Olfa Nasraoui Myra Spiliopoulou
Jaideep Srivastava Bamshad Mobasher
Brij Masand (Eds.)

Advances in Web Mining and Web Usage Analysis

LNAI 4811

8th International Workshop
on Knowledge Discovery on the Web, WebKDD 2006
Philadelphia, PA, USA, August 2006, Revised Papers

Springer

Olfa Nasraoui Myra Spiliopoulou
Jaideep Srivastava Bamshad Mobasher
Brij Masand (Eds.)

Advances in Web Mining and Web Usage Analysis

8th International Workshop
on Knowledge Discovery on the Web, WebKDD 2006
Philadelphia, PA, USA, August 20, 2006
Revised Papers

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Olfa Nasraoui
University of Louisville
Louisville, KY 40292, USA
E-mail: olfa.nasraoui@louisville.edu

Myra Spiliopoulou
Otto-von-Guericke-Universität Magdeburg
39106 Magdeburg, Germany
E-mail: myra@iti.cs.uni-magdeburg.de

Jaideep Srivastava
University of Minnesota
Minneapolis, MN 55455, USA
E-mail: srivasta@cs.umn.edu

Bamshad Mobasher
DePaul University
Chicago, IL 60604, USA
E-mail: mobasher@cs.depaul.edu

Brij Masand
Data Miners Inc.
Boston, MA 02114, USA
E-mail: brij@data-miners.com

Library of Congress Control Number: 2007941802

CR Subject Classification (1998): I.2, H.2.8, H.3-5, K.4, C.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-77484-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-77484-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12210404 06/3180 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 4811

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Preface

This book contains the postworkshop proceedings with selected revised papers from the 8th international workshop on knowledge discovery from the Web, WEBKDD 2006. The WEBKDD workshop series has taken place as part of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) since 1999.

The discipline of data mining delivers methodologies and tools for the analysis of large data volumes and the extraction of comprehensible and non-trivial insights from them. Web mining, a much younger discipline, concentrates on the analysis of data pertinent to the Web. Web mining methods are applied on usage data and Web site content; they strive to improve our understanding of how the Web is used, to enhance usability and to promote mutual satisfaction between e-business venues and their potential customers.

In the last few years, the interest for the Web as a medium for communication, interaction and business has led to new challenges and to intensive, dedicated research. Many of the infancy problems in Web mining have been solved by now, but the tremendous potential for new and improved uses, as well as misuses, of the Web are leading to new challenges.

The theme of the WebKDD 2006 workshop was "Knowledge Discovery on the Web", encompassing lessons learned over the past few years and new challenges for the years to come. While some of the infancy problems of Web analysis have been solved and proposed methodologies have reached maturity, the reality poses new challenges: The Web is evolving constantly; sites change and user preferences drift. And, most of all, a Web site is more than a see-and-click medium; it is a venue where a user interacts with a site owner or with other users, where group behavior is exhibited, communities are formed and experiences are shared.

The WebKDD 2006 workshop invited research results in all areas of Web mining and Semantic Web mining, with an emphasis on a seven years' update: What are the lessons learned on algorithms, semantics, data preparation, data integration and applications of the Web? How do new technologies, like adaptive mining methods, stream mining algorithms and techniques for the Grid, apply to Web mining? What new challenges are posed by new forms of data, especially flat texts, documents, pictures and streams, as well as the emergence of Web communities? How do we study the evolution of the Web and its effects on searching and browsing behavior? Which lessons have we learned about usability, e-commerce applications, personalization, recommendation engines, Web marketplaces, Web search, Web security, and misuse and abuse of the Web and its services? WebKDD 2006 attempted to address these challenging questions, with an emphasis on expanding the horizon of traditional Web mining to embrace and keep up with recent and emerging trends and emphasis on the Web

domain, such as mining search engine queries, mining Web evolution, robustness of recommender systems, and mining blogs for sentiment analysis.

In the first paper, "Adaptive Web site Design using Caching Algorithms", Justin Brickell, Inderjit S. Dhillon, and Dharmendra S. Modha present improved online algorithms for shortcut link selection that are based on a novel analogy drawn between shortcutting and caching. In the same way that cache algorithms predict which memory pages will be accessed in the future, the proposed algorithms predict which Web pages will be accessed in the future. These algorithms are efficient and can consider accesses over a long period of time, but give extra weight to recent accesses. Experiments show significant improvement in the utility of shortcut links selected by the proposed algorithm as compared to those selected by existing algorithms.

