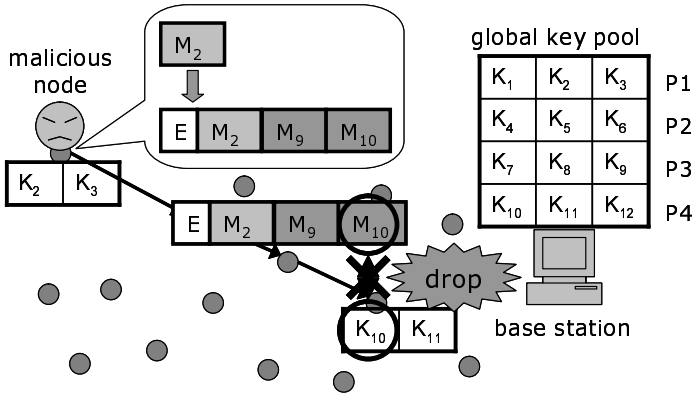


Fig. 2. Case that a false report with forged MACs from a malicious node is dropped by the intermediate forwarding node

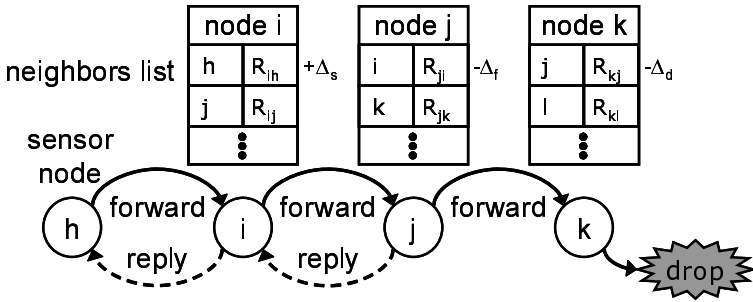


Fig. 3. An immunity-based scheme for detecting compromised nodes with SEF

$$R_{ji}(t) = \begin{cases} R_{ji}(t-1) + \Delta_s & \text{if node } j \text{ receives the reply from} \\ & \text{next node } k \\ R_{ji}(t-1) - \Delta_f & \text{if node } j \text{ does not receive the reply} \\ & \text{from next node } k \\ R_{ji}(t-1) - \Delta_d & \text{if node } j \text{ drops the report using SEF} \end{cases} \quad (1)$$

If credibility $R_{ji}(t)$ is over 1 (under 0), it is set to 1 (0). The initial value of credibility $R_{ji}(0)$ is 1. The values of the parameters Δ_s , Δ_f and Δ_d should be chosen through mathematical analysis and simulation. For example, in Fig. 3, node i increases the credibility R_{ih} of the previous node h because the reply from next node j can be received. However, node j decreases the credibility R_{ji} of the previous node i because next node k drops the event report using SEF and does not reply to node j . Since node k filters out the report by itself, the credibility R_{kj} of the previous node j also decreases.

Only the credibility update process can not achieve the identification of compromised nodes. For instance, in Fig. 3, if node h is compromised, false reports

are still forwarded toward node k . Therefore each node uses the updated credibility as the probability of next communication. In the same example, node i has adversely higher probability of receiving the report from compromised node h because of the increase of the credibility R_{ih} . However, since node j has lower probability of receiving the report from node i , node i may fail to communicate with node j at next stage, and then the credibility R_{ih} of the previous node h in the neighbors list of node i decreases. Although the credibility R_{kj} of the previous node j in the neighbors list of node k decreases at first, if node j sends legitimate reports received from the other previous nodes to node k , the credibility R_{kj} can be recovered. By iterating the credibility update and the communication based on the updated credibility, our scheme will be expected to inhibit neighbor nodes of compromised nodes from forwarding false reports.

4 Conclusions

In this paper, we proposed an immunity-based scheme for identifying compromised nodes combining with SEF in wireless sensor networks. It is important that the proposed scheme is additionally combined with an authentication mechanism for higher security level. Since we are now carrying out some simulations to evaluate the performance of our scheme, so that we will show the results at conference site.

