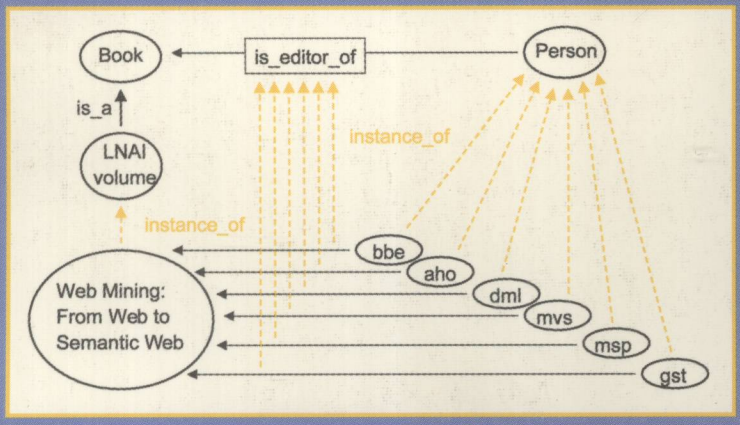
Bettina Berendt Andreas Hotho
Dunja Mladenic Maarten van Someren
Myra Spiliopoulou Gerd Stumme (Eds.)

LNAI 3209 Hot Topics

Web Mining: From Web to Semantic Web

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



Springer

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

Lecture Notes in Artificial Intelligence 3209

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

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.

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004.

Vol. 3245: E. Suzuki, S. Arikawa (Eds.), Discovery Science. XIV, 430 pages. 2004.

Vol. 3244: S. Ben-David, J. Case, A. Maruoka (Eds.), Algorithmic Learning Theory. XIV, 505 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004.

Vol. 3215: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVII, 906 pages. 2004.

Vol. 3214: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1302 pages. 2004.

Vol. 3213: M.G. Negoita, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems. LVIII, 1280 pages. 2004.

Vol. 3209: B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou, G. Stumme (Eds.), Web Mining: From Web to Semantic Web. IX, 201 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004.

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004.

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004.

Vol. 3171: A.L.C. Bazzan, S. Labidi (Eds.), Advances in Artificial Intelligence – SBIA 2004. XVII, 548 pages. 2004.

Vol. 3159: U. Visser, Intelligent Information Integration for the Semantic Web. XIV, 150 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004.

Vol. 3071: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World. XIII, 409 pages. 2004.

Vol. 3070: L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Artificial Intelligence and Soft Computing - ICAISC 2004. XXV, 1208 pages. 2004.

Vol. 3068: E. André, L. Dybkjær, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems. XII, 324 pages. 2004.

Vol. 3067: M. Dastani, J. Dix, A. El Fallah-Seghrouchni (Eds.), Programming Multi-Agent Systems. X, 221 pages. 2004.

Vol. 3066: S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymała-Busse (Eds.), Rough Sets and Current Trends in Computing. XX, 853 pages. 2004.

Vol. 3065: A. Lomuscio, D. Nute (Eds.), Deontic Logic in Computer Science. X, 275 pages. 2004.

Vol. 3060: A.Y. Tawfik, S.D. Goodwin (Eds.), Advances in Artificial Intelligence. XIII, 582 pages. 2004.

Vol. 3056: H. Dai, R. Srikant, C. Zhang (Eds.), Advances in Knowledge Discovery and Data Mining. XIX, 713 pages. 2004.

Vol. 3055: H. Christiansen, M.-S. Hacid, T. Andreasen, H.L. Larsen (Eds.), Flexible Query Answering Systems. X, 500 pages. 2004.

Vol. 3040: R. Conejo, M. Urretavizcaya, J.-L. Pérez-de-la-Cruz (Eds.), Current Topics in Artificial Intelligence. XIV, 689 pages. 2004.

Vol. 3035: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XII, 326 pages. 2004.

Vol. 3034: J. Favela, E. Menasalvas, E. Chávez (Eds.), Advances in Web Intelligence. XIII, 227 pages. 2004.