In the second paper,"Incorporating Usage Information into Average-Clicks Algorithm", Kalyan Beemanapalli, Ramya Rangarajan, and Jaideep Srivastava present an extension to the Average-Clicks Algorithm, called "Usage Aware Average-Clicks," where the static Web link structure graph is combined with the dynamic Usage Graph (built using the information available from the Web logs) to assign different weights to links on a Web page and hence capture the user's intuition of distance between two Web pages more accurately. This method has been used as a new metric to calculate the page similarities in a recommendation engine to improve its predictive power.

In "Nearest-Biclusters Collaborative Filtering", Panagiotis Symeonidis, Alexandros Nanopoulos, Apostolos Papadopoulos, and Yannis Manolopoulos use biclustering to disclose the duality between users and items in Nearest-neighbor Collaborative Filtering, by grouping them in both dimensions simultaneously. A novel nearest-biclusters algorithm is proposed, that uses a new similarity measure that achieves partial matching of users' preferences. Performance evaluation results are offered, which show that the proposed method improves substantially the performance of the CF process.

In "Fast Categorization of Web Documents Represented by Graphs", Alex Markov, Mark Last, and Abraham Kandel address the limitations of the vector-space model of information retrieval. This traditional model does not capture important structural information, such as the order and proximity of word occurrence, the location of a word within the document, or mark-up information. Three new hybrid approaches to Web document classification are presented, built upon both graph and vector space representations, thus preserving the benefits and discarding the limitations of each. The hybrid methods outperform, in most cases, vector-based models using two model-based classifiers (C4.5 decision-tree algorithm and probabilistic Naïve Bayes) on several benchmark Web document collections.

In "Leveraging Structural Knowledge for Hierarchically Informed Keyword Weight Propagation in the Web," Jong Wook Kim and K. Selcuk Candan elaborate on indexing Web documents that have non-atomic structures, such as navigational/semantic hierarchies on the Web. A novel keyword and keyword weight

propagation technique is proposed to properly enrich the data nodes in structured content. The approach first relies on understanding the context provided by the relative content relationships between entries in the structure, and then leveraging this information for relative-content preserving keyword propagation. Experiments show a significant improvement in precision with the proposed keyword propagation algorithm.

In the paper "How to Define Searching Sessions on Web Search Engines," Bernard J. Jansen, Amanda Spink, and Vinish Kathuria investigate three methods for defining a session on Web search engines. The authors examine 2,465,145 interactions from 534, 507 Web searchers, and compare defining sessions using: (1) Internet Protocol address and cookie; (2) Internet Protocol address, cookie, and a temporal limit on intra-session interactions; and (3) Internet Protocol address, cookie, and query reformulation patterns. Research results show that defining sessions by query reformulation along with Internet Protocol address and cookie, provides the best measure, resulting in an 82% increase in the number of sessions; while for all methods, mean session length was fewer than three queries and the mean session duration was less than 30 minutes. Implications are that unique sessions may be a better indicator than the common industry metric of unique visitors for measuring search traffic.

In the paper "Incorporating Concept Hierarchies into Usage Mining Based Recommendations," Amit Bose, Kalyan Beemanapalli, Jaideep Srivastava, and Sigal Sahar address the limitation of most recommendation models in their ability to use domain knowledge in the form of conceptual and structural characteristics of a Web site. Conceptual content organization can play an important role in the quality of recommendations, and forms the basis of resources like Google Directory, Yahoo Directory and Web-content management systems. The authors propose a novel technique to incorporate the conceptual characteristics of a Web site into a usage-based recommendation model. The authors use a framework based on biological sequence alignment. Similarity scores play a crucial role in such a construction, and a scoring system that is generated from the Web site's concept hierarchy is introduced. These scores fit seamlessly with other quantities used in similarity calculation like browsing order and time spent on a page. Additionally they demonstrate a simple, extensible system for assimilating more domain knowledge. Experimental results illustrate the benefits of using a concept hierarchy.

