Enrique Herrera-Viedma, Gabriella Pasi, Fabio Crestani (Eds.)

Soft Computing in Web Information Retrieval

Studies in Fuzziness and Soft Computing, Volume 197

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. 181. Nadia Nedjah,
Luiza de Macedo Mourelle
Fuzzy Systems Engineering, 2005
ISBN 3-540-25322-X

Vol. 182. John N. Mordeson,
Kiran R. Bhutani, Azriel Rosenfeld
Fuzzy Group Theory, 2005
ISBN 3-540-25072-7

Vol. 183. Larry Bull, Tim Kovacs (Eds.)
Foundations of Learning Classifier Systems,
2005
ISBN 3-540-25073-5

Vol. 184. Barry G. Silverman, Ashlesha Jain,
Ajita Ichalkaranje, Lakhmi C. Jain (Eds.)
*Intelligent Paradigms for Healthcare
Enterprises*, 2005
ISBN 3-540-22903-5

Vol. 185. Spiros Sirmakessis (Ed.)
Knowledge Mining, 2005
ISBN 3-540-25070-0

Vol. 186. Radim Bělohlávek, Vilém
Vychodil
Fuzzy Equational Logic, 2005
ISBN 3-540-26254-7

Vol. 187. Zhong Li, Wolfgang A. Halang,
Guanrong Chen (Eds.)
*Integration of Fuzzy Logic and Chaos
Theory*, 2006
ISBN 3-540-26899-5

Vol. 188. James J. Buckley, Leonard J.
Jowers
Simulating Continuous Fuzzy Systems, 2006
ISBN 3-540-28455-9

Vol. 189. Hans Bandemer
Mathematics of Uncertainty, 2006
ISBN 3-540-28457-5

Vol. 190. Ying-ping Chen
*Extending the Scalability of Linkage
Learning Genetic Algorithms*, 2006
ISBN 3-540-28459-1

Vol. 191. Martin V. Butz
*Rule-Based Evolutionary Online Learning
Systems*, 2006
ISBN 3-540-25379-3

Vol. 192. Jose A. Lozano, Pedro Larrañaga,
Iñaki Inza, Endika Bengoetxea (Eds.)
Towards a New Evolutionary Computation,
2006
ISBN 3-540-29006-0

Vol. 193. Ingo Glöckner
Fuzzy Quantifiers: A Computational Theory,
2006
ISBN 3-540-29634-4

Vol. 194. Dawn E. Holmes, Lakhmi C. Jain
(Eds.)
Innovations in Machine Learning, 2006
ISBN 3-540-30609-9

Vol. 195. Zongmin Ma
*Fuzzy Database Modeling of Imprecise and
Uncertain Engineering Information*, 2006
ISBN 3-540-30675-7

Vol. 196. James J. Buckley
Fuzzy Probability and Statistics, 2006
ISBN 3-540-30841-5

Vol. 197. Enrique Herrera-Viedma, Gabriella
Pasi, Fabio Crestani (Eds.)
*Soft Computing in Web Information
Retrieval*, 2006
ISBN 3-540-31588-8

Enrique Herrera-Viedma
Gabriella Pasi
Fabio Crestani
(Eds.)

Soft Computing in Web Information Retrieval

Models and Applications

Professor Enrique Herrera-Viedma

Department of Computer Science and A.I
E.T.S.I. Informatica
University of Granada
C/Periodista Daniel
Saucedo Aranda s/n
Granada, Spain
E-mail: viedma@decsai.ugr.es

Professor Fabio Crestani

Department of Computer
and Information Sciences
University of Strathclyde
Livingstone Tower
26 Richmond Street
Glasgow G1 1XH
Scotland, UK
E-mail: F.Crestani@cis.strath.ac.uk

Professor Gabriella Pasi

Università degli Studi di Milano Bicocca
Department of Informatics
Systems and Communication (DISCO)
Via Bicocca degli Arcimboldi
8 (Edificio U7)
20126 Milano (ITALY)
E-mail: pasi@disco.unimib.it

Library of Congress Control Number: 2005938670

ISSN print edition: 1434-9922

ISSN electronic edition: 1860-0808

ISBN-10 3-540-31588-8 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-31588-9 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com
© Springer-Verlag Berlin Heidelberg 2006
Printed in The Netherlands

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: by the authors and TechBooks using a Springer L^AT_EX macro package

Printed on acid-free paper SPIN: 11370697 89/TechBooks 5 4 3 2 1 0

Preface

The *World Wide Web*, or simply the Web, is a popular and interactive medium to collect, disseminate, and access an increasingly huge amount of information. Nowadays, *information access on the Web* is the main problem of the so called Web Information Retrieval (IR). The Web represents a new framework that is rather different from with respect to the traditional IR framework and sheds new and difficult challenges.

