Lecture Notes in Artificial Intelligence 4198

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Olfa Nasraoui Osmar Zaïane
Myra Spiliopoulou Bamshad Mobasher
Brij Masand Philip S. Yu (Eds.)

Advances in Web Mining and Web Usage Analysis

7th International Workshop
on Knowledge Discovery on the Web, WebKDD 2005
Chicago, IL, USA, August 21, 2005
Revised Papers

Springer

Series Editors

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

Volume Editors

Olfa Nasraoui
Speed School of Engineering, Louisville KY 40292
E-mail: olfa.nasraoui@louiswville.edu

Osmar Zaïane
University of Alberta, Edmonton, AB, T6G2E8, Canada
E-mail: zaiane@ualberta.ca

Myra Spiliopoulou
Otto-von-Guericke-University Magdeburg, Germany
E-mail: myra@iti.cs.uni-magdeburg.de

Bamshad Mobasher
School of Computer Science, Chicago, IL 60604, USA
E-mail: mobasher@cs.depaul.edu

Brij Masand
Data Miners Inc., , Boston, MA 02114, USA
E-mail: brij@data-miners.com

Philip S. Yu
IBM T. J. Inc., N.Y. 10598, USA
E-mail: psyu@us.ibm.com

Library of Congress Control Number: 2006933535

CR Subject Classification (1998): I.2, H.2.8, H.3-5, K.4, C.2

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-46346-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-46346-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11891321 06/3142 5 4 3 2 1 0

Preface

This book contains the postworkshop proceedings of the 7th International Workshop on Knowledge Discovery from the Web, WEBKDD 2005. The WEBKDD workshop series takes place as part of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) since 1999.

The discipline of data mining delivers methodologies and tools for the analysis of large data volumes and the extraction of comprehensible and non-trivial insights from them. Web mining, a much younger discipline, concentrates on the analysis of data pertinent to the Web. Web mining methods are applied on usage data and Web site content; they strive to improve our understanding of how the Web is used, to enhance usability and to promote mutual satisfaction between e-business venues and their potential customers.

In the last years, the interest for the Web as medium for communication, interaction and business has led to new challenges and to intensive, dedicated research. Many of the infancy problems in Web mining have now been solved but the tremendous potential for new and improved uses, as well as misuses, of the Web are leading to new challenges.

The theme of the WebKDD 2005 workshop was "Taming Evolving, Expanding and Multi-faceted Web Clickstreams." While solutions on some of the infancy problems of Web analysis have reached 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 2005 addressed these emerging aspects of Web reality.

In the first paper, *Mining Significant Usage Patterns from Clickstream Data*, Lu, Dunham, and Meng propose a technique to generate significant usage patterns (SUP) and use it to acquire significant "user preferred navigational trails." The technique uses pipelined processing phases including sub-abstraction of sessionized Web clickstreams, clustering of the abstracted Web sessions, concept-based abstraction of the clustered sessions, and SUP generation. Using this technique, valuable customer behavior information can be extracted by Web site practitioners. Experiments conducted using J.C.Penney Web log data demonstrate that SUPs of different types of customers are distinguishable and interpretable.

In the second paper, *Using and Learning Semantics in Frequent Subgraph Mining*, Berendt addresses the need for incorporating background knowledge into graph mining and for studying patterns at different levels of abstraction, by using taxonomies in mining and extending frequency / support measures by the notion of context-induced interestingness. Semantics are used as well as learned in this process, and a visualization tool is used to allow the user to navigate through detail-and-context views of taxonomy context, pattern context, and transaction context. A case study of a real-life Web site shows the advantages of the proposed solutions.

In the third paper, *Overcoming Incomplete User Models in Recommendation Systems via an Ontology*, Schickel and Faltings propose a new method that extends the utility model and assumes that the structure of user preferences follows an ontology of product attributes. Using the MovieLens data, their experiments show that real user preferences indeed closely follow an ontology based on movie attributes. Furthermore, a recommender based just on a single individual's preferences and this ontology performs better than collaborative filtering, with the greatest differences when few data about the user are available. This points the way to how proper inductive bias (in the form of an ontology) can be used for significantly more powerful recommender systems in the future.