In the paper "A Random-Walk-Based Scoring Algorithm Applied to Recommender Engines," Augusto Pucci, Marco Gori, and Marco Maggini present "ItemRank," a random-walk-based scoring algorithm, which can be used to rank products according to expected user preferences, in order to recommend top-rank items to potentially interested users. The authors tested their algorithm on the MovieLens data set, which contains data collected from a popular recommender system on movies, and compared ItemRank with other state-of-the-art ranking techniques, showing that ItemRank performs better than the other techniques, while being less complex than other algorithms with respect to memory usage

and computational cost. The paper also presents an analysis that helps to discover some intriguing properties of the MovieLens data set, that has been widely exploited as a benchmark for evaluating recently proposed approaches to recommender system.

In "Towards a Scalable k-NN CF Algorithm: Exploring Effective Applications of Clustering," Al Mamunur Rashid, Shyong K. Lam, Adam LaPitz, George Karypis, and John Riedl address the need for specially designed CF algorithms that can gracefully cope with the vast size of the data representing customers and items in typical e-commerce systems. Many algorithms proposed thus far, where the principal concern is recommendation quality, may be too expensive to operate in a large-scale system. The authors propose ClustKNN, a simple and intuitive algorithm that is well suited for large data sets. The method first compresses data tremendously by building a straightforward but efficient clustering model. Recommendations are then generated quickly by using a simple Nearest Neighbor-based approach. The feasibility of ClustKNN is demonstrated both analytically and empirically, and a comparison with a number of other popular CF algorithms shows that, apart from being highly scalable and intuitive, ClustKNN provides very good recommendation accuracy.

In "Detecting Profile Injection Attacks in Collaborative Filtering: A Classification-Based Approach," Chad Williams, Bamshad Mobasher, Robin Burke, and Runa Bhaumik address the vulnerability of Collaborative recommender systems to profile injection attacks. By injecting a larger number of biased profiles into a system, attackers can manipulate the predictions of targeted items. To decrease this risk, researchers have begun to study mechanisms for detecting and preventing profile injection attacks. In this paper, the authors extend their previous work that proposed several attributes for attack detection and for classification of attack profiles, through a more detailed analysis of the informativeness of these attributes as well as an evaluation of their impact at improving the robustness of recommender systems.

In "Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection", Kathleen T. Durant and Michael D. Smith investigate data mining techniques that can automatically identify the political *sentiment* of Web log posts, and thus help bloggers categorize and filter this exploding information source. They illustrate the effectiveness of supervised learning for sentiment classification on Web log posts, showing that a Naïve Bayes classifier coupled with a forward feature selection technique can on average correctly predict a postings sentiment 89.77% of the time. It significantly outperforms Support Vector Machines at the 95% confidence level with a confidence interval of [1.5, 2.7]. The feature selection technique provides on average an 11.84% and a 12.18% increase for Naïve Bayes and Support Vector Machines results, respectively. Previous sentiment classification research achieved an 81% accuracy using Naïve Bayes and 82.9% using SVMs on a movie domain corpus.

In "Analysis of Web Search Engine Query Session and Clicked Documents," David Nettleton, Liliana Calderón-Benavides, and Ricardo Baeza-Yates present

the analysis of a Web search engine query log from two different perspectives: the query session and the clicked document. In the query session perspective, they process and analyze a Web search engine query and click data for the query session (query + clicked results) conducted by the user. They initially state some hypotheses for possible user types and quality profiles for the user session, based on descriptive variables of the session. In the clicked document perspective, they repeat the process from the perspective of the documents (URL's) selected. They also initially define possible document categories and select descriptive variables to define the documents. They apply a systematic data mining process to click data, contrasting non- supervised (Kohonen) and supervised (C4.5) methods to cluster and model the data, in order to identify profiles and rules which relate to theoretical user behavior and user session "quality," from the point of view of user session, and to identify document profiles which relate to theoretical user behavior, and document (URL) organization, from the document perspective.

In "Understanding Content Reuse on the Web: Static and Dynamic Analyses", Ricardo Baeza-Yates, Álvaro Pereira, and Nivio Ziviani present static and dynamic studies of duplicate and near-duplicate documents in the Web. The static and dynamic studies involve the analysis of similar content among pages within a given snapshot of the Web and how pages in an old snapshot are reused to compose new documents in a more recent snapshot. With experiments using four snapshots of the Chilean Web, they identify duplicates (in the static study) in both parts of the Web graph – reachable (connected by links) and unreachable components (unconnected) – aiming to identify where duplicates occur more frequently. They show that the number of duplicates in the Web seems to be much higher than previously reported (about 50% higher) and in their data the duplicated in the unreachable Web is 74.6% higher than the number of duplicates in the reachable component of the Web graph. In the dynamic study, they show that some of the old content is used to compose new pages. If a page in a newer snapshot has content of a page in an older snapshot, they consider that the source is a parent of the new page. They state the hypothesis that people use search engines to find pages and republish their content as a new document, and present evidence that this happens for part of the pages that have parents. In this case, part of the Web content is biased by the ranking function of search engines.

