Olfa Nasraoui Osmar Zaïane
Myra Spiliopoulou Bamshad Mobasher
Brij Masand Philip S. Yu (Eds.)

Advances in Web Mining and Web Usage Analysis

LNAI 4198

7th International Workshop
on Knowledge Discovery on the Web, WebKDD 2005
Chicago, IL, USA, August 2005, Revised Papers

Springer

Olfa Nasraoui Osmar Zaïane
Myra Spiliopoulou Bamshad Mobasher
Brij Masand Philip S. Yu (Eds.)

Advances in Web Mining and Web Usage Analysis

7th International Workshop
on Knowledge Discovery on the Web, WebKDD 2005
Chicago, IL, USA, August 21, 2005
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
Speed School of Engineering, Louisville KY 40292
E-mail: olfa.nasraoui@louiswville.edu

Osmar Zaïane
University of Alberta, Edmonton, AB, T6G2E8, Canada
E-mail: zaiane@ualberta.ca

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

Bamshad Mobasher
School of Computer Science, 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

Philip S. Yu
IBM T. J. Inc., N.Y. 10598, USA
E-mail: psyu@us.ibm.com

Library of Congress Control Number: 2006933535

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 3-540-46346-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-46346-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11891321 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 4198

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Preface

This book contains the postworkshop proceedings of the 7th International Workshop on Knowledge Discovery from the Web, WEBKDD 2005. The WEBKDD workshop series takes 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 years, the interest for the Web as 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 now been solved 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 2005 workshop was "Taming Evolving, Expanding and Multi-faceted Web Clickstreams." While solutions on some of the infancy problems of Web analysis have reached maturity, the reality poses new challenges: Most of the solutions on Web data analysis assume a static Web, in which a solitary user interacts with a Web site. It is prime time to depart from such simplifying assumptions and conceive solutions that are closer to Web reality: The Web is evolving constantly; sites change and user preferences drift. Clickstream data that form the basis of Web analysis are, obviously, streams rather than static datasets. 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. Furthermore, the inherent and increasing heterogeneity of the Web has required Web-based applications to more effectively integrate a variety of types of data across multiple channels and from different sources in addition to usage, such as content, structure, and semantics. A focus on techniques and architectures for more effective exploitation and mining of such multi-faceted data is likely to stimulate a next generation of intelligent applications. Recommendation systems form a prominent application area of Web analysis. One of the emerging issues in this area is the vulnerability of a Web site and its users towards abuse and offence. "How should an intelligent recommender system be designed to resist various malicious manipulations, such as schilling attacks that try to alter user ratings to influence the recommendations?" This motivates the need to study and design robust recommender systems. WebKDD 2005 addressed these emerging aspects of Web reality.

In the first paper, *Mining Significant Usage Patterns from Clickstream Data*, Lu, Dunham, and Meng propose a technique to generate significant usage patterns (SUP) and use it to acquire significant "user preferred navigational trails." The technique uses pipelined processing phases including sub-abstraction of sessionized Web clickstreams, clustering of the abstracted Web sessions, concept-based abstraction of the clustered sessions, and SUP generation. Using this technique, valuable customer behavior information can be extracted by Web site practitioners. Experiments conducted using J.C.Penney Web log data demonstrate that SUPs of different types of customers are distinguishable and interpretable.

In the second paper, *Using and Learning Semantics in Frequent Subgraph Mining*, Berendt addresses the need for incorporating background knowledge into graph mining and for studying patterns at different levels of abstraction, by using taxonomies in mining and extending frequency / support measures by the notion of context-induced interestingness. Semantics are used as well as learned in this process, and a visualization tool is used to allow the user to navigate through detail-and-context views of taxonomy context, pattern context, and transaction context. A case study of a real-life Web site shows the advantages of the proposed solutions.

In the third paper, *Overcoming Incomplete User Models in Recommendation Systems via an Ontology*, Schickel and Faltings propose a new method that extends the utility model and assumes that the structure of user preferences follows an ontology of product attributes. Using the MovieLens data, their experiments show that real user preferences indeed closely follow an ontology based on movie attributes. Furthermore, a recommender based just on a single individual's preferences and this ontology performs better than collaborative filtering, with the greatest differences when few data about the user are available. This points the way to how proper inductive bias (in the form of an ontology) can be used for significantly more powerful recommender systems in the future.