The fourth paper, *Data Sparsity Issues in the Collaborative Filtering Framework* by Grcar, Fortuna, Mladenic, and Grobelnik gives an overview of collaborative filtering approaches, and presents experimental results that compare the k-nearest neighbor (kNN) algorithm with support vector machines (SVM) in the collaborative filtering framework using data sets with different properties. Experiments on two standard, publicly available data sets and a real-life corporate data set that does not fit the profile of ideal data for collaborative filtering lead the authors to conclude that the quality of collaborative filtering recommendations is highly dependent on the sparsity of available data. Furthermore, they show that kNN is dominant on data sets with relatively low sparsity while SVM-based approaches may perform better on highly sparse data.

The fifth paper focuses on the multi-faceted aspect of Web personalization. In *USER: User-Sensitive Expert Recommendations for Knowledge-Dense Environments*, DeLong, Desikan, and Srivastava address the challenge of making relevant recommendations given a large, knowledge-dense Web site and a non-expert user searching for information. They propose an approach to provide recommendations to non-experts, helping them understand what they need to

know, as opposed to what is popular among other users. The approach is user-sensitive in that it adopts a "model of learning" whereby the user's context is dynamically interpreted as they browse, and then leveraging that information to improve the recommendations.

In the sixth paper, *Analysis and Detection of Segment-Focused Attacks Against Collaborative Recommendation*, Mobasher, Burke, Williams, and Bhaumik examine the vulnerabilities that have recently been identified in collaborative filtering recommender systems. These vulnerabilities mostly emanate from the open nature of such systems and their reliance on user-specified judgments for building profiles. Hence, attackers can easily introduce biased data in an attempt to force the system to "adapt" in a manner advantageous to them. The authors explore an attack model that focuses on a subset of users with similar tastes and show that such an attack can be highly successful against both user-based and item-based collaborative filtering. They also introduce a detection model that can significantly decrease the impact of this attack.

The seventh paper, *Adaptive Web Usage Profiling*, by Suryavanshi, Shiri, and Mudur, addresses the challenge of maintaining profiles so that they dynamically adapt to new interests and trends. They present a new profile maintenance scheme, which extends the relational fuzzy subtractive clustering (RFSC) technique and enables efficient incremental update of usage profiles. An impact factor is defined whose value can be used to decide the need for recompilation. The results from extensive experiments on a large real dataset of Web logs show that the proposed maintenance technique, with considerably reduced computational costs, is almost as good as complete remodeling.

In the eigth paper, *On Clustering Techniques for Change Diagnosis in Data Streams*, Aggarwal and Yu, address the challenge of exploring the underlying changing trends in data streams that are generated by applications which are time-changing in nature. They explore and survey some of their recent methods for change detection, particularly methods that use clustering in order to provide a concise understanding of the underlying trends. They discuss their recent techniques which use micro-clustering in order to diagnose the changes in the underlying data, and discuss the extension of this method to text and categorical data sets as well community detection in graph data streams.

In *Personalized Search Results with User Interest Hierarchies Learnt from Bookmarks*, Kim and Chan propose a system for personalized Web search that incorporates an individual user's interests when deciding relevant results to return. They propose a method to (re)rank the results from a search engine using a learned user profile, called a user interest hierarchy (UIH), from Web pages that are of interest to the user. The users interest in Web pages will be determined implicitly, without directly asking the user. Experimental results indicate that the personalized ranking methods, when used with a popular search engine, can yield more potentially interesting Web pages for individual users.

We would like to thank the authors of all submitted papers. Their creative efforts have led 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 (in alphabetical order): 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. O. Nasraoui gratefully acknowledges the support of the US National Science Foundation as part of NSF CAREER award IIS-0133948.

June 2006

Olfa Nasraoui
Osmar Zaane
Myra Spiliopoulou
Bamshad Mobasher
Philip Yu
Brij Masand

Table of Contents