Lecture Notes in Artificial Intelligence 4811

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Olfa Nasraoui   Myra Spiliopoulou
Jaideep Srivastava   Bamshad Mobasher
Brij Masand (Eds.)

Advances in Web Mining and Web Usage Analysis

8th International Workshop
on Knowledge Discovery on the Web, WebKDD 2006
Philadelphia, PA, USA, August 20, 2006
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
University of Louisville
Louisville, KY 40292, USA
E-mail: olfa.nasraoui@louisville.edu

Myra Spiliopoulou
Otto-von-Guericke-Universität Magdeburg
39106 Magdeburg, Germany
E-mail: myra@iti.cs.uni-magdeburg.de

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

Bamshad Mobasher
DePaul University
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

Library of Congress Control Number: 2007941802

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 3-540-77484-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-77484-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12210404 06/3180 5 4 3 2 1 0

Preface

This book contains the postworkshop proceedings with selected revised papers from the 8th international workshop on knowledge discovery from the Web, WEBKDD 2006. The WEBKDD workshop series has taken 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 few years, the interest for the Web as a 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 been solved by now, 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 2006 workshop was "Knowledge Discovery on the Web", encompassing lessons learned over the past few years and new challenges for the years to come. While some of the infancy problems of Web analysis have been solved and proposed methodologies have reached maturity, the reality poses new challenges: The Web is evolving constantly; sites change and user preferences drift. 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.

The WebKDD 2006 workshop invited research results in all areas of Web mining and Semantic Web mining, with an emphasis on a seven years' update: What are the lessons learned on algorithms, semantics, data preparation, data integration and applications of the Web? How do new technologies, like adaptive mining methods, stream mining algorithms and techniques for the Grid, apply to Web mining? What new challenges are posed by new forms of data, especially flat texts, documents, pictures and streams, as well as the emergence of Web communities? How do we study the evolution of the Web and its effects on searching and browsing behavior? Which lessons have we learned about usability, e-commerce applications, personalization, recommendation engines, Web marketplaces, Web search, Web security, and misuse and abuse of the Web and its services? WebKDD 2006 attempted to address these challenging questions, with an emphasis on expanding the horizon of traditional Web mining to embrace and keep up with recent and emerging trends and emphasis on the Web domain, such as mining search engine queries, mining Web evolution, robustness of recommender systems, and mining blogs for sentiment analysis.

In the first paper, “Adaptive Web site Design using Caching Algorithms”, Justin Brickell, Inderjit S. Dhillon, and Dharmendra S. Modha present improved online algorithms for shortcut link selection that are based on a novel analogy drawn between shortcutting and caching. In the same way that cache algorithms predict which memory pages will be accessed in the future, the proposed algorithms predict which Web pages will be accessed in the future. These algorithms are efficient and can consider accesses over a long period of time, but give extra weight to recent accesses. Experiments show significant improvement in the utility of shortcut links selected by the proposed algorithm as compared to those selected by existing algorithms.

In the second paper,“Incorporating Usage Information into Average-Clicks Algorithm”, Kalyan Beemanapalli, Ramya Rangarajan, and Jaideep Srivastava present an extension to the Average-Clicks Algorithm, called “Usage Aware Average-Clicks,” where the static Web link structure graph is combined with the dynamic Usage Graph (built using the information available from the Web logs) to assign different weights to links on a Web page and hence capture the user’s intuition of distance between two Web pages more accurately. This method has been used as a new metric to calculate the page similarities in a recommendation engine to improve its predictive power.

In “Nearest-Biclusters Collaborative Filtering”, Panagiotis Symeonidis, Alexandros Nanopoulos, Apostolos Papadopoulos, and Yannis Manolopoulos use biclustering to disclose the duality between users and items in Nearest-neighbor Collaborative Filtering, by grouping them in both dimensions simultaneously. A novel nearest-biclusters algorithm is proposed, that uses a new similarity measure that achieves partial matching of users’ preferences. Performance evaluation results are offered, which show that the proposed method improves substantially the performance of the CF process.