The fourth paper, *Data Sparsity Issues in the Collaborative Filtering Framework* by Grcar, Fortuna, Mladenic, and Grobelnik gives an overview of collaborative filtering approaches, and presents experimental results that compare the k-nearest neighbor (kNN) algorithm with support vector machines (SVM) in the collaborative filtering framework using data sets with different properties. Experiments on two standard, publicly available data sets and a real-life corporate data set that does not fit the profile of ideal data for collaborative filtering lead the authors to conclude that the quality of collaborative filtering recommendations is highly dependent on the sparsity of available data. Furthermore, they show that kNN is dominant on data sets with relatively low sparsity while SVM-based approaches may perform better on highly sparse data.

The fifth paper focuses on the multi-faceted aspect of Web personalization. In *USER: User-Sensitive Expert Recommendations for Knowledge-Dense Environments*, DeLong, Desikan, and Srivastava address the challenge of making relevant recommendations given a large, knowledge-dense Web site and a non-expert user searching for information. They propose an approach to provide recommendations to non-experts, helping them understand what they need to

know, as opposed to what is popular among other users. The approach is user-sensitive in that it adopts a "model of learning" whereby the user's context is dynamically interpreted as they browse, and then leveraging that information to improve the recommendations.

In the sixth paper, *Analysis and Detection of Segment-Focused Attacks Against Collaborative Recommendation*, Mobasher, Burke, Williams, and Bhaumik examine the vulnerabilities that have recently been identified in collaborative filtering recommender systems. These vulnerabilities mostly emanate from the open nature of such systems and their reliance on user-specified judgments for building profiles. Hence, attackers can easily introduce biased data in an attempt to force the system to "adapt" in a manner advantageous to them. The authors explore an attack model that focuses on a subset of users with similar tastes and show that such an attack can be highly successful against both user-based and item-based collaborative filtering. They also introduce a detection model that can significantly decrease the impact of this attack.

The seventh paper, *Adaptive Web Usage Profiling*, by Suryavanshi, Shiri, and Mudur, addresses the challenge of maintaining profiles so that they dynamically adapt to new interests and trends. They present a new profile maintenance scheme, which extends the relational fuzzy subtractive clustering (RFSC) technique and enables efficient incremental update of usage profiles. An impact factor is defined whose value can be used to decide the need for recompilation. The results from extensive experiments on a large real dataset of Web logs show that the proposed maintenance technique, with considerably reduced computational costs, is almost as good as complete remodeling.

In the eigth paper, *On Clustering Techniques for Change Diagnosis in Data Streams*, Aggarwal and Yu, address the challenge of exploring the underlying changing trends in data streams that are generated by applications which are time-changing in nature. They explore and survey some of their recent methods for change detection, particularly methods that use clustering in order to provide a concise understanding of the underlying trends. They discuss their recent techniques which use micro-clustering in order to diagnose the changes in the underlying data, and discuss the extension of this method to text and categorical data sets as well community detection in graph data streams.

In *Personalized Search Results with User Interest Hierarchies Learnt from Bookmarks*, Kim and Chan propose a system for personalized Web search that incorporates an individual user's interests when deciding relevant results to return. They propose a method to (re)rank the results from a search engine using a learned user profile, called a user interest hierarchy (UIH), from Web pages that are of interest to the user. The users interest in Web pages will be determined implicitly, without directly asking the user. Experimental results indicate that the personalized ranking methods, when used with a popular search engine, can yield more potentially interesting Web pages for individual users.

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 2005. 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): Charu Aggarwal, Sarabjot S. Anand, Jonathan Becher, Bettina Berendt, Ed Chi, Robert Cooley, Wei Fan, Joydeep Ghosh, Marco Gori, Fabio Grandi, Dimitrios Gunopulos, George Karypis, Raghu Krishnapuram, Ravi Kumar, Vipin Kumar, Mark Levene, Ee-Peng Lim, Bing Liu, Huan Liu, Stefano Lonardi, Ernestina Menasalvas, Rajeev Motwani, Alex Nanopoulos, Jian Pei, Rajeev Rastogi, Jaideep Srivastava, and Mohammed Zaki. O. Nasraoui gratefully acknowledges the support of the US National Science Foundation as part of NSF CAREER award IIS-0133948.

June 2006

Olfa Nasraoui
Osmar Zaane
Myra Spiliopoulou
Bamshad Mobasher
Philip Yu
Brij Masand

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 4248: S. Staab, V. Svátek (Eds.), Engineering Knowledge in the Age of the Semantic Web. XIV, 400 pages. 2006.

Vol. 4223: L. Wang, L. Jiao, G. Shi, X. Li, J. Liu (Eds.), Fuzzy Systems and Knowledge Discovery. XXVIII, 1335 pages. 2006.

Vol. 4213: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Knowledge Discovery in Databases: PKDD 2006. XXII, 660 pages. 2006.

Vol. 4212: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Machine Learning: ECML 2006. XXIII, 851 pages. 2006.

