WebKDD 2004 – Web Mining and Web Usage Analysis Post-Workshop report

Olfa Nasraoui
Computer Engineering and Computer Science department
Speed School of Engineering
University of Louisville
Louisville KY 40292
olfa.nasraoui@louisville.edu

Bamshad Mobasher
School of Computer Science, Telecommunications & Information Systems
DePaul University
243 S. Wabash Ave. Chicago, IL 60604
mobasher@cti.depaul.edu

Brij Masand
Data Miners, Inc
77 North Washington Street, 2nd Floor
Boston , MA 02114
brij@data-miners.com

Bing Liu
Department of Computer Science
University of Illinois at Chicago, 851 S. Morgan (M/C 152), Chicago, IL 60607-7053
liub@cs.uic.edu

ABSTRACT

In this report, we summarize the contents and outcomes of the recent WebKDD 2004 workshop that was held in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25, 2004, in Seattle, Washington. We also reflect on the trend in participation levels in the WebKDD series of workshops over the last six years, and indicate new directions in Web mining research as reflected in the latest workshop.

Keywords

Web mining, profiling, personalization, click-stream analysis, recommender systems.

1. INTRODUCTION

The birth of the World Wide Web in the nineties and the dot.com bubble that followed have contributed to an initial proliferation of research and interest in Web mining. Despite the "dot.com crash," interest in Web mining both in the research community as well as in industry has persisted and flourished. In particular, in the last few years we have witnessed an increasing level of participation in the WebKDD workshops [12-16] (see Table 1). Figure 1 shows the number of submissions, a reasonable gauge for the active interest in Web mining, which is displayed for each year since the beginning of WebKDD in 1999. The submission trend closely follows the boom in this area that followed the .com bubble from 1999 to 2000, while reflecting the burst in this bubble soon after this period, sinking to its lowest level in 2001. Subsequent years witnessed a reversal in this sharp decrease, hinting at a recovering interest in Web mining.

Table 1. Past WebKDD workshop themes and participation numbers

year	Theme	submissions	accepted	attendees
1999	Workshop on Web Usage Analysis and User Profiling	23	10	56
2000	Web Mining for E-Commerce -- Challenges and Opportunities	31	13	85 (110 requested)
2001	Mining Log Data Across All Customer Touch Points	18	9	52 (by invitation only)
2002	Web Mining for Usage Patterns & User Profiles	23	10	50
2003	Workshop on Web mining as a Premise to to Effective and Intelligent Web Applications	22	9	(45-50) during Jim Sterne's tal*k*
2004	Web Mining and Web Usage Analysis	28	10	33-55

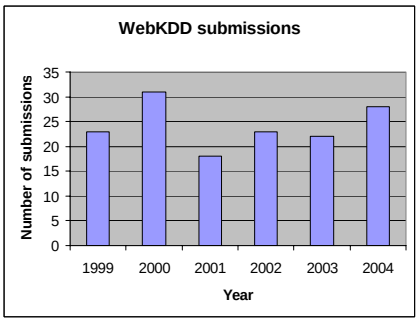

Figure 1. Number of submissions to the WebKDD workshop since its start in 1999

2. THEME

The Web is a live environment that gestates and drives a wide spectrum of applications in which a user interacts with a company, a governmental authority, a non-governmental organization or other non-profit institution or other users. User preferences and expectations, together with usage patterns, form the basis for personalized, user-friendly and business-optimal services. Key Web business metrics enabled by proper data capture and processing are essential to run an effective business or service. Enabling technologies include data mining, scalable warehousing and preprocessing, sequence discovery, real time processing, document classification, user modeling and quality evaluation models for them. Recipient technologies that demand for user profiling and usage patterns include recommendation systems, Web analytics applications, application servers coupled with content management systems and fraud detectors.

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. The development and application of Web mining techniques in the context of Web content, Web usage, and Web structure data has already resulted in dramatic improvements in a variety of Web applications, from search engines, Web agents, and content management systems, to Web analytics and personalization services. A focus on techniques and architectures for more effective integration and mining of content, usage, and structure data from different sources is likely to lead to the next generation of more useful and more intelligent applications.

