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

LNAI 4289

Semantics, Web and Mining

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

Springer

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

Lecture Notes in Artificial Intelligence 4289

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

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

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 4304: A. Sattar, B.-H. Kang (Eds.), AI 2006: Advances in Artificial Intelligence. XXVII, 1303 pages. 2006.

Vol. 4293: A. Gelbukh, C.A. Reyes-Garcia (Eds.), MICAI 2006: Advances in Artificial Intelligence. XXVIII, 1232 pages. 2006.

Vol. 4289: M. Ackermann, B. Berendt, M. Grobelnik, A. Hotho, D. Mladenič, G. Semeraro, M. Spiliopoulou, G. Stumme, V. Svátek, M. van Someren (Eds.), Semantics, Web and Mining. X, 197 pages. 2006.

Vol. 4285: Y. Matsumoto, R. Sproat, K.-F. Wong, M. Zhang (Eds.), Computer Processing of Oriental Languages. XXXX, 55000 pages. 2006.

Vol. 4274: Q. Huo, B. Ma, E.-S. Chng, H. Li (Eds.), Chinese Spoken Language Processing. XXXX, 8000 pages. 2006.

Vol. 4265: L. Todorovski, N. Lavrač, K.P. Jantke (Eds.), Discovery Science. XIV, 384 pages. 2006.

Vol. 4264: J.L. Balcázar, P.M. Long, F. Stephan (Eds.), Algorithmic Learning Theory. XIII, 393 pages. 2006.

Vol. 4259: S. Greco, Y. Hata, S. Hirano, M. Inuiguchi, S. Miyamoto, H.S. Nguyen, R. Słowiński (Eds.), Rough Sets and Current Trends in Computing. XXII, 951 pages. 2006.

Vol. 4253: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part III. XXXII, 1301 pages. 2006.

Vol. 4252: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part II. XXXIII, 1335 pages. 2006.

Vol. 4251: B. Gabrys, R.J. Howlett, L.C. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems, Part I. LXVI, 1297 pages. 2006.

Vol. 4248: S. Staab, V. Svátek (Eds.), Managing Knowledge in a World of Networks. XIV, 400 pages. 2006.

Vol. 4246: M. Hermann, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. XIII, 588 pages. 2006.

Vol. 4223: L. Wang, L. Jiao, G. Shi, X. Li, J. Liu (Eds.), Fuzzy Systems and Knowledge Discovery. XXVIII, 1335 pages. 2006.

Vol. 4213: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Knowledge Discovery in Databases: PKDD 2006. XXII, 660 pages. 2006.

Vol. 4212: J. Fürnkranz, T. Scheffer, M. Spiliopoulou (Eds.), Machine Learning: ECML 2006. XXIII, 851 pages. 2006.

Vol. 4211: P. Vogt, Y. Sugita, E. Tuci, C.L. Nehaniv (Eds.), Symbol Grounding and Beyond. VIII, 237 pages. 2006.

Vol. 4203: F. Esposito, Z.W. Raś, D. Malerba, G. Semeraro (Eds.), Foundations of Intelligent Systems. XVIII, 767 pages. 2006.

Vol. 4201: Y. Sakakibara, S. Kobayashi, K. Sato, T. Nishino, E. Tomita (Eds.), Grammatical Inference: Algorithms and Applications. XII, 359 pages. 2006.

Vol. 4200: I.F.C. Smith (Ed.), Intelligent Computing in Engineering and Architecture. XIII, 692 pages. 2006.

Vol. 4198: O. Nasraoui, O. Zaïane, M. Spiliopoulou, B. Mobasher, B. Masand, P.S. Yu (Eds.), Advances in Web Mining and Web Usage Analysis. IX, 177 pages. 2006.

Vol. 4196: K. Fischer, I.J. Timm, E. André, N. Zhong (Eds.), Multiagent System Technologies. X, 185 pages. 2006.