Vol. 4211: P. Vogt, Y. Sugita, E. Tuci, C. Nehaniv (Eds.), Symbol Grounding and Beyond. VIII, 237 pages. 2006.

Vol. 4203: F. Esposito, Z.W. Raś, D. Malerba, G. Semeraro (Eds.), Foundations of Intelligent Systems. XVIII, 767 pages. 2006.

Vol. 4201: Y. Sakakibara, S. Kobayashi, K. Sato, T. Nishino, E. Tomita (Eds.), Grammatical Inference: Algorithms and Applications. XII, 359 pages. 2006.

Vol. 4198: O. Nasraoui, O. Zaïane, M. Spiliopoulou, B. Mobasher, B. Masand, P. S. Yu (Eds.), Advances in Web Mining and Web Usage Analysis. IX, 177 pages. 2006.

Vol. 4196: K. Fischer, I.J. Timm, E. André, N. Zhong (Eds.), Multiagent System Technologies. X, 185 pages. 2006.

Vol. 4188: P. Sojka, I. Kopeček, K. Pala (Eds.), Text, Speech and Dialogue. XIV, 721 pages. 2006.

Vol. 4183: J. Euzenat, J. Domingue (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 291 pages. 2006.

Vol. 4180: M. Kohlhase, OMDoc – An Open Markup Format for Mathematical Documents [version 1.2]. XIX, 428 pages. 2006.

Vol. 4177: R. Marín, E. Onaindía, A. Bugarín, J. Santos (Eds.), Current Topics in Artificial Intelligence. XIII, 621 pages. 2006.

Vol. 4160: M. Fisher, W.v.d. Hoek, B. Konev, A. Lisitsa (Eds.), Logics in Artificial Intelligence. XII, 516 pages. 2006.

Vol. 4155: O. Stock, M. Schaerf (Eds.), Reasoning, Action and Interaction in AI Theories and Systems. XVIII, 343 pages. 2006.

Vol. 4149: M. Klusch, M. Rovatsos, T.R. Payne (Eds.), Cooperative Information Agents X. XII, 477 pages. 2006.

Vol. 4139: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala, Advances in Natural Language Processing. XVI, 771 pages. 2006.

Vol. 4133: J. Gratch, M. Young, R. Aylett, D. Ballin, P. Olivier (Eds.), Intelligent Virtual Agents. XIV, 472 pages. 2006.

Vol. 4130: U. Furbach, N. Shankar (Eds.), Automated Reasoning. XV, 680 pages. 2006.

Vol. 4120: J. Calmet, T. Ida, D. Wang (Eds.), Artificial Intelligence and Symbolic Computation. XIII, 269 pages. 2006.

Vol. 4114: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence, Part II. XXVII, 1337 pages. 2006.

Vol. 4108: J.M. Borwein, W.M. Farmer (Eds.), Mathematical Knowledge Management. VIII, 295 pages. 2006.

Vol. 4106: T.R. Roth-Berghofer, M.H. Göker, H. A. Güvenir (Eds.), Advances in Case-Based Reasoning. XIV, 566 pages. 2006.

Vol. 4099: Q. Yang, G. Webb (Eds.), PRICAI 2006: Trends in Artificial Intelligence. XXVIII, 1263 pages. 2006.

Vol. 4095: S. Nolfi, G. Baldassarre, R. Calabretta, J.C. T. Hallam, D. Marocco, J.-A. Meyer, O. Miglino, D. Parisi (Eds.), From Animals to Animats 9. XV, 869 pages. 2006.

Vol. 4093: X. Li, O.R. Zaïane, Z. Li (Eds.), Advanced Data Mining and Applications. XXI, 1110 pages. 2006.

Vol. 4092: J. Lang, F. Lin, J. Wang (Eds.), Knowledge Science, Engineering and Management. XV, 664 pages. 2006.

Vol. 4088: Z.-Z. Shi, R. Sadananda (Eds.), Agent Computing and Multi-Agent Systems. XVII, 827 pages. 2006.

Vol. 4087: F. Schwenker, S. Marinai (Eds.), Artificial Neural Networks in Pattern Recognition. IX, 299 pages. 2006.

Vol. 4068: H. Schärfe, P. Hitzler, P. Øhrstrøm (Eds.), Conceptual Structures: Inspiration and Application. XI, 455 pages. 2006.

Vol. 4065: P. Perner (Ed.), Advances in Data Mining. XI, 592 pages. 2006.

Vol. 4062: G. Wang, J.F. Peters, A. Skowron, Y. Yao (Eds.), Rough Sets and Knowledge Technology. XX, 810 pages. 2006.