References

1. Yick, J., Mukherjee, B., Ghosal, D.: Wireless Sensor Network Survey. *Computer Networks* 52, 2292–2330 (2008)
2. Zhu, S., Setia, S., Jajodia, S., Ning, P.: An Interleaved Hop-by-Hop Authentication Scheme for Filtering of Injected False Data in Sensor Networks. In: *IEEE Symposium on Security and Privacy*, pp. 259–271 (2004)
3. Ye, F., Luo, H., Lu, S., Zhang, L.: Statistical En-Route Filtering of Injected False Data in Sensor Networks. *IEEE Journal on Selected Areas in Communications* 23(4), 839–850 (2005)
4. Yu, Z., Guan, Y.: A Dynamic En-route Scheme for Filtering False Data Injection in Wireless Sensor Networks. In: *Proceedings of the 25th IEEE Conference on Computer Communications, INFOCOM* (2006)
5. Kim, M.S., Cho, T.H.: A Multipath En-Route Filtering Method for Dropping Reports in Sensor Networks. *IEICE Transactions on Information and Systems* E90-D(12), 2108–2109 (2007)
6. Ishida, Y.: Fully Distributed Diagnosis by PDP Learning Algorithm: Towards Immune Network PDP Model. In: *Proc. of IJCNN*, pp. 777–782 (1990)
7. Jerne, N.: The Immune System. *Scientific American* 229-1, 52–60 (1973)
8. Watanabe, Y., Ishida, Y.: Migration Strategies of Immunity-based Diagnostic Nodes for Wireless Sensor Network. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) *KES 2006. LNCS (LNAI)*, vol. 4252, pp. 131–138. Springer, Heidelberg (2006)
9. Ye, F., Zhong, G., Lu, S., Zhang, L.: GRAdient Broadcast: A Robust Data Delivery Protocol for Large Scale Sensor Networks. *ACM Wireless Networks* 11(3), 285–298 (2005)

Author Index

- Abe, Akinori III-307
Abecker, Andreas I-660
Abdul Maulud, Khairul Nizam IV-22
Abe, Hidenao III-297
Abe, Jair Minoro III-123, III-133,
III-143, III-154, III-164,
III-200
Abu Bakar, Azuraliza IV-22
Adachi, Yoshinori III-63, III-81
Adam, Giorgos III-389
Aguilera, Felipe II-591
Ahmad Basri, Noor Ezlin IV-22
Ain, Qurat-ul I-340
Akama, Seiki III-133, III-143,
III-164, III-200
Albusac, J. IV-347
Alechina, Natasha IV-41
Alonso-Betanzos, Amparo I-168
Alosefer, Yaser IV-556
Álvarez, Héctor II-581
Alvarez, Héctor II-591
Alvez, Carlos E. II-44
Alvi, Atif IV-576
An, Dongchan II-302
Aoki, Kumiko II-143
Aoki, Masato IV-153
Apostolakis, Ioannis III-23
Appice, Annalisa III-339
Aritsugi, Masayoshi IV-220
Asano, Yu I-649
Ashida, Masaya II-611
Asimakis, Konstantinos III-389
Aspragathos, Nikos II-341
Ayes, Gareth IV-566
Azpeitia, Eneko II-495
Azuma, Haruka III-273

Baba, Norio III-555
Baba, Takahiro III-207
Babensyshev, S. II-224
Babensyshev, Sergey I-230
Baig, Abdul Rauf I-61
Bajo, Javier IV-318
Baralis, Elena III-418

Bardone, Emanuele III-331
Barry, Dana M. IV-200
Bath, Peter II-163
Baumgartner Jr., William A. IV-420
Belohlavek, Radim I-471
Belša, Igor II-21
Ben Hassine, Mohamed Ali I-532
Bermejo-Alonso, Julita I-522
Bhatti, Asim I-5
Biernacki, Pawel I-350, I-360
Biscarri, Félix I-410
Biscarri, Jesús I-410
Bishop, Christopher I-3
Blašković, Bruno II-292
Bluemke, Ilona II-82
Boochs, Frank I-576
Borzemski, Leszek II-505
Bouamama, Sadok II-312
Bouché, P. IV-32
Bouras, Christos III-379, III-389
Bourgoin, Steve IV-410
Bratosin, Carmen I-41
Bravo-Marquez, Felipe II-93
Bretonnel Cohen, K. IV-420
Bruno, Giulia III-418
Buckingham, Christopher D. IV-88
Bukatović, Martin I-432
Bumbaru, Severin I-188
Byrne, Caroline IV-365

Cambria, Erik IV-385
Carpio Valadez, Juan Martín II-183
Carrella, Stefano II-361
Caspar, Joachim IV-402
Cavallaro, Alexander I-290
Chakkour, Fairouz IV-586
Champesme, Marc II-351
Chang, Jae-Woo I-511
Charton, Eric IV-410
Chiusano, Silvia III-418
Chowdhury, Nihad Karim I-511
Ciampi, Anna III-339
Cîrlugea, Mihaela IV-613
Ciruela, Sergio IV-70

