Data Mining and Knowledge Discovery, 6, 5–8, 2002

Web Mining

RON KOHAVI ronnyk@CS.Stanford.EDU
Blue Martini Software, 2600 Campus Dr., San Mateo, CA 94403, USA

BRIJ MASAND bmasand@verilytics.com
Verilytics, 1 Wayside Road, Burlington, MA 01803, USA

MYRA SPILIOPOULOU myra@ebusiness.hhl.de
Leipzig Graduate School of Management (HHL), Department of E-Business, Jahnallee 59,
D-04109 Leipzig, Germany

JAIDEEP SRIVASTAVA srivasta@cs.umn.edu
Computer Science & Engineering, 4-192 EECS Building, 200 Union Street SE, University of Minnesota
Minneapolis, MN 55455, USA

Editors: Kohavi, Masand, Spiliopoulou, Srivastava

Received August 3, 2001; Revised November 9, 2001

The ease and speed with which business transactions can be carried out over the Web has been a key driving force in the rapid growth of electronic commerce. In addition, customer interactions, including personalized content, e-mail campaigns, online customer service, and online surveys provide new channels of communication that were not previously available or were very inefficient. The Web is revolutionizing the way businesses interact with each other (B2B) and with each customer (B2C). It has introduced entirely new ways of doing commerce, including auctions and reverse auctions, micro-segmented offers, dynamic pricing, and up-to-date content. It also made it imperative for organizations and companies to optimize their electronic business.

Knowledge about the customer is fundamental for the establishment of viable e-commerce solutions. As described by Jeff Bezos, CEO of Amazon.com, and mentioned by Joseph Pine in his book *The Experience Economy* (Pine et al.), customer experience is the key to building customer loyalty in an on-line store because leaving the store is only one click away. Web mining for e-commerce is the application of mining techniques to acquire this knowledge for improving e-commerce. The use of mining techniques in e-commerce can help improve cross-sells, up-sells, assortments shown, ads shown. In addition, clickstream collection allows for unprecedented measurement of site activities, conversion rates, and the effect of action (Kohavi, 2001).

WEBKDD 2000 was the second workshop,[1] held in conjunction with the Sixth ACM SIGKDD International Conference on Knowledge Discovery in Databases (KDD), dedicated to the challenges of web mining. In response to call for papers, WEBKDD 2000 received 31 contributions. Each was reviewed by at least three program committee members. Seven submissions were selected for presentation as long papers, and six as short papers reporting on good ideas at a rather preliminary phase. The authors of all papers accepted at WEBKDD 2000 were invited to prepare and submit an extended version of their work for consideration in this special issue. All papers were reviewed by at least three reviewers. Based on their ratings and comments, four papers were selected. The main selection criteria was the maturity of the work and its impact.

Being able to analyze click-stream data provides an unprecedented opportunity to understand in detail the process leading up to a buy/not buy decision vs. just recording the final outcome—as is the case with point-of-sale data. Clickstream data is over 95% of all data collected in most large-scale e-commerce environments, and contains a wealth of knowledge embedded in it. Tan and Kumar's *Modeling of Web Robot Navigational Patterns* addresses the challenging and commercially important problem of separating the site visits of web robots from humans. This is crucial for at least two reasons: (1) as competitive pressures increase, commerce sites would like to block robots that collect sensitive information, and (2) accurate modeling of human users' e-commerce behavior requires that web robot accesses be filtered out. Berendt's paper, titled *Web Usage Mining, Site Semantics, and the Support of Navigation*, provides a general overview of the issues in click-stream analysis, and shows how the mined knowledge can be used for supporting site navigation. Mobasher, Dai, Luo, Nakagawa, Sun, and Wiltshire's paper, titled *Discovery of Aggregate Usage Profiles for Web Personalization*, describe how usage data from web logs can be analyzed/mined to build user profiles, and how these could be use to enhance the user's browsing experience. Lin, Alvarez and Ruiz's *Collaborative Recommendation via Adaptive Association Rule Mining* describe an approach to using association rules for collaborative recommendation. An algorithm for mining association rules, which uses characteristics of a target user to improve the efficiency of the mining process, is introduced.

WEBKDD 2000 turned out to be a very successful workshop. More than 110 people showed interest in the workshop and over 85 attended it. The quality of papers was excellent, the discussion was lively, and a number of interesting directions of research were identified. This is a strong endorsement of the level of interest in this rapidly emerging field of inquiry. We are pleased to present to the readers the very best papers from WEBKDD 2000. Collectively, this represents work that we believe will have significant impact on the future of web mining.[2]

We wish to thank our reviewers, especially Paul Alpar, Alex Buechner, Jonathan Becher, Michael Berry, Alan Broder, Robert Cooley, Brij Masand, Brad Miller, Günter Mueller, Maurice Mulvenna, Carsten Pohle, Myra Spiliopoulou, Jaideep Srivastava, and Alex Tuzhilin.

Notes

1. The first workshop in this series was held in conjunction with SIGKDD 1999, and its proceedings have been published in Masand and Spiliopoulou (2000).
2. Interested readers are referred to Srivastava et al. (2000) and Kosala and Blockeel (2000) as recent surveys of the field of web mining.