Vol. 4049: S. Parsons, N. Maudet, P. Moraitis, I. Rahwan (Eds.), Argumentation in Multi-Agent Systems. XIV, 313 pages. 2006.

Vol. 4048: L. Goble, J.-J.C.. Meyer (Eds.), Deontic Logic and Artificial Normative Systems. X, 273 pages. 2006.

Vol. 4045: D. Barker-Plummer, R. Cox, N. Swoboda (Eds.), Diagrammatic Representation and Inference. XII, 301 pages. 2006.

Vol. 4031: M. Ali, R. Dapoigny (Eds.), Advances in Applied Artificial Intelligence. XXIII, 1353 pages. 2006.

Vol. 4029: L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada (Eds.), Artificial Intelligence and Soft Computing – ICAISC 2006. XXI, 1235 pages. 2006.

Vol. 4027: H.L. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreasen, H. Christiansen (Eds.), Flexible Query Answering Systems. XVIII, 714 pages. 2006.

Vol. 4021: E. André, L. Dybkjær, W. Minker, H. Neumann, M. Weber (Eds.), Perception and Interactive Technologies. XI, 217 pages. 2006.

Vol. 4020: A. Bredenfeld, A. Jacoff, I. Noda, Y. Takahashi (Eds.), RoboCup 2005: Robot Soccer World Cup IX. XVII, 727 pages. 2006.

Vol. 4013: L. Lamontagne, M. Marchand (Eds.), Advances in Artificial Intelligence. XIII, 564 pages. 2006.

Vol. 4012: T. Washio, A. Sakurai, K. Nakajima, H. Takeda, S. Tojo, M. Yokoo (Eds.), New Frontiers in Artificial Intelligence. XIII, 484 pages. 2006.

Vol. 4008: J.C. Augusto, C.D. Nugent (Eds.), Designing Smart Homes. XI, 183 pages. 2006.

Vol. 4005: G. Lugosi, H.U. Simon (Eds.), Learning Theory. XI, 656 pages. 2006.

Vol. 3978: B. Hnich, M. Carlsson, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. VIII, 179 pages. 2006.

Vol. 3963: O. Dikenelli, M.-P. Gleizes, A. Ricci (Eds.), Engineering Societies in the Agents World VI. XII, 303 pages. 2006.

Vol. 3960: R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, M.C. Dias (Eds.), Computational Processing of the Portuguese Language. XII, 274 pages. 2006.

Vol. 3955: G. Antoniou, G. Potamias, C. Spyropoulos, D. Plexousakis (Eds.), Advances in Artificial Intelligence. XVII, 611 pages. 2006.

Vol. 3949: F. A. Savacı (Ed.), Artificial Intelligence and Neural Networks. IX, 227 pages. 2006.

Vol. 3946: T.R. Roth-Berghofer, S. Schulz, D.B. Leake (Eds.), Modeling and Retrieval of Context. XI, 149 pages. 2006.

Vol. 3944: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (Eds.), Machine Learning Challenges. XIII, 462 pages. 2006.

Vol. 3930: D.S. Yeung, Z.-Q. Liu, X.-Z. Wang, H. Yan (Eds.), Advances in Machine Learning and Cybernetics. XXI, 1110 pages. 2006.

Vol. 3918: W.K. Ng, M. Kitsuregawa, J. Li, K. Chang (Eds.), Advances in Knowledge Discovery and Data Mining. XXIV, 879 pages. 2006.

Vol. 3913: O. Boissier, J. Padget, V. Dignum, G. Lindemann, E. Matson, S. Ossowski, J.S. Sichman, J. Vázquez-Salceda (Eds.), Coordination, Organizations, Institutions, and Norms in Multi-Agent Systems. XII, 259 pages. 2006.

Vol. 3910: S.A. Brueckner, G.D.M. Serugendo, D. Hales, F. Zambonelli (Eds.), Engineering Self-Organising Systems. XII, 245 pages. 2006.

Vol. 3904: M. Baldoni, U. Endriss, A. Omicini, P. Torroni (Eds.), Declarative Agent Languages and Technologies III. XII, 245 pages. 2006.

Vol. 3900: F. Toni, P. Torroni (Eds.), Computational Logic in Multi-Agent Systems. XVII, 427 pages. 2006.

Vol. 3899: S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. XIV, 216 pages. 2006.

Vol. 3898: K. Tuyls, P.J. 't Hoen, K. Verbeeck, S. Sen (Eds.), Learning and Adaption in Multi-Agent Systems. X, 217 pages. 2006.

Vol. 3891: J.S. Sichman, L. Antunes (Eds.), Multi-Agent-Based Simulation VI. X, 191 pages. 2006.