Vol. 3030: P. Giorgini, B. Henderson-Sellers, M. Winikoff (Eds.), Agent-Oriented Information Systems. XIV, 207 pages. 2004.

Vol. 3029: B. Orchard, C. Yang, M. Ali (Eds.), Innovations in Applied Artificial Intelligence. XXI, 1272 pages. 2004.

Vol. 3025: G.A. Vouros, T. Panayiotopoulos (Eds.), Methods and Applications of Artificial Intelligence. XV, 546 pages. 2004.

Vol. 3020: D. Polani, B. Browning, A. Bonarini, K. Yoshida (Eds.), RoboCup 2003: Robot Soccer World Cup VII. XVI, 767 pages. 2004.

Vol. 3012: K. Kurumatani, S.-H. Chen, A. Ohuchi (Eds.), Multi-Agnets for Mass User Support. X, 217 pages. 2004.

Vol. 3010: K.R. Apt, F. Fages, F. Rossi, P. Szeredi, J. Váncza (Eds.), Recent Advances in Constraints. VIII, 285 pages. 2004.

Vol. 2990: J. Leite, A. Omicini, L. Sterling, P. Torroni (Eds.), Declarative Agent Languages and Technologies. XII, 281 pages. 2004.

Vol. 2980: A. Blackwell, K. Marriott, A. Shimojima (Eds.), Diagrammatic Representation and Inference. XV, 448 pages. 2004.

Vol. 2977: G. Di Marzo Serugendo, A. Karageorgos, O.F. Rana, F. Zambonelli (Eds.), Engineering Self-Organising Systems. X, 299 pages. 2004.

Vol. 2972: R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, H. Sossa (Eds.), MICAI 2004: Advances in Artificial Intelligence. XVII, 923 pages. 2004.

Vol. 2969: M. Nickles, M. Rovatsos, G. Weiss (Eds.), Agents and Computational Autonomy. X, 275 pages. 2004.

Vol. 2961: P. Eklund (Ed.), Concept Lattices. IX, 411 pages. 2004.

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. X, 301 pages. 2004.

Vol. 2930: F. Winkler (Ed.), Automated Deduction in Geometry. VII, 231 pages. 2004.

Vol. 2926: L. van Elst, V. Dignum, A. Abecker (Eds.), Agent-Mediated Knowledge Management. XI, 428 pages. 2004.

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004.

Vol. 2915: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction. XIII, 558 pages. 2004.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. XX, 512 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. XVI, 1075 pages. 2003.

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. XV, 504 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003.

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. X, 215 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003.

Vol. 2871: N. Zhong, Z.W. Raś, S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems. XV, 697 pages. 2003.

Vol. 2854: J. Hoffmann, Utilizing Problem Structure in Planing. XIII, 251 pages. 2003.

Vol. 2843: G. Grieser, Y. Tanaka, A. Yamamoto (Eds.), Discovery Science. XII, 504 pages. 2003.

Vol. 2842: R. Gavaldá, K.P. Jantke, E. Takimoto (Eds.), Algorithmic Learning Theory. XI, 313 pages. 2003.

Vol. 2838: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Knowledge Discovery in Databases: PKDD 2003. XVI, 508 pages. 2003.

Vol. 2837: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Machine Learning: ECML 2003. XVI, 504 pages. 2003.

Vol. 2835: T. Horváth, A. Yamamoto (Eds.), Inductive Logic Programming. X, 401 pages. 2003.

Vol. 2821: A. Günter, R. Kruse, B. Neumann (Eds.), KI 2003: Advances in Artificial Intelligence. XII, 662 pages. 2003.

Vol. 2807: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. XIII, 426 pages. 2003.

Vol. 2801: W. Banzhaf, J. Ziegler, T. Christaller, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life. XVI, 905 pages. 2003.

Vol. 2797: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. XII, 281 pages. 2003.

Vol. 2792: T. Rist, R.S. Aylett, D. Ballin, J. Rickel (Eds.), Intelligent Virtual Agents. XV, 364 pages. 2003.

Vol. 2782: M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (Eds.), Cooperative Information Agents VII. XI, 345 pages. 2003.