- Clive, Barry IV-633
 Cocea, Mihaela II-103, II-124
 Condell, Joan IV-430
 Cooper, Jerry IV-497
 Corchado, Emilio S. IV-318
 Corchado, Juan M. IV-318
 Coyne, Bob IV-375
 Cruz, Christophe I-576
 Csipkes, Doris IV-603
 Csipkes, Gabor IV-603
 Cui, Hong IV-506
 Culham, Alastair IV-517
 Čupić, Marko I-100
 Cuzzocrea, Alfredo III-426
 Czarnecki, Adam II-533
- d'Amato, Claudia III-359
 Dąbrowska-Kubik, Katarzyna III-369
 Dalbello Bašić, Bojana I-100, II-31
 Da Silva Filho, João Inácio III-154
 de Faria, Ricardo Coelho III-174
 Debenham, John I-220
 Deb Nath, Rudra Pratap I-511
 Delgado, Miguel IV-70, IV-337
 Dembitz, Šandor II-292
 Dengel, Andreas I-290
 Dentsoras, Argyris II-331
 De Paz, Juan F. IV-318
 Detyniecki, Marcin I-544
 Di Bitonto, Pierpaolo II-64
 Diniz, Janaina A.S. III-182
 Dino Matijaš, Vatroslav I-100
 Dolog, Peter III-398
 Domenici, Virna C. III-418
 Doña, J.M. III-445
 do Prado, Hércules Antonio III-174
 Duarte, Abraham II-183
- Eckl, Chris IV-385
 Ercan, Tuncay II-253, II-263
 Ernst, Patrick I-290
 Ezawa, Hiroyasu IV-280
- Fadzli, Syed Abdullah IV-240
 Fahlman, Scott E. II-193
 Fanizzi, Nicola III-359
 Farago, Paul IV-623
 Farquad, M.A.H. I-461
 Feng, Haifeng I-544
 Fernández-Breis, Jesualdo Tomás I-597
- Fernandez-Canque, Hernando IV-603
 Fernández de Alba, José M. IV-328
 Ferneda, Edilson III-174, III-182
 Ferro, Alfredo III-438
 Festila, Lelia IV-623
 Festilä, Lelia IV-613
 Figueiredo, Adelaide III-182
 Flann, Christina IV-497
 Fontenla-Romero, Óscar I-168
 Forge, David II-351
 Fortino, Giancarlo I-240
 Fraire Huacuja, Héctor Joaquín II-183
 Fujii, Satoru III-483, III-519
 Fujiki, Takatoshi III-509
 Fujiwara, Minoru IV-163
 Fukuda, Taro III-473
 Fukui, Shinji III-89
 Fukumi, Minoru III-612
 Fukumura, Yoshimi II-143,
 IV-190, IV-200
 Fulcher, John II-454
 Furutani, Michiko III-307
 Furutani, Yoshiyuki III-307
- G.V.R., Kiran II-11
 Gaál, Balázs I-607
 Gabbar, Hossam A. II-427
 Gagnon, Michel IV-410
 Gartiser, N. IV-32
 Gharahbagh, Abdorreza Alavi I-331
 Ghofrani, Sedigheh I-331
 Gibbins, Nicholas IV-594
 Giddy, Jonathan IV-485
 Giugno, Rosalba III-438
 Glaser, Hugh IV-594
 Gledec, Gordán II-292
 Glez-Morcillo, C. IV-347
 Goesele, Michael IV-402
 Golemanova, Emilia II-253, II-263
 Golemanov, Tzanko II-253, II-263
 Gomez, Juan Carlos I-566
 Gómez-Ruiz, J. III-445
 Gómez Zuluaga, Giovanni II-601
 Gonda, Yuuji IV-210
 Gonzaga Martins, Helga III-154
 Gonzalez B., Juan J. II-203
 González, Yanira I-51
 Görg, Carsten IV-420
 Gotoda, Naka II-620, IV-145
 Graczyk, Magdalena I-111

- Graña, M. IV-80
 Grauer, Manfred II-399
 Greaves, David IV-576
 Grillo, Nicola III-426
 Grimnes, Gunnar I-290
 Grosvenor, Roger II-371
 Grzech, Adam II-523
 Guerrero, Juan I. I-410
 Guerrero, Luis A. II-93, II-591
 Gurevych, Iryna IV-402
 Gutierrez-Santos, Sergio II-124