Vol. 4188: P. Sojka, I. Kopeček, K. Pala (Eds.), Text, Speech and Dialogue. XV, 721 pages. 2006.

Vol. 4183: J. Euzenat, J. Domingue (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 291 pages. 2006.

Vol. 4180: M. Kohlhase, OMDoc – An Open Markup Format for Mathematical Documents [version 1.2]. XIX, 428 pages. 2006.

Vol. 4177: R. Marín, E. Onaindía, A. Bugarín, J. Santos (Eds.), Current Topics in Artificial Intelligence. XV, 482 pages. 2006.

Vol. 4160: M. Fisher, W. van der Hoek, B. Konev, A. Lisitsa (Eds.), Logics in Artificial Intelligence. XII, 516 pages. 2006.

Vol. 4155: O. Stock, M. Schaerf (Eds.), Reasoning, Action and Interaction in AI Theories and Systems. XVIII, 343 pages. 2006.

Vol. 4149: M. Klusch, M. Rovatsos, T.R. Payne (Eds.), Cooperative Information Agents X. XII, 477 pages. 2006.

Vol. 4140: J.S. Sichman, H. Coelho, S.O. Rezende (Eds.), Advances in Artificial Intelligence - IBERAMIA-SBIA 2006. XXIII, 635 pages. 2006.

Vol. 4139: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (Eds.), Advances in Natural Language Processing. XVI, 771 pages. 2006.

Vol. 4133: J. Gratch, M. Young, R. Aylett, D. Ballin, P. Olivier (Eds.), Intelligent Virtual Agents. XIV, 472 pages. 2006.

Vol. 4130: U. Furbach, N. Shankar (Eds.), Automated Reasoning. XV, 680 pages. 2006.

Vol. 4120: J. Calmet, T. Ida, D. Wang (Eds.), Artificial Intelligence and Symbolic Computation. XIII, 269 pages. 2006.

Vol. 4114: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence, Part II. XXVII, 1337 pages. 2006.

Vol. 4108: J.M. Borwein, W.M. Farmer (Eds.), Mathematical Knowledge Management. VIII, 295 pages. 2006.

Vol. 4106: T.R. Roth-Berghofer, M.H. Göker, H.A. Güvenir (Eds.), Advances in Case-Based Reasoning. XIV, 566 pages. 2006.

Vol. 4099: Q. Yang, G. Webb (Eds.), PRICAI 2006: Trends in Artificial Intelligence. XXVIII, 1263 pages. 2006.

Vol. 4095: S. Nolfi, G. Baldassarre, R. Calabretta, J.C.T. Hallam, D. Marocco, J.-A. Meyer, O. Miglino, D. Parisi (Eds.), From Animals to Animats 9. XV, 869 pages. 2006.

Vol. 4093: X. Li, O.R. Zaïane, Z. Li (Eds.), Advanced Data Mining and Applications. XXI, 1110 pages. 2006.

Vol. 4092: J. Lang, F. Lin, J. Wang (Eds.), Knowledge Science, Engineering and Management. XV, 664 pages. 2006.

Vol. 4088: Z.-Z. Shi, R. Sadananda (Eds.), Agent Computing and Multi-Agent Systems. XVII, 827 pages. 2006.

Vol. 4087: F. Schwenker, S. Marinai (Eds.), Artificial Neural Networks in Pattern Recognition. IX, 299 pages. 2006.

Vol. 4068: H. Schärfe, P. Hitzler, P. Øhrstrøm (Eds.), Conceptual Structures: Inspiration and Application. XI, 455 pages. 2006.

Vol. 4065: P. Perner (Ed.), Advances in Data Mining. XI, 592 pages. 2006.

Vol. 4062: G. Wang, J.F. Peters, A. Skowron, Y. Yao (Eds.), Rough Sets and Knowledge Technology. XX, 810 pages. 2006.

Vol. 4049: S. Parsons, N. Maudet, P. Moraitis, I. Rahwan (Eds.), Argumentation in Multi-Agent Systems. XIV, 313 pages. 2006.