Vol. 2780: M. Dojat, E. Keravnou, P. Barahona (Eds.), Artificial Intelligence in Medicine. XIII, 388 pages. 2003.

Vol. 2777: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. XIV, 746 pages. 2003.

Vol. 2752: G.A. Kaminka, P.U. Lima, R. Rojas (Eds.), RoboCup 2002: Robot Soccer World Cup VI. XVI, 498 pages. 2003.

Vol. 2741: F. Baader (Ed.), Automated Deduction – CADE-19. XII, 503 pages. 2003.

Vol. 2705: S. Renals, G. Grefenstette (Eds.), Text- and Speech-Triggered Information Access. VII, 197 pages. 2003.

Vol. 2703: O.R. Zaïane, J. Srivastava, M. Spiliopoulou, B. Masand (Eds.), WEBKDD 2002 - MiningWeb Data for Discovering Usage Patterns and Profiles. IX, 181 pages. 2003.

Vol. 2700: M.T. Pazienza (Ed.), Extraction in the Web Era. XIII, 163 pages. 2003.

Vol. 2699: M.G. Hinchey, J.L. Rash, W.F. Truszkowski, C.A. Rouff, D.F. Gordon-Spears (Eds.), Formal Approaches to Agent-Based Systems. IX, 297 pages. 2002.

Table of Contents

A Roadmap for Web Mining: From Web to Semantic Web

Bettina Berendt[1], Andreas Hotho[2], Dunja Mladenic[3], Maarten van Someren[4], Myra Spiliopoulou[5], and Gerd Stumme[2]

[1] Institute of Information Systems, Humboldt University Berlin, Germany. berendt@wiwi.hu-berlin.de
[2] Chair of Knowledge & Data Engineering, University of Kassel, Germany, {hotho,stumme}@cs.uni-kassel.de
[3] Jozef Stefan Institute, Ljubljana, Slovenia, Dunja.Mladenic@ijs.si
[4] Social Science Informatics, University of Amsterdam, The Netherlands, maarten@swi.psy.uva.nl
[5] Institute of Technical and Business Information Systems, Otto–von–Guericke–University Magdeburg, Germany, myra@iti.cs.uni-magdeburg.de

1 Introduction

The purpose of Web mining is to develop methods and systems for discovering models of objects and processes on the World Wide Web and for web-based systems that show adaptive performance. Web Mining integrates three parent areas: Data Mining (we use this term here also for the closely related areas of Machine Learning and Knowledge Discovery), Internet technology and World Wide Web, and for the more recent Semantic Web. The World Wide Web has made an enormous amount of information electronically accessible. The use of email, news and markup languages like HTML allow users to publish and read documents at a world-wide scale and to communicate via chat connections, including information in the form of images and voice records. The HTTP protocol that enables access to documents over the network via Web browsers created an immense improvement in communication and access to information. For some years these possibilities were used mostly in the scientific world but recent years have seen an immense growth in popularity, supported by the wide availability of computers and broadband communication. The use of the internet for other tasks than finding information and direct communication is increasing, as can be seen from the interest in "e-activities" such as e-commerce, e-learning, e-government, e-science.

Independently of the development of the Internet, Data Mining expanded out of the academic world into industry. Methods and their potential became known outside the academic world and commercial toolkits became available that allowed applications at an industrial scale. Numerous industrial applications have shown that models can be constructed from data for a wide variety of industrial problems (e.g. [1,2]).

The World-Wide Web is an interesting area for Data Mining because huge amounts of information are available. Data Mining methods can be used to analyse the behaviour of individual users, access patterns of pages or sites, properties of collections of documents. Almost all standard data mining methods are designed for data that are organised as multiple "cases" that are comparable and can be viewed as instances of a single pattern,

B. Berendt et al. (Eds.): EWMF 2003, LNAI 3209, pp. 1–22, 2004.

for example patients described by a fixed set of symptoms and diseases, applicants for loans, customers of a shop. A "case" is typically described by a fixed set of features (or variables). Data on the Web have a different nature. They are not so easily comparable and have the form of free text, semi-structured text (lists, tables) often with images and hyperlinks, or server logs. The aim to learn models of documents has given rise to the interest in Text Mining [3]: methods for modelling documents in terms of properties of documents. Learning from the hyperlink structure has given rise to graph-based methods, and server logs are used to learn about user behavior.

