Lecture Notes in Artificial Intelligence 2703

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Osmar R. Zaïane Jaideep Srivastava
Myra Spiliopoulou Brij Masand (Eds.)

WEBKDD 2002 – Mining Web Data for Discovering Usage Patterns and Profiles

4th International Workshop
Edmonton, Canada, July 23, 2002
Revised Papers

Springer

Series Editors

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

Volume Editors

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

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

Myra Spiliopoulou
Otto-von-Guericke University of Magdeburg, Faculty of Computer Science
Institute of Technical and Business Information Systems
P.O. Box 4120, 39016 Magdeburg, Germany
E-mail: myra@iti.cs.uni-magdeburg.de

Brij Masand
Data Miners Inc.
77 North Washington Street, 9th Floor, Boston, MA 02114, USA
E-mail: brij@data-miners.com

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

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

ISSN 0302-9743
ISBN 3-540-20304-4 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg NewYork
a member of BertelsmannSpringer Science+Business Media GmbH

www.springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin GmbH
Printed on acid-free paper SPIN: 10928684 06/3142 5 4 3 2 1 0

Preface

1 Workshop Theme

Data mining as a discipline aims to relate the analysis of large amounts of user data to shed light on key business questions. Web usage mining in particular, a relatively young discipline, investigates methodologies and techniques that address the unique challenges of discovering insights from Web usage data, aiming to evaluate Web usability, understand the interests and expectations of users and assess the effectiveness of content delivery. The maturing and expanding Web presents a key driving force in the rapid growth of electronic commerce and a new channel for content providers. Customized offers and content, made possible by discovered knowledge about the customer, are fundamental for the establishment of viable e-commerce solutions and sustained and effective content delivery in noncommercial domains. Rich Web logs provide companies with data about their online visitors and prospective customers, allowing microsegmentation and personalized interactions.

While Web mining as a domain is several years old, the challenges that characterize data analysis in this area continue to be formidable. Though preprocessing data routinely takes up a major part of the effort in data mining, Web usage data presents further challenges based on the difficulties of assigning data streams to unique users and tracking them over time. New innovations are required to reliably reconstruct sessions, to ascertain similarity and differences between sessions, and to be able to segment online users into relevant groups. While Web usage data is large in volume, recommender system approaches still suffer from the challenges of compiling enough data to correlate user preferences to products. Intelligent use of domain abstractions that can help characterize how to group Web pages and incorporating knowledge of Web site structure into Web data analysis are new areas that can take Web usage mining to the next level. This workshop addresses advances along these lines as demonstrated by the papers included in this volume.

WEBKDD 2002 was the fourth in the WEBKDD series of workshops, with special emphasis on mining Web data for discovering usage patterns and profiles. WEBKDD 1999 focused on the aspects of Web mining related to user profiling, WEBKDD 2000 focused on Web Mining for E-Commerce, and WEBKDD 2001 focused on mining Web log data across all customer touchpoints.

The KDD community responded very enthusiastically to the WEBKDD 2002 workshop. More than 50 people attended the workshop, which brought together practitioners, tool vendors and researchers interested in Web mining. The paper presentations were divided into three sessions, titled "Categorization of Users and Usage," "Prediction and recommendation," and "Evaluation of algorithms." A total of 23 papers were submitted to WEBKDD 2002, of which 10 were selected for presentation at the workshop – a 44% acceptance rate. The authors of the

papers presented at WEBKDD 2002 were invited to submit extended versions of their papers for this special issue. A second round of review was carried out for each paper, and the revised and enhanced versions of all 10 papers are included in this book. In the next section we summarize each paper.

The final versions of the workshop papers can be found in the online repository of published papers: http://www.acm.org/sigkdd/proceedings/webkdd02/.

2 Papers

The first presentation, by Chi, Rosien and Heer investigated whether major groupings of user traffic on a Web site can be discovered in an automated fashion. In their paper "LumberJack: Intelligent Discovery and Analysis of Web User Traffic Composition," they describe how using multiple features from the user session data, automated clustering analysis can achieve high classification accuracy.