Vol. 4048: L. Goble, J.-J.C.. Meyer (Eds.), Deontic Logic and Artificial Normative Systems. X, 273 pages. 2006.

Vol. 4045: D. Barker-Plummer, R. Cox, N. Swoboda (Eds.), Diagrammatic Representation and Inference. XII, 301 pages. 2006.

Vol. 4031: M. Ali, R. Dapoigny (Eds.), Advances in Applied Artificial Intelligence. XXIII, 1353 pages. 2006.

Vol. 4029: L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada (Eds.), Artificial Intelligence and Soft Computing – ICAISC 2006. XXI, 1235 pages. 2006.

Vol. 4027: H.L. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreasen, H. Christiansen (Eds.), Flexible Query Answering Systems. XVIII, 714 pages. 2006.

Vol. 4021: E. André, L. Dybkjær, W. Minker, H. Neumann, M. Weber (Eds.), Perception and Interactive Technologies. XI, 217 pages. 2006.

Vol. 4020: A. Bredenfeld, A. Jacoff, I. Noda, Y. Takahashi (Eds.), RoboCup 2005: Robot Soccer World Cup IX. XVII, 727 pages. 2006.

Vol. 4013: L. Lamontagne, M. Marchand (Eds.), Advances in Artificial Intelligence. XIII, 564 pages. 2006.

Vol. 4012: T. Washio, A. Sakurai, K. Nakajima, H. Takeda, S. Tojo, M. Yokoo (Eds.), New Frontiers in Artificial Intelligence. XIII, 484 pages. 2006.

Vol. 4008: J.C. Augusto, C.D. Nugent (Eds.), Designing Smart Homes. XI, 183 pages. 2006.

Vol. 4005: G. Lugosi, H.U. Simon (Eds.), Learning Theory. XI, 656 pages. 2006.

Vol. 3978: B. Hnich, M. Carlsson, F. Fages, F. Rossi (Eds.), Recent Advances in Constraints. VIII, 179 pages. 2006.

Vol. 3963: O. Dikenelli, M.-P. Gleizes, A. Ricci (Eds.), Engineering Societies in the Agents World VI. XII, 303 pages. 2006.

Vol. 3960: R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, M.C. Dias (Eds.), Computational Processing of the Portuguese Language. XII, 274 pages. 2006.

Vol. 3955: G. Antoniou, G. Potamias, C. Spyropoulos, D. Plexousakis (Eds.), Advances in Artificial Intelligence. XVII, 611 pages. 2006.

Vol. 3949: F.A. Savacı (Ed.), Artificial Intelligence and Neural Networks. IX, 227 pages. 2006.

Vol. 3946: T.R. Roth-Berghofer, S. Schulz, D.B. Leake (Eds.), Modeling and Retrieval of Context. XI, 149 pages. 2006.

Vol. 3944: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (Eds.), Machine Learning Challenges. XIII, 462 pages. 2006.

Vol. 3937: H. La Poutré, N.M. Sadeh, S. Janson (Eds.), Agent-Mediated Electronic Commerce. X, 227 pages. 2006.

Vol. 3932: B. Mobasher, O. Nasraoui, B. Liu, B. Masand (Eds.), Advances in Web Mining and Web Usage Analysis. X, 189 pages. 2006.

Vol. 3930: D.S. Yeung, Z.-Q. Liu, X.-Z. Wang, H. Yan (Eds.), Advances in Machine Learning and Cybernetics. XXI, 1110 pages. 2006.

Vol. 3918: W.-K. Ng, M. Kitsuregawa, J. Li, K. Chang (Eds.), Advances in Knowledge Discovery and Data Mining. XXIV, 879 pages. 2006.

