Lecture Notes in Artificial Intelligence 2356

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell, and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Ron Kohavi Brij M. Masand
Myra Spiliopoulou Jaideep Srivastava (Eds.)

WEBKDD 2001 – Mining Web Log Data Across All Customers Touch Points

Third International Workshop
San Francisco, CA, USA, August 26, 2001
Revised Papers

Springer

Volume Editors

Ron Kohavi
Blue Martini Software
2600 Campus Drive, San Mateo, CA 94403, USA
E-mail: ronnyk@cs.stanford.edu

Brij M. Masand
Data Miners Inc.
77 North Washington Street, Boston, MA 02114, USA
E-mail: bmasand@alum.mit.edu

Myra Spiliopoulou
Leipzig Graduate School of Management
Jahnallee 59, 04109 Leipzig, Germany
E-mail: myra@ebusiness.hhl.de

Jaideep Srivastava
University of Minnesota, 4-192 EECS Building
200 Union St SE, Minneapolis, MN 55455
E-mail: srivasta@cs.umn.edu

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Mining web log data across all customers touch points : third international workshop ; revised papers ; proceedings / WEBKDD 2001, San Francisco, CA, USA, August 26, 2001. Ron Kohave ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2356 : Lecture notes in artificial intelligence)
ISBN 3-540-43969-2

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

ISSN 0302-9743
ISBN 3-540-43969-2 Springer-Verlag Berlin Heidelberg New York

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

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN 10870091 06/3142 5 4 3 2 1 0

Preface

Workshop Theme

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, and online feedback provide new channels of communication that were not previously available or were very inefficient.

The Web presents a key driving force in the rapid growth of electronic commerce and a new channel for content providers. Knowledge about the customer is fundamental for the establishment of viable e-commerce solutions. Rich web logs provide companies with data about their customers and prospective customers, allowing micro-segmentation and personalized interactions. Customer acquisition costs in the hundreds of dollars per customer are common, justifying heavy emphasis on correct targeting. Once customers are acquired, customer retention becomes the target. Retention through customer satisfaction and loyalty can be greatly improved by acquiring and exploiting knowledge about these customers and their needs.

Although web logs are the source for valuable knowledge patterns, one should keep in mind that the Web is only one of the interaction channels between a company and its customers. Data obtained from conventional channels provide invaluable knowledge on existing market segments, while mobile communication adds further customer groups. In response, companies are beginning to integrate multiple sources of data including web, wireless, call centers, and brick-and-mortar store data into a single data warehouse that provides a multifaceted view of their customers, their preferences, interests, and expectations.

Web mining for e-commerce is the application of web mining techniques to acquire this knowledge for e-commerce. Typical concerns in e-commerce include improved cross-sells, up-sells, personalized ads, targeted assortments, improved conversion rates, and measurements of the effectiveness of actions.

WEBKDD 2001 was the third in the WEBKDD series of workshops, devoted to mining web data. WEBKDD'99 focused on the aspects of web mining related to user profiling, and WEBKDD 2000 focused on Web Mining for E-Commerce.

The URL http://robotics.stanford.edu/~ronnyk/WEBKDD2001 contains the final versions of the workshop papers and the slide presentations.

Papers

The KDD community responded very enthusiastically to the WEBKDD 2001 workshop, and we received a number of requests for attendance. Attendance to the workshop was by invitation only, and requests for attendance were accompanied by short CVs, to allow for discussions among participants with appropriate

background and interest. A total of 52 people attended the workshop, which brought together practitioners, tool vendors, and researchers interested in web mining. The paper presentation was divided into three sessions, titled "Predicting User Accesses", "Recommender Systems and Access Modeling", and "Acquiring and Modeling Data and Patterns". A total of 18 papers were submitted to WEBKDD 2001, of which 9 were selected for presentation at the workshop – making it a 50% acceptance rate. The authors of the papers presented at WEBKDD 2001 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 seven papers are included in this book. In this section we summarize each paper.