Vol. 3890: S.G. Thompson, R. Ghanea-Hercock (Eds.), Defence Applications of Multi-Agent Systems. XII, 141 pages. 2006.

Vol. 3885: V. Torra, Y. Narukawa, A. Valls, J. Domingo-Ferrer (Eds.), Modeling Decisions for Artificial Intelligence. XII, 374 pages. 2006.

Vol. 3881: S. Gibet, N. Courty, J.-F. Kamp (Eds.), Gesture in Human-Computer Interaction and Simulation. XIII, 344 pages. 2006.

Vol. 3874: R. Missaoui, J. Schmidt (Eds.), Formal Concept Analysis. X, 309 pages. 2006.

Vol. 3873: L. Maicher, J. Park (Eds.), Charting the Topic Maps Research and Applications Landscape. VIII, 281 pages. 2006.

Vol. 3864: Y. Cai, J. Abascal (Eds.), Ambient Intelligence in Everyday Life. XII, 323 pages. 2006.

Vol. 3863: M. Kohlhase (Ed.), Mathematical Knowledge Management. XI, 405 pages. 2006.

Vol. 3862: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), Programming Multi-Agent Systems. XIV, 267 pages. 2006.

Vol. 3849: I. Bloch, A. Petrosino, A.G.B. Tettamanzi (Eds.), Fuzzy Logic and Applications. XIV, 438 pages. 2006.

Vol. 3848: J.-F. Boulicaut, L. De Raedt, H. Mannila (Eds.), Constraint-Based Mining and Inductive Databases. X, 401 pages. 2006.

Vol. 3847: K.P. Jantke, A. Lunzer, N. Spyratos, Y. Tanaka (Eds.), Federation over the Web. X, 215 pages. 2006.

Vol. 3835: G. Sutcliffe, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIV, 744 pages. 2005.

Vol. 3830: D. Weyns, H. V.D. Parunak, F. Michel (Eds.), Environments for Multi-Agent Systems II. VIII, 291 pages. 2006.

Vol. 3817: M. Faundez-Zanuy, L. Janer, A. Esposito, A. Satue-Villar, J. Roure, V. Espinosa-Duro (Eds.), Nonlinear Analyses and Algorithms for Speech Processing. XII, 380 pages. 2006.

Vol. 3814: M. Maybury, O. Stock, W. Wahlster (Eds.), Intelligent Technologies for Interactive Entertainment. XV, 342 pages. 2005.

Vol. 3809: S. Zhang, R. Jarvis (Eds.), AI 2005: Advances in Artificial Intelligence. XXVII, 1344 pages. 2005.

¥359.00元

Table of Contents

Mining Significant Usage Patterns from Clickstream Data 1
Lin Lu, Margaret Dunham, Yu Meng

Using and Learning Semantics in Frequent Subgraph Mining 18
Bettina Berendt

Overcoming Incomplete User Models in Recommendation Systems Via an Ontology 39
Vincent Schickel-Zuber, Boi Faltings

Data Sparsity Issues in the Collaborative Filtering Framework 58
Miha Grčar, Dunja Mladenič, Blaž Fortuna, Marko Grobelnik

USER: User-Sensitive Expert Recommendations for Knowledge-Dense Environments 77
Colin DeLong, Prasanna Desikan, Jaideep Srivastava

Analysis and Detection of Segment-Focused Attacks Against Collaborative Recommendation 96
Bamshad Mobasher, Robin Burke, Chad Williams, Runa Bhaumik

Adaptive Web Usage Profiling 119
Bhushan Shankar Suryavanshi, Nematollaah Shiri, Sudhir P. Mudur

On Clustering Techniques for Change Diagnosis in Data Streams 139
Charu C. Aggarwal, Philip S. Yu

Personalized Search Results with User Interest Hierarchies Learnt from Bookmarks 158
Hyoung-rae Kim, Philip K. Chan

Author Index 177

Mining Significant Usage Patterns from Clickstream Data*

Lin Lu, Margaret Dunham, and Yu Meng

Department of Computer Science and Engineering
Southern Methodist University
Dallas, Texas 75275-0122, USA
{llu, mhd, ymeng}@engr.smu.edu

Abstract. Discovery of usage patterns from Web data is one of the primary purposes for Web Usage Mining. In this paper, a technique to generate Significant Usage Patterns (SUP) is proposed and used to acquire significant "user preferred navigational trails". The technique uses pipelined processing phases including sub-abstraction of sessionized Web clickstreams, clustering of the abstracted Web sessions, concept-based abstraction of the clustered sessions, and SUP generation. Using this technique, valuable customer behavior information can be extracted by Web site practitioners. Experiments conducted using Web log data provided by J.C.Penney demonstrate that SUPs of different types of customers are distinguishable and interpretable. This technique is particularly suited for analysis of dynamic websites.