Vol. 3913: O. Boissier, J. Padget, V. Dignum, G. Lindemann, E. Matson, S. Ossowski, J.S. Sichman, J. Vázquez-Salceda (Eds.), Coordination, Organizations, Institutions, and Norms in Multi-Agent Systems. XII, 259 pages. 2006.

Vol. 3910: S.A. Brueckner, G.D.M. Serugendo, D. Hales, F. Zambonelli (Eds.), Engineering Self-Organising Systems. XII, 245 pages. 2006.

Vol. 3904: M. Baldoni, U. Endriss, A. Omicini, P. Torroni (Eds.), Declarative Agent Languages and Technologies III. XII, 245 pages. 2006.

Vol. 3900: F. Toni, P. Torroni (Eds.), Computational Logic in Multi-Agent Systems. XVII, 427 pages. 2006.

Vol. 3899: S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search. XIV, 216 pages. 2006.

Vol. 3898: K. Tuyls, P.J. 't Hoen, K. Verbeeck, S. Sen (Eds.), Learning and Adaption in Multi-Agent Systems. X, 217 pages. 2006.

Table of Contents

EWMF Papers

KDO Papers on KDD for Ontology

KDO Papers on Ontology for KDD

A Website Mining Model Centered on User Queries

Ricardo Baeza-Yates[1,2,3] and Barbara Poblete[1,2]

[1] Web Research Group, Technology Department,
University Pompeu Fabra, Barcelona, Spain
[2] Center for Web Research, CS Department
University of Chile, Santiago, Chile
[3] Yahoo! Research, Barcelona, Spain
{ricardo.baeza, barbara.poblete}@upf.edu

Abstract. We present a model for mining user queries found within the access logs of a website and for relating this information to the website's overall usage, structure and content. The aim of this model is to discover, in a simple way, valuable information to improve the quality of the website, allowing the website to become more intuitive and adequate for the needs of its users. This model presents a methodology of analysis and classification of the different types of queries registered in the usage logs of a website, such as queries submitted by users to the site's internal search engine and queries on global search engines that lead to documents in the website. These queries provide useful information about topics that interest users visiting the website and the navigation patterns associated to these queries indicate whether or not the documents in the site satisfied the user's needs at that moment.

1 Introduction

The Web has been characterized by its rapid growth, massive usage and its ability to facilitate business transactions. This has created an increasing interest for improving and optimizing websites to fit better the needs of their visitors. It is more important than ever for a website to be found easily in the Web and for visitors to reach effortlessly the contents they are looking for. Failing to meet these goals can result in the loss of many potential clients.

Web servers register important data about the usage of a website. This information generally includes visitors navigational behavior, the queries made to the website's internal search engine (if one is available) and also the queries on external search engines that resulted in requests of documents from the website, queries that account for a large portion of the visits of most sites on the Web. All of this information is provided by visitors implicitly and can hold the key to significantly optimize and enhance a website, thus improving the "quality" of that site, understood as *"the conformance of the website's structure to the intuition of each group of visitors accessing the site"* [1].

Most of the queries related to a website represent actual information needs of the users that visit the site. However, user queries in Web mining have been

M. Ackermann et al. (Eds.): EWMF/KDO 2005, LNAI 4289, pp. 1–17, 2006.

studied mainly with the purpose of enhancing website search, and not with the intention of discovering new data to increase the quality of the website's contents and structure. For this reason in this paper we present a novel model that mines queries found in the usage logs of a website, classifying them into different categories based in navigational information. These categories differ according to their importance for discovering new and interesting information about ways to improve the site. Our model also generates a visualization of the site's content distribution in relation to the link organization between documents, as well as the URLs selected due to queries. This model was mostly designed for websites that register traffic from internal and/or external search engines, even if this is not the main mechanism of navigation in the site. The output of the model consists of several reports from which improvements can be made to the website.