WebKDD 2004 is the sixth of a successful series of workshops on knowledge discovery on the Web. The WebKDD'04 workshop continued its tradition of serving as a bridge between academia and industry by bringing together practitioners and researchers from both areas in order to foster the exchange of ideas and the dissemination of emerging solutions for intelligent Web-based applications using Web usage, structure and content mining

3. SUBMISSIONS

28 papers were submitted to WebKDD 2004 and were strictly reviewed by at least three reviewers from the WebKDD 2004 program committee. Due to the keen competition, only 10 papers were accepted for the workshop. Two of the accepted papers, with primary topics relating more directly to the Semantic Web, were selected to be presented in a joint session with a new workshop, the KDD Workshop on Mining for and from the Semantic Web (MSW04). The joint session also includes an additional paper submitted and accepted at the MSW04, for a total of three papers in this session, which brings together topics of interest common to both the WebKDD 2004 and MSW04 workshops. The workshop papers are available at http://maya.cs.depaul.edu/webkdd04/.

According to the topics, the papers were grouped into four categories that would form the structure of the workshop sessions. These categories are *(i)* Web Usage Analysis and User Modeling, *(ii)* Web Personalization and Recommender Systems, *(iii)* Search Engine Personalization, and *(iv)* Semantic Web Mining.

We would like to thank the authors of all submitted papers. Their creative efforts have supported the competitive technical program of WebKDD 2004. We would also like to express our gratitude to the members of the program committee for their vigilant and timely reviews.

4. WORKSHOP

The workshop attracted an audience consisting of between 30 and 55 participants. In addition to several people from academia, the workshop attracted several participants from the industry including Microsoft, Amazon, AT&T, Google, SAS, DaimlerChrysler, Accenture, buy.com, shopping.com, and a few other dot.com companies. The paper presentations were segmented into four areas according to their main topics, leading to the following sessions.

2.1 Session 1: Web Usage Analysis and User Modeling

The first three papers deal with *Web Usage Analysis and User Modeling*. In *Sequential Analysis for Learning Modes of Browsing* [6], Hooker and Finkelman present a mathematical framework for directly learning a user's mode of browsing during a given session. This framework is inspired by sequential analysis in the setting of educational testing. They demonstrate its feasibility and utility in the context of click-stream data and explore the range of models and variations that this framework makes available. In *Mining Temporally Evolving Graphs* [5], Desikan and Srivastava address the limited 'data-centric' point of view of most previous Web Mining research by examining another dimension of Web Mining, namely the temporal dimension. They highlight the significance of studying the evolving nature of Web graphs, and classify the approach to such problems at three levels of analysis: single node, sub-graphs and whole graphs. They provide a framework to approach problems in this kind of analysis and identify interesting problems at each level. In *A Formal Model of the ETL*

Process for OLAP-Based Web Usage Analysis [8], Maier addresses the laborious and time-consuming task of populating a data warehouse, to be used for sophisticated analysis of the Web channel in a multi-channel environment of an organization. To this end, is proposed a logical object-oriented relational data storage model, which simplifies modeling the ETL process and supports direct deployment within a WUSAN (Web USage ANalyis) system.

2.2 Session 2: Web Personalization and Recommender Systems

Three papers focus on *Web Personalization and Recommender Systems*. In *Using Distinctive Information Channels for a Mission-based Web Recommender System* [7], Li and Zaiane advocate the use of additional information channels such as the content of visited pages and the connectivity between web pages, as an alternative to using only one information channel, namely the web access history. They propose the concept of "*missions*", which are identified by different channels, to help in better modeling users' concurrent information needs. The combination of three channels is shown to improve the quality of the recommendations. In *Complete this Puzzle: A Connectionist Approach to Accurate Web Recommendations based on a Committee of Predictors* [10], Nasraoui and Pavuluri present a *Context Ultra-Sensitive Approach to personalization based on two-step Recommender systems (CUSA-2-step-Rec)*. The approach relies on a committee of profile-specific neural networks. Similar to the task of completing the missing pieces of a puzzle, each neural network is trained to predict the missing URLs of several complete ground-truth sessions from a given profile. The approach outperforms nearest profile and K-Nearest Neighbors based collaborative filtering. In *Collaborative Quality Filtering: Establishing Consensus or Recovering Ground Truth?* [11], Traupman and Wilensky present an algorithm based on factor analysis for performing collaborative quality filtering (CQF). Unlike previous approaches to CQF, which estimate the consensus opinion of a group of reviewers, their algorithm uses a generative model of the review process to estimate the latent intrinsic quality of the items under review. The results of their tests suggest that asymptotic consensus, which purports to model peer review, is in fact not recovering the ground truth quality of reviewed items.

