Lecture Notes in Artificial Intelligence 4289

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Markus Ackermann Bettina Berendt
Marko Grobelnik Andreas Hotho
Dunja Mladenič Giovanni Semeraro
Myra Spiliopoulou Gerd Stumme
Vojtěch Svátek Maarten van Someren (Eds.)

Semantics, Web and Mining

Joint International Workshops, EWMF 2005 and KDO 2005
Porto, Portugal, October 3 and 7, 2005
Revised Selected Papers

Springer

Volume Editors

Markus Ackermann
University of Leipzig, E-mail: markus.ackermann@rz.uni-leipzig.de

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

Marko Grobelnik
J. Stefan Institute, Ljubljana, E-mail: marko.grobelnik@ijs.si

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

Dunja Mladenič
J. Stefan Institute, Ljubljana, E-mail: dunja.mladenic@ijs.si

Giovanni Semeraro
University of Bari, E-mail: semeraro@di.uniba.it

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

Gerd Stumme
University of Kassel, E-mail: stumme@cs.uni-kassel.de

Vojtěch Svátek
University of Economics, Prague, E-mail: svatek@vse.cz

Maarten van Someren
University of Amsterdam, E-mail: maarten@science.uva.nl

Library of Congress Control Number: 2006936937

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

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-47697-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-47697-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11908678 06/3142 5 4 3 2 1 0

Preface

Finding knowledge – or meaning – in data is the goal of every knowledge discovery effort. Subsequent goals and questions regarding this knowledge differ among knowledge discovery (KD) projects and approaches. One central question is whether and to what extent the meaning extracted from the data is expressed in a formal way that allows not only humans but also machines to understand and re-use it, i.e., whether the semantics are formal semantics. Conversely, the input to KD processes differs between KD projects and approaches. One central question is whether the background knowledge, business understanding, etc. that the analyst employs to improve the results of KD is a set of natural-language statements, a theory in a formal language, or somewhere in between. Also, the data that are being mined can be more or less structured and/or accompanied by formal semantics.

These questions must be asked in every KD effort. Nowhere may they be more pertinent, however, than in KD from Web data ("Web mining"). This is due especially to the vast amounts and heterogeneity of data and background knowledge available for Web mining (content, link structure, and usage), and to the re-use of background knowledge and KD results over the Web as a global knowledge repository and activity space. In addition, the (Semantic) Web can serve as a publishing space for the results of knowledge discovery from other resources, especially if the whole process is underpinned by common ontologies.

We have explored this close connection in a series of workshops at the European Conference on Machine Learning / Principles and Practice of Knowledge Discovery from Databases (ECML/PKDD) conference series (Semantic Web Mining, 2001, 2002) and in the selection of papers for the post-proceedings of the European Web Mining Forum 2003 Workshop (published as the Springer LNCS volume Web Mining: From Web to Semantic Web in 2004). We have also investigated the uses of ontologies (as the most commonly used type of formal semantics) in KD in the Knowledge Discovery and Ontologies workshop in 2004.

In 2005, we organized, in two partly overlapping teams and again at ECML/-PKDD, a workshop on Web mining (European Web Mining Forum) and a workshop on Knowledge Discovery and Ontologies. The submissions, and in particular the highest-quality accepted contributions, convinced us that the specific importance of semantics for Web mining continues to hold. We therefore decided to prepare a joint publication of the best papers from the two workshops that presented a variety of ways in which semantics can be understood and brought to bear on Web data. In addition, we included a particularly fitting contribution from KDO 2004, by Vanzin and Becker. The result of our selection, the reviewers' comments, and the authors' revision and extension of their workshop papers is this book.

Paper summaries

To emphasize the common themes, we will give a combined summary of the contributions in this volume. To make it easier to understand the papers in the organizational context for which they were written and in which they were discussed, we have ordered them by workshop in the table of contents.

Understanding the Web and supporting its users was addressed in the papers of both workshops: KDO 2005 and EWMF 2005. The invited contribution of Eirinaki, Mavroeidis, Tsatsaronis, and Vazirgiannis elaborates on the role of semantics for Web personalization. Degemmis, Lops, and Semeraro concentrate on learning user profiles with help of a rich taxonomy of terms, WordNet. The subject of building ontologies and taxonomies is pursued in the papers of Bast, Dupret, Majumdar, and Piwowarski and of Fortuna, Mladenič, and Grobelnik. The former proposes a mechanism that extracts a term taxonomy from Web documents using Principal Component Analysis. Fortuna et al. present OntoGen, a tool implementing an approach to semi-automatic topic ontology construction that uses Latent Semantic Indexing and K-means clustering to discover topics from document collections, while a support vector machine is used to support the user in naming the constructed ontology concepts.

