Lecture Notes in Artificial Intelligence 3932

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Bamshad Mobasher Olfa Nasraoui
Bing Liu Brij Masand

Advances in Web Mining and Web Usage Analysis

6th International Workshop
on Knowledge Discovery on the Web, WebKDD 2004
Seattle, WA, USA, August 22-25, 2004
Revised Selected Papers

Springer

Series Editors

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

Authors

Bamshad Mobasher
DePaul University
School of Computer Science
243 S. Wabash Ave., Chicago, IL 60604, USA
E-mail: mobasher@cs.depaul.edu

Olfa Nasraoui
University of Louisville
Computer Science and Computer Engineering
Louisville, KY 40292
E-mail: olfa.nasraoui@louisville.edu

Bing Liu
University of Illinois at Chicago
851 South Morgan Street, Chicago, IL 60607-7053, USA
E-mail: liub@cs.uic.edu

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

Library of Congress Control Number: 2006934110

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 3-540-47127-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-47127-1 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11899402 06/3142 5 4 3 2 1 0

Preface

The Web is a live environment that manages and drives a wide spectrum of applications in which a user may interact with a company, a governmental authority, a non-governmental organization or other non-profit institution or other users. User preferences and expectations, together with usage patterns, form the basis for personalized, user-friendly and business-optimal services. Key Web business metrics enabled by proper data capture and processing are essential to run an effective business or service. Enabling technologies include data mining, scalable warehousing and preprocessing, sequence discovery, real time processing, document classification, user modeling and quality evaluation models for them. Recipient technologies required for user profiling and usage patterns include recommendation systems, Web analytics applications, and application servers, coupled with content management systems and fraud detectors.

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. The development and application of Web mining techniques in the context of Web content, Web usage, and Web structure data has already resulted in dramatic improvements in a variety of Web applications, from search engines, Web agents, and content management systems, to Web analytics and personalization services. A focus on techniques and architectures for more effective integration and mining of content, usage, and structure data from different sources is likely to lead to the next generation of more useful and more intelligent applications.

This volume is comprised of revised and extended chapters based on WebKDD 2004, the sixth of a successful series of workshops on Knowledge Discovery on the Web. The WebKDD 2004 workshop was held in conjunction with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), August 22-25, 2004, in Seattle, Washington. As in the five preceding years, the WebKDD 2004 workshop continued its tradition of serving as a bridge between academia and industry by bringing together practitioners and researchers from both areas in order to foster the exchange of ideas and the dissemination of emerging solutions for intelligent Web-based applications using Web usage, structure and content mining. In addition to participants from academia, the workshop attracted many participants from industry including Microsoft, Amazon, AT&T, Google, SAS, DaimlerChrysler, Accenture, buy.com, shopping.com, and others.

WebKDD 2004 included a joint session (comprised of three papers) with a new workshop, the KDD Workshop on Mining for and from the Semantic Web (MSW04). The *Semantic Web Mining* portion of this book includes selected papers from both workshops.

About This Volume

Together the chapters span four complementary areas related to Web mining:

1. Web Usage Analysis and User Modeling
2. Web Personalization and Recommender Systems
3. Search Personalization
4. Semantic Web Mining

Web Usage Analysis and User Modeling. The first two chapters deal with Web Usage Analysis and User Modeling. In "Mining Temporally Changing Web Usage Graphs", Desikan and Srivastava address the limited "data-centric" point of view of most previous Web mining research by examining another dimension of Web mining, namely the temporal dimension. They highlight the significance of studying the evolving nature of Web graphs, and classify the approach to such problems at three levels of analysis: single node, sub-graphs and whole graphs. They provide a framework to approach problems in this kind of analysis and identify interesting problems at each level. In "Improving the Web Usage Analysis Process: A UML Model of the ETL Process", Maier addresses the laborious and time-consuming task of populating a data warehouse, to be used for sophisticated analysis of the Web channel in a multi-channel environment of an organization. To this end, a logical object-oriented relational data storage model is proposed, which simplifies modeling the ETL process and supports direct deployment within a WUSAN (Web USage ANalyis) system.

