WebKDD 2005 – 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

Osmar R. Zaïane
Department of Computing Science, University of Alberta
Edmonton, Alberta, Canada, T6G 2E8
zaiane@cs.ualberta.ca

Myra Spiliopoulou
Otto-von-Guericke-University Magdeburg,
Faculty of Computer Science
D-39016 Magdeburg, Germany
myra@iti.cs.uni-magdeburg.de

Bamshad Mobasher
School of Computer Science, Telecommunications & Information Systems, DePaul University
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

Philip S. Yu
IBM T. J. Watson
P.O. Box 218, Yorktown Heights, N.Y. 10598
psyu@us.ibm.com

ABSTRACT

In this report, we summarize the contents and outcomes of the recent WebKDD 2005 workshop on Web Mining and Web Usage Analysis that was held in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24, 2005, in Chicago, Illinois. The theme of this workshop was "*Taming Evolving, Expanding and Multi-faceted Web Clickstreams*". We also reflect on possible new directions in Web mining research as reflected by the discussions and the talks during the workshop.

Keywords

Web mining, profiling, personalization, clickstream analysis, recommender systems, mining evolving web data.

1. INTRODUCTION

WebKDD 2005 is the seventh of a successful series of workshops on knowledge discovery on the Web. The WebKDD'05 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.

2. THEME

The theme of this year's WebKDD workshop was "*Taming Evolving, Expanding and Multi-faceted Web Clickstreams*". While solutions on some of the infancy problems of Web analysis reach 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'05 addressed these emerging aspects of Web reality.

3. SUBMISSIONS

17 papers were submitted to WebKDD 2005 and were strictly reviewed by at least three reviewers from the WebKDD 2005 program committee. After review, only 9 papers were accepted for the workshop. They are available at http://db.cs.ulberta.ca/webkdd05/proceedings.html.
According to the topics, the papers were grouped into four categories: (i) Recommendation systems and Personalization, (ii) Web Usage Mining and Profiling, and (iii) Pattern Importance and Semantics.

4. WORKSHOP

The workshop attracted interest from a large number of conference participants. Over 100 conference attendees had pre-registered for WebKDD'05 before it started, and at any given time throughout the day, there were between 30 and 50 participants attending the talks. In addition to several people from academia, one third of the workshop participants came from the industry, including Microsoft, Amazon, Yahoo, E-Bay, Siemens, Overstock.com, and a few other companies. The paper presentations were segmented into three sessions according to their main topics, as described below. In addition to the talks by the authors of the accepted papers, we had an interesting and timely invited talk by Charu Aggarwal from IBM Research, who brought some perspectives from the area of *mining evolving data streams* in support of the workshop theme of this year.

2.1 Invited Talk: *"On Change Diagnosis and Monitoring in Data Streams"*

Charu Aggarwal, from IBM T.J. Watson center, talked about understanding, visualizing, and diagnosing the evolution of temporal trends in data streams based on the concept of velocity density estimation. Velocity density estimation is used to create temporal velocity profiles and spatial velocity profiles at periodic instants in time. After further discussing the problem of data stream evolution in the context of other kinds of data such as graphs or community detection, clustering and classification, Charu also discussed the importance of using the evolution process in order to improve the effectiveness of data mining algorithms.

Session 1: Recommendation systems and Personalization

The three papers of this session dealt with *Recommendation systems and Personalization*. The first paper focused on *The Multi-Faceted Aspect of Web Personalization*. In *USER (User Sensitive Expert Recommendation): What Non-Experts NEED to Know*, DeLong, Desikan, and Srivastava addressed the problem of providing expert-driven recommendations to non-experts, helping them understand what they *need* to know, as opposed to what is popular among other users. The approach adopts a 'model of learning' in which user questions are dynamically interpreted as the user navigates through the Web site and are then used to formulate recommendations. The next paper dealt with *Search Personalization*. In *Personalized Ranking of Search Results with Learned User Interest Hierarchies from Bookmarks*: Kim and Chan proposed to learn a user profile, called a user interest hierarchy (UIH), from web pages that are of interest to the user, as implicitly inferred from their *Bookmark* collections. Using the UIH, they studied methods that (re-)rank the results of a search engine to improve relevance. The third paper was in the area of *Secure Web Personalization*: In *Effective Attack Models for Shilling Item-Based Collaborative Filtering Systems*, Mobasher, Burke, Bhaumik, and Williams examined several attack models and recommendation techniques relative to the costs and benefits of mounting an attack. They then presented a new attack model that focuses on a subset of users with similar tastes and showed that such an attack can be highly successful against an item-based collaborative filtering algorithm.