The main contributions of our model for improving a website are: *to mine user queries within a website's usage logs*, *obtain new interesting contents to broaden the current coverage of certain topics in the site*, *suggest changes or additions to words in the hyperlink descriptions*, and at a smaller scale *suggest to add new links between related documents and revise links between unrelated documents in a site.*

We have implemented this model and applied it to different types of websites, ranging from small to large, and in all cases the model helps to point out ways to improve the site, even if this site does not have an internal search engine. We have found our model specially useful on large sites, in which the contents have become hard to manage for the site's administrator.

This paper is organized as follows. Section 2 presents related work and section 3 our model. Section 4 gives an overview of our evaluation and results. The last section presents our conclusions and future work.

2 Related Work

Web mining [2] is the process of discovering patterns and relations in Web data. Web mining generally has been divided into three main areas: *content mining*, *structure mining* and *usage mining*. Each one of these areas are associated mostly, but not exclusively, to these three predominant types of data found in a website:

Content: The "real" data that the website was designed to give to its users. In general this data consists mainly of text and images.

Structure: This data describes the organization of the content within the website. This includes the organization inside a Web page, internal and external links and the site hierarchy.

Usage: This data describes the use of the website, reflected in the Web server's access logs, as well as in logs for specific applications.

Web usage mining has generated a great amount of commercial interest [3,4]. The analysis of Web server logs has proven to be valuable in discovering many

issues, such as: if a document has never been visited it may have no reason to exist, or on the contrary, if a very popular document cannot be found from the top levels of a website, this might suggest a need for reorganization of its link structure.

There is an extensive list of previous work using Web mining for improving websites, most of which focuses on supporting adaptive websites [5] and automatic personalization based on Web Mining [6]. Amongst other things, using analysis of frequent navigational patterns and association rules, based on the pages visited by users, to find interesting rules and patterns in a website [1,7,8,9,10]. Other research targets mainly modeling of user sessions, profiles and cluster analysis [11,12,13,14,15].

Queries submitted to search engines are a valuable tool for improving websites and search engines. Most of the work in this area has been directed at using queries to enhance website search [16] and to make more effective global Web search engines [17,18,19,20]. In particular, in [21] chains (or sequences) of queries with similar information needs are studied to learn ranked retrieval functions for improving Web search. Queries can also be studied to improve the quality of a website. Previous work on this subject include [22] which proposed a method for analyzing similar queries on Web search engines, the idea is to find new queries that are similar to ones that directed traffic to a website and later use this information to improve the website. Another kind of analysis based on queries, is presented in [23] and consists of studying queries submitted to a site's internal search engine, and indicates that valuable information can be discovered by analyzing the behavior of users in the website after submitting a query. This is the starting point of our work.

3 Model Description

In this section we will present the description of our model for mining website usage, content and structure, centered on queries. This model performs different mining tasks, using as input the website's access logs, its structure and the content of its pages. These tasks also includes data cleaning, session identification, merging logs from several applications and removal of robots amongst other things which we will not discuss in depth at this moment, for more details please refer to [24,25,26]. The following concepts are important to define before presenting our model:

Session: A session is a sequence of document accesses registered for one user in the website's usage logs within a maximum time interval between each request. This interval is set by default to 30 minutes, but can be changed to any other value considered appropriate for a website [24]. Each user is identified uniquely by the `IP` and `User-Agent`.

Queries: A query consists of a set of one or more keywords that are submitted to a search engine and represents an information need of the user generating that query.

Information Scent: IS [27] indicates how well a word, or a set of words, describe a certain concept in relation to other words with the same semantics. For example, polysemic words (words with more than one meaning) have less IS due to their ambiguity.

In our model the structure of the website is obtained from the links between documents and the content is the text extracted from each document. The aim of this model is to generate information that will allow to improve the structure and contents of a website, and also to evaluate the interconnections amongst documents with similar content.

For each query that is submitted to a search engine, a page with results is generated. This page has links to documents that the search engine considers appropriate for the query. By reviewing the brief abstract of each document displayed (which allows the user to decide roughly if a document is a good match for his or her query) the user can choose to visit zero or more documents from the results page. Our model analyzes two different types of queries, that can be found in a website's access registries. These queries are:

External queries: These are queries submitted on Web search engines, from which users selected and visited documents in a particular website. They can be discovered from the log's `referer` field.