In their paper "Mining eBay: Bidding Strategies and Shill Detection," Shah, Joshi, Sureka, and Wurman investigate bidding strategies in online auction markets. Using online auction data from eBay for video game console auctions, they propose new attributes of bidding engagements and rules for classifying strategies. Through the analysis of auctions where multiple sessions for individuals are tracked over time, analysis of the bidding sessions led them to identify some known and new bidding behaviors, including clustering of strategies. Among the bidding strategies they identify is shilling behaviour (where there is a fake orchestrated sequence of bids to drive up the price).

In their presentation on "Automatic Categorization of Web Pages and User Clustering by Using Mixtures of Hidden Markov Models," Ypma and Heskes describe an EM algorithm for training the mixture of HMMs and also show how to use prior knowledge to help the learning process. They test their algorithm on both artificial data where they demonstrate that the correct patterns are being learnt, as well as on real data from a commercial Web site.

In their presentation on "Web Usage Mining by Means of Multidimensional Sequence Alignment Methods," Hay, Wets and Vanhoof explore the analysis of Web usage sequences by using a multidimensional sequence alignment method. They illustrate how navigation patterns are discovered by aligning sequences by using not only page sequence but also the time visited per page. They demonstrate empirical test results of the new algorithm, MDSAM, which show discovered user profiles.

In their paper "A Customizable Behavior Model for Temporal Prediction of Web User Sequences," Frias-Martinez and Karamcheti model the behavior of users for the prediction of sequences of Web user access. They propose a new model that allows customizing of the sequential nature of the preceding (known) patterns as well as the predicted patterns. Their algorithm also calculates a measure of the gap between the antecedent and the consequent sequences.

In his paper "Coping with Sparsity in a Recommender System," André Bergholz addresses the important problem of sparseness in the data used for

recommender systems. Typically there are too few ratings, resulting in limited correlations between users. Bergholz explores two innovative approaches for resolving this problem. The first one uses transitive correlations between existing users, which adds new ratings without much computational expense. The second approach uses a ratings agent that actually assigns ratings in accordance with some predefined preferences, resulting in increased coverage but some impact on system performance.

In their presentation "On the Use of Constrained Associations for Web Log Mining," Yang and Parthasarthy investigated the problem of scalability when mining for association rules for predicting patterns of Web usage access. They describe a new approach for mining such rules which introduces constraints in the way the rules are discovered. They explore the hypothesis that considering recent pages is more important in predicting consequent pages. This ordering and temporal constraint results in simpler and fewer rules, thus enabling faster online deployment.

In their paper "Mining WWW Access Sequence by Matrix Clustering," Oyanagi, Kubota and Nakase present a novel approach to sequence mining using matrix clustering. Their method, which results in generalized sequence patterns, decomposes a sequence into a set of sequence elements, each of which corresponds to an ordered pair of items. Then matrix clustering is applied to extract a cluster of similar sequences. The resulting sequence elements are then combined into generalized sequences. The authors demonstrate the discovery of long sequences in actual Web logs.

In their presentation on "Comparing Two Recommender Algorithms with the Help of Recommendations by Peers," Geyer-Schulz and Hahsler presented a framework for evaluating the effectiveness of recommender systems. They investigate whether the recommendations produced by algorithms such as those using frequent itemsets are useful for the social process of making further recommendations to others. They use two algorithms, one using association rules and another using repeat-buying theory from marketing research to compare how well the recommendations produced by brokers in an information market firm compare with those that are produced by the algorithm. They find that the recommendations compare quite favorably, with a high degree of accuracy and precision, and that both algorithms perform similarly when tuned appropriately on real data.

