Lecture Notes in Artificial Intelligence 3209

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Bettina Berendt Andreas Hotho
Dunja Mladenic Maarten van Someren
Myra Spiliopoulou Gerd Stumme (Eds.)

Web Mining: From Web to Semantic Web

First European Web Mining Forum, EWMF 2003
Cavtat-Dubrovnik, Croatia, September 22, 2003
Invited and Selected Revised Papers

Springer

Series Editors

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

Volume Editors

Bettina Berendt
Humboldt University Berlin, Institute of Information Systems
E-mail: berendt@wiwi.hu-berlin.de

Andreas Hotho
University of Kassel, Department of Mathematics and Informatics
E-mail: hotho@cs.uni-kassel.de

Dunja Mladenic
J. Stefan Institute, Ljubljana, Slovenia
and Carnegie Mellon University, Pittsburgh, USA
E-mail: Dunja.Mladenic@ijs.si

Maarten van Someren
University of Amsterdam, Department of Social Science Informatics
E-mail: maarten@swi.psy.uva.nl

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

Gerd Stumme
University of Kassel, Department of Mathematics and Computer Science
E-mail: stumme@cs.uni-kassel.de

Library of Congress Control Number: 2004112647

CR Subject Classification (1998): I.2, H.2.8, H.3, H.4, H.5.2-4, K.4

ISSN 0302-9743
ISBN 3-540-23258-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 11321798 06/3142 5 4 3 2 1 0

Preface

In the last years, research on Web mining has reached maturity and has broadened in scope. Two different but interrelated research threads have emerged, based on the dual nature of the Web:

- The Web is a practically infinite collection of documents: The acquisition and exploitation of information from these documents asks for intelligent techniques for information categorization, extraction and search, as well as for adaptivity to the interests and background of the organization or person that looks for information.
- The Web is a venue for doing business electronically: It is a venue for interaction, information acquisition and service exploitation used by public authorities, non-governmental organizations, communities of interest and private persons. When observed as a venue for the achievement of business goals, a Web presence should be aligned to the objectives of its owner and the requirements of its users. This raises the demand for understanding Web usage, combining it with other sources of knowledge inside an organization, and deriving lines of action.

The birth of the Semantic Web at the beginning of the decade led to a coercion of the two threads in two aspects: (i) the extraction of semantics from the Web to build the Semantic Web; and (ii) the exploitation of these semantics to better support information acquisition and to enhance the interaction for business and non-business purposes. Semantic Web mining encompasses both aspects from the viewpoint of knowledge discovery.

The *Web Mining Forum* initiative is motivated by the insight that knowledge discovery on the Web from the viewpoint of hyperarchive analysis and from the viewpoint of interaction among persons and institutions are complementary, both for the familiar, conventional Web and for the Semantic Web. The Web Mining Forum was launched in September 2002 as an initiative of the KDNet Network of Excellence [1]. It encompasses an information portal and discussion forum for researchers who specialize in data mining on data *from* and on data *about* the Web/Semantic Web and its usage. In its function as an information portal, it focusses on the announcement of events associated with knowledge discovery and the Web, on the collection of datasets for the evaluation of Web mining algorithms and on the specification of a common terminology. In its function as a discussion forum, it initiated the "European Web Mining Forum" Workshop (EWMF 2003) during the ECML/PKDD conference in Cavtat, Croatia.

EWMF 2003 was the follow-up workshop of the Semantic Web Mining workshop that took place during ECML/PKDD 2002, and also built upon the tradition of the WEBKDD workshop series that has taken place during the ACM SIGKDD conference since 1999.

The EWMF 2003 workshop hosted eight regular papers and two invited talks, by Sarabjot Sing Anand (University of Ulster) and by Rayid Ghani (Accenture). The presentations were organized into four sessions followed by a plenary discussion. Following the well-accepted tradition of the WEBKDD series, a postworkshop proceedings volume was prepared. It consists of extended versions of six of the papers and is further extended

[1] Funded by the EU 5th Framework Programme under grant IST-2001-33086

by four invited papers and a roadmap describing our vision of the future of Semantic Web mining.