1 Introduction

The detailed records of Web data, such as Web server logs and referrer logs, provide enormous amounts of user information. Hidden in these data is valuable information that implies users' preferences and motivations for visiting a specific website. Research in Web Usage Mining (WUM) is to uncover such kind of information [10]. WUM is a branch of Web mining. By applying data mining techniques to discover useful knowledge of user navigation patterns from Web data, WUM is aimed at improving the Web design and developing corresponding applications to better cater to the needs of both users and website owners [20]. A pioneer work proposed by Nasraoui, *et al.*, used a concept hierarchy directly inferred from the website structure to enhance web usage mining [25, 26]. The idea is to segment Web logs into sessions, determine the similarity/distance among the sessions, and cluster the session data into the optimal number of components in order to obtain typical session profiles of users. Our work will extend to analyzing dynamic websites.

A variety of usage patterns have been investigated to examine the Web data from different perspectives and for various purposes. For instance, the maximal frequent forward sequence mines forward traversal patterns which are maximal and with backward traversals removed [9], the maximal frequent sequence examines the sequences

* This work is supported by the National Science Foundation under Grant No. IIS-0208741.

O. Nasraoui et al. (Eds.): WebKDD 2005, LNAI 4198, pp. 1 – 17, 2006.

that have a high frequency of occurrence as well as being maximal in length [24], sequential patterns explore the sequences with certain a support that are maximal [1], and user preferred navigational trails extract user preferred navigation paths [5] [6].

In this paper, a new data mining methodology that involves exploring the Significant Usage Patterns (SUP) is introduced. SUPs are paths that correspond to clusters of user sessions. A SUP may have specific beginning and/or ending states, and its corresponding normalized product of probabilities along the path satisfies a given threshold. SUP is a variation of "user preferred navigational trail" [5] [6]. Compared with earlier work, SUP differs in the following four aspects:

1. SUP is extracted from clusters of abstracted user sessions.
2. Practitioners may designate the beginning and/or ending Web pages of preferences before generating SUPs. For example, you may only want to see sequences that end on a purchase page.
3. SUPs are patterns with normalized probability, making it easy for practitioners to determine the probability threshold to generate corresponding patterns.
4. SUP uses a unique two-phase abstraction technique (see sections 3.1 & 3.3).
5. SUP is especially useful in analysis of dynamic websites.

We assume that the clickstream data has already been sessionized. The focus of this paper will be on abstracting the Web clickstream, clustering of abstracted sessions and generation of SUPs.

The rest of the paper is organized as follows. Section 2 discusses the related work. The methodology related to the alignment, abstraction, and clustering of Web sessions is provided in Section 3. Section 4 gives the analysis of experimental results performed using Web log data provided by J. C. Penney. Finally, conclusive discussions and perspectives for future research will be presented.

2 Related Work

Work relevant to the three main steps involved in mining SUPs: URL abstraction, clustering sessions of clickstream data, and generating usage patterns, are discussed in detail in the following subsections. We conclude each subsection with a brief examination of how our work fits into the literature.

2.1 URL Abstraction

URL abstraction is the process of generalizing URLs into higher level groups. Page-level aggregation is important for user behavior analysis [20]. In addition, it may lead to much more meaningful clustering results [4]. Since behavior patterns in user sessions consist of a sequence of low level page views, there is no doubt that patterns discovered using exact URLs will give fewer matches among user sessions, than those where abstraction of these pages is performed. Web page abstraction allows the discovery of correlations among user sessions with frequent occurrences at an abstract concept level. These frequent occurrences may not be frequent when viewed at the precise page level. In addition, many pages in a specific web site may be semantically

equivalent (such as all colors/sizes of the same dress) which makes web page generalization not only possible, but also desirable.

In [4], concept-category of page hierarchy was introduced, in which web pages were grouped into categories, based on proper analytics and/or metadata information. Since this approach categorizes web pages using only the top-most level of the page hierarchy, it could be viewed as a simple version of generalization-based clustering. A generalization-based page hierarchy was described in [11]. According to this approach, each page was generalized to its higher level. For instance, pages under /school/department/courses would be categorized to "department" pages and pages under /school/department would be classified as "school" pages. Spiliopoulou et al. employed a content-based taxonomy of web site abstraction, in which taxonomy was defined according to a task-based model and each web page was mapped to one of the taxonomy's concepts [22]. In [18], pages were generalized to three categories, namely administrative, informational, and shopping pages, to describe an online nutrition supply store.

