Robert Bembenik, Łukasz Skonieczny, Henryk Rybiński, and Marek Niezgódka (Eds.)

Intelligent Tools for Building a Scientific Information Platform

Studies in Computational Intelligence, Volume 390

Editor-in-Chief
Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Further volumes of this series can be found on our homepage: springer.com

Vol. 367. Gabriel Luque and Enrique Alba
Parallel Genetic Algorithms, 2011
ISBN 978-3-642-22083-8

Vol. 368. Roger Lee (Ed.)
Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2011, 2011
ISBN 978-3-642-22287-0

Vol. 369. Dominik Ryżko, Piotr Gawrysiak, Henryk Rybiński, and Marzena Kryszkiewicz (Eds.)
Emerging Intelligent Technologies in Industry, 2011
ISBN 978-3-642-22731-8

Vol. 370. Alexander Mehler, Kai-Uwe Kühnberger, Henning Lobin, Harald Lüngen, Angelika Storrer, and Andreas Witt (Eds.)
Modeling, Learning, and Processing of Text Technological Data Structures, 2011
ISBN 978-3-642-22612-0

Vol. 371. Leonid Perlovsky, Ross Deming, and Roman Ilin (Eds.)
Emotional Cognitive Neural Algorithms with Engineering Applications, 2011
ISBN 978-3-642-22829-2

Vol. 372. António E. Ruano and Annamária R. Várkonyi-Kóczy (Eds.)
New Advances in Intelligent Signal Processing, 2011
ISBN 978-3-642-11738-1

Vol. 373. Oleg Okun, Giorgio Valentini, and Matteo Re (Eds.)
Ensembles in Machine Learning Applications, 2011
ISBN 978-3-642-22909-1

Vol. 374. Dimitri Plemenos and Georgios Miaoulis (Eds.)
Intelligent Computer Graphics 2011, 2011
ISBN 978-3-642-22906-0

Vol. 375. Marenglen Biba and Fatos Xhafa (Eds.)
Learning Structure and Schemas from Documents, 2011
ISBN 978-3-642-22912-1

Vol. 376. Toyohide Watanabe and Lakhmi C. Jain (Eds.)
Innovations in Intelligent Machines – 2, 2011
ISBN 978-3-642-23189-6

Vol. 377. Roger Lee (Ed.)
Software Engineering Research, Management and Applications 2011, 2011
ISBN 978-3-642-23201-5

Vol. 378. János Fodor, Ryszard Klempous, and Carmen Paz Suárez Araujo (Eds.)
Recent Advances in Intelligent Engineering Systems, 2011
ISBN 978-3-642-23228-2

Vol. 379. Ferrante Neri, Carlos Cotta, and Pablo Moscato (Eds.)
Handbook of Memetic Algorithms, 2011
ISBN 978-3-642-23246-6

Vol. 380. Anthony Brabazon, Michael O'Neill, and Dietmar Maringer (Eds.)
Natural Computing in Computational Finance, 2011
ISBN 978-3-642-23335-7

Vol. 381. Radosław Katarzyniak, Tzu-Fu Chiu, Chao-Fu Hong, and Ngoc Thanh Nguyen (Eds.)
Semantic Methods for Knowledge Management and Communication, 2011
ISBN 978-3-642-23417-0

Vol. 382. F.M.T. Brazier, Kees Nieuwenhuis, Gregor Pavlin, Martijn Warnier, and Costin Badica (Eds.)
Intelligent Distributed Computing V, 2011
ISBN 978-3-642-24012-6

Vol. 383. Takayuki Ito, Minjie Zhang, Valentin Robu, Shaheen Fatima, and Tokuro Matsuo (Eds.)
New Trends in Agent-Based Complex Automated Negotiations, 2012
ISBN 978-3-642-24695-1

Vol. 384. Daphna Weinshall, Jörn Anemüller, and Luc van Gool (Eds.)
Detection and Identification of Rare Audiovisual Cues, 2012
ISBN 978-3-642-24033-1

Vol. 385. Alex Graves
Supervised Sequence Labelling with Recurrent Neural Networks, 2012
ISBN 978-3-642-24796-5