The role of semantic information in improving personalized recommendations is discussed by Mobasher et al. in [7]: They elaborate on collaborative filtering and stress the importance of item-based recommendations in dealing with scalability and sparsity problems. Semantic information on the items, extracted with the help of domain-specific ontologies, is combined with user-item mappings and serves as basis for the formulation of recommendations, thus increasing prediction accuracy and demonstrating robustness over sparse data. Approaches for the extraction of semantic information appear in [4, 6,9]. Rayiv Ghani elaborates on the extraction of semantics features from product descriptions with text mining techniques, with the goal of enriching the (Web) transaction data [4]. The method has been implemented in a system for personalized product recommendations but is also appropriate for further applications like store profiling and demand forecasting. Mladenic and Grobelnik discuss the automated mapping of Web pages onto an ontology with the help of document classification techniques [6]. They focus on skewed distributions and propose a solution on the basis of multiple independent classifiers that predict the probability with which a document belongs to each class. Sigletos et al. study the extraction of information from multiple Web sites and the disambiguation of extracted facts [9] by combining the induction of wrappers and the discovery of named entities.

Personalization through recommendation mechanisms is the subject of several contributions. While the emphasis of [7] is on individual users, [8] elaborates on user communities. In the paper of Pierrakos et al., community models are built on the basis of usage data and of a concept hierarchy derived through content-based clustering of the documents in the collection [8]. The induction of user models is also studied by Esposito et al. in [3]: The emphasis of their work is on the evaluation of two user profiling methods in terms of classification accuracy and performance. Evaluation is also addressed by van Someren et al., who concentrate on recommendation strategies [10]: They observe that current systems optimize the quality of single recommendations and argue that this strategy is suboptimal with respect to the ultimate goal of finding the desired information in a minimal number of steps.

Evaluation from the viewpoint of deploying Web mining results is studied by Anand et al. in [1]. They elaborate on modelling and measuring the effectiveness of the interaction between business venues and the visitors of their Web sites and propose the development of scenaria, on the basis of which effectiveness should be evaluated. Architectures for the knowledge discovery, evaluation *and* deployment are described in [1] and [5]. While Anand et al. focus on scenario-based deployment [1], Menasalvas et al. stress the existence of multiple viewpoints and goals of deployment and propose a method for assessing the value of a session for each viewpoint [5]. Finally, the paper of Baron and Spiliopoulou elaborates on one of the effects of deployment, the change in the patterns derived during knowledge discovery [2]: The authors model patterns as temporal objects and propose a method for the detection of changes in the statistics of association rules over a Web-server log.

Acknowledgments

This volume owes much to the engagement of many scientists. The editors are indebted to the PC members of the EWMF 2003 workshop

Ed Chi (Xerox Parc, Palo Alto, USA)
Ronen Feldman (Bar-Ilan University, Ramat Gan, Israel)
Marko Grobelnik (J. Stefan Institute, Ljubljana, Slovenia)
Oliver Günther (Humboldt University Berlin, Germany)
Stefan Haustein (Universität Dortmund, Germany)
Jörg-Uwe Kietz (kdlabs AG, Zuerich, Switzerland)
Ee-Peng Lim (Nanyang Technological University, Singapore)
Alexander Maedche (Robert Bosch GmbH, Stuttgart, Germany)
Brij Masand (Data Miners, Boston, USA)
Yannis Manolopoulos (Aristotle University, Greece)
Ryszard S. Michalski (George Mason University, USA)
Bamshad Mobasher (DePaul University, Chicago, USA)
Claire Nedellec (Université Paris Sud, France)
George Paliouras (National Centre for Scientific Research "Demokritos", Athens, Greece)
Jian Pei (Simon Fraser University, Canada)
John R. Punin (Rensselaer Polytechnic Institute, Troy, NY, USA)
Jaideep Srivastava (University of Minnesota, Minneapolis, USA)
Rudi Studer (Universität Karlsruhe, Germany)
Stefan Wrobel (Fraunhofer Institute for Autonomous Intelligent Systems, Sankt Augustin, Germany)
Mohammed Zaki (Rensellaer Polytechnic Institute, USA)
Osmar Zaiane (University of Alberta, Edmonton, Canada)