We would like to thank the authors of all submitted papers. Their creative efforts have led to a rich set of good contributions for WebKDD 2006. We would also like to express our gratitude to the members of the Program Committee for their vigilant and timely reviews, namely (in alphabetical order): Corin Anderson, Ricardo A. Baeza-Yates, Bettina Berendt, Zheng Chen, Ed H. Chi, Brian D. Davison, Wei Fan, Fabio Grandi, Michael Hahsler, Xin Jin, Thorsten Joachims, George Karypis, Ravi Kumar, Vipin Kumar, Mark Last, Mark Levene, Ee-Peng Lim, Huan Liu, Stefano Lonardi, Alexandros D. Nanopoulos, Georgios Paliouras, Aniruddha G. Pant, Jian Pei, Ellen Spertus, Andrew Tomkins, and Mohammed

J. Zaki. O. Nasraoui gratefully acknowledges the support of the US National Science Foundation as part of NSF CAREER award IIS-0133948.

September 2007

Olfa Nasraoui
Myra Spiliopoulou
Jaideep Srivastava
Bamshad Mobasher
Brij Masand

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 4874: J. Neves, M.F. Santos, J.M. Machado (Eds.), Progress in Artificial Intelligence. XVIII, 704 pages. 2007.

Vol. 4869: F. Botana, T. Recio (Eds.), Automated Deduction in Geometry. X, 213 pages. 2007.

Vol. 4850: M. Lungarella, F. Iida, J. Bongard, R. Pfeifer (Eds.), 50 Years of Artificial Intelligence. X, 399 pages. 2007.

Vol. 4845: N. Zhong, J. Liu, Y. Yao, J. Wu, S. Lu, K. Li (Eds.), Web Intelligence Meets Brain Informatics. XI, 516 pages. 2007.

Vol. 4830: M.A. Orgun, J. Thornton (Eds.), AI 2007: Advances in Artificial Intelligence. XIX, 841 pages. 2007.

Vol. 4828: M. Randall, H.A. Abbass, J. Wiles (Eds.), Progress in Artificial Life. XII, 402 pages. 2007.

Vol. 4827: A. Gelbukh, Á.F. Kuri Morales (Eds.), MICAI 2007: Advances in Artificial Intelligence. XXIV, 1234 pages. 2007.

Vol. 4826: P. Perner, O. Salvetti (Eds.), Advances in Mass Data Analysis of Signals and Images in Medicine, Biotechnology and Chemistry. X, 183 pages. 2007.

Vol. 4819: T. Washio, Z.-H. Zhou, J.Z. Huang, X. Hu, J. Li, C. Xie, J. He, D. Zou, K.-C. Li, M.M. Freire (Eds.), Emerging Technologies in Knowledge Discovery and Data Mining. XIV, 675 pages. 2007.

Vol. 4811: O. Nasraoui, M. Spiliopoulou, J. Srivastava, B. Mobasher, B. Masand (Eds.), Advances in Web Mining and Web Usage Analysis. XII, 247 pages. 2007.

Vol. 4798: Z. Zhang, J.H. Siekmann (Eds.), Knowledge Science and Engineering and Management. XVI, 669 pages. 2007.

Vol. 4795: F. Schilder, G. Katz, J. Pustejovsky (Eds.), Annotating, Extracting and Reasoning about Time and Events. VII, 141 pages. 2007.

Vol. 4790: N. Dershowitz, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIII, 562 pages. 2007.

Vol. 4788: D. Borrajo, L. Castillo, J.M. Corchado (Eds.), Current Topics in Artificial Intelligence. XI, 280 pages. 2007.

Vol. 4775: A. Esposito, M. Faundez-Zanuy, E. Keller, M. Marinaro (Eds.), Verbal and Nonverbal Communication Behaviours. XII, 325 pages. 2007.

