Lecture Notes in Artificial Intelligence 5439

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Haizheng Zhang Myra Spiliopoulou
Bamshad Mobasher C. Lee Giles
Andrew McCallum Olfa Nasraoui
Jaideep Srivastava John Yen (Eds.)

Advances in Web Mining and Web Usage Analysis

9th International Workshop
on Knowledge Discovery on the Web, WebKDD 2007
and 1st International Workshop
on Social Networks Analysis, SNA-KDD 2007
San Jose, CA, USA, August 12-15, 2007
Revised Papers

Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Haizheng Zhang
One Microsoft Way, Redmond, WA, USA - E-mail: hazhan@microsoft.com

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

Bamshad Mobasher
DePaul University, Chicago, IL, USA - E-mail: mobasher@cti.depaul.edu

C. Lee Giles
Pennsylvania State University, University Park, PA, USA - E-mail: giles@ist.psu.edu

Andrew McCallum
University of Massachusetts, Amherst, MA, USA - E-mail: mccallum@cs.umass.edu

Olfa Nasraoui
University of Louisville, Louisville, KY, USA - E-mail: olfa.nasraoui@louisville.edu

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

John Yen
Pennsylvania State University, Unversity Park, PA, USA
E-mail: jyen@ist.psu.edu

Library of Congress Control Number: 2009921448

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 3-642-00527-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-00527-5 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12624300 06/3180 5 4 3 2 1 0

Preface

This year's volume of Advances in Web Mining and Web Usage Analysis contains the postworkshop proceedings of a joint event, the 9th International Workshop on Knowledge Discovery from the Web (WEBKDD 2007) and the First SNA-KDD Workshop on Social Network Analysis (SNA-KDD 2007). The joint workshop on Web Mining and Social Network Analysis took place at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). It attracted 23 submissions, of which 14 were accepted for presentation at the workshop. Eight of them have been extended for inclusion in this volume.

WEBKDD is one of the most traditional workshops of the ACM SIGKDD international conference, under the auspices of which it has been organized since 1999. The strong interest for knowledge discovery in the Web, fostered not least by WEBKDD itself, has led to solutions for many problems in the Web's premature era. In the meanwhile, the Web has stepped into a new era, where it is experienced as a *social medium*, fostering interaction among people, enabling and promoting the sharing of knowledge, experiences and applications, characterized by group activities, community formation, and evolution. The design of Web 2.0 reflects the social character of the Web, bringing new potential and new challenges. The 9th WEBKDD was devoted to the challenges and opportunities of mining for the social Web and promptly gave rise to the joint event with the First Workshop on Social Network Analysis (SNA-KDD).

Social network research has advanced significantly in the last few years, strongly motivated by the prevalence of online social websites and a variety of large-scale offline social network systems. These social network systems are usually characterized by complex network structures and by rich contextual information. Researchers are interested in identifying common static topological properties of these networks, as well as the dynamics pertaining to formation and evolution. Social network analysis becomes necessary in an increasing number of application domains, including community discovery, recommendation systems, and information retrieval.

The objective of the joint WEBKDD/SNA-KDD 2007 workshop was to foster the study and interchange of ideas for the analysis and understanding of the social Web as the largest example of a social network.

Social networking on the Web is a phenomenon of scientific interest per se; there is demand for flexible and robust community discovery technologies, but also for interdisciplinary research on the rules and behavioral patterns that emerge and characterize community formation and evolution. The social flair of the Web poses new challenges and brings new opportunities for the individual. Among other things, the need for information now encompasses more than the traditional plain document search, as people started getting informed in blogs, as well as contributing with content, ratings, and recommendations to the

satisfaction of the information needs of others. Data miners are expected to deliver solutions for searching, personalizing, understanding, and protecting these social structures, bearing in mind their diversity and their scale.

The WEBKDD/SNA-KDD workshop invited research results on the emerging trends and industry needs associated with the traditional Web, the social Web, and other forms of social networking systems. This included data mining advances on the discovery and analysis of communities, on personalization for solitary activities (like search) and social activities (like discovery of potential friends), and on the analysis of user behavior in social structures (like blogs).

In the first paper *Spectral Clustering in Social Networks*, Miklós Kurucz, András A. Benczúr, Károly Csalogány, and László Lukács study large graphs of interconnected entities like phonecall networks and graphs of linked Web pages. They study the potential of spectral clustering for the identification of modular and homogeneous clusters and propose heuristics that alleviate shortcomings of the basis method and yield better results with respect to homogeneity and to the distribution of cluster sizes.

In the second paper *Looking for Great Ideas: Analyzing the Innovation Jam*, Wojciech Gryc, Mary Helander, Rick Lawrence, Yan Liu, Claudia Perlich, Chandan Reddy, and Saharon Rosset of IBM T.J. Watson Research Center report on methods for the analysis of the *Innovation Jam*. IBM introduced this online discussion forum in 2006, with the objective of providing a platform where new ideas were fostered and discussed among IBM employees and some external participants. The authors report on their findings about the activities and the social formations within this forum, and about their methods for analyzing the graph structure and the contributed content.