The subject of evaluating the performance of such semi-automatic ontology enhancement tools for topic discovery is studied by Spiliopoulou, Schaal, Müller, and Brunzel. Topic discovery in the Web with semantic networks is also the subject of the contribution by Kiefer, Stein, and Schlieder, who concentrate on the visibility of topics. The incorporation of semantics into the mining process is studied in the work of Svátek, Rauch, and Ralbovský on ontology-enhanced association mining, while Vanzin and Becker elaborate on the role of ontologies in interpreting Web usage patterns.

The retrieval of information from the Web is another topic that was studied in both workshops. Baeza-Yates and Poblete examine the mining of user queries made in a Web site, while Stein and Hess consider information retrieval in trust-enhanced document networks. Information retrieval from the Web is the subject of the webTopic approach proposed by Escudeiro and Jorge, who concentrate on persistent information needs that require the regular retrieval of documents on specific topics. Document classification is a further powerful means towards the same objective. The classification of Web documents is addressed by Utard and Fürnkranz, who focus on the information in hyperlinks and in the texts around them.

August 2006

EWMF and KDO 2005
Workshop Chairs

Organization

EWMF 2005 and KDO 2005 were organized as part of the 16th European Conference on Machine Learning (ECML) and the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD).

EWMF Workshop Chairs

Bettina Berendt	Institute of Information Systems Humboldt University Berlin, Germany
Andreas Hotho	Knowledge and Data Engineering Group University of Kassel, Germany
Dunja Mladenič	J. Stefan Institute Ljubljana, Slovenia
Giovanni Semeraro	Department of Informatics University of Bari, Italy
Myra Spiliopoulou	Faculty of Computer Science Otto-von-Guericke-Univ. Magdeburg, Germany
Gerd Stumme	Knowledge and Data Engineering Group University of Kassel, Germany
Maarten van Someren	Informatics Institute University of Amsterdam, Netherlands

EWMF Program Committee

Sarabjot Singh Anand	University of Warwick, UK
Mathias Bauer	DFKI, Germany
Stephan Bloehdorn	University of Karlsruhe, Germany
Janez Brank	J. Stefan Institute, Slovenia
Marko Grobelnik	J. Stefan Institute, Slovenia
Haym Hirsh	Rutgers University, USA
Ernestina Menasalvas	Universidad Politecnica de Madrid, Spain
Bamshad Mobasher	DePaul University, USA
Ion Muslea	Language Weaver, Inc., USA
Michael J. Pazzani	University of California, Irvine, USA
Lars Schmidt-Thieme	University of Freiburg, Germany
Steffen Staab	University of Koblenz-Landau, Germany

EWMF Additional Reviewers

P. Basile (University of Bari, Italy)
M. Degemmis (University of Bari, Italy)
P. Lops (University of Bari, Italy)

EWMF Sponsoring Institutions

EU Network of Excellence PASCAL
Pattern Analysis, Statistical Modelling, and Computational Learning

KDO Workshop Chairs

Markus Ackermann	Dept. of Natural Language Processing, Institute for Computer Science University of Leipzig, Germany
Bettina Berendt	Institute of Information Systems Humboldt University Berlin, Germany
Marko Grobelnik	J. Stefan Institute Ljubljana, Slovenia
Vojtěch Svátek	University of Economics Prague, Czech Republic

KDO Program Committee

Nathalie Assenac-Gilles	IRIT, Toulouse, France
Chris Biemann	University of Leipzig, Germany
Philipp Cimiano	AIFB, University of Karlsruhe, Germany
Martine Collard	University of Nice, France
Andreas Hotho	University of Kassel, Germany
François Jacquenet	University of Saint-Etienne, France
Alípio Jorge	University of Porto, Portugal
Nada Lavrač	Jožef Stefan Institute, Ljubljana, Slovenia
Bernardo Magnini	ITC-IRST, Trento, Italy
Bamshad Mobasher	DePaul University, USA
Gerhard Paaß	Fraunhofer AIS, St. Augustin, Germany
John Punin	Oracle Corporation, USA
Massimo Ruffolo	ICAR-CNR and EXEURA, Italy
Michael Sintek	DFKI, Kaiserslautern, Germany

Table of Contents

EWMF Papers

KDO Papers on KDD for Ontology

KDO Papers on Ontology for KDD