Vol. 4772: H. Prade, V.S. Subrahmanian (Eds.), Scalable Uncertainty Management. X, 277 pages. 2007.

Vol. 4766: N. Maudet, S. Parsons, I. Rahwan (Eds.), Argumentation in Multi-Agent Systems. XII, 211 pages. 2007.

Vol. 4755: V. Corruble, M. Takeda, E. Suzuki (Eds.), Discovery Science. XI, 298 pages. 2007.

Vol. 4754: M. Hutter, R.A. Servedio, E. Takimoto (Eds.), Algorithmic Learning Theory. XI, 403 pages. 2007.

Vol. 4737: B. Berendt, A. Hotho, D. Mladenic, G. Semeraro (Eds.), From Web to Social Web: Discovering and Deploying User and Content Profiles. XI, 161 pages. 2007.

Vol. 4733: R. Basili, M.T. Pazienza (Eds.), AI*IA 2007: Artificial Intelligence and Human-Oriented Computing. XVII, 858 pages. 2007.

Vol. 4724: K. Mellouli (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty. XV, 914 pages. 2007.

Vol. 4722: C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, D. Pelé (Eds.), Intelligent Virtual Agents. XV, 425 pages. 2007.

Vol. 4720: B. Konev, F. Wolter (Eds.), Frontiers of Combining Systems. X, 283 pages. 2007.

Vol. 4702: J.N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, A. Skowron (Eds.), Knowledge Discovery in Databases: PKDD 2007. XXIV, 640 pages. 2007.

Vol. 4701: J.N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, A. Skowron (Eds.), Machine Learning: ECML 2007. XXII, 809 pages. 2007.

Vol. 4696: H.-D. Burkhard, G. Lindemann, R. Verbrugge, L.Z. Varga (Eds.), Multi-Agent Systems and Applications V. XIII, 350 pages. 2007.

Vol. 4694: B. Apolloni, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. XXIX, 1126 pages. 2007.

Vol. 4693: B. Apolloni, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. XXXII, 1380 pages. 2007.

Vol. 4692: B. Apolloni, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LV, 882 pages. 2007.

Vol. 4687: P. Petta, J.P. Müller, M. Klusch, M. Georgeff (Eds.), Multiagent System Technologies. X, 207 pages. 2007.

Vol. 4682: D.-S. Huang, L. Heutte, M. Loog (Eds.), Advanced Intelligent Computing Theories and Applications. XXVII, 1373 pages. 2007.

Vol. 4676: M. Klusch, K.V. Hindriks, M.P. Papazoglou, L. Sterling (Eds.), Cooperative Information Agents XI. XI, 361 pages. 2007.

Vol. 4667: J. Hertzberg, M. Beetz, R. Englert (Eds.), KI 2007: Advances in Artificial Intelligence. IX, 516 pages. 2007.

Vol. 4660: S. Džeroski, L. Todorovski (Eds.), Computational Discovery of Scientific Knowledge. X, 327 pages. 2007.

Vol. 4659: V. Mařík, V. Vyatkin, A.W. Colombo (Eds.), Holonic and Multi-Agent Systems for Manufacturing. VIII, 456 pages. 2007.

Vol. 4651: F. Azevedo, P. Barahona, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. VIII, 185 pages. 2007.

Vol. 4648: F. Almeida e Costa, L.M. Rocha, E. Costa, I. Harvey, A. Coutinho (Eds.), Advances in Artificial Life. XVIII, 1215 pages. 2007.

Vol. 4635: B. Kokinov, D.C. Richardson, T.R. Roth-Berghofer, L. Vieu (Eds.), Modeling and Using Context. XIV, 574 pages. 2007.

Vol. 4632: R. Alhajj, H. Gao, X. Li, J. Li, O.R. Zaïane (Eds.), Advanced Data Mining and Applications. XV, 634 pages. 2007.

Vol. 4629: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. XVII, 663 pages. 2007.

Vol. 4626: R.O. Weber, M.M. Richter (Eds.), Case-Based Reasoning Research and Development. XIII, 534 pages. 2007.

Vol. 4617: V. Torra, Y. Narukawa, Y. Yoshida (Eds.), Modeling Decisions for Artificial Intelligence. XII, 502 pages. 2007.