2.2 Session 2: Web Usage Mining and Profiling

Three papers focused on *Web Usage Mining and Profiling*. The first paper, *kNN Versus SVM in the Collaborative Filtering Framework*, Grcar, Fortuna, Mladenic, and Grobelnik presented empirical results comparing the k-Nearest Neighbors (kNN) algorithm with Support Vector Machines (SVM) in the collaborative filtering framework using data sets with different properties. The authors concluded that the quality of collaborative filtering recommendations is highly dependent on the quality of the data. The next paper addressed the problem of *Mining Evolving Web Usage Data*. In *Incremental Relational Fuzzy Subtractive Clustering for Dynamic Web Usage Profiling*, Suryavanshi, Shiri, and Mudur presented a web usage profile maintenance scheme using the Relational Fuzzy Subtractive Clustering algorithm (RFSC) that can add new usage data to an existing model without the expense associated with frequent remodeling. They defined a quantitative measure that indicates when remodeling should be performed to avoid a degradation of the model. In *Discovery of Significant Usage Patterns from Clusters of Clickstream Data*, Lu, Dunham, and Meng presented a variation of the "user preferred navigational trail" called Significant Usage Pattern (SUP). SUPs are patterns that are extracted from clustered abstracted clickstream data, with a higher normalized probability of occurrence. After grouping Web sessions into clusters, sessions are abstracted again using a concept-based abstraction approach and a first-order Markov model is built for each cluster of sessions.

2.3 Session 3: Pattern Importance and Semantics

The last group of papers dealt with *Pattern Importance and Semantics*. In *The Semantics of Frequent Subgraphs: Mining and Navigation Pattern Analysis*, Berendt introduced the AP-IP (Abstract Pattern – Individual Pattern) mining problem, and presents the AP-IP algorithm that solves it. AP-IP uses a taxonomy and searches for frequent patterns at an abstract level, but also returns the individual (but infrequent) subgraphs that constitute this pattern. She also presented a procedure for mining patterns at the concept level. In *Heterogeneous Attribute Utility Model: A New Approach for Modeling User Profiles for Recommendation Systems*, Schickel and Faltings made the case that both Collaborative filtering and Preference-based techniques currently require more data about a customer than is usually available to make accurate recommendations. They proposed a new method, called Heterogeneous Attribute Utility Model (HAUM), where the structure of user preferences is assumed to follow an ontology of product attributes.

5. CONCLUSIONS AND FUTURE DIRECTIONS

In addition to the discussions that followed each presentation, there was an exciting discussion among all the participants in the workshop just before closing the WebKDD'05 workshop. These discussions helped raise several questions and bring to light several interesting future directions and challenges in the area of Web mining. In particular, participants discussed the effect of the quality of the ontology on the final recommendations when *semantics* are taken into account. More specifically, at what particular level of abstraction, do we obtain optimal results, and how the results compare depending on whether the ontology is hand crafted or learned automatically? The submissions on the Semantic Web and the discussion thread on semantics of web data indicate great interest and many ongoing efforts in this area.

The discussions also brought out the issue of *scalability* of web usage mining techniques and the difficulty in handling evolving web data. Some have complained from the lack of benchmark data sets to study and test tasks that are related to mining *evolving* patterns, while others raised the problem of *validation* in the context of mining evolving web data. We hope that the discussions and this report will encourage researchers to explore some of these challenges, and perhaps present some of their answers in future workshops.

6. ACKNOWLEDGMENTS

We would like to thank the authors of all submitted papers. Their creative efforts have lead 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: *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.* We are grateful to *Stefano Lonardi* for allowing us to use a photo that he took in Venice (see Figure 1) for the WebKDD'05 brochure and proceedings cover. O. Nasraoui gratefully acknowledges the support of *NSF* as part of NSF CAREER award IIS-0133948.

Picture taken by Stefano Lonardi in Venice and used for the WebKDD'05 brochure.

Figure 1: resemblance between the light streaks and web clickstreams.

7. REFERENCES

[1] Bettina Berendt, "The semantics of frequent subgraphs: Mining and navigation pattern analysis", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[2] Colin DeLong, Prasanna Desikan, Jaideep Srivastava, "USER (User Sensitive Expert Recommendation): What Non-Experts NEED to Know", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[3] Miha Grcar, Blaz Fortuna, Dunja Mladenic, Marko Grobelnik, "kNN Versus SVM in the Collaborative Filtering Framework", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[4] Hyoung-rae Kim, Philip K. Chan, "S Personalized Ranking of Search Results with Learned User Interest Hierarchies from Bookmarks", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[5] Lin Lu, Margaret Dunham, Yu Meng, "Discovery of Significant Usage Patterns from Clusters of Clickstream Data", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[6] Bamshad Mobasher, Robin Burke, Runa Bhaumik, Chad Williams, "Effective Attack Models for Shilling Item-Based Collaborative Filtering System", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[7] Vincent Schickel, Boi Faltings, "Heterogeneous Attribute Utility Model: A new approach for Modeling User Profiles for Recommendation Systems", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[8] Bhushan Shankar Suryavanshi, Nematollaah Shiri, Sudhir P. Mudur, "Incremental Relational Fuzzy Subtractive Clustering for Dynamic Web Usage Profiling", In Proc. of *WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis*, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

[9] Sujatha Upadhyaya, Saleena N, Sreenivasa Kumar, "Realising OWL Individuals in XML Data", *In Proc. of WebKDD 2005: KDD Workshop on Web Mining and*

Web Usage Analysis, in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), August 21-24 2005, Chicago, IL.