 Hacid, Mohand-Said III-426
 Hagita, Norihiro III-307
 Håkansson, Anne II-273, IV-60,
 IV-98, IV-124
 Halabi, Ammar IV-527
 Haller, Heiko I-660
 Hamaguchi, Takashi II-381, II-417
 Hamdan, Abdul Razak I-491
 Hanabusa, Hisatomo IV-308
 Handa, Hisashi III-555
 Hangos, Katalin M. II-389
 Hanser, Eva IV-430
 Harada, Kouji III-637
 Hardisty, Alex IV-485
 Hartung, Ronald IV-124
 Hartung, Ronald L. II-273
 Hasegawa, Mikio IV-271
 Hasegawa, Naoki IV-190
 Hashimoto, Yoshihiro II-417
 Hashizume, Aoi II-135
 Haskkour, Nadia IV-586
 Hattori, Akira IV-290
 Havasi, Catherine IV-385
 Hayami, Haruo IV-290
 Hayashi, Hidehiko IV-475
 Hayashi, Yuki IV-153
 Heap, Marshall J. IV-517
 Hernández, Carlos I-522
 Hintea, Sorin IV-603, IV-613, IV-623
 Hiratsuka, Yoshimune III-315
 Hirokawa, Masakazu I-148
 Hirokawa, Sachio III-207
 Hirschberg, Julia IV-375
 Hocenski, Željko I-300
 Höppner, F. I-442
 Horák, Aleš I-432
 Horiguchi, Ryota IV-308
 Hosseini, Mohammad Mehdi I-331

 Huang, Houkuan II-1
 Hunger, A. II-114
 Hunter, Lawrence E. IV-420
 Hussain, Amir IV-385

 Iftikhar, Nadeem III-349
 Igarashi, Masao III-622
 Iijima, Chie III-264
 Iijima, Morihisa IV-308
 Ikeda, Mitsuru IV-163
 Inoue, Akiya III-225
 Inoue, Etsuko III-509
 Inuzuka, Nobuhiro III-72
 Iribe, Yurie II-143, IV-173
 Ishida, Keisuke IV-190
 Ishida, Yoshiteru III-628, III-637,
 III-645, III-652
 Ishii, Naohiro III-97, III-104, III-113
 Islim, Ahmed-Derar IV-527
 Isokawa, Teijiro III-592
 Iswandy, Kuncup II-361
 Itoh, Toshiaki II-381
 Itokawa, Tsuyoshi IV-220
 Ito, Momoyo III-612
 Itou, Junko III-473, III-527
 Ivan, Lavallée I-452
 Iwahori, Yuji III-63, III-81, III-89
 Iwashita, Motoi III-225
 Iwazaki, Tomonori IV-190

 Jabeen, Hajira I-61
 Jaffar, M. Arfan I-340
 Jain, Lakhmi C. II-454
 Jakobović, Domagoj I-100
 Jamil, Hasan III-408
 Jantan, Hamidah I-491
 Jascanu, Nicolae I-188
 Jascanu, Veronica I-188
 Jezic, Gordan I-261
 Jia, Dawei I-5
 Jiao, Roger I-131
 Jimbo, Takashi III-97
 Jimenez, L. IV-347
 Jimenez-Molina, Angel II-54
 Jing, Liping II-1
 Jin, Zhe III-464
 Ji, Xiaofei I-369
 Jones, Andrew C. IV-485