Vol. 4612: I. Miguel, W. Ruml (Eds.), Abstraction, Reformulation, and Approximation. XI, 418 pages. 2007.

Vol. 4604: U. Priss, S. Polovina, R. Hill (Eds.), Conceptual Structures: Knowledge Architectures for Smart Applications. XII, 514 pages. 2007.

Vol. 4603: F. Pfenning (Ed.), Automated Deduction – CADE-21. XII, 522 pages. 2007.

Vol. 4597: P. Perner (Ed.), Advances in Data Mining. XI, 353 pages. 2007.

Vol. 4594: R. Bellazzi, A. Abu-Hanna, J. Hunter (Eds.), Artificial Intelligence in Medicine. XVI, 509 pages. 2007.

Vol. 4585: M. Kryszkiewicz, J.F. Peters, H. Rybinski, A. Skowron (Eds.), Rough Sets and Intelligent Systems Paradigms. XIX, 836 pages. 2007.

Vol. 4578: F. Masulli, S. Mitra, G. Pasi (Eds.), Applications of Fuzzy Sets Theory. XVIII, 693 pages. 2007.

Vol. 4573: M. Kauers, M. Kerber, R. Miner, W. Windsteiger (Eds.), Towards Mechanized Mathematical Assistants. XIII, 407 pages. 2007.

Vol. 4571: P. Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition. XIV, 913 pages. 2007.

Vol. 4570: H.G. Okuno, M. Ali (Eds.), New Trends in Applied Artificial Intelligence. XXI, 1194 pages. 2007.

Vol. 4565: D.D. Schmorrow, L.M. Reeves (Eds.), Foundations of Augmented Cognition. XIX, 450 pages. 2007.

Vol. 4562: D. Harris (Ed.), Engineering Psychology and Cognitive Ergonomics. XXIII, 879 pages. 2007.

Vol. 4548: N. Olivetti (Ed.), Automated Reasoning with Analytic Tableaux and Related Methods. X, 245 pages. 2007.

Vol. 4539: N.H. Bshouty, C. Gentile (Eds.), Learning Theory. XII, 634 pages. 2007.

Vol. 4529: P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyk, W. Pedrycz (Eds.), Foundations of Fuzzy Logic and Soft Computing. XIX, 830 pages. 2007.

Vol. 4520: M.V. Butz, O. Sigaud, G. Pezzulo, G. Baldassarre (Eds.), Anticipatory Behavior in Adaptive Learning Systems. X, 379 pages. 2007.

Vol. 4511: C. Conati, K. McCoy, G. Paliouras (Eds.), User Modeling 2007. XVI, 487 pages. 2007.

Vol. 4509: Z. Kobti, D. Wu (Eds.), Advances in Artificial Intelligence. XII, 552 pages. 2007.

Vol. 4496: N.T. Nguyen, A. Grzech, R.J. Howlett, L.C. Jain (Eds.), Agent and Multi-Agent Systems: Technologies and Applications. XXI, 1046 pages. 2007.

Vol. 4483: C. Baral, G. Brewka, J. Schlipf (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 327 pages. 2007.

Vol. 4482: A. An, J. Stefanowski, S. Ramanna, C.J. Butz, W. Pedrycz, G. Wang (Eds.), Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. XIV, 585 pages. 2007.

Vol. 4481: J. Yao, P. Lingras, W.-Z. Wu, M.S. Szczuka, N.J. Cercone, D. Ślęzak (Eds.), Rough Sets and Knowledge Technology. XIV, 576 pages. 2007.

Vol. 4476: V. Gorodetsky, C. Zhang, V.A. Skormin, L. Cao (Eds.), Autonomous Intelligent Systems: Multi-Agents and Data Mining. XIII, 323 pages. 2007.

Vol. 4460: S. Aguzzoli, A. Ciabattoni, B. Gerla, C. Manara, V. Marra (Eds.), Algebraic and Proof-theoretic Aspects of Non-classical Logics. VIII, 309 pages. 2007.

Vol. 4457: G.M.P. O'Hare, A. Ricci, M.J. O'Grady, O. Dikenelli (Eds.), Engineering Societies in the Agents World VII. XI, 401 pages. 2007.

Vol. 4456: Y. Wang, Y.-m. Cheung, H. Liu (Eds.), Computational Intelligence and Security. XXIII, 1118 pages. 2007.