2.3 Session 3: Search Engine Personalization

Two papers deal with *Search Engine Personalization*. In *Spying Out Real User Preferences for Metasearch Engine Personalization* [4], Deng, Chai, Tan, Ng, and Lee propose a learning technique called "*Spy Naive Bayes*" (*SpyNB*) to identify the user preference pairs generated from clickthrough data. They then employ a ranking SVM algorithm to build a metasearch engine optimizer. Their empirical results on a metasearch engine prototype, comprising MSNSearch, Wisenut and Overture, show that, compared with no learning, SpyNB can significantly improve the average ranks of users' clicks. In *Personalizing PageRank Based on Domain Profiles* [1], Aktas, Nacar, and Menczer introduce a methodology for personalizing PageRank vectors based on URL features such as Internet domains. Users specify interest profiles as binary feature vectors where a feature corresponds to a DNS tree node. Then, given a profile vector, a weighted PageRank can be computed, assigning a weight to each URL based on the match between the URL and the profile features. Preliminary results show that Personalized PageRank performed favorably compared to pure similarity based ranking and traditional PageRank.

2.4 Session 4: Semantic Web Mining (Joint Session with MSW'04 Workshop)

Finally, three papers form the core of the joint session on *Semantic Web Mining*. The first paper was submitted and accepted by the MSW04 workshop, while the last two papers were submitted and accepted by the WebKDD 2004 workshop. Bloehdorn and Hotho, in *Boosting for Text Classification with Semantic Features* [2], propose an enhancement of the classical term stem based document representation through higher semantic concepts extracted from background knowledge. Boosting, a successful machine learning technique is used for classification, and comparative experimental evaluations show consistent improvement of the results. Meo, Lanzi, Matera, and Esposito, in *Integrating Web Conceptual Modeling and Web Usage Mining* [9], present a case study regarding the application of the inductive database approach to the analysis of Web logs and to enable the rapid customization of the mining procedures following the Web developers' need. They integrate the user request information with meta-data concerning the Web site structure into rich XML Web logs, called "*conceptual logs*", produced by Web applications specified with the WebML conceptual model. Then, they apply a data mining language (MINE RULE) to conceptual logs in order to identify different types of patterns, such as recurrent navigation paths, page contents most frequently visited, and anomalies. In *Discovering Links between Lexical and Surface Features in Questions and Answers* [3], Chakrabarti presents a data-driven approach, assisted by machine learning to build question answering information retrieval systems that return short passages or direct answers to questions, rather than URLs pointing to whole pages. Learning is based on a simple conditional exponential model over a pair of feature vectors, one derived from the question and the other derived from a candidate passage. Using this model, candidate passages are filtered, and substantial improvements are obtained in the mean rank at which the first answer is found. The model parameters distill and reveal linguistic artifacts coupling questions and their answers, which can be used for better annotation and indexing.

5. CONCLUSIONS AND FUTURE DIRECTIONS

The WebKDD'04 workshop helped bring to light several interesting future directions and challenges in the area of Web mining. In particular, we note an increasing interest in exploring Web mining and Web data along three dimensions: *temporal*, *breadth*, and *granularity*. Some of the emerging directions include:

- mining evolving usage patterns

- mining usage patterns at higher level concepts such as task goals or missions that may be different even in the same session,
- considering usage data at increasing levels of detail or granularity, for example by associating a session to one or more profiles and involving this information in the early learning stages of recommender systems,
- developing formal models for warehouses that can support the web mining process
- developing user-friendly systems based on previously developed data mining tools (ex: MINERULE),
- mining huge amounts of data or scalability
- exploring innovative ways to model web usage by borrowing ideas from other domains, such as testing in education.

6. ACKNOWLEDGMENTS

We are grateful to past WebKDD organizers, in particular Jaideep Srivastava, Osmar Zaiane, and Myra Spiliopoulou, for providing us with some of the previous year statistics. O. Nasraoui gratefully acknowledges the support of NSF as part of NSF CAREER award IIS-0133948.

7. REFERENCES

[1] M. S. Aktas, M. A. Nacar, and F. Menczer, "Personalizing PageRank Based on Domain Profiles", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[2] S. Bloehdorn and A. Hotho. "Boosting for Text Classification with Semantic Features", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[3] S. Chakrabarti, "Discovering Links between Lexical and Surface Features in Questions and Answers", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[4] L. Deng, X. Chai, Q. Tan, W. Ng, and D. Lee, "Spying Out Real User Preferences for Metasearch Engine Personalization", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[5] P. Desikan and J. Srivastava, "Mining Temporally Evolving Graphs", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[6] G. Hooker and M. Finkelman, "Detail Sequential Analysis for Learning Modes of Browsing", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[7] J. Li and O. R. Zaiane, "Using Distinctive Information Channels for a Mission-based Web Recommender System", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[8] T. Maier, "A Formal Model of the ETL Process for OLAP-Based Web Usage Analysis", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[9] R. Meo, P. Lanzi, M. Matera, and R. Esposito, "Integrating Web Conceptual Modeling and Web Usage Mining", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[10] O. Nasraoui and M. Pavuluri, "Complete this Puzzle: A Connectionist Approach to Accurate Web Recommendations based on a Committee of Predictors", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[11] J. Traupman and R. Wilensky, "Collaborative Quality Filtering: Establishing Consensus or Recovering Ground Truth?", In Proc. of *WebKDD 2004: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25 2004, Seattle, WA.