In her paper titled "Detail and Context in Web Usage Mining: Coarsening and Visualizing Sequences" [1], Berendt illustrates the power of using visualization to better understand the results of web usage mining. Specifically, Berendt shows how visualization can be helpful in understanding long patterns, with little expected structure, that have been mined from the usage logs. Concept hierarchies are presented as a basic method for aggregating web pages, and interval-based coarsening as an approach to representing sequences at different levels of abstraction. A tool, called STRATDYN is described, which uses chi-square test and coarsened stratograms for analyzing differences in support and confidence values. Stratograms with uniform or differential coarsening provide various detail and context views of actual and intended web usage. Relationship to the measures of support and confidence, and methods of analyzing generalized sequences are shown. A case study of an agent-supported e-commerce shopping scenario is used to illustrate the framework.

In "A Customer Purchase Incidence Model Applied to Recommender Services" [2], Geyer-Schulz, Hahsler, and Jahn show how Ehrenberg's theory of repeat-buying can be adapted to the web-based buying environment. Ehrenberg's theory has been successful in describing regularities in a large number of consumer goods and markets. The authors apply the same to show that regularities exist in electronic markets as well, and purchase incidence models provide a theoretically sound basis for recommender and alert services. An empirical validation of the approach, based on data collected from the University of Vienna, is provided.

"A Cube Model and Cluster Analysis for Web Access Sessions" [3] by Huang, Ng, Ching, Ng, and Cheung, show how the application of data cube model and cluster analysis techniques can help in extracting useful e-commerce patterns from web usage data. The cube model organizes session data into three dimensions. The COMPONENT dimension represents a session as a sequence of ordered components, in which the i-th component represents the i-th page visit of the session. Each component is represented by a number of attributes including page ID, category, and time spent. The ATTRIBUTE dimension describes the attributes associated with each component, while the SESSION dimension indexes individual sessions. Irregular sessions are converted into a regular data structure, so that data mining algorithms can be more easily applied. The k-modes algorithms, designed for clustering categorical data and a clustering technique using

Markov transition probabilities is used for the clustering of sequences. An experimental validation of the technique is presented.

In their paper "Exploiting Web Log Mining for Web Cache Enhancement" [4], Nanopoulos, Katsaros, and Manolopoulos show how knowledge gained form mining web usage logs can be used to enhance the performance of better web cache management, thereby reducing the latency perceived by a web user. This is of great value in an e-commerce scenario from a customer experience perspective. The key idea is to use the knowledge extracted to perform better pre-fetching, based on the development of a good page access prediction model. The proposed scheme achieves a good balance between caching and pre-fetching. The pre-fetched documents are placed in a dedicated part of the cache, to avoid the drawback of replacing requested documents with the ones whose access is only speculative. Experimental evaluation of the proposed scheme is presented.

Tan and Kumar, in their paper entitled "Mining Indirect Associations in Web Data" [5], introduce the concept of indirect associations and show how it is applicable to web usage mining. An indirect association between items A and B is said to exist if the direct association between them is quite weak, but there exists item-set X such that the association between A and X is strong and the association between X and B is also strong. The application of this idea to web usage data enables the identification of groups of users with distinct interests. Such patterns may not be discernible using traditional approaches, unless these user groups are known a priori. The approach is validated using data from an academic site as well as a commercial site.

In "A Framework for Efficient and Anonymous Web Usage Mining Based on Client Side Tracking" [6], Shahabi and Bannaei-Kashani describe an approach to using web usage mining for personalization applications. The claim is that for on-line and anonymous personalization to be effective, web usage mining must be carried out in real-time, and with high accuracy. In addition, the approach must allow a tradeoff between accuracy and speed. They introduce the distributed web-tracking approach for accurate, efficient, and scalable collection of usage data. An approach called Feature Matrices (FM) model is proposed to capture and analyze usage patterns. With FM various features of the usage data can be captured with flexible precision, so that accuracy and scalability can be traded off based on application requirements. Additionally, low model complexity allows FM to adapt to user behavior changes in real time. A new similarity measure, designed for capturing partial navigational patterns, is presented. Experimental validation with synthetic and real-life data sets is presented.

In their paper "LOGML: Log Markup Language for Web Usage Mining" [7], Punin, Krishnamoorthy, and Zaki introduce an XML-based language for representing objects of use in the web log mining domain. As per the authors, while extracting simple information from web logs is straightforward, identifying complex knowledge is very challenging. In addition, data cleaning and preprocessing are very complex and demanding tasks. This paper presents two new ideas, namely XGMML and LOGML, to ease this task. The former is a graph description language, while the latter is a web-log report description language.