Vol. 386. Marek R. Ogiela and Lakhmi C. Jain (Eds.)
Computational Intelligence Paradigms in Advanced Pattern Classification, 2012
ISBN 978-3-642-24048-5

Vol. 387. David Alejandro Pelta, Natalio Krasnogor, Dan Dumitrescu, Camelia Chira, and Rodica Lung (Eds.)
Nature Inspired Cooperative Strategies for Optimization (NICSO 2011), 2011
ISBN 978-3-642-24093-5

Vol. 388. Tiansi Dong
Recognizing Variable Environments, 2012
ISBN 978-3-642-24057-7

Vol. 389. Patricia Melin
Modular Neural Networks and Type-2 Fuzzy Systems for Pattern Recognition, 2012
ISBN 978-3-642-24138-3

Vol. 390. Robert Bembenik, Łukasz Skonieczny, Henryk Rybiński, and Marek Niezgódka (Eds.)
Intelligent Tools for Building a Scientific Information Platform, 2012
ISBN 978-3-642-24808-5

Robert Bembenik, Łukasz Skonieczny,
Henryk Rybiński, and Marek Niezgódka (Eds.)

Intelligent Tools for Building a Scientific Information Platform

Springer

Editors

Dr. Robert Bembenik
Institute of Computer Science
Faculty of Electronics and
Information Technology
Warsaw University of Technology
Ul. Nowowiejska 15/19
00-665 Warsaw
Poland
E-mail: R.Bembenik@ii.pw.edu.pl

Prof. Henryk Rybiński
Institute of Computer Science
Faculty of Electronics and
Information Technology
Warsaw University of Technology
Ul. Nowowiejska 15/19
00-665 Warsaw
Poland
E-mail: H.Rybinski@ii.pw.edu.pl

Dr. Łukasz Skonieczny
Institute of Computer Science
Faculty of Electronics and
Information Technology
Warsaw University of Technology
Ul. Nowowiejska 15/19
00-665 Warsaw
Poland
E-mail: L.Skonieczny@ii.pw.edu.pl

Prof. Marek Niezgódka
Interdisciplinary Centre for Mathematical and
Computational Modelling (ICM)
University of Warsaw
Ul. Pawińskiego 5a
02-106 Warsaw
Poland
E-mail: M.Niezgodka@icm.edu.pl

ISBN 978-3-642-24808-5 e-ISBN 978-3-642-24809-2

DOI 10.1007/978-3-642-24809-2

Studies in Computational Intelligence ISSN 1860-949X

Library of Congress Control Number: 2011939766

Typesetting: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

SYNAT is a program funded by the National Centre for Research and Development in Poland (NCBiR), with the main objective to set up an ICT platform for the national system of repositories that will cover content areas of science and humanities. A network of 16 academic partners have committed to implement SYNAT's concept in the form of a universal open knowledge infrastructure for the information society in Poland. Beyond the system development, a comprehensive portfolio of research problems is addressed by the network partners.

The program is scheduled for the period of three years, initiated in 2010. In view of the limited implementation time, the primary goal of the program consists in meeting the challenges of global digital information revolution viewed from the perspective of Poland. So not only the access to knowledge, at any scale, but also the forms of visibility gain for all categories of the publications and other forms documenting results of the research are foreseen.

Novel algorithms, their implementations and practical use are expected to result from SYNAT's activities. A broad range of new functionalities will get introduced and implemented upon validating their real practical features for scalability and interoperability. The result of a one year research comprises a number of works in the areas of, *inter alia*, artificial intelligence, knowledge discovery and data mining, information retrieval and natural language processing, addressing the problems of implementing intelligent tools for building a scientific information platform.

The idea of this book is based on the very successful SYNAT Project Conference (January, 2011) and the SYNAT Workshop accompanying the 19th International Symposium on Methodologies for Intelligent Systems (ISMIS 2011). During the Conference, research areas for the purpose of the platform have been outlined. The Workshop was a place of discussions on proposed solutions based on intelligent tools able to significantly improve the quality of the planned scientific information services.

The papers included in this volume cover the following topics: The SYNAT Project Concepts, Semantic Clustering, Ontology-based Systems, Multimedia Data Processing. We will now briefly summarize contents of the particular chapters.

Chapter I, "The SYNAT Project Concepts", is dealing with issues and problems related to constructing a large IT software system capable of collecting, storing and searching information in an intelligent way.