The Web presents particular characteristics that limit the existing IR technologies and determine the need to design new information access technologies:

- the Web is possibly the biggest and dynamic information resource existing.
- the Web presents a structure of linked pages.
- the Web is growing and updating at a very high rate.
- the Web is very heterogeneous.
- last but not least, imprecision and vagueness characterize several tasks in Web IR, such as assessing the relevance of Web pages, dealing with the multimedia nature of information, identifying spam problem, discovering deception, etc.

Furthermore, due to this complexity, any major advance in the field of information access on the Web requires the application of intelligent techniques. In fact, several authors suggest to proceed towards the *Web Intelligence* by incorporating and embedding some form of intelligence (such as learning capabilities, and tolerance to uncertainty, vagueness and imprecision) in Web technologies.

Soft Computing (SC) techniques constitute a synergy of methodologies (fuzzy logic, neural networks, probabilistic reasoning, rough-set theory, evolutionary computing and parts of machine learning theory) useful for solving problems requiring some form of intelligence. The basis of SC is its tolerance to imprecision, uncertainty, partial truth, and approximation. Because of these properties SC can provide very powerful tools for modelling the different activities related with the information access problem. In a previous book edited in this series, titled “*Soft Computing in Information Retrieval*,

Techniques and Applications", F. Crestani and G. Pasi (Eds.), collected a selection of SC-based approaches to the traditional IR. The present edited volume focuses on the use of the SC techniques for improving information access in Web IR.

This book presents some recent works on the application of SC techniques in information access on the Web. The book comprises 15 chapters from internationally known researchers. It is divided in four parts reflecting the areas of research of the presented works.

The first part focuses on the use of SC in *Document Classification*. The chapter by Bordogna and Pasi proposes a hierarchical fuzzy clustering algorithm for dynamically supporting document filtering that performs a fuzzy hierarchical categorization of documents allowing updating as new documents are fed. The chapter by de Campos, Fernández-Luna, and Huete presents a theoretical framework for classifying Web pages in a hierarchical directory using the Bayesian Network formalism which is able to perform multi-label text categorization in a category tree in a Web framework. The chapter by Loia and Senatore describes a customized system for information discovery based on fuzzy clustering of RDF-based documents which are classified in terms of the semantics of their metadata. The chapter by Zhang, Fan, Chen, Fox, Gonçalves, Cristo, and Pável Calado defines a Genetic Programming Approach for Combining Structural and Citation-Based Evidence for Text Classification in Web Digital Libraries.

The second part presents experiences on the development of the *Semantic Web* using SC techniques. The chapter by Ceravolo, Damiani, and Viviani describes a complete approach for developing a Trust Layer service, aimed at improving the quality of automatically generated semantic Web-style metadata and based on non-intrusive collection of user feedback. The chapter by Herrera-Viedma, Peis, and Morales-del-Castillo defines a model of a Web fuzzy linguistic multi-agent system that combines the use of Semantic Web technologies together with the application of user profiles to carry out its information access processes. The chapter by Barriocanal and Sicilia proposes a first fuzzy approach for the design of ontology-based browsers. The chapter by Loiseau, Boughanem, and Prade presents an evaluation method of term-based queries using possibilistic ontologies which allows to retrieve information containing terms that may not match exactly those of the query. This tool can be applied in both textual information retrieval and data base management.

The third part shows different SC approaches to the *Web Information Retrieval*. The chapter by Dominich, Skrop, and Tuza presents a unified formal framework for three major methods used for Web retrieval tasks: PageRank, HITS, I²R. It is based on the Artificial Neural Networks and the generic network equation. It was shown that the PageRank, HITS and I²R methods can be formally obtained from the generic equation as different particular cases by making certain assumptions reflecting the corresponding underlying paradigm. The chapter by Losada, Díaz-Hermida, and Bugarín carries out an empirical study that demonstrates the usefulness of the semi-fuzzy quantifiers

for improving the query languages in information retrieval. In particular, it is shown that fuzzy quantifiers-based IR models are competitive with respect to models such as the vector-space model. The chapter by Martín-Bautista, Sánchez, Serrano, and Vila describes a query refinement technique based on fuzzy association rules that helps the user to search information and improve the Web information retrieval. The chapter by Valverde-Albacete proposes a formal model of the