and the reviewers of the papers in this volume

Philipp Cimiano (Universität Karlsruhe, Germany)
Marko Grobelnik (J. Stefan Institute, Ljubljana, Slovenia)
Dimitrios Katsaros (Aristotle University, Greece)
Jörg-Uwe Kietz (kdlabs AG, Zuerich, Switzerland)
Ee-Peng Lim (Nanyang Technological University, Singapore)
Zehua Liu (Nanyang Technological University, Singapore)
Brij Masand (Data Miners, Boston, USA)
Yannis Manolopoulos (Aristotle University, Greece)
Bamshad Mobasher (DePaul University, Chicago, USA)
Alexandros Nanopoulos (Aristotle University, Greece)
George Paliouras (National Centre for Scientific Research "Demokritos", Athens, Greece)
Ljiljana Stojanovic (FZI Forschungszentrum Informatik, Germany)
Rudi Studer (Universität Karlsruhe, Germany)
Stefan Wrobel (Fraunhofer AIS and Univ. of Bonn, Germany)
Mohammed Zaki (Rensselaer Polytechnic Institute, USA)
Osmar Zaiane (University of Alberta, Edmonton, Canada)

for the involvement and effort they contributed to guarantee a high scientific niveau for both the workshop and the follow-up proceedings.

We would like to thank the organizers of ECML/PKDD 2003 for their support in the organization of the EWMF 2003 workshop. Last but foremost, we are indebted to the KDNet network of excellence for the funding of the Web Mining Forum and for the financial support of the EWMF 2003 workshop, and especially to Ina Lauth from the Fraunhofer Institute for Autonomous Intelligent Systems (AIS), the KDNet project coordinator, for her intensive engagement and support in the establishment of the Web Mining Forum and in the organization of the EWMF 2003.

The EWMF workshop chairs

Bettina Berendt, Humboldt Universität zu Berlin (Germany)
Andreas Hotho, Universität Kassel (Germany)
Dunja Mladenic, J. Stefan Institute (Slovenia)
Maarten van Someren, University of Amsterdam (The Netherlands)
Myra Spiliopoulou, Otto-von-Guericke-Universität Magdeburg (Germany)
Gerd Stumme, Universität Kassel (Germany)

References

1. S.S. Anand, M. Mulvenna, and K. Chevalier. On the deployment of Web usage mining. (invited paper)
2. S. Baron and M. Spiliopoulou. Monitoring the evolution of Web usage patterns.
3. F. Esposito, G. Semeraro, S. Ferilli, M. Degemmis, N. Di Mauro, T. Basile, and P. Lops. Evaluation and validation of two approaches to user profiling.
4. R. Ghani. Mining the Web to add semantics to retail data mining. (invited paper)
5. E. Menasalvas, S. Millán, M. Pérez, E. Hochsztain, V. Robles, O. Marbán, A. Tasistro, and J. Peña. An approach to estimate user sessions value dealing with multiple viewpoints and goals.
6. D. Mladenić and M. Grobelnik. Mapping documents onto a Web page ontology. (invited paper)
7. B. Mobasher, X. Jin, and Y. Zhou. Semantically enhanced collaborative filtering on the Web. (invited paper)
8. D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, and M. Dikaiakos. Web community directories: A new approach to Web personalization.
9. G. Sigletos, G. Paliouras, C.D. Spyropoulos, and M. Hatzopoulos. Mining Web sites using wrapper induction, named-entities and post-processing.
10. M. van Someren, V. Hollink, and S. ten Hagen. Greedy recommending is not always optimal.

Table of Contents