The web robot of the WWWPal system is used to generate the XGMML graph of a web site. Web-log reports, in LOGML format, are generated from web usage data, and the XGMML graph of a site. Examples illustrate how the combination of XGMML and LOGML can be used to easily capture the web mining process for a particular web site. This make it easy to reuse the various components of the analysis – providing much better leverage for the effort spent in creating the analysis.

Conclusion

WEBKDD 2001 turned out to be a very successful workshop by all measures. A number of people showed interest in the workshop and over 50 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.

Acknowledgements: We would like to acknowledge the Program Committee members of WEBKDD 2001 who invested their time in carefully reviewing papers for this volume: Jonathan Becher (Accrue/Neovista Software, Inc.), Bettina Berendt, (HU Berlin, Germany), Ed Chi (Xerox Parc, USA), Robert Cooley (KXEN, USA), Johannes Gehrke (Cornell Univ., USA), Joydeep Ghosh (Univ. of Texas, USA), Oliver Guenther (HU Berlin, Germany), Vipin Kumar, (AHPCRC-University of Minnesota, USA), Yannis Manolopoulos (Aristotle Univ., Greece), Llew Mason (Blue Martini Software, USA), Ann Milley (SAS Institute Inc., USA), Bamshad Mobasher (De Paul Univ., USA), Rajeev Rastogi (Bell Labs, Lucent, USA), Alex Tuzhilin (NYU/Stern School of Business, USA), Mohammed Zaki (Rensellaer Polytechnic Institute, USA) and Zijian Zheng, Blue Martini Software, USA.

We would also like to thank others that contributed to WEBKDD 2001, including the original PC members who reviewed the first set of workshop papers. We are grateful to the KDD 2001 organizing committee, especially Roberto Bayardo (workshops chair) and Jeonghee Yi (Registration 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 the WEBKDD research to continue.

June 2002

Ronny Kohavi
Brij Masand
Jaideep Srivastava
Myra Spiliopoulou

References

1. B. Berendt, “Detail and Context in Web Usage Mining: Coarsening and Visualizing Sequences”, pp. 1–24, this book.
2. Geyer-Schulz, Hahsler, and Jahn “A Customer Purchase Incidence Model Applied to Recommender Services”, pp. 25–47, this book.
3. Huang, Ng, Ching, Ng, and Cheung, “A Cube Model and Cluster Analysis for Web Access Sessions”, pp. 48–67. This book.
4. Nanopoulos, Katsaros, and Manolopoulos, “Exploiting Web Log Mining for Web Cache Enhancement”, pp. 68–87, this book.
5. Tan and Kumar, “Mining Indirect Associations in Web Data”, pp. 145–166, this book.
6. Shahabi and Bannaei-Kashani, “A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking”, pp. 113–144, this book.
7. Punin, Krishnamoorthy, and Zaki, “LOGML: Log Markup Language for Web Usage Mining”, pp. 88–112, this book.

Table of Contents

Detail and Context in Web Usage Mining:
Coarsening and Visualizing Sequences 1
Bettina Berendt

A Customer Purchase Incidence Model Applied
to Recommender Services 25
Andreas Geyer-Schulz, Michael Hahsler, and Maximillian Jahn

A Cube Model and Cluster Analysis for Web Access Sessions 48
Joshua Zhexue Huang, Michael Ng, Wai-Ki Ching, Joe Ng, and David Cheung

Exploiting Web Log Mining for Web Cache Enhancement 68
Alexandros Nanopoulos, Dimitrios Katsaros, and Yannis Manolopoulos

LOGML: Log Markup Language for Web Usage Mining 88
John R. Punin, Mukkai S. Krishnamoorthy, and Mohammed J. Zaki

A Framework for Efficient and Anonymous Web Usage Mining
Based on Client-Side Tracking 113
Cyrus Shahabi and Farnoush Banaei-Kashani

Mining Indirect Associations in Web Data 145
Pang-Ning Tan and Vipin Kumar

Author Index 167