The Semantic Web is a recent initiative, inspired by Tim Berners-Lee [4], to take the World-Wide Web much further and develop in into a distributed system for knowledge representation and computing. The aim of the Semantic Web is to not only support access to information "on the Web" by direct links or by search engines but also to support its *use*. Instead of searching for a document that matches keywords, it should be possible to combine information to answer questions. Instead of retrieving a plan for a trip to Hawaii, it should be possible to automatically construct a travel plan that satisfies certain goals and uses opportunities that arise dynamically. This gives rise to a wide range of challenges. Some of them concern the infrastructure, including the interoperability of systems and the languages for the exchange of information rather than data. Many challenges are in the are of knowledge representation, discovery and engineering. They include the extraction of knowledge from data and its representation in a form understandable by arbitrary parties, the intelligent questioning and the delivery of answers to problems as opposed to conventional queries and the exploitation of formerly extracted knowledge in this process. The ambition of representing content in a way that can be understood and consumed by an arbitrary reader leads to issues in which cognitive sciences and even philosophy are involved, such as the understanding of an asset's intended meaning.

The Semantic Web proposes several additional innovative ideas to achieve this:

Standardised format. The Semantic Web proposes standards for uniform metalevel description language for representation formats. Besides acting as a basis for exchange, this language supports representation of knowledge at multiple levels. For example, text can be *annotated* with a formal representation of it. The natural language sentence "Amsterdam is the capital of the Netherlands", for instance, can be annotated such that the annotation formalises knowledge that is implicit in the sentence, e.g. Amsterdam can be annotated as "city", Netherlands as "country" and the sentence with the structured "capital-of(Amsterdam, Netherlands)". Annotating textual documents (and also images and possibly audio and video) thus enables a combination of textual and formal representations of knowledge. A small step further is to store the annotated text items in a structured database or knowledge base.

Standardised vocabulary and knowledge. The Semantic Web encourages and facilitates the formulation of shared vocabularies and shared knowledge in the form of ontologies: if knowledge about university courses is to be represented and shared, it is useful to define and use a common vocabulary and common basic knowledge. The Semantic Web aims to collect this in the form of ontologies and make them available

for modelling new domains and activities. This means that a large amount of knowledge will be structured, formalised and represented to enable automated access and use.

Shared services. To realise the full Semantic Web, beside static structures also "Web services" are foreseen. Services mediate between requests and applications and make it possible to automatically invoke applications that run on different systems.

In this chapter, we concentrate on one thread of challenges associated with the Semantic Web, those that can be addressed with knowledge discovery techniques, putting the emphasis on the transition from Web Mining to mining the Semantic Web and on the role of ontologies and information extraction for this transition. Section 2 summarises the more technical aspects of the Semantic Web, in particular the main representation languages, section 3 summarises basic concepts from Data Mining, section 4 reviews the main developments in the application of Data Mining to the World Wide Web, section 5 extends this to the combination of Data Mining and the Semantic Web and section 6 reviews developments that are expected in the near future and issues for research and development. Each section has the character of a summary and includes references to more detailed discussions and explanations. This chapter summarises and extends [5], [6] and [7].

2 Languages for the Semantic Web

The Semantic Web requires a language in which information can be represented. This language should support (a) knowledge representation and reasoning (including information retrieval but ultimately a wide variety of tasks), (b) the description of document content, (c) the exchange of the documents and the incorporated knowledge and (d) standardisation. The first two aspects demand adequate expressiveness. The last two aspects emphasise that the Semantic Web, like the Web, should be a medium for the exchange of a wide variety of objects and thus allow for ease-of-use and for agreed-upon protocols. Naturally enough, the starting point for describing the Semantic Web has been XML. However, XML has not been designed with the intention to express or exchange knowledge. In this section, we review three W3C initiatives, XML, RDF(S) and OWL and their potential for the Semantic Web.