The third paper *Segmentation and Automated Social Hierarchy Detection through Email Network Analysis* by Germán Creamer, Ryan Rowe, Shlomo Hershkop, and Salvatore J. Stolfo studies the potential of data mining in corporate householding. The task is the identification of patterns of communication and the ranking of relationships among persons that communicate electronically, in particular through email. The authors have analyzed the Enron mailserver log and compared their findings with the human-crafted knowledge about the relationships of major players in that corporation.

The fourth paper *Mining Research Communities in Bibliographical Data* by Osmar R. Zaïane, Jiyang Chen, and Randy Goebel studies the implicit relationships among entities in a bibliographic database. Bibliographic data are of paramount importance for a research community, but the understanding of the underlying social structure is not straightforward. The authors have studied the DBLP database and designed the *DBConnect* tool. *DBConnect* uses random walks to identify interconnected nodes, derive relationships among the individuals/authors that correspond to these nodes, and even formulate recommendations about research cooperations among network members.

In the fifth paper *Dynamics of a Collaborative Rating System*, Kristina Lerman studies the Web as a participatory medium, in which users contribute, distribute, and evaluate information and she investigates collaborative decision

taking in the news aggregator platform Digg. Decision taking refers to the selection of the front-page stories featured regularly by Digg. This selection is based on the preferences of individual users, so the author investigates how a user influences other users and how this influence changes when a user contributes new content and obtains new friends.

In the sixth paper *Applying Link-Based Classification to Label Blogs*, Smitri Bhagat, Graham Cormode, and Irina Rozenbaum study the challenge of object labeling in blogs, thereby exploiting the links used by bloggers to connect related contents. They model this task as a graph labeling problem, for which they propose generic solutions. They then apply these solutions to the issue of blog labeling, whereby they are not only considering content but also the profiles of the bloggers themselves, attempting to assess the similarity of bloggers with respect to specific properties, such as age and gender, by studying the graph structure in which they participate.

In the seventh paper *Why We Twitter: An Analysis of a Microblogging Community*, Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng study the microblogging platform Twitter to understand the motives of users who choose microblogging for communication and information sharing. They identify four categories of microblogger intention, as well as different user roles within Twitter. They stress the differences between blogging and microblogging and compare the statistics of traffic in Twitter with those of blogs and other social networks.

In the last paper *A Recommender System Based on Local Random Walks and Spectral Methods*, Zeinab Abbassi and Vahab S. Mirrokni study interlinked blogs and propose a recommendation system for blogs that exploits this link structure. They observe the blogs as nodes of a social network, design a metric of similarity among them and devise also a *personalized* rank vector that expresses the relevance among nodes in the social network. They analyze the blog network, identify connected and strongly connected components and propose two algorithms that use this structure to formulate recommendations to a user.

August 2007

Haizheng Zhang
Myra Spiliopoulou
Bamshad Mobasher
C. Lee Giles
Andrew McCallum
Olfa Nasraoui
Jaideep Srivastava
John Yen

Organization

Workshop Chairs

Haizheng Zhang	Microsoft, USA
Myra Spiliopoulou	Otto von Guericke University Magdeburg, Germany
Lee Giles	Pennsylvania State University, USA
Andrew McCallum	University of Massachusetts, Amherst, USA
Bamshad Mobasher	DePaul University, USA
Olfa Nasraoui	University of Louisville, USA
Jaideep Srivastava	University of Minnesota, USA
John Yen	Pennsylvania State University, USA

Program Committee

Lada Adamic	University of Michigan
Sarabjot S. Anand	University of Warwick
Ricardo Baeza-Yates	Yahoo Research & Univ. Pompeu Fabra-Barcelona
Arindam Banerjee	University of Minnesota
Bettina Berendt	HU Berlin
Ed Chi	Xerox PARC
Tina Eliassi-Rad	Lawrence Livermore National Laboratory
Lise Getoor	University of Maryland
Joydeep Ghosh	University of Texas
Mark K. Goldberg	Rensselaer Polytechnic Institute
Andreas Hotho	University of Kassel
David Jensen	University of Massachusetts, Amherst
Ke Ke	Central Washington University
Ravi Kumar	Yahoo Research
Mark Last	Ben-Gurion University
Victor Lesser	University of Massachusetts, Amherst
Jure Leskovec	Carnegie Mellon University
Mark Levene	Birkbeck University of London
Ee-Peng Lim	Nanyang Tech. University, Singapore
Huan Liu	Arizona State University
Sanjay Kumar Madria	University of Missouri-Rolla
Ernestina Menasalvas	University Polytecnica Madrid, Spain
Dunja Mladenic	J. Stefan Institute, Slovenia
Alex Nanopoulos	Aristotle University, Greece
Seung-Taek Park	Yahoo! Research

Srinivasan Parthasarathy	Ohio State University
Jian Pei	Simon Fraser University, Canada
Xiaodan Song	NEC Labs America
Chris Volinsky	AT&T Labs-Research
Stefan Wrobel	Fraunhofer IAIS
Xifeng Yan	IBM Research
Mohammed Zaki	Rensselaer Polytechnic Institute
Alice Zheng	Carnegie Mellon University

Table of Contents