In our study, two different abstraction strategies are applied to user sessions before and after the clustering process. User sessions are sub-abstracted before applying the clustering algorithm in order to make the sequence alignment approach used in clustering more meaningful. After clustering user sessions, a concept-based abstraction approach is applied to user sessions in each cluster, which allows more insight into the SUPs associated with each cluster. Both abstraction techniques are based on a user provided site concept hierarchy.

2.2 Clustering User Sessions of Clickstream Data

In order to mine useful information concerning user navigation patterns from clickstream data, it is appropriate to first cluster user sessions. The purpose of clustering is to find groups of users with similar preferences and objectives for visiting a specific website. Actually, the knowledge of user groups with similar behavior patterns is extremely valuable for e-commerce applications. With this knowledge, domain experts can infer user demographics in order to perform market segmentations [20].

Various approaches have been introduced in the literature to cluster user sessions [4] [7] [11] [16] [23]. [7] used a mixture of first-order Markov chains to partition user sessions with similar navigation patterns into the same cluster. In [11], page accesses in each user session were substituted by a generalization-based page hierarchy scheme. Then, generalized sessions were clustered using a hierarchical clustering algorithm, BIRCH.

Banerjee et al. developed an algorithm that combined both the time spent on a page and Longest Common Subsequences (LCS) to cluster user sessions [4]. The LCS algorithm was first applied on all pairs of user sessions. After each LCS path was compacted using a concept-category of page hierarchy, similarities between LCS paths were computed as a function of the time spent on the corresponding pages in the paths weighted by a certain factor. Then, an abstract similarity graph was constructed for the set of sessions to be clustered. Finally, a graph partition algorithm, called Metis, was used to segment the graph into clusters.

The clustering approach discussed in [16] [23] was based on the sequence alignment method. They took the order of page accesses within the session into consideration when computing the similarities between sessions. More specifically, they used the idea of sequence alignment widely adopted in bio-informatics to measure the similarity between sessions. Then, sessions were clustered according to their similarities. In [16], Ward's clustering method [15] was used, and [23] applied three clustering algorithms, ROCK [13], CHAMELEON [17], and TURN [12].

The clustering approach used in our work is based on [16] [23], however, a sub-abstraction is first conducted before the similarities among Web pages are measured. Then the Needleman-Wunsch global alignment algorithm [21] is used to align the abstracted sessions based on the pagewise similarities. By performing a global alignment of sessions, the similarity matrix can be obtained. Finally, the nearest neighbor clustering algorithms is applied to cluster the user sessions based on the similarity matrix.

2.3 Generating Usage Patterns

Varieties of browsing patterns have been investigated to examine Web data from different perspectives and for various purposes. These patterns include the maximal frequent forward sequence [9], the maximal frequent sequence [24], the sequential pattern [1], and user preferred navigational trail [5] [6].

Table 1. A comparison of various popular usage patterns

	Clustering	Abstraction	Beginning/ending Web page(s)	Normalized
Sequential Pattern	N	Y*	N	-
Maximal Frequent Sequence	N	N	N	-
Maximal Frequent Forward Sequence	N	N	N	-
User Preferred Navigational Trail	N	N	N	N
Significant Usage Pattern	Y	Y	Y	Y

* Abstraction may be applied to some of the patterns, i.e. in [2], but not all, i.e. in [19].

The usage pattern proposed in [5] [6] are the most related to our research. [5] proposed a data-mining model to extract the higher probability trails which represent user preferred navigational paths. In that paper, user sessions were modeled as a Hypertext Probabilistic Grammar (HPG), which can be viewed as an absorbing Markov chain, with two additional states, start (S) and finish (F). The set of strings generated from HPG with higher probability are considered as preferred navigation trails of users. The depth first search algorithm was used to generate the trails given specific support and confidence thresholds. Support and confidence thresholds were used to control

the quality and quantity of trails generated by the algorithm. In [6], it was proved that the average complexity of the depth first search algorithm used to generate the higher probability trails is linear to the number of web pages accessed.

In our approach, SUPs are trails with high probability extracted from each of the clusters of abstracted user sessions. We use the normalized probabilities of occurrence to make the probabilities of SUPs insensitive to the length of the sessions. In addition, SUPs may begin and/or end with specific Web pages of user preferences. Table 1 provides a comparison of SUPs with other patterns.

3 Methodology

Our technique uses pipelined processing phases including sub-abstraction of sessionized Web clickstream, clustering of the abstracted Web sessions, concept-based abstraction of the clustered sessions and SUP generation.