2.1 XML

XML (Extensible mark-up language) was designed as a language for mark-up or annotation of documents. An XML object is a labeled tree and consists of objects with attributes and values that can themselves be XML objects. Beside annotation for formatting, XML allows the definition of any kind of annotation, thus opening the way to annotation with ontologies and to use as data model for arbitrary information. This makes it extensible, unlike its ancestors like HTML.

XML Schema allows the definition of grammars for valid XML documents, and the reference to "name spaces", sets of labels that can be accessed via the internet. XML can also be used as a scheme for structured databases. The value of an attribute can be

text but it can also be an element of a limited set or a number. XML is only an abstract data format.

Furthermore, XML does not include any procedural component. Tools have been developed for search and retrieval in XML trees. Tools can create formatted output from formatting annotations but in general any type of operation is possible. When tools are integrated in the Web and can be called from outside they are called "services". This creates a very flexible representation format that can be used to represent information that is partially structured.

Details about XML can be found in many books, reports and Web pages. In the context of the Semantic Web, the most important role for XML is that it provides a simple standard abstract data model that can be used to access both (annotated) documents and structured data (for example tables) and that it can be used as a representation for ontologies. However, XML and XML schema were designed to describe the structure of text documents, like HTML, Word, StarOffice, or LaTeXdocuments. It is possible to define tags in XML to carry meta data but these tags may not have a well-defined meaning. XML helps organizing documents by providing a formal syntax for annotation. Erdmann [8] provides a detailed analysis of the capabilities of XML, the shortcomings of XML concerning semantics and possible solutions. For Web Mining the standardisation created by XML simplifies the development of generic systems that learn from data on the web.

2.2 RDF(S)

The *Resource Description Framework (RDF)* is, according to the W3C recommendation [9], "a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web."

RDF documents consist of three types of entities: resources, properties, and statements. Resources may be Web pages, parts or collections of Web pages, or any (real-world) objects which are not directly part of the World-Wide Web. In RDF, resources are always addressed by URIs, Universal Resource Identifiers, a generalisation of URLs that includes services besides locations. Properties are specific attributes, characteristics, or relations describing resources. A resource together with a property having a value for that resource form an RDF statement. A value is either a literal, a resource, or another statement. Statements can thus be considered as object–attribute–value triples.

The data model underlying RDF is basically a directed labeled graph. RDF Schema defines a simple modeling language on top of RDF which includes classes, is-a relationships between classes and between properties, and domain/range restrictions for properties. XML provides the standard syntax for RDF and RDF Schema.

Summarising, RDF and RDF Schema provide base support for the specification of semantics *and* use the widespread XML as syntax. However, the expressiveness is limited, disallowing the specification of facts that one is bound to expect, given the long tradition of database schema theory. They include the notion of key, as in relational databases, as well as factual assertions, e.g. stating that each print of this book can be either hardcover or softcover but not both. The demand for supporting more expressive semantics and reasoning is addressed in languages like DAML, OIL and the W3C recommendation OWL described below. More information on RDF(S) can be found on

the W3C website (`www.w3.org`) and many books. As with XML, the standardisation provided by RDF(S) simplifies development and application of Web Mining.

2.3 OWL

Like RDF and RDF Schema, OWL is a W3C recommendation, intended to support more elaborate semantics. OWL includes elements from description logics and provides many constructs for the specification of semantics, including conjunction and disjunction, existentially and universally quantified variables and property inversion. Using these constructs, a reasoning module can make logical inferences and derive knowledge that was previously only implicit in the data. Using OWL for the Semantic Web implies that an application could invoke such a reasoning module and acquire inferred knowledge rather than simply retrieve data.