In their paper on "The Impact of Site Structure and Web Environment on Session Reconstruction in Web Usage Analysis," Berendt, Mobasher, Nakagawa and Spiliopoulou investigate the reliability of session reconstruction, and continue their work on characterizing the issues in Web preprocessing. They investigate factors such as the presence and absence of cookies, server-side session information, effect of site structure, etc., to evaluate errors introduced in session reconstruction and its subsequent impact on predictions for Web personalization. They specifically analyze data from frame-based and frame-free versions of a site with respect to the quality of session reconstruction using different heuristics,

and suggest that Web site characteristics can be a guide to the use of specific heuristics.

3 Conclusion

In its fourth year as a workshop, WEBKDD 2002 continued to show the sustained interest in this area and turned out to be a very successful workshop. About 50 people attended it, roughly divided evenly between industry and academia. It was an interactive workshop with a lively exchange between presenters and attendees. This year, apart from the continuing themes of new approaches to analyze and cluster sessions, there were more papers on recommender systems. We hope to be more inclusive in future years to expand the range of research topics included in the workshop.

4 Acknowledgements

We would like to acknowledge the Program Committee members of WEBKDD 2002 who invested their time in carefully reviewing papers for this volume: Jonathan Becher (Accrue/Neovista Software, Inc.), Bettina Berendt (HU Berlin, Germany), Alex Buechner (Lumio Ltd. UK), Ed Chi (Xerox Parc, USA), Robert Cooley (KXEN, USA), Wolfgang Gaul (University Karlsruhe, Germany), Oliver Guenther (HU Berlin, Germany), Ronny Kohavi (Blue Martini Software, USA), Vipin Kumar (AHPCRC-University of Minnesota, USA), Ee-Peng Lim (Chinese University of Hong Kong, China), Sanjay Kumar Madria (University of Missouri-Rolla), Yannis Manolopoulos (Aristotle Univ., Greece), Bamshad Mobasher (De Paul Univ., USA), Jian Pei (SUNY Buffalo, USA), Alex Tuzhilin (NYU/Stern School of Business, USA), Terry Woodfield (SAS Institute, Inc.), and Mohammed Zaki (Rensselaer Polytechnic Institute, USA).

We would also like to thank others who contributed to WEBKDD 2002, including the original PC members who reviewed the first set of workshop papers. We are grateful to the KDD 2002 organizing committee, especially Renée Miller, University of Toronto (Workshops Chair), Jörg Sander, University of Alberta, Canada (Registration Chair) and Mario Nascimento, University of Alberta, Canada (Local Arrangements Chair), for their help in bringing the WEBKDD community together. Finally we would like to thank the many participants who brought their ideas, research and enthusiasm to the workshop and proposed many new directions for WEBKDD research.

June 2003

Osmar R. Zaïane
Jaideep Srivastava
Myra Spiliopoulou
Brij Masand

Table of Contents

LumberJack: Intelligent Discovery and Analysis of Web User Traffic Composition 1
Ed H. Chi, Adam Rosien, Jeffrey Heer

Mining eBay: Bidding Strategies and Shill Detection 17
Harshit S. Shah, Neeraj R. Joshi, Ashish Sureka, Peter R. Wurman

Automatic Categorization of Web Pages and User Clustering with Mixtures of Hidden Markov Models 35
Alexander Ypma, Tom Heskes

Web Usage Mining by Means of Multidimensional Sequence Alignment Methods 50
Birgit Hay, Geert Wets, Koen Vanhoof

A Customizable Behavior Model for Temporal Prediction of Web User Sequences 66
Enrique Frías-Martínez, Vijay Karamcheti

Coping with Sparsity in a Recommender System 86
André Bergholz

On the Use of Constrained Associations for Web Log Mining 100
Hui Yang, Srinivasan Parthasarathy

Mining WWW Access Sequence by Matrix Clustering 119
Shigeru Oyanagi, Kazuto Kubota, Akihiko Nakase

Comparing Two Recommender Algorithms with the Help of Recommendations by Peers 137
Andreas Geyer-Schulz, Michael Hahsler

The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis 159
Bettina Berendt, Bamshad Mobasher, Miki Nakagawa, Myra Spiliopoulou

Author Index 181