- Kambayashi, Yasushi I-198
 Kamide, Norihiro I-178, II-153
 Kanda, Taki II-477
 Kanematsu, Hideyuki IV-200
 Karadgi, Sachin II-399
 Karmacharya, Ashish I-576
 Kasabov, Nikola I-1
 Kastania, Anastasia N. III-43, III-53
 Kasugai, Kunio III-81
 Katarzyniak, Radosław I-271
 Katoh, Takashi III-455
 Kavakli, Manolya II-214
 Kawaguchi, Masashi III-97
 Kawakatsu, Hidefumi III-281
 Kawano, Hiromichi III-225
 Kelly, Michael IV-11, IV-135
 Khalid, Marzuki I-69, II-464
 Kholod, Marina III-273
 Kim, Daewoong IV-261
 Kim, Hakin II-302
 Kimura, Makito III-264
 Kimura, Naoki II-381, II-409
 Kipsang Choge, Hillary III-612
 Kitasuka, Teruaki IV-220
 Klawonn, Frank I-141, II-244
 Ko, In-Young II-54
 Kodama, Issei IV-475
 Koffa, Antigoni III-53
 Kogawa, Keisuke III-555
 Kohlert, Christian I-321
 Kohlert, Michael I-321
 Kojima, Masanori III-572
 Kojiri, Tomoko IV-153
 Kolp, Manuel I-209
 Komine, Noriyuki III-545, III-572
 König, Andreas I-321, II-361
 Koroušić Seljak, Barbara I-587
 Kou, Tekishi II-409
 Kouno, Shouji III-225
 Kountchev, Roumen III-133, III-215
 Koziarkiewicz-Hetmańska, Adrianna
 I-281
 Kozmann, György I-607
 Kratchanov, Kostadin II-253, II-263
 Krišto, Ivan II-21
 Kubo, Masao IV-298
 Kumakawa, Toshiro III-315
 Kunifuji, Susumu IV-457
 Kunimune, Hisayoshi IV-210
 Kurahashi, Wataru III-89
 Kurdi, Mohamed-Zakaria IV-527
 Kurosawa, Takeshi III-225
 Kusaka, Mariko III-555
 Kuwahara, Daiki IV-465
 Lambert-Torres, Germano III-154
 Lasota, Tadeusz I-111
 Laterza, Maria II-64
 Lawrenz, Wolfhard II-244
 Lawrynowicz, Agnieszka III-359
 Le, D.-L. II-114
 Lee, Huey-Ming II-438
 Lee, Hyun-Jo I-511
 Lensch, Hendrik P.A. IV-402
 León, Carlos I-410
 León, Coromoto I-51
 Lesot, Marie-Jeanne I-544
 Lewandowski, Andrzej I-311
 L'Huillier, Gaston II-93, II-581
 Li, Chunping I-131
 Li, Kai IV-173
 Li, Yibo I-369
 Li, You II-445
 Lin, Lily II-438
 Liu, Honghai I-369
 Liu, Jin I-379
 Liu, Jing II-214
 Liu, Kecheng I-554
 Liu, Lucing III-207
 Liu, Xiaofan IV-41
 Liu, Yang I-90
 Liu, Ying I-131
 Logan, Brian IV-41
 Lopes, Helder F.S. III-164
 López, Juan C. II-193
 Loukakos, Panagiotis I-481
 Lovrek, Ignac I-251
 Ludwig, Simone A. IV-536
 Lukose, Dickson I-627
 Lunney, Tom IV-430
 Luz, Saturnino IV-394
 Ma, Minhua IV-430
 Macía, I. IV-80
 Mackin, Kenneth J. III-622
 Maddouri, Mondher I-121
 Maeda, Kaoru I-639
 Maehara, Chihiro IV-261
 Maeno, Hiroshi IV-163
 Maezawa, Toshiki II-645

- Magnani, Lorenzo III-331
 Magoulas, George D. II-103, II-124
 Mahanti, Ambuj II-282
 Maheswaran, Ravi II-163
 Mahoto, Naeem A. III-418
 Majumder, Sandipan II-282
 Mák, Erzsébet I-607
 Makino, Toshiyuki III-72
 Małachowski, Bartłomiej IV-180
 Malerba, Donato III-339
 Mancilla-Amaya, Leonardo II-553
 Marc, Bui I-452
 Marín, Nicolás IV-70
 Markos, Panagiotis II-331
 Marteau, Pierre-François I-420
 Martínez-Romero, Marcos II-74
 Martínez-Costa, Catalina I-597
 Martínez F., José A. II-173, II-203
 Marzani, Franck I-576
 Masoodian, Masood IV-394
 Mat Ali, Nazmona I-554
 Matic, Tomislav I-300
 Matsubara, Takashi IV-298
 Matsuda, Noriyuki II-637
 Matsui, Nobuyuki III-592
 Matsumoto, Hideyuki II-417
 Matsuoka, Rumiko III-307
 Matsushita, Kotaro III-622
 Matsuura, Kenji II-620, IV-145
 Mattila, Jorma K. IV-108
 Maus, Heiko I-639
 Mc Kevitt, Paul IV-430
 McMeekin, Scott G. IV-633
 Meddouri, Nida I-121
 Mehmood, Irfan I-340
 Mehmood, Rashid IV-566, IV-576
 Mello, Bernardo A. III-182
 Menárguez-Tortosa, Marcos I-597
 Merlo, Eduardo II-581, II-591
 Metz, Daniel II-399
 Millán, Rocío I-410
 Millard, Ian C. IV-594
 Minaduki, Akinori IV-475
 Miñarro-Giménez, José Antonio I-597
 Mineno, Hiroshi II-135, III-535
 Miranda, Gara I-51
 Mishina, Takashi III-493
 Misue, Kazuo IV-440
 Mitsubishi, Takashi III-281
 Miura, Hajime IV-190
 Miura, Hirokazu II-637
 Miura, Motoki IV-457, IV-465
 Miyachi, Taizo II-645
 Miyaji, Isao III-483
 Miyoshi, Masato III-612
 Mizuno, Tadanori III-572
 Mizuno, Shinji II-143
 Mizuno, Tadanori II-135, III-535
 Mizutani, Masashi I-198
 Moens, Marie-Francine I-566
 Mohd Yatid, Moonyati Binti III-473
 Molina, José Manuel IV-357
 Möller, Manuel I-290
 Molnar, Goran I-100
 Monedero, Iñigo I-410
 Moradian, Esmiralda IV-98, IV-124
 Morihiro, Koichiro III-592
 Morii, Fujiki I-390
 Morita, Hiroki III-572
 Moulianitis, Vassilis II-341
 Moya, Francisco II-193
 Muhammad Fuad, Muhammad Marwan I-420
 Muhammad-Sukki, Firdaus IV-633
 Mukai, Naoto IV-280
 Müller, Ulf II-399
 Munemori, Jun III-473, III-527
 Munteanu, Cristian R. II-74
 Murakami, Akira III-315
 Murat, Ahat I-452
 Musa, Zalili II-454
 Nabi, Zubair IV-576
 Nahavandi, Saeid I-5
 Nakada, Toyohisa IV-449
 Nakagawa, Masaru III-509
 Nakahara, Takanobu III-244, III-273
 Nakamatsu, Kazumi III-123, III-133, III-143, III-164, III-200, III-215
 Namiki, Junji III-562
 Naqi, Syed M. I-340
 Naruse, Keitaro IV-298
 Nasri, Chaker Abidi I-532
 Nauck, Detlef I-141
 Naveen, Nekuri I-80
 Németh, Erzsébet II-389
 Németh, Istvänné I-607
 Nguyen, A.-T. II-114
 Nguyen, D.-T. II-114

