Wykorzystanie sieci neuronowych i algorytmów genetycznych w analizie i kategoryzacji dokumentów naukowych

Autoreferat rozprawy doktorskiej

Błażej Zyglarski
Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika
blazej.zyglarski@mat.umk.pl

11 czerwca 2010

1 Wprowadzenie

Przedstawione w niniejszej rozprawie wyniki dotyczą badań nad metodami porównywania tekstów naukowych, ich kategoryzacji oraz automatycznego wyznaczania słów kluczowych dla tych tekstów.

W epoce łatwego dostępu do powszechnie obecnej informacji, odbiorcy skazani są na kontakt z dużymi ilościami danych. W ciągu ostatnich lat ilość dostępnych informacji znacząco wzrosła. Przyczynił się do tego gwałtowny rozwój internetu i ogólny pęd do digitalizacji wszelkich dokumentów.

W większości przypadków wśród wielu informacji bezwartościowych dla odbiorcy znajdują się nieliczne posiadające wartość. Tyczy się to szczególnie wyników zwracanych przez wszelkiego rodzaju wyszukiwarki. Odbiorca musi więc przebrnąć przez gąszcz danych zanim znajdzie to, czego potrzebuje. W szczególności dotyczy to wszelkiego rodzaju tekstów naukowych, których znaczne ilości można znaleźć w internecie. Wybranie tych, które dotyczą żądanej dziedziny jest wysoce pracochłonne.

W popularnych dzisiaj wyszukiwarkach internetowych, w większości przypadków, odsetek poprawnych wyników jest niewielki - ilość dokumentów i danych dostępnych

w internecie jest tak ogromna, że praktycznie niemożliwe jest zindeksowanie i szczegółowe zbadanie ich wszystkich (co jest pokazane w Albert *et al.* (1999)). Użytkownik musi więc samodzielnie przeglądać szereg wyników wyszukiwań, tracąc cenny czas.

Również pozyskanie danych do samodzielnej analizy wymaga czasu - zasada działania popularnych serwisów wyszukiwawczych wymaga interakcji z użytkownikiem. Aby otrzymać żądany rezultat, użytkownik musi dostarczyć odpowiednie zapytanie. Języki zapytań są zwykle proste, aczkolwiek pozwalają na tworzenie konstrukcji, które mogą zostać użyte do bardziej skomplikowanych wyszukiwań. Aktualnie dostępne systemy wyszukiwawcze ograniczają się zwykle do sprawdzenia występowania kilku podanych przez pytającego słów i na tej podstawie zwracają wyniki. Taka technologia wyparła w ciągu ostatnich lat inne mechanizmy wyszukiwania informacji, jednak bardziej skomplikowane wyszukiwanie wciąż wymaga od użytkownika myślenia i tworzenia trudnych zapytań. Kolejne utrudnienie pojawia się, gdy użytkownik chce znaleźć dodatkowe informacje na temat poruszany przez jakieś dokumenty, które już posiada. Wymaga to od niego przeczytania posiadanych dokumentów i sformułowania zapytań na podstawie ich zawartości.

Badania przeprowadzone w niniejszej pracy skupiają się na tekstach, które nie mają wyznaczonych słów kluczowych, podanych cytowań, czy wspólnych autorów. Na podstawie znanych metod analizy czystego tekstu (Rajman & Besançon (1997)), bez korzystania z metadanych, w ramach tej pracy zaproponowano nową metodę porównywania dokumentów. Metoda ta polega na połączeniu trzech różnych sposobów porównywania tekstów w celu zwrócenia uwagi na różne jego właściwości. Do podziału tekstów na kategorie zastosowano samoorganizujące się sieci Kohonena. Opracowano również nową metodę wyznaczania słów kluczowych w dokumentach, korzystającą z łańcuchów Markowa i sieci Kohonena oraz z informacji zdobytych podczas porównywania dokumentów.