Vol. 4455: S. Muggleton, R. Otero, A. Tamaddoni-Nezhad (Eds.), Inductive Logic Programming. XII, 456 pages. 2007.

Vol. 4452: M. Fasli, O. Shehory (Eds.), Agent-Mediated Electronic Commerce. VIII, 249 pages. 2007.

Vol. 4451: T.S. Huang, A. Nijholt, M. Pantic, A. Pentland (Eds.), Artifical Intelligence for Human Computing. XVI, 359 pages. 2007.

Vol. 4442: L. Antunes, K. Takadama (Eds.), Multi-Agent-Based Simulation VII. X, 189 pages. 2007.

Vol. 4441: C. Müller (Ed.), Speaker Classification II. X, 309 pages. 2007.

Vol. 4438: L. Maicher, A. Sigel, L.M. Garshol (Eds.), Leveraging the Semantics of Topic Maps. X, 257 pages. 2007.

Vol. 4434: G. Lakemeyer, E. Sklar, D.G. Sorrenti, T. Takahashi (Eds.), RoboCup 2006: Robot Soccer World Cup X. XIII, 566 pages. 2007.

Vol. 4429: R. Lu, J.H. Siekmann, C. Ullrich (Eds.), Cognitive Systems. X, 161 pages. 2007.

Table of Contents

Adaptive Website Design Using Caching Algorithms

Justin Brickell[1], Inderjit S. Dhillon[1], and Dharmendra S. Modha[2]

[1] The University of Texas at Austin, Austin, TX, USA
[2] IBM Almaden Research Center, San Jose, CA, USA

Abstract. Visitors enter a website through a variety of means, including web searches, links from other sites, and personal bookmarks. In some cases the first page loaded satisfies the visitor's needs and no additional navigation is necessary. In other cases, however, the visitor is better served by content located elsewhere on the site found by navigating links. If the path between a user's current location and his eventual goal is circuitous, then the user may never reach that goal or will have to exert considerable effort to reach it. By mining site access logs, we can draw conclusions of the form "users who load page p are likely to later load page q." If there is no direct link from p to q, then it is advantageous to provide one. The process of providing links to users' eventual goals while skipping over the in-between pages is called *shortcutting*. Existing algorithms for shortcutting require substantial offline training, which make them unable to adapt when access patterns change between training sessions. We present improved online algorithms for shortcut link selection that are based on a novel analogy drawn between shortcutting and caching. In the same way that cache algorithms predict which memory pages will be accessed in the future, our algorithms predict which web pages will be accessed in the future. Our algorithms are very efficient and are able to consider accesses over a long period of time, but give extra weight to recent accesses. Our experiments show significant improvement in the utility of shortcut links selected by our algorithm as compared to those selected by existing algorithms.

1 Introduction

As websites increase in complexity, they run headfirst into a fundamental tradeoff: the more information that is available on the website, the more difficult it is for visitors to pinpoint the specific information that they are looking for. A well-designed website limits the impact of this tradeoff, so that even if the amount of information is increased significantly, locating that information becomes only marginally more difficult. Typically, site designers ease information overload by organizing the site content into a hierarchy of topics, and then providing a navigational tree that allows visitors to descend into the hierarchy and find the information they are looking for. In their paper on adaptive website design [12], Perkowitz and Etzioni describe these static, master-designed websites as "fossils cast in HTML." They claim that a site designer's *a priori* expectations for how a

O. Nasraoui et al. (Eds.): WebKDD 2006, LNAI 4811, pp. 1–20, 2007.

site will be used and navigated are likely to inaccurately reflect actual usage patterns, especially as the site adds new content over time. As it is infeasible for even the most dedicated site designer to understand the goals and access patterns of all site visitors, Perkowitz and Etzioni proposed building websites that mine their own access logs in order to automatically determine helpful self-modifications.

One example of a helpful modification is *shortcutting*, in which links are added between unlinked pages in order to allow visitors to reach their intended destinations with fewer clicks. Typically a limit N is imposed on the maximum number of outgoing shortcuts on any one particular page. The shortcutting problem can then be thought of as an optimization problem to choose the N shortcuts per page that minimize the number of clicks needed for future visitors to reach their goal pages. These shortcuts may be modified at any time based on past accesses in order to account for anticipated changes in the access patterns of future visitors. Finding an optimal solution to this problem would require an exact knowledge of the future and a precise way of determining each user's goal. However, shortcutting algorithms must provide shortcuts in an on-line framework, so the shortcuts must be chosen without knowledge of future accesses. Rather than solving the optimization problem exactly, shortcutting algorithms use heuristics and analyze past accesses in order to provide good shortcuts.