- Nguyen, Ngoc Thanh I-281
 Niimura, Masaaki IV-210
 Nishi, Daisuke II-637
 Nishide, Tadashi III-473
 Nishihara, Takanao II-645
 Nishihara, Yoko III-315
 Nishimura, Haruhiko III-592
 Nishino, Kazunori II-143
 Niskanen, Vesa A. IV-116
 Niwa, Takahito III-113
 Noda, Masaru II-381
 Noguchi, Daijiro IV-163
 Nonaka, Yuki IV-271
 Nunohiro, Eiji III-622
- Obembe, Olufunmilayo IV-88
 Obermöller, Nils II-244
 Oehlmann, Ruediger III-290
 O'Grady, Michael J. IV-365
 O'Hare, Gregory M.P. IV-365
 Ohsawa, Yukio III-315
 Oikawa, Ryotaro I-198
 Okada, Masashi III-104
 Okada, Yoshihiro IV-251
 Okada, Yousuke III-113
 Okajima, Seiji IV-251
 Okamoto, Takeshi III-628
 Okumura, Noriyuki IV-51
 Oliver, José L. I-31
 Oltean, Gabriel IV-623
 Omitola, Tope IV-594
 Omori, Yuichi IV-271
 Onn, Kow Weng I-627
 Onogi, Manabu III-113
 Ooshaksaraie, Leila IV-22
 Orlewicz, Agnieszka II-82
 Orłowski, Aleksander II-515
 Orłowski, Cezary II-533, II-543, II-571
 Othman, Zulaiha Ali I-491
 Otsuka, Shinji II-620
 Ounelli, Habib I-532
 Oyama, Tadahiro III-612
 Ozaki, Masahiro III-63
 Ozell, Benoit IV-410
- Palenzuela, Carlos II-495
 Paloc, C. IV-80
 Park, Jong Geol III-622
 Park, Seog II-302
- Pavón, Juan IV-328
 Pazos, Alejandro II-74
 Pazos R., Rodolfo II-183
 Pazos R., Rodolfo A. II-173, II-203
 Pedersen, Torben Bach III-349
 Peláez, J.I. III-445
 Pérez O., Joaquín II-173
 Pereira, Javier II-74
 Peter, S. I-442
 Petre, Emil II-234
 Petric, Ana I-261
 Petrigni, Caterina III-418
 Pham, Tuan D. I-379
 Pintér, Balázs I-607
 Podobnik, Vedran I-251
 Porto-Díaz, Iago I-168
 Pouloupoulos, Vassilis III-389
 Pratim Sanyal, Partha IV-506
 Prickett, Paul II-371
 Pudi, Vikram II-11
 Pu, Fei IV-135
 Puga Soberanes, Héctor José II-183
 Puglisi, Piera Laura III-438
 Pulvirenti, Alfredo III-438
- Raghavendra Rao, C. I-80
 Raja, Hardik IV-485
 Raju, S. Bapi I-461
 Rambousek, Adam I-432
 Rambow, Owen IV-375
 Ramirez-Iniguez, Roberto IV-633
 Rana, Omer F. IV-546, IV-556
 Rango, Francesco I-240
 Ravi, V. I-80, I-461
 Ray, Sanjog II-282
 Read, Simon II-163
 Reicher, Tomislav II-21
 Renaud, D. IV-32
 Resta, Marina III-583
 Richards, Kevin IV-497
 Ríos, Sebastián A. II-93, II-581, II-591
 Rodríguez, Manuel I-522
 Rodríguez, Sara IV-318
 Rogers, Bill IV-394
 Rojas P., Juan C. II-203
 Rojtberg, Pavel IV-402
 Roos, Stefanie IV-536
 Ros, María IV-337
 Roselli, Teresa II-64