2 Główne wyniki rozprawy

Główne wyniki niniejszej pracy zostały przedstawione w trzech rozdziałach:

2.1 Rozdział 1. Metody analizy tekstu

Metody analizy danych (wskazywane między innymi w Mannila *et al.* (1994)), które sprawdzają się dla danych konkretnie zdefiniowanych (jak na przykład ściśle określone zbiory danych z bioinformatyki (Lord *et al.* (2003)), czy posiadające precyzyjną strukturę strony www (Cruz *et al.* (1998))) nie są wystarczająco elastyczne dla danych nie posiadających ustalonej struktury. Dlatego prowadzonych jest wiele

badań nad metodami pozwalającymi określić tę strukturę automatycznie, jak Ferrer i Cancho & Solé (2001), czy Nahm & Mooney (2000).

W rozdziale tym przedstawiono znane metody analizy tekstu i porównywania dokumentów (Cavnar & Trenkle (1994), Arimura *et al.* (2000)). Zdefiniowano odległości pomiędzy dokumentami (bazujące na częstości słów, n-gramów i złożoności Kołmogorowa) i na ich podstawie wprowadzono oryginalny sposób wyznaczania powiązań, opierający się na kilku wcześniej przedstawionych miarach i spostrzeżeniu, że podobieństwo dokumentów naukowych jest związane z wieloma ich cechami, możliwymi do sprawdzenia w różny sposób. Sposób ten polega na wykorzystaniu wielu różnych miar odległości pomiędzy dokumentami do ich porównywania i kojarzenia tylko dokumentów, które są podobne względem wszystkich sprawdzanych cech. W rozdziale tym opisano również zaprojektowane w ramach niniejszej pracy szybkie algorytmy obliczania zaproponowanych miar oparte o drzewa tekstów.

2.2 Rozdział 2. Kategoryzacja tekstów

Głównym wynikiem tego rozdziału jest nowa metoda kategoryzacji dokumentów (przez kategoryzację należy rozumieć przydzielanie konkretnych tekstów do konkretnych kategorii (Frank *et al.* (2000))), bazująca na łatwo wykrywalnych zależnościach pomiędzy nimi oraz na dynamicznych sieciach neuronowych typu Kohonena (Kohonen & Honkela (2007), Kohonen (1984), Kohonen & Somervuo (2002)). Jako podstawowy algorytm kategoryzacji przyjęto algorytm samoorganizujących się sieci Kohonena dla elementów symbolicznych (Cilibrasi & Vitányi (2005)), który w ramach niniejszej pracy dostosowano do potrzeb kategoryzacji dokumentów, po raz pierwszy wykorzystując jako bazę zdefiniowane w poprzednim rozdziale odległości. Pokazano, że wykorzystanie wszystkich odległości zaprezentowanych w rozdziale 1 gwarantuje dobrą kategoryzację.

W ramach niniejszej pracy zaproponowano również modyfikację sieci neuronowych Kohonena, wprowadzając dynamiczną możliwość zmiany ich struktury, poprzez usuwanie zbędnych węzłów (Jankowski (2003)) i zastąpienie standardowego sąsiedztwa węzłów przez uznanie za sąsiadów węzłów o prototypach kategorii najbliższych względem prototypu węzła rozpatrywanego, oraz wprowadzono modyfikację kroku zmiany prototypów węzłów poprzez zastąpienie wyznaczania uogólnionej mediany algorytmem aproksymacyjnym opartym o ideę algorytmów genetycznych.

Jak wynika z Tan (1999) *text mining* jest w obecnym czasie uważany za ważniejszy od *data miningu*, działającego zwykle na bazach danych posiadających dobrze zdefiniowaną strukturę, i znajduje szerokie zastosowania w biznesie i przedsięwzięciach komercyjnych. Właśnie brak zdefiniowanej struktury języka naturalnego jest

największym problemem utrudniającym wybieranie z tekstu istotnych danych.

Metoda ta pozwala efektywnie wyznaczyć zbiory dokumentów, które są podobne do siebie treściowo i strukturalnie.

2.3 Rozdział 3. Wyznaczanie słów kluczowych

W tym rozdziale podjęto dyskusję na temat sposobów wyznaczania słów kluczowych.

Głównym wynikiem jest zaprojektowany w ramach niniejszej pracy nowy sposób wyznaczania słów kluczowych, wykorzystujący kategoryzację słów w dokumencie opartą o sieci neuronowe Kohonena. Sieci te zostały wzmocnione poprzez sprawdzanie częściowych wyników kategoryzacji w oparciu o dane zgromadzone podczas porównywania dokumentów. Dodatkowo został zaproponowany algorytm wykorzystujący łańcuchy Markowa (Bremaud (2001)), zainspirowany algorytmem PageRank (Page *et al.* (1999)), wyznaczający wstępny ranking słów na podstawie ich położenia w dokumencie. Przedstawione wyniki składają się na nową, oryginalną metodę wyznaczania słów kluczowych w dokumentach naukowych.