References

Kohavi, R. 2001. Mining e-commerce data: The Good, the Bad, and the Ugly (invited industrial track talk). In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, F. Provost and R. Srikant (Eds.). Aug. 2001. http://robotics.Stanford.EDU/users/ronnyk/goodBadUglyKDDI-track.pdf

Kosala, R. and Blockeel, H. 2000. Web mining research: A survey, ACM SIGKDD Explorations, 2(1). http://www.acm.org/sigkdd/explorations/issue2-1.htm

Masand, B. and Spiliopoulou, M. (Eds.). 2000. Advances in Web Usage Mining and User Profiling: Proceedings of the WEBKDD'99 Workshop, Springer Verlag, July 2000. LNAI, Vol. 1836. (BibTeX Entry)

Pine, B.J., Gilmore, J.H., and Pine, B.J. II. The Experience Economy. Harvard Business School Pr; ISBN: 0875848192, http://www.amazon.com/exec/obidos/ASIN/0875848192/ref=sc_b_1/103-2009916-9046229

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. 2000. Web usage mining: Discovery and applications of web usage patterns from web data, ACM SIGKDD Explorations, 1(2). http://www.acm.org/sigkdd/explorations/issue1-2.htm

Ron Kohavi is the senior director of data mining at Blue Martini Software, where he heads the engineering group responsible for the data collection, analysis, visualization, reporting, and campaign management modules in Blue Martini's applications. Prior to joining Blue Martini, Kohavi managed the MineSet project, Silicon Graphics' award-winning product for data mining and visualization. He joined Silicon Graphics after getting a Ph.D. in Machine Learning from Stanford University, where he led the MLC++ project, the Machine Learning library in C++ now used in MineSet and at Blue Martini Software. Kohavi received his BA from the Technion, Israel. He co-chaired KDD 99's industrial track with Jim Gray and the KDD Cup 2000 with Carla Brodley. He was an invited speaker at the National Academy of Engineering in 2000, a keynote speaker at PAKDD 2001, and an invited speaker at KDD 2001's industrial track. He co-chaired WEBKDD 2000 and co-taught with Jon Becher a tutorial on e-commerce and clickstream mining at the SIAM Data Mining conference in 2001. He co-edited with Foster Provost the special issue of the journal Machine Learning on Applications of Machine Learning and the special issue of the Data Mining and Knowledge Discovery journal on Applications of Data Mining to Electronic Commerce, now available as a book. He is a member of the editorial board for the Data Mining and Knowledge Discovery Journal from its inception and served as a member of the editorial board for the journal of Machine Learning from 1997 to 1999.

Brij Masand serves as architect, Intelligent Systems, at Verilytics and leads a engineering team to design and develop intelligent agents using real time analytics for the financial services, consumer finance and supply chain markets. Mr. Masand received his M.S. in EECS from MIT in 1982 and a B. Tech in EE from IIT, New Delhi, India in 1979. He has 18 years of research and product development experience in the areas of AI, data mining, text mining, real time analytics and intelligent agents. Prior to joining Verilytics Brij led data mining teams at Redwood Investments, where he developed intelligent agents for the financial markets to track both text based and numeric information, and at GTE Laboratories, where he led a team to build the largest predictive medeling system at GTE to model its cellular customers and launched their web analytics initiative. Prior to that at Thinking Machines, he applied memory based reasoning (MBR), using parallel processing technologies, to text classification and helped build Darwin, a data mining product for massively parallel architectures.

Mr. Masand has numerous publications in eh areas of data mining and AI and he also holds two patents for text processing. He has served on the program committee for KDD for several years and co-founded the WEBKDD series of workshops (WEBKDD'99). His current interest include intelligent text processing to reduce information overload, time series modeling, real time web analytics and meta data architectures to support real time intelligent response."

Myra Spiliopoulou received her M.Sc. in Mathematics and the Ph.D. degree in computer science from the University of Athens, Greece, in 1986 and 1992, respectively. Between 1987–1994, she worked as a research assistant in the Department of Informatics, University of Athens, and was involved in national and European projects on parallel databases, hypermedia and multimedia modelling and querying, and on computer-aided education. Between 1994 and 2000, she was with the institute of information systems in the Humboldt University Berlin, Germany. In September 2000, she joined the faculty of computer science in the University of Magdeburg, Germany, as a guest associate professor for one semester. Since April 2001, she is professor of e-business in the Leipzig Graduate School of Management, Germany.

Her research interests cover several KDD areas. A major research area concerns the analysis of web usage data, including web usage mining, site evaluation and e-metrics for web sites. She also investigates aspects of text analysis and XML DTD extraction with mining techniques. She considers data and text mining as one phase of a complex process, and investigates methodologies for the support of the whole process, in particular pre-mining techniques for data preparation and post-mining techniques for the evaluation and the lifelong monitoring of the results. In her work, she addresses both the algorithmic aspects of knowledge discovery and its practical applications in areas such as electronic commerce and knowledge management.

Jaideep Srivastava received his Ph.D. from the University of California—Berkeley in 1988, and has been on the faculty of the University of Minnesota, where he is a Full Professor. For over 15 years he has been active as a researcher, educator, consultant, and invited speaker, in the areas of databases, artificial intelligence, and multimedia. He has established and led a database and multimedia research laboratory, which has graduated 15 Ph.D. and 36 M.S. students, and in the process published over 135 papers in journals and conferences. Throughout his career Dr. Srivastava has had an active collaboration with the industry, both for collaborative research and technology transfer.

In a two year sabbatical between 1999 and 2001, Dr. Srivastava spent time at Amazon.com (www.amazon.com) as the Chief Data Mining Architect, at Yodlee Inc. (www.yodlee.com) as Director—Data Analytics, and at Chingari Inc. (www.chingari.com) as the Chief Technology Officer. This wide-ranging industry experience has provided Dr. Srivastava a unique perspective on the application of various computer science technologies in the Internet economy.

Dr. Srivastava is an often-invited participant in technical as well as technology strategy forums. He has given more than a hundred talks in various industry, academic, and government forums. He has served on the program committee of a number of conferences, and on the editorial board of various journals. He has also served in an advisory role to the governments of India and Chile on various software technologies. Dr. Srivastava is a member of the ACM, and a senior member of the IEEE.