- Rossano, Veronica II-64
 Rostanin, Oleg I-639
 Roussetot, F. IV-32
 Rouveyrol, Claire III-81
 Rózewski, Przemysław IV-180
 Ruhlmann, Laurent IV-410
 Russo, Wilma I-240
 Rybakov, Vladimir I-230, II-224, III-323
 Rygielski, Piotr II-523
- Sadanandan, Arun Anand I-627
 Sadek, Jawad IV-586
 Saha, Sourav II-282
 Said, Fouchal I-452
 Sakamoto, Ryuuki III-501
 Salem, Ziad IV-586
 San, Tay Cheng I-69
 Sánchez-Pi, Nayat IV-357
 Sanín, Cesar II-553, II-601
 Sanin, Cesar II-563
 Santaolaya S., René II-203
 Santofimia, María J. II-193
 Sanz, Ricardo I-522
 Sasaki, Takuya III-455
 Sato, Hiroshi IV-298
 Sato, Hitomi IV-290
 Sawamoto, Jun III-455
 Sawaragi, Tetsuo I-2
 Schäfer, Walter II-399
 Schwarz, Katharina IV-402
 Segawa, Norihisa III-455
 Segura, Carlos I-51
 Seifert, Sascha I-290
 Selişteanu, Dan II-234
 Şendrescu, Dorin II-234
 Seta, Kazuhisa IV-163
 Setchi, Rossitza I-481, I-617, IV-240
 Shadbolt, Nigel IV-594
 Shankar, Ravi II-11
 Shi, Lei I-617
 Shida, Haruki III-628
 Shidama, Yasunari III-281
 Shima, Takahiro III-519
 Shimoda, Toshifumi II-143
 Shimogawa, Shinsuke III-225
 Shiraishi, Yoh III-493
 Siddiqui, Raees II-371
 Sidirokastriti, Sofia III-43
 Sidorova, Natalia I-41
- Sierra, Carles I-220
 Šilić, Artur II-21, II-31
 Sintek, Michael I-290
 Sitek, Tomasz II-571
 Skorupa, Grzegorz I-271
 Sofiane, Benamor I-452
 Sohn, So Young IV-200
 Soldano, Henry II-351
 Sproat, Richard IV-375
 Srivastava, Muni S. III-7
 Stasko, John IV-420
 Stewart, Brian G. IV-633
 Suchacka, Grażyna II-505
 Sugihara, Taro IV-457
 Sugino, Eiji III-455
 Sugiyama, Takeshi III-15
 Sun, Fan I-90
 Sunayama, Wataru III-235
 Suzuki, Kenji I-148
 Suzuki, Nobuo III-1
 Suzuki, Takeshi II-645
 Suzuki, Takeshi I-639
 Suzuki, Yu IV-440
 Szczerbicki, Edward II-515, II-553,
 II-563, II-601
 Szlachetko, Bogusław I-311
 Szolga, Lorant Andras IV-613
- Taguchi, Ryosuke IV-200
 Takahashi, Hiroataka IV-190
 Takahashi, Megumi III-519
 Takahashi, Osamu III-493
 Takai, Keiji III-254
 Takano, Shigeru IV-251
 Takeda, Kazuhiro II-381, II-417
 Takeda, Kosuke III-501
 Takeda, Masaki III-555
 Takeshima, Ryo IV-230
 Taki, Hirokazu II-611, II-637
 Takimoto, Munehiro I-198
 Tamura, Yukihiro III-235
 Tanabe, Kei-ichi III-645
 Tanaka, Jiro IV-440
 Tanaka, Takushi III-190
 Tanaka, Toshio II-620
 Tanaka-Yamawaki, Mieko III-602
 Tanaka, Yuzuru I-14, I-649
 Telec, Zbigniew I-111
 Tenorio, E. III-445

- Tipney, Hannah IV-420
 Tokuda, Mitsuhiro IV-465
 Tominaga, Yuuki IV-210
 Tomiyama, Yuuki III-235
 Torii, Ippei III-104, III-113
 Toro, Carlos II-495
 Torres, Claudio Rodrigo III-154
 Tortosa, Leandro I-31
 Tóth, Attila II-389
 Tran, V.-H. II-114
 Trawiński, Bogdan I-111
 Tschumitschew, Katharina I-141, II-244
 Tsogkas, Vassilis III-379
 Tsuchiya, Seiji I-400, IV-1
 Tsuda, Kazuhiko III-1
 Tsuge, Satoru III-612
 Tsuge, Yoshifumi II-409
 Tsumoto, Shusaku III-297

 Uemura, Yuki IV-220
 Ueta, Tetsushi IV-145
 Uno, Takeaki III-244
 Ushiana, Taketoshi IV-261

 Valencia-García, Rafael I-597
 Vallejo, D. IV-347
 Valsamos, Harry II-341
 van der Aalst, Wil I-41
 Vaquero, Javier II-495
 Varlamis, Iraklis III-23, III-33
 Vassányi, István I-607
 Vázquez A., Graciela II-173
 Vázquez-Naya, José M. II-74
 Vecchietti, Aldo R. II-44
 Velásquez, Juan D. II-93, II-581
 Ventos, Véronique II-351
 Verspoor, Karin IV-420
 Vicent, José F. I-31
 Vila, Amparo IV-337
 Villanueva, David Terán II-183
 Vychodil, Vilem I-471

 Wada, Yuji III-455
 Wakayama, Yuki IV-251
 Walters, Simon II-322
 Wan, Chang I-501
 Wan, Jie IV-365
 Wang, Bo II-445
 Watabe, Hirokazu I-400, IV-1

 Watada, Junzo II-445, II-454,
 II-485
 Watanabe, Takashi III-123
 Watanabe, Toyohide IV-153, IV-230
 Watanabe, Yuji III-660
 Watanabe, Yuta IV-475
 Wautelet, Yves I-209
 Whitaker, Roger I-4
 White, Richard J. IV-485
 Willett, Peter II-163
 Wilton, Aaron IV-497
 Woodham, Robert J. III-81, III-89
 Wu, Dan IV-60, IV-124

 Xu, Guandong III-398

 Yaakob, Shamshul Bahar II-485
 Yada, Katsutoshi III-244, III-254,
 III-273
 Yamada, Kunihiro III-483, III-535,
 III-545, III-562, III-572
 Yamada, Takayoshi I-158
 Yamaguchi, Takahira III-264
 Yamaguchi, Takashi III-622
 Yamamoto, Hidehiko I-158
 Yamamura, Mariko III-1, III-7
 Yamazaki, Atsuko K. II-630
 Yamazaki, Makoto IV-190
 Yanagihara, Hirokazu III-7
 Yanagisawa, Yukio III-622
 Yano, Yoneo II-620, IV-145
 Yaoi, Takumu III-483
 Yasue, Kizuki II-409
 Yatsugi, Kotaro IV-261
 Yonekura, Naohiro III-97
 Yoshida, Koji III-519
 Yoshida, Kouji III-483, III-572
 Yoshihara, Yuriko III-555
 Yoshihiro, Takuya III-509
 Yoshimura, Eriko I-400, IV-1
 Yu, Chunshui IV-506
 Yu, Jian II-1
 Yuizono, Takaya III-464
 Yukawa, Takashi IV-190
 Yun, Jiali II-1
 Yunfei, Zeng III-63
 Yunus, Mohd. Ridzuan II-464
 Yusa, Naoki III-572
 Yusof, Rubiyah I-69, II-464

Yusof, Yuhanis IV-546
Yuuki, Osamu III-535, III-562

Zamora, Antonio I-31
Zanni-Merk, C. IV-32
Zhang, Haoxi II-563

Zhang, Hui I-131
Zhang, Yan IV-11, IV-135
Zhang, Yanchun III-398
Zhou, Yi IV-135
Ziólkowski, Artur II-543
Zong, Yu III-398