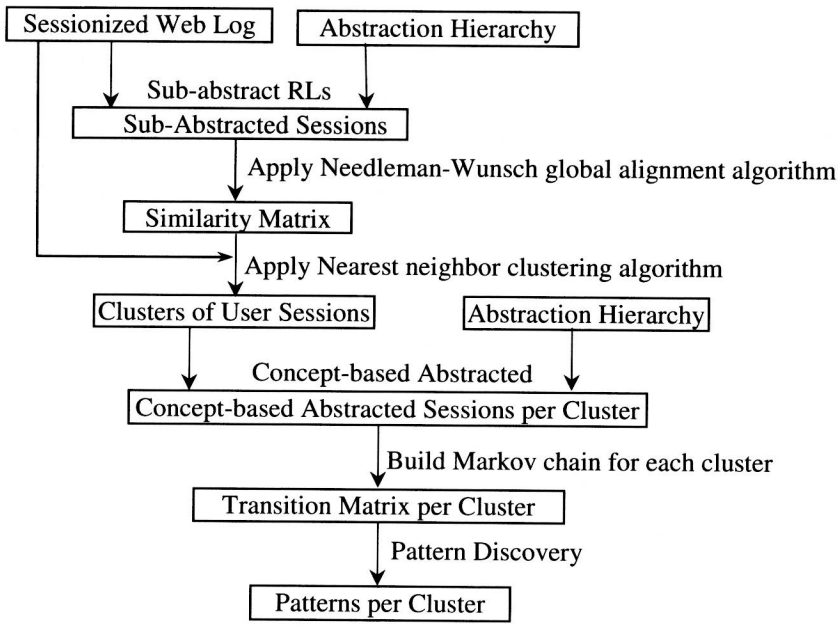

Fig. 1. Logic flow to generate SUPs

To generate SUPs, first, a sequence alignment [16] [23] approach based on the Needleman-Wunsch global alignment algorithm [21] is applied to the sessionized abstracted clickstream data to compute the similarities between each pair of sessions. This approach preserves the sequential relationship between sessions, and reflects the characteristics of chronological sequential order associated with the clickstream data. Based on the pairwise alignment results, a similarity matrix is constructed and then

original un-abstracted sessions are grouped into clusters according to their similarities. By applying clustering on sessions, we are more likely to discover the common and useful usage patterns associated with each cluster. Then, the original Web sessions are abstracted again using a concept-based abstraction approach and then a first order Markov chain is built for each cluster of sessions. Finally, the SUPs with a normalized product of probability along the path that is greater than a given threshold are extracted from each cluster based on their corresponding Markov chain. This process is illustrated in Fig 1. A more detailed description of each step is provided in the following subsections.

3.1 Create Sub-abstracted Sessions

In this study, we assume that the Web data has already been cleansed and sessionized. Detailed techniques for preprocessing the Web data can be found in [8].

A Web session is a sequence of Web pages accessed by a single user. However, for the sequence alignment result to be more meaningful, we abstract the pages to produce sub-abstracted sessions. We use the term "sub-abstracted" instead of "abstracted" session, because we do not use a typical abstraction approach, but rather a concept-based abstraction hierarchy, e.g., Department, Category, and Item in e-commerce Web site, plus some specific information, such as Department ID, Category ID in the abstracted session. Thus parts of the Web page URL are abstracted and some are not. With this approach, we preserve certain information to make Web page similarity comparison more meaningful for that session alignment described below. A URL in a session is mapped to a sub-abstracted URL as follows:

URL -> {<Concept hierarchy keyword> <Unique ID> <|>}

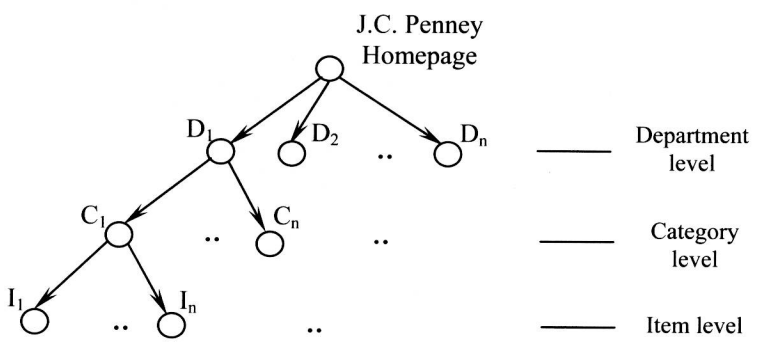

Fig. 2. Hierarchy of J.C. Penney Web site

Example 1. Based on the hierarchical structure of J.C. Penney's Web site, each Web page access in the session sequence is abstracted into three levels of hierarchy, as shown in Fig 2, where D, C, I are the initials for Department, Category, and Item respectively, 1, 2, …, n represent IDs, and vertical bar | is used to separate different levels in the hierarchy.