In this paper, we draw a novel analogy between shortcutting algorithms, which maintain an active set of shortcuts on each page, and caching algorithms, which maintain an active set of items in cache. The goal of caching algorithms—maximizing the fraction of future memory accesses for items in the cache—is analogous to the goal of shortcutting algorithms. The main contribution of this paper is the CACHECUT algorithm for shortcutting. By using replacement policies developed for caching applications, CACHECUT is able to run with less memory than other shortcutting algorithms, while producing better results. A second contribution is the FRONTCACHE algorithm, which uses similar caching techniques in order to select pages for promotion on the front page.

The remainder of this paper is organized as follows. In Section 2 we discuss related work in adaptive website problems. In Section 3 we give definitions for terms that are used throughout the paper. Section 4 gives a formulation of the shortcutting problem and presents two shortcutting algorithms from existing literature. In Section 5 we detail our CACHECUT algorithm for shortcutting, and in section 6 we describe the FRONTCACHE algorithm for promoting pages with links on the front page. Section 7 describes our experimental setup and the results of our experiments. Finally, in Section 8, we offer some concluding thoughts and suggest directions for future work.

2 Related Work

Perkowitz and Etzioni [11] issued the original challenge to the AI community to build adaptive web sites that learn visitor access patterns from the access log in order to automatically improve their organization and presentation. Their follow-up paper [12] presented several *global* adaptations that affect the presentation of

the website to all users. One adaptation from their paper is "index page synthesis," in which new pages are created containing collections of links to related but currently unlinked pages. In his thesis [10], Perkowitz presents the shortcutting problem as a global adaptive problem, in which links are added to each page to ease the browsing experience of all site visitors. Ramakrishnan *et al.* [13] have also done work in global adaptation; they observe that frustrated users who cannot find the content they are looking for are apt to use the "back" button. The authors scan the access log looking for these "backtracks" to identify documents that are misclassified in the site hierarchy, and correct these misclassifications.

Other work has explored adaptations that are *individual*, rather than global; sometimes this is referred to as *personalization*. It is increasingly common for portals to allow users to manually customize portions of their front pages [14]. For instance, a box with local weather information can be provided based on zip code information stored in a client cookie. The Newsjunkie system [7] provides personalized newsfeeds to users based on their news preferences. Personalization is easy when users provide both their identity and their desired customizations, but more difficult when the personalization must take place automatically without explicit management on the part of the user. The research community has made some stabs at the more difficult problem. Anderson and Horvitz [2] automatically generate a personal web page that contains all of the content that the target user visits during a typical day of surfing. Frayling *et al.* [9] improve the "back" button so that it jumps to key pages in the navigation session. Eirinaki and Vazirgiannis [6] give a survey of the use of web mining for personalization.

Operating at a level between global adaptations and individual adaptations are *group* adaptations. The mixture-model variants of the MinPath algorithm [1] are examples of group-based shortcutting algorithms. When suggesting shortcuts to a website visitor, they first classify that visitor based on browsing behavior, and then provide shortcuts that are thought to be useful to that class of visitors. Classifying users requires examining the "trails" or "clickstreams" in the access log, which are the sequences of pages accessed by individual visitors. Other researchers have investigated trails without the intention of adapting a website. Banerjee and Ghosh [3] use trails to cluster users. Cooley *et al.* [5] discover association rules to find correlations such as "60% of clients who accessed page A also accessed page B." Yang *et al.* [15] conduct temporal event prediction, in which they also estimate *when* the client is likely to access B.

Our work follows the global model of shortcutting [10], in which shortcutting is viewed as a global adaptation that adds links to each page that are the same for every visitor. Like Perkowitz' algorithm, when choosing shortcuts for a page p we pay close attention to the number of times other pages q are accessed *after* p within a trail; however, our algorithm provides improvements in the form of reduced memory requirements and higher-quality shortcuts. A related work by Yang and Zhang [16] sought to create an improved replacement policy for website caching by analyzing the access log. In contrast, our work uses existing caching policies to create an improved website.