- **Hung Son Nguyen, Dominik Ślęzak, Andrzej Skowron, and Jan G. Bazan** ("Semantic Search and Analytics over Large Repository of Scientific Articles") present an architecture of the system aimed at searching and synthesizing information within document repositories originating from different sources, i.e., with documents provided not necessarily in the same format and the same level of detail. The system is expected to provide domain knowledge interfaces enabling the internally implemented algorithms to identify relationships between documents (as well as authors, institutions etc.), and concepts (such as, e.g., areas of science) extracted from various types of knowledge bases. The system should be scalable by means of scientific content storage, performance of analytic processes, and speed of search.
- **Linh Anh Nguyen and Hung Son Nguyen** ("On Designing the SONCA System") present ideas and proposals for the SONCA system. The main idea is to allow combination of metadata-based search, syntactic keyword-based search and semantic search, and to use ranks of objects. Semantic search criteria may be keywords, concepts, or objects (for checking similarity). Search criteria based on metadata play the role of exact restrictions, while syntactic keywords and semantic search criteria are fuzzy restrictions. To enable metadata-based search, an appropriate document representation is used. To enable syntactic keyword-based search, each document (object) is stored together with information about the terms occurring in its text attributes. Authors provide an abstract model for the SONCA system, an instantiation of that model, some ideas for the user interface of SONCA as well as proposals for increasing efficiency of the query answering process.
- **Piotr Gawrysiak, Dominik Ryżko, Przemysław Więch, and Marek Kozłowski** ("Retrieval and Management of Scientific Information from Heterogeneous Sources") describe the process of automated retrieval and management of scientific information from various sources, including the Internet. They describe application of semantic methods in different phases of the process. The system envisaged in the project is a scientific digital library, with automated retrieval and hosting capabilities. An overall architecture for the system is proposed.
- **Marcin Kowalski, Dominik Ślęzak, Krzysztof Stencel, Przemysław Pardel, Marek Grzegorowski, and Michał Kijowski** ("RDBMS Model for Scientific Articles Analytics") present a relational database schema aimed at efficient storage and querying of parsed scientific articles, as well as entities referring to researchers, institutions, scientific areas, etc. An important

requirement in front of the proposed model is to operate with various types of entities, but with no increase of schema's complexity. Another aspect is to store detailed information about parsed articles in order to conduct advanced analytics in combination with the domain knowledge about scientific topics, by means of standard SQL and RDBMS management. The overall goal is to enable offline, possibly incremental computation of semantic indexes supporting end users via other modules, optimized for fast search and not necessarily for fast analytics.

Chapter II, "Semantic Clustering", deals with semantic clustering of documents as well as problems of document representation and document representation formats facilitating such clustering.

- **Marcin Szczuka, Andrzej Janusz, and Kamil Herba** ("Semantic Clustering of Scientific Articles with Use of DBpedia Knowledge Base") present a case study of semantic clustering of scientific articles related to Rough Sets. The proposed method groups the documents on the basis of their content and with assistance of DBpedia knowledge base. The text corpus is first treated with Natural Language Processing tools in order to produce vector representations of the content and then matched against a collection of concepts retrieved from DBpedia. As a result, a new representation is constructed that better reflects the semantics of the texts. With this new representation, the documents are hierarchically clustered in order to form partition of papers that share semantic relatedness. An assessment of clustering quality by human experts, compared to traditional approach, is presented.
- **S. Hoa Nguyen, Wojciech Świeboda, and Grzegorz Jaśkiewicz** ("Extended Document Representation for Search Result Clustering") discuss a framework of document description extension which utilizes domain knowledge and semantic similarity. The idea is based on application of Tolerance Rough Set Model, semantic information extracted from a source text and domain ontology to approximate concepts associated with documents and to enrich the vector representation.
- **Paweł Betliński, Paweł Gora, Kamil Herba, Trung Tuan Nguyen, and Sebastian Stawicki** ("Semantic Recognition of Digital Documents") describe document representation format together with a proof of concept of the system converting scientific articles in PDF format into this representation. Another topic presented in the article is an experiment with clustering documents by style.