[12] B. Masand and M. Spiliopoulou, "WEBKDD'99: workshop on Web usage analysis and user profiling", ACM SIGKDD Explorations Newsletter, Vol. 1, No. 2, 108 – 111, Jan. 2000.

[13] M. Spiliopoulou**,** J. Srivastava, R. Kohavi, and B. Masand, **"**WEBKDD 2000 - Web Mining for E-Commerce", ACM SIGKDD Explorations Newsletter, Vol. 1, No. 2, 106-107, Dec. 2000.

[14] R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, Eds., "WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points", Revised Papers. Springer, LNCS series, 2002.

[15] B. M. Masand, M. Spiliopoulou, J. Srivastava, and O. R. Zaiane, "WEBKDD 2002 - Web Mining for Usage Patterns & Profiles", ACM SIGKDD Explorations Newsletter**,** Vol. 4, No. 2, 125-127, Dec. 2002.

[16] R. Kohavi, B. Liu, B. M. Masand, J. Srivastava, and O. R. Zaiane, "WEBKDD 2003 - Web Mining as Premise to Intelligent Web Applications", Workshop Proceedings Notes**,** http://www.acm.org/sigs/sigkdd/proceedings/webkdd03.

About the authors:

Olfa Nasraoui is the Endowed Chair of E-commerce and the Director of the Knowledge Discovery and Intelligent Web Applications Lab at the University of Louisville. She received her Ph.D. in Computer Engineering and Computer Science from the University of Missouri-Columbia in 1999. From 2000 to 2004, she was an Assistant Professor at the University of Memphis. Her research activities include Data Mining, Web mining, Personalization, and Computational Intelligence. She has served on the organizing and program committees of several conferences and workshops, and is the recipient of the National Science Foundation CAREER Award for outstanding young scientists.
(http:/www.louisville.edu/~o0nasr01)

Bamshad Mobasher is an Associate professor of Computer Science and the director of the Center for Web Intelligence (CWI) at DePaul University. He received his PhD from Iowa State University in 1994. His research areas include data mining, Web mining, intelligent agents, and computational logic. He has published more than 70 scientific articles in these areas. As the director of the CWI, Dr. Mobasher directs research in Web mining, Web analytics, user modeling, and personalization; and he oversees several NSF or industry funded projects. Dr. Mobasher has served as an organizer and on the program committees of numerous conferences and workshops, including, the recently held WebKDD workshop on Web Mining and Web Usage Analysis at the 2004 ACM SIGKDD conference in Seattle. He will be the local arrangements chair for the 2005 ACM SIGKDD conference (KDD'05) to be held in Chicago during August 2005.
(http://maya.cs.depaul.edu/~mobasher)

Brij Masand is a partner at Data Miners's Inc. He was formerly head of the data mining group at GTE Labs, where he pioneered web usage mining for analyzing behavior of on-line yellow pages users. He has more than 15 years of experience in applying machine learning technologies to data mining, web usage mining, text mining and intelligent agents and has published numerous papers on these subjects. He has also done extensive work in implementing reliable web usage metrics and applying survival analysis techniques for business applications such as modeling churn and other time-to-event predictions. Brij has an MS in EECS from MIT.
(http://www.data-miners.com/brij/welcome.html)

Bing Liu is an associate professor at the Department of Computer Science, University of Illinois at Chicago (UIC). He received his PhD degree in Artificial Intelligence from University of Edinburgh. Before joining UIC in Aril 2002, he was with National University of Singapore. His current research interests include data mining, Web and text mining, and machine learning. Since 1996, he has been active in data mining research, and has published many papers in leading conferences and journals related to data mining, Web mining and Artificial Intelligence. He served (or serves) in technical program committees of many data mining and Web related international conferences. He is currently an associate editor of IEEE Transactions on Knowledge and Data Engineering. He also serves on the editorial boards of two other international journals related to data analysis and Web technology.
(http://www.cs.uic.edu/~liub)