Prezentowany algorytm wyznacza słowa najczęściej występujące w tekście, zwracając uwagę na to, w jakim sąsiedztwie się one znajdują. Sąsiedztwo to ma wpływ na końcową ocenę przydatności wybranych słów. Dla każdej grupy lokalnie bliskich słów (znajdujących się w pobliżu siebie i w jednej wynikowej kategorii) wyznacza się jej pozycję rankingową. Na podstawie tego rankingu i częstości występowania słów wyznacza się końcową listę wynikową zawierającą większość słów kluczowych.

Dodatkowo, na końcową ocenę każdego słowa mają wpływ odległości dokumentu analizowanego od innych dokumentów z repozytorium zawierających to słowo.

Kluczowym elementem prezentowanych rozważań było wykorzystanie danych zdobytych podczas analizy i porównywania pozostałych tekstów. Okazuje się, że im więcej wiedzy algorytm zdobył (tzn. im więcej dokumentów przeanalizował), tym dokładniej może wyznaczać słowa kluczowe dla kolejnych dokumentów. Dzięki tworzeniu grup lokalnie bliskich słów, algorytm potrafi zidentyfikować i umieścić w jednej kategorii wyrażenia kluczowe składające się z więcej niż jednego słowa.

3 Uwagi końcowe

Uzyskane w ramach niniejszej rozprawy wyniki zostały opublikowane w następujących pracach:

- „*Scientific Documents Management System. Application of Kohonen Neural Networks with Reinforcement in Keywords Extraction*" (Zyglarski & Bała

(2009))

- „*Neural Networks Aided Automatic Keywords Selection*" (Zyglarski & Bała (2010b))
- „*Web Services Based Scientific Article Manager*" (Zyglarski *et al.* (2008))
- „*Document Management System based on Neural Networks*" (Zyglarski (2009)).

Przygotowane zostały również kolejne prace:

- „*Genetic Algorithms and Dynamic Neural Networks in Data Categorization*" (Zyglarski (2010))[1],
- „*Keywords Extraction. Selecting Keywords in Natural Language Texts with Markov Chains and Neural Networks*" (Zyglarski & Bała (2010a))[2].

Wyniki pracy zostały zaprezentowane na następujących konferencjach:

- Information Systems Architecture and Technology (ISAT) 2008,
- Sejmik Młodych Informatyków 2009,
- IC3K Knowledge Management and Information Sharing Conference 2009,

oraz będą prezentowane na konferencjach:

- Sejmik Młodych Informatyków 2010.

Bibliografia

Albert, Réka, Jeong, Hawoong, & Barabási, Albert-László. 1999. Diameter of the world-wide web. *Science*, **401**(Septmeber), 130–131.

Arimura, Hiroki, Abe, Junichiro, Fujino, Ryoichi, Sakamoto, Hiroshi, Shimozono, Shinichi, & Arikawa, Setsuo. 2000. Text data mining: Discovery of important keywords in the cyberspace. *Digital libraries: Research and practice, kyoto international conference on*, **0**, 220. doi:http://doi.ieeecomputersociety.org/10.1109/DLRP.2000.942178.

[1]Praca przyjęta do publikacji na SMI 2010.
[2]Praca zgłoszona do publikacji.

Bremaud, Pierre. 2001. *Markov chains: Gibbs fields, monte carlo simulation, and queues*. Corrected edn. Springer-Verlag New York Inc. Dostępne na: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0387985093.

Cavnar, William B., & Trenkle, John M. 1994. N-gram-based text categorization. *Pages 161–175 of: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Dostępne na: citeseer.ist.psu.edu/68861.html.

Cilibrasi, Rudi, & Vitányi, Paul M. B. 2005. Clustering by compression. *Ieee transactions on information theory*, **51**(4), 1523–1545.