In “Fast Categorization of Web Documents Represented by Graphs”, Alex Markov, Mark Last, and Abraham Kandel address the limitations of the vector-space model of information retrieval. This traditional model does not capture important structural information, such as the order and proximity of word occurrence, the location of a word within the document, or mark-up information. Three new hybrid approaches to Web document classification are presented, built upon both graph and vector space representations, thus preserving the benefits and discarding the limitations of each. The hybrid methods outperform, in most cases, vector-based models using two model-based classifiers (C4.5 decision-tree algorithm and probabilistic Naïve Bayes) on several benchmark Web document collections.

In “Leveraging Structural Knowledge for Hierarchically Informed Keyword Weight Propagation in the Web,” Jong Wook Kim and K. Selcuk Candan elaborate on indexing Web documents that have non-atomic structures, such as navigational/semantic hierarchies on the Web. A novel keyword and keyword weight propagation technique is proposed to properly enrich the data nodes in structured content. The approach first relies on understanding the context provided by the relative content relationships between entries in the structure, and then leveraging this information for relative-content preserving keyword propagation. Experiments show a significant improvement in precision with the proposed keyword propagation algorithm.

In the paper "How to Define Searching Sessions on Web Search Engines," Bernard J. Jansen, Amanda Spink, and Vinish Kathuria investigate three methods for defining a session on Web search engines. The authors examine 2,465,145 interactions from 534, 507 Web searchers, and compare defining sessions using: (1) Internet Protocol address and cookie; (2) Internet Protocol address, cookie, and a temporal limit on intra-session interactions; and (3) Internet Protocol address, cookie, and query reformulation patterns. Research results show that defining sessions by query reformulation along with Internet Protocol address and cookie, provides the best measure, resulting in an 82% increase in the number of sessions; while for all methods, mean session length was fewer than three queries and the mean session duration was less than 30 minutes. Implications are that unique sessions may be a better indicator than the common industry metric of unique visitors for measuring search traffic.

In the paper "Incorporating Concept Hierarchies into Usage Mining Based Recommendations," Amit Bose, Kalyan Beemanapalli, Jaideep Srivastava, and Sigal Sahar address the limitation of most recommendation models in their ability to use domain knowledge in the form of conceptual and structural characteristics of a Web site. Conceptual content organization can play an important role in the quality of recommendations, and forms the basis of resources like Google Directory, Yahoo Directory and Web-content management systems. The authors propose a novel technique to incorporate the conceptual characteristics of a Web site into a usage-based recommendation model. The authors use a framework based on biological sequence alignment. Similarity scores play a crucial role in such a construction, and a scoring system that is generated from the Web site's concept hierarchy is introduced. These scores fit seamlessly with other quantities used in similarity calculation like browsing order and time spent on a page. Additionally they demonstrate a simple, extensible system for assimilating more domain knowledge. Experimental results illustrate the benefits of using a concept hierarchy.

In the paper "A Random-Walk-Based Scoring Algorithm Applied to Recommender Engines," Augusto Pucci, Marco Gori, and Marco Maggini present "ItemRank," a random-walk-based scoring algorithm, which can be used to rank products according to expected user preferences, in order to recommend top-rank items to potentially interested users. The authors tested their algorithm on the MovieLens data set, which contains data collected from a popular recommender system on movies, and compared ItemRank with other state-of-the-art ranking techniques, showing that ItemRank performs better than the other techniques, while being less complex than other algorithms with respect to memory usage and computational cost. The paper also presents an analysis that helps to discover some intriguing properties of the MovieLens data set, that has been widely exploited as a benchmark for evaluating recently proposed approaches to recommender system.

In “Towards a Scalable k-NN CF Algorithm: Exploring Effective Applications of Clustering,” Al Mamunur Rashid, Shyong K. Lam, Adam LaPitz, George Karypis, and John Riedl address the need for specially designed CF algorithms that can gracefully cope with the vast size of the data representing customers and items in typical e-commerce systems. Many algorithms proposed thus far, where the principal concern is recommendation quality, may be too expensive to operate in a large-scale system. The authors propose ClustKNN, a simple and intuitive algorithm that is well suited for large data sets. The method first compresses data tremendously by building a straightforward but efficient clustering model. Recommendations are then generated quickly by using a simple Nearest Neighbor-based approach. The feasibility of ClustKNN is demonstrated both analytically and empirically, and a comparison with a number of other popular CF algorithms shows that, apart from being highly scalable and intuitive, ClustKNN provides very good recommendation accuracy.