Web Personalization and Recommender Systems. Three chapters focus on Web Personalization and Recommender Systems. In "Mission-Based Navigational Behavior Modeling for Web Recommender Systems", Zaiane, Li, and Hayward advocate the use of additional information channels such as the content of visited pages and the connectivity between Web pages, as an alternative to using only one information channel, namely the Web access history. They propose the concept of "missions", which are identified by different channels, to help in better modeling users concurrent information needs. The combination of three channels is shown to improve the quality of the recommendations. In "Complete this Puzzle: A Connectionist Approach to Accurate Web Recommendations based on a Committee of Predictors", Nasraoui and Pavuluri present a Context Ultra-Sensitive Approach to personalization based on two-step Recommender systems (CUSA-2-step-Rec). The approach relies on a committee of profile-specific neural networks. Similar to the task of completing the missing pieces of a puzzle, each neural network is trained to predict the missing URLs of several complete ground-truth sessions from a given profile. The approach outperforms nearest profile and K-Nearest Neighbors based collaborative filtering. In "Collaborative Quality Filtering: Establishing Consensus or Recovering Ground Truth?", Traupman and Wilensky present an algorithm based on factor analysis for performing Collaborative Quality Filtering (CQF). Unlike previous

approaches to CQF, which estimate the consensus opinion of a group of reviewers, their algorithm uses a generative model of the review process to estimate the latent intrinsic quality of the items under review. The results of their tests suggest that asymptotic consensus, which purports to model peer review, is in fact not recovering the ground truth quality of reviewed items.

Search Personalization. Two chapters deal with Search Engine Personalization. In "Spying Out Accurate User Preferences for Search Engine Adaptation", Deng et al. propose a learning technique called "Spy Naive Bayes" (SpyNB) to identify the user preference pairs generated from clickthrough data. They then employ a ranking SVM algorithm to build a metasearch engine optimizer. Their empirical results on a metasearch engine prototype, comprising MSNSearch, Wisenut and Overture, show that, compared with no learning, SpyNB can significantly improve the average ranks of users' clicks. In "Using Hyperlink Features to Personalize Web Search", Aktas, Nacar, and Menczer introduce a methodology for personalizing PageRank vectors based on hyperlink features such as anchor terms or URL tokens. Users specify interest profiles as binary feature vectors where a feature corresponds to a DNS tree node. Then, given a profile vector, a weighted PageRank can be computed, assigning a weight to each URL based on the match between the URL and the profile features. Preliminary results show that Personalized PageRank performed favorably compared with pure similarity based ranking and traditional PageRank.

Semantic Web Mining. Four chapters in the book relate to Semantic Web Mining. In "Discovering Links Between Lexical and Surface Features in Questions and Answers", Chakrabarti presents a data-driven approach, assisted by machine learning, to build question answering information retrieval systems that return short passages or direct answers to questions, rather than URLs pointing to whole pages. Learning is based on a simple log-linear model over a pair of feature vectors, one derived from the question and the other derived from a candidate passage. Using this model, candidate passages are filtered, and substantial improvements are obtained in the mean rank at which the first answer is found. The model parameters reveal linguistic artifacts coupling questions and their answers, which can be used for better annotation and indexing. Meo et al., in "Integrating Web Conceptual Modeling and Web Usage Mining", present a case study regarding the application of the inductive database approach to the analysis of Web logs to enable the rapid customization of the mining procedures following the Web developers' needs. They integrate the user request information with meta-data concerning the Web site structure into rich XML Web logs, called "conceptual logs", produced by Web applications specified with the WebML conceptual model. Then, they apply a data mining language (MINE RULE) to conceptual logs in order to identify different types of patterns, such as recurrent navigation paths, page contents most frequently visited, and anomalies. Bloehdorn and Hotho, in "Boosting for Text Classification with Semantic Features", propose an enhancement of the classical term stem based document

representation through higher semantic concepts extracted from background knowledge. Boosting, a successful machine learning technique, is used for classification, and comparative experimental evaluations show consistent improvement of the results. In "Markov Blankets and Meta-Heuristics Search: Sentiment Extraction from Unstructured Texts", Airoldi, Bai, and Padman address the problem of extracting sentiments (positive versus negative comments) from unstructured text by proposing a two-stage Bayesian algorithm that can capture the dependencies among words, and find a vocabulary that is efficient for the purpose of extracting sentiments. Their work has potential to mine on-line opinions from the Internet and learn customers' preferences for economic or marketing research, or for leveraging a strategic advantage.

January 2006

Bamshad Mobasher
Olfa Nasraoui
Bing Liu
Brij Masand

Table of Contents

Web Usage Analysis and User Modeling

Web Personalization and Recommender Systems

Search Personalization

Semantic Web Mining