Chapter III, "Ontology-based Systems", is concerned primarily with semantics and ontologies. These notions are employed to the construction of semantically-driven knowledge bases.

- **Piotr Wasilewski** ("Towards Semantic Evaluation of Information Retrieval") discusses fundamentals of semantic evaluation of information retrieval systems. Semantic evaluation is understood in two ways. Semantic evaluation *sensu stricto* consists of automatic global methods of information retrieval evaluation which are based on knowledge representation systems. Semantic

evaluation *sensu largo* includes also evaluation of retrieved results presented using new methods and comparing them to previously used ones. The paper focuses on sematic relevance of documents, both binary and graded, together with semantic ranking of documents. Various types of semantic value and semantic relevance are proposed and also some semantic versions of information retrieval evaluation measures are given.

- **Anna Wróblewska, Teresa Podsiadły-Marczykowska, Robert Bembenik, Grzegorz Protaziuk, and Henryk Rybiński** ("Methods and Tools for Ontology Building, Learning and Integration - Application in the SYNAT Project") started building an experimental platform where different kinds of stored knowledge will be modeled with the use of ontologies, in particular a system ontology, accompanied by domain ontologies. The system ontology defines "system domain" (a kind of meta knowledge) for the scientific community, covering concepts and activities related to the scientific community. The paper makes a contribution to understanding semantically modeled knowledge and its incorporation into the SYNAT project. The authors present a review of ontology building, learning, and integration methods and their potential application in the project.
- **Cezary Mazurek, Krzysztof Sielski, Maciej Stroiński, Justyna Walkowska, Marcin Werla, and Jan Węglarz** ("Transforming a Flat Metadata Schema to a Semantic Web Ontology: The Polish Digital Libraries Federation and CIDOC CRM Case Study") describe the transformation of the metadata schema used by the Polish Digital Libraries Federation, to the CIDOC CRM model implemented in OWL as Erlangen CRM. In the paper the authors identify a number of problems that are common to all such transformations and propose solutions. They also present statistics concerning the mapping process and the resulting knowledge base.
- **Krzysztof Goczyła, Aleksander Waloszek, Wojciech Waloszek, and Teresa Zawadzka** ("Modularized Knowledge Bases Using Contexts, Conglomerates and a Query Language") present a novel approach to design and development of a modularized knowledge base. The approach is oriented towards decomposition of a knowledge base into logical components, called contexts, and further, into semantic components called conglomerates. The paper shows how contexts and conglomerates concepts can work in harmony to create a maintainable knowledge base. A thorough discussion of related work is also given.

Chapter IV, "Multimedia Data Processing", is devoted to the data mining approach for processing multimedia data like sound and images. Both hardware and software solutions are discussed.

- **Maciej Wielgosz, Ernest Jamro, Dominik Żurek, and Kazimierz Wiatr** ("FPGA Implementation of the Selected Parts of the Fast Image Segmentation") present preliminary hardware implementation of a SVM (Support Vector Machine) algorithm (in FPGA). The work is primarily focused on the FPGA implementation aspects of the algorithm as well as on comparison of

the hardware and software performance. The approach provides a high performance of the hardware classification module.

- **Rafał Frączek and Bogusław Cyganek** ("Evaluation of Image Descriptors for Retrieval of Similar Images") address the issue of searching for similar images in a repository. The contained images are annotated with help of the sparse descriptors. In the presented research different color and edge histogram descriptors were used. To measure distances among images the sets of their descriptors are compared. For this purpose different similarity measures were employed. The results of these experiments, as well as discussion of the advantages and limitations of different combinations of methods for retrieval of similar images are presented.
- **Andrzej Czyżewski, Adam Kupryjanow, and Bożena Kostek** ("Online Sound Restoration for Digital Library Applications") discuss a sound restoration system conceived and engineered at the Multimedia Systems Department of the Gdansk University of Technology with regard to the principles of its design, features of operation and the achieved results. The system has been designed so that (1) no special sound restoration software is needed to perform audio restoration; (2) no skills in digital signal processing are required from the user.

Chapter V, "Intelligent Systems, Tools and Applications", deals with the development and applications of various tools which might be used as a part of the scientific information platform.