In “Detecting Profile Injection Attacks in Collaborative Filtering: A Classification-Based Approach,” Chad Williams, Bamshad Mobasher, Robin Burke, and Runa Bhaumik address the vulnerability of Collaborative recommender systems to profile injection attacks. By injecting a larger number of biased profiles into a system, attackers can manipulate the predictions of targeted items. To decrease this risk, researchers have begun to study mechanisms for detecting and preventing profile injection attacks. In this paper, the authors extend their previous work that proposed several attributes for attack detection and for classification of attack profiles, through a more detailed analysis of the informativeness of these attributes as well as an evaluation of their impact at improving the robustness of recommender systems.

In “Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection”, Kathleen T. Durant and Michael D. Smith investigate data mining techniques that can automatically identify the political *sentiment* of Web log posts, and thus help bloggers categorize and filter this exploding information source. They illustrate the effectiveness of supervised learning for sentiment classification on Web log posts, showing that a Naïve Bayes classifier coupled with a forward feature selection technique can on average correctly predict a postings sentiment 89.77% of the time. It significantly outperforms Support Vector Machines at the 95% confidence level with a confidence interval of $[1.5, 2.7]$. The feature selection technique provides on average an 11.84% and a 12.18% increase for Naïve Bayes and Support Vector Machines results, respectively. Previous sentiment classification research achieved an 81% accuracy using Naïve Bayes and 82.9% using SVMs on a movie domain corpus.

In “Analysis of Web Search Engine Query Session and Clicked Documents,” David Nettleton, Liliana Calderón-Benavides, and Ricardo Baeza-Yates present the analysis of a Web search engine query log from two different perspectives: the query session and the clicked document. In the query session perspective, they process and analyze a Web search engine query and click data for the query session (query + clicked results) conducted by the user. They initially state some hypotheses for possible user types and quality profiles for the user session, based on descriptive variables of the session. In the clicked document perspective, they repeat the process from the perspective of the documents (URL's) selected. They also initially define possible document categories and select descriptive variables to define the documents. They apply a systematic data mining process to click data, contrasting non- supervised (Kohonen) and supervised (C4.5) methods to cluster and model the data, in order to identify profiles and rules which relate to theoretical user behavior and user session "quality," from the point of view of user session, and to identify document profiles which relate to theoretical user behavior, and document (URL) organization, from the document perspective.

In "Understanding Content Reuse on the Web: Static and Dynamic Analyses", Ricardo Baeza-Yates, Álvaro Pereira, and Nivio Ziviani present static and dynamic studies of duplicate and near-duplicate documents in the Web. The static and dynamic studies involve the analysis of similar content among pages within a given snapshot of the Web and how pages in an old snapshot are reused to compose new documents in a more recent snapshot. With experiments using four snapshots of the Chilean Web, they identify duplicates (in the static study) in both parts of the Web graph – reachable (connected by links) and unreachable components (unconnected) – aiming to identify where duplicates occur more frequently. They show that the number of duplicates in the Web seems to be much higher than previously reported (about 50% higher) and in their data the duplicated in the unreachable Web is 74.6% higher than the number of duplicates in the reachable component of the Web graph. In the dynamic study, they show that some of the old content is used to compose new pages. If a page in a newer snapshot has content of a page in an older snapshot, they consider that the source is a parent of the new page. They state the hypothesis that people use search engines to find pages and republish their content as a new document, and present evidence that this happens for part of the pages that have parents. In this case, part of the Web content is biased by the ranking function of search engines.

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 2006. 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): Corin Anderson, Ricardo A. Baeza-Yates, Bettina Berendt, Zheng Chen, Ed H. Chi, Brian D. Davison, Wei Fan, Fabio Grandi, Michael Hahsler, Xin Jin, Thorsten Joachims, George Karypis, Ravi Kumar, Vipin Kumar, Mark Last, Mark Levene, Ee-Peng Lim, Huan Liu, Stefano Lonardi, Alexandros D. Nanopoulos, Georgios Paliouras, Aniruddha G. Pant, Jian Pei, Ellen Spertus, Andrew Tomkins, and Mohammed

J. Zaki. O. Nasraoui gratefully acknowledges the support of the US National Science Foundation as part of NSF CAREER award IIS-0133948.

September 2007

Olfa Nasraoui
Myra Spiliopoulou
Jaideep Srivastava
Bamshad Mobasher
Brij Masand

Table of Contents