About the authors:

Olfa Nasraoui is the Endowed Chair of E-commerce and the Director of the Knowledge Discovery and Web Mining Lab at the University of Louisville, where she is also Assistant professor in Computer Science and Engineering. She received her Ph.D. in Computer Engineering and Computer Science from the University of Missouri-Columbia in 1999. From 2000 to 2005, she was an Assistant Professor at the University of Memphis. Her research activities include Data Mining, in particular, Web mining and mining evolving data streams, Personalization, and Computational Intelligence. She has served on the organizing and program committees of several conferences and workshops, including WebKDD 2004 and WebKDD 2005. She is the recipient of the National Science Foundation CAREER Award for outstanding young scientists. Her research is funded mainly by NSF and by NASA. She is a member of IEEE and ACM.
(http:/www.louisville.edu/~o0nasr01)

Osmar R. Zaïane is an Associate Professor in Computing Science at the University of Alberta, Canada. Dr. Zaiane joined the University of Alberta in July of 1999 after obtaining his Ph.D. from Simon Fraser University, Canada, under the supervision of Dr. Jiawei Han. His Ph.D. thesis work focused on web mining and multimedia data mining. He has research interests in novel data mining algorithms, web mining, text mining, image mining, and information retrieval. He has published more than 70 papers in refereed international conferences and journals, and taught on all six continents. Osmar Zaïane was the co-chair of the ACM SIGKDD International Workshop on Multimedia Data Mining in 2000, 2001 and 2002 as well as co-Chair of the ACM SIGKDD WebKDD workshop in 2002, 2003 and 2005. He was guest-editor of the special issue on multimedia data mining of the journal of Intelligent Information Systems (Kluwer), and wrote multiple book chapters on multimedia mining and web mining. He has been an ACM Member since 1986. Osmar Zaïane is the ACM SIGKDD Explorations Associate Editor and Associate Editor of the International Journal of Internet Technology and Secured Transactions.
(http://www.cs.ualberta.ca/~zaiane/)

Myra Spiliopoulou is Professor at the Faculty of Computer Science,Otto-von-Guericke-University Magdeburg,Germany. Her research is on the development and enhancement of knowledge discovery methods for person-web interaction, for text analysis and for knowledge management. Her research on web usage mining and data preparation, text mining, temporal mining and pattern evolution has appeared in journals, conferences, books and workshops. She is regular reviewer in many data mining conferences, including KDD, ECML/PKDD and SIAM Data Mining, of the IEEE TKDE Journal and of many workshops in subjects associated to web mining. She has co-organized most of the WEBKDD workshops in the KDD conference series since 1999, In Europe, she has launched the Web Mining Forum initiative under the KDNet Network of Excellence and organized the workshop "European Web Mining Forum" in ECML/PKDD'2003 and '2005. In the ECML/PKDD conference series she has also given several tutorials on web mining since 1999. In March 2005, she organized the 29th Annual Conference of the German Classification Society (GfKl'2005) in Magdeburg.
(http://omen.cs.uni-magdeburg.de/itikmd)

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 2005 ACM SIGKDD conference in Seattle. He was the local arrangements chair for the 2005 ACM SIGKDD conference (KDD'05) 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)

Philip S. Yu is with IBM Thomas J. Watson Research Center and currently manager of the Software Tools and Techniques group. He recieved his PhD from Stanford Universithy. Dr. Yu is a Fellow of the ACM and the IEEE. He is an associate editor of ACM Transactions on Internet Technology. He is a member of the IEEE Data Engineering steering committee and is also on the steering committee of IEEE Conference on Data Mining. He was the Editor-in-Chief of IEEE Transactions on Knowledge and Data Engineering (2001-2004). Dr. Yu has published more than 430 papers in refereed journals and conferences. He holds or has applied for more than 250 US patents. He received an Outstanding Contributions Award from IEEE Intl. Conference on Data Mining in 2003 and also an IEEE Region 1 Award for "promoting and perpetuating numerous new electrical engineering concepts" in 1999. Dr. Yu is an IBM Master Inventor.
(http://www.research.ibm.com/people/p/psyu/)