- **Łukasz Brocki, Krzysztof Marasek, and Danijel Koržinek** ("Connectionist Language Model for Polish") describe a connectionist language model, which may be used as an alternative to the well known n-gram models. A comparison experiment between n-gram and connectionist language models is performed on a Polish text corpus. Statistical language modeling is based on estimating a joint probability function of a sequence of words in a given language. This task is made problematic due to a phenomenon known commonly as the "curse of dimensionality". In the presented experiments, perplexity is used as a measure of language model quality.
- **Henryk Krawczyk and Marek Downar** ("Commonly Accessible Web Service Platform - Wiki-WS") present a SOA-enabled platform - Wiki-WS - that empowers users to deploy, modify, discover and invoke web services. The main concept of the Wiki-WS platform is searching and the invocation of web services written by different workgroups in different technologies deployed on different servers. Wiki-based web service code modification allows engineers from any place to construct and to implement components, which are mature and ready to use. There is also included presentation of sample scenarios of Wiki-WS usage and advantages derived from its deployment.

This book could not have been completed without the help of many people. We would like to thank all the authors for their contribution to the book and their effort in addressing reviewers' and editorial feedback. We would also like to thank reviewers and all program committee members of the SYNAT Workshop. Finally, we would like to thank Bożenna Skalska for her administrative work.

July 2011
Warszawa

Robert Bembenik
Łukasz Skonieczny
Henryk Rybiński
Marek Niezgódka

Contents

Chapter 1: The SYNAT Project Concepts

Semantic Search and Analytics over Large Repository of Scientific Articles 1
Hung Son Nguyen, Dominik Ślęzak, Andrzej Skowron, Jan G. Bazan

On Designing the SONCA System 9
Linh Anh Nguyen, Hung Son Nguyen

Retrieval and Management of Scientific Information from Heterogeneous Sources 37
Piotr Gawrysiak, Dominik Ryżko, Przemysław Więch, Marek Kozłowski

RDBMS Model for Scientific Articles Analytics 49
Marcin Kowalski, Dominik Ślęzak, Krzysztof Stencel, Przemysław Pardel, Marek Grzegorowski, Michał Kijowski

Chapter 2: Semantic Clustering

Semantic Clustering of Scientific Articles with Use of DBpedia Knowledge Base 61
Marcin Szczuka, Andrzej Janusz, Kamil Herba

Extended Document Representation for Search Result Clustering 77
S. Hoa Nguyen, Wojciech Świeboda, Grzegorz Jaśkiewicz

Semantic Recognition of Digital Documents 97
Paweł Betliński, Paweł Gora, Kamil Herba, Trung Tuan Nguyen, Sebastian Stawicki

Chapter 3: Ontology-Based Systems

Towards Semantic Evaluation of Information Retrieval 107
Piotr Wasilewski

Methods and Tools for Ontology Building, Learning and Integration – Application in the SYNAT Project 121
Anna Wróblewska, Teresa Podsiadły-Marczykowska, Robert Bembenik, Grzegorz Protaziuk, Henryk Rybiński

Transforming a Flat Metadata Schema to a Semantic Web Ontology: The Polish Digital Libraries Federation and CIDOC CRM Case Study 153
Cezary Mazurek, Krzysztof Sielski, Maciej Stroiński, Justyna Walkowska, Marcin Werla, Jan Węglarz

Modularized Knowledge Bases Using Contexts, Conglomerates and a Query Language 179
Krzysztof Goczyła, Aleksander Waloszek, Wojciech Waloszek, Teresa Zawadzka

Chapter 4: Multimedia Data Processing

FPGA Implementation of the Selected Parts of the Fast Image Segmentation 203
Maciej Wielgosz, Ernest Jamro, Dominik Żurek, Kazimierz Wiatr

Evaluation of Image Descriptors for Retrieval of Similar Images 217
Rafał Frączek, Bogusław Cyganek

Online Sound Restoration for Digital Library Applications 227
Andrzej Czyżewski, Adam Kupryjanow, Bożena Kostek

Chapter 5: Intelligent Systems, Tools and Applications

Connectionist Language Model for Polish 243
Łukasz Brocki, Krzysztof Marasek, Danijel Koržinek

Commonly Accessible Web Service Platform — Wiki-WS 251
Henryk Krawczyk, Marek Downar

Author Index 265