However, the expressiveness of OWL comes at a high cost. First, OWL contains constructs that make it undecidable. Second, reasoning is not efficient. Third, the expressiveness is achieved by increased complexity, so that ease-of-use and intuitiveness are no more given. These observations lead to two variations of OWL, *OWL DL* (stands for OWL Description Logic) and *OWL Lite*, which disallow the constructs that make the original *OWL Full* undecidable and at the same time aim for more efficient reasoning and for higher ease-of-use. To this end, OWL DL is more expressive than OWL Lite, while OWL Lite is even more restricted but easier to understand and to implement.

In terms of standardisation, it should be recalled that RDF and RDF Schema use XML as their syntax. OWL Full is upward compatible with RDF. This desirable aspect does not hold for OWL DL and OWL Lite. A legal OWL DL document is also a legal RDF document but not vice versa. This implies that reasoning and the targeted knowledge extraction are limited to the set of documents supporting OWL DL (resp. OWL Lite), while other documents, even if RDF Schema, cannot be taken into account in the reasoning process. For the transition of the Web to the Semantic Web, this is a more serious caveat than for other environments (e.g. institutional information sources) which need ontological support. More information on OWL can be found on the W3C website and many books.

The development of OWL and its application is still in an early stage. If it leads to the availability of large knowledge bases via the internet, this will increase the relevance of knowledge-intensive Data Mining methods, that combine data with prior (OWL) knowledge.

2.4 Ontologies

Beside the formal languages to be used for the Semantic Web there is the ambition to develop ontologies for general use. There are in practice two types of ontologies. The first type uses a small number of relations between concepts, usually the subclass relation and sometimes the part-of relation. Popular and commonly used are ontologies of Web documents, such as DMoz or Yahoo!, where the documents are hierarchically organized based on the content. For each content topic, there is an ontology node, with more general topics placed higher in the hierarchy. For instance, one of the top level topics in DMoz is "Computers" that has as one of the subtopics "Data Formats. Under it, there is a subtopic

"Markup Languages" that has "XML" as one of its subtopics. There are several hundred documents assigned to the node on "XML" or some of its subnodes.[1] Each Web document is very briefly described and this description together with the hyperlink to the document is placed into one or more ontology nodes. For instance, one item in the "XML" node is a hyperlink to W3C page on XML, http://www.w3.org/XML/, with the associated brief description: "Extensible Markup Language (XML) - Main page for World Wide Web Consortium (W3C) XML activity and information". We can say that here each concept (topic in this case) in the ontology is described by a set of Web documents and their corresponding short descriptions with hyperlinks. The only kind of relations that appear in such ontologies are implicit relations between more specific topic, that is a "subtopic of" a more general topic while the more general topic is a "supertopic of" a more specific topic.

The other kind of ontologies are rich with relations but have a rather limited description of concepts consisting usually of a few words. A well known example of a general, manually constructed ontology is the semantic network WordNet [10] with 26 different relations (e.g., hypernym, synonym). For instance, concepts such as "bird" and "animal" are connected with the relation "is a kind of", concepts "bird" and "wing" are connected with the relation "has part".

3 Data Mining

Before considering what the Semantic Web means with respect to Data Mining, we briefly review the main tasks that are studied in Data Mining. Data Mining methods construct models of data. These models can be used for prediction or explanation of observations or for adaptive behaviour. Reviews of the main methods can be found in textbooks such as [11,12,13]. The main tasks are classification, rule discovery, event prediction and clustering.

3.1 Classification

Classification methods construct models that assign a class to a new object on the basis of its description. A wide range of models can be constructed. In this context an important property of classification methods is the form in which objects are given to the data miner and the form of the models. Most learning methods take as input object descriptions in the form of attribute-value pairs where the scales of the variables are nominal or numerical. One class of methods, relational learning or Inductive Logic Programming, see for example [14], takes input in the form of relational structures that describe multiple objects with relations between them creating general models over structures.

Classification methods vary in the type of model that they construct. Decision tree learners construct models basically in the form of rules. A condition in a rule is a constraint on the value of a variable. Usually constraints have the form of identity (e.g. colour = red) or an interval on a scale (age > 50). The consequent of a rule is a class. Decision trees have a variable at each node and a partitioning of the values of this

[1] See http://dmoz.org/Computers/Data_Formats/Markup_Languages/XML/