Internal queries: These are queries submitted to a website's internal search box. Additionally, external queries that are specified by users for a particular site, will be considered as internal queries for that site. For example, Google.com queries that include `site:example.com` are internal queries for the website `example.com`. In this case we can have queries without clicked results.

Figure 1 (left) shows the description of the model, which gathers information about internal and external queries, navigational patterns and links in the website to discover IS that can be used to improve the site's contents. Also the link and content data from the website is analyzed using clustering of similar documents and connected components. These procedures will be explained in more detail in the following subsections.

3.1 Navigational Model

By analyzing the navigational behaviors of users within a website, during a period of time, the model can classify documents into different types, such as: *documents reached without a search*, *documents reached from internal queries* and *documents reached from external queries*. We define these types of documents as follows:

Documents reached Without a Search (DWS): These are documents that, throughout the course of a session, were reached by browsing and without the interference of a search (in a search engine internal or external to the website). In other words, documents reached from the results page of a search

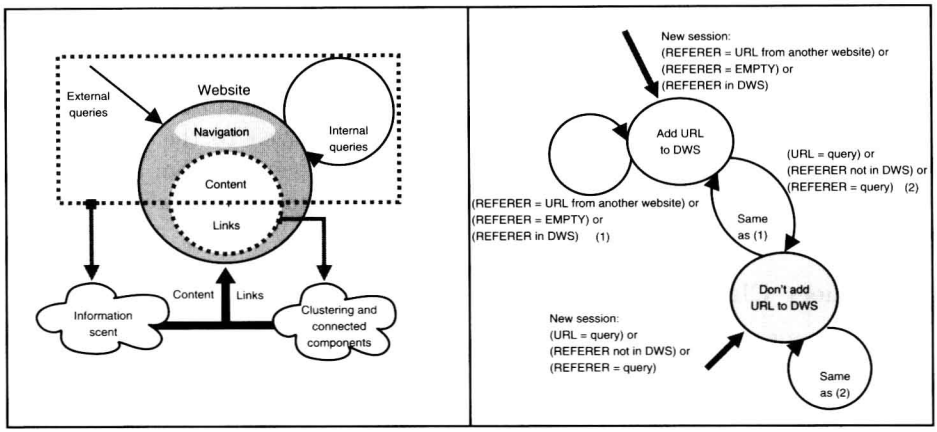

Fig. 1. Model description *(left)* and heuristic for DWS *(right)*

engine and documents attained from those results, are *not* considered in this category. Any document reached from documents visited previously to the use of a search engine will be considered in this category.

Documents reached from Internal Queries (DQ_i): These are documents that, throughout the course of a session, were reached by the user as a direct result of an *internal query*.

Documents reached from External Queries (DQ_e): These are documents that, throughout the course of a session, were reached by the user as a direct result of an *external query*.

For future references we will drop the subscript for DQ_i and DQ_e and will refer to these documents as *DQ*.

It is important to observe that DWS and DQ are *not disjoint sets of documents*, because in one session a document can be reached using a search engine (therefore belonging to DQ) and in a different session it can also be reached without using a search engine. The important issue then, is to register *how many times* each of these different events occur for each document. We will consider the frequency of each event directly proportional to that event's significance for improving a website. The classification of documents into these three categories will be essential in our model for discovering useful information from queries in a website.

Heuristic to Classify Documents. Documents belonging to DQ sets can be discovered directly by analyzing the referer URL in an HTTP request to see if it is equal to the results page of a search engine (internal or external). In these cases only the *first occurrence* of each requested document in a session is classified. On the other hand, documents in DWS are more difficult to classify, due to the fact that backward and forward navigation in the browser's cached history of previously visited documents is not registered in web servers usage