Cruz, Isabel F., Borisov, Slava, Marks, Michael A., & Webb, Timothy R. 1998. Measuring structural similarity among Web documents: Preliminary results. *Lecture notes in computer science*, **1375**, 513–?? Dostępne na: http://link.springer-ny.com/link/service/series/0558/bibs/1375/13750513.htm;http://link.springer-ny.com/link/service/series/0558/papers/1375/13750513.pdf.

Ferrer i Cancho, Ramon, & Solé, Ricard V. 2001. The small-world of human language. *Proceedings of the royal society of london b*, **268**(1482), 2261–2265.

Frank, Eibe, Chui, Chang, & Witten, Ian H. 2000. Text categorization using compression models. *Pages 200–209 of: In proceedings of dcc-00, ieee data compression conference, snowbird, us*. IEEE Computer Society Press.

Jankowski, Norbert. 2003. *Ontogeniczne sieci neuronowe. o sieciach zmieniających swoją strukturę*. Akademicka Oficyna Wydawnicza EXIT.

Kohonen, T. 1984. *Self-organization and associative memory (2nd edition)*. Berlin: Springer-Verlag.

Kohonen, Teuvo, & Honkela, Timo. 2007. Kohonen network. *Scholarpedia*, **2**(1), 1568. Dostępne na: http://www.scholarpedia.org/article/Kohonen_network.

Kohonen, Teuvo, & Somervuo, Panu. 2002. How to make large self-organizing maps for nonvectorial data. *Neural networks*, **15**(8-9), 945–952.

Lord, Phillip, Wroe, Chris, Stevens, Robert, Goble, Carole, Miles, Simon, Moreau, Luc, Decker, Keith, Payne, Terry, & Papay, Juri. 2003. Semantic and personalised service discovery. Department of Mathematics and Computing Science, Saint Mary's University. Dostępne na: http://eprints.ecs.soton.ac.uk/9455/1/WIKGGI03Semantic.pdf;http://eprints.ecs.soton.ac.uk/9455/.

Mannila, Heikki, Toivonen, Hannu, & Verkamo, A. Inkeri. 1994 (July). Efficient algorithms for discovering association rules. *Pages 181–192 of:* Fayyad, Usama M., & Uthurusamy, Ramasamy (eds), *Aaai workshop on knowledge discovery in databases (kdd-94)*. Dostępne na: ftp://ftp.cs.helsinki.fi/pub/Reports/by_Project/PMDM/Efficient_Algorithms_for_Discovering_Association_Rules.ps.gz.

Nahm, Un Yong, & Mooney, Raymond J. 2000. A mutually beneficial integration of data mining and information extraction. *Pages 627–632 of: Aaai/iaai*. AAAI Press / The MIT Press.

Page, Lawrence, Brin, Sergey, Motwani, Rajeev, & Winograd, Terry. 1999 (November). *The pagerank citation ranking: Bringing order to the web.* Technical Report 1999-66. Stanford InfoLab. Previous number = SIDL-WP-1999-0120. Dostępne na: http://ilpubs.stanford.edu:8090/422/.

Rajman, Martin, & Besançon, E. 1997. Natural language techniques for text mining applications. *Page 50 of: Ds-7*.

Tan, Ah-hwee. 1999. *Text mining: The state of the art and the challenges*. Dostępne na: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.6973;http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf.

Zyglarski, Błażej. 2009. Document management system based on neural networks. *Polish Journal of Environmental Studies*, **18**(3B), 416–420.

Zyglarski, Błażej. 2010. Genetic algorithms and dynamic neural networks in data categorization. *In: Proceedings of SMI 2010*.

Zyglarski, Błażej, & Bała, Piotr. 2009. Scientific documents management system. Application of Kohonen neural networks with reinforcement in keywords extraction. *Pages 55–62 of: In proceedings of IC3K 2009*. INSTICC.

Zyglarski, Błażej, & Bała, Piotr. 2010a. *Keywords extraction. selecting keywords in natural language texts with Markov chains and neural networks*.

Zyglarski, Błażej, & Bała, Piotr. 2010b. Neural networks aided automatic keywords selection. *Communications in Computer and Information Science*.

Zyglarski, Błażej, Bała, Piotr, & Schreiber, Tomasz. 2008. Web services based scientific article manager. *Pages 205–215 of: Information Systems Architecture and Technology, Web Information Systems: Models, Concepts and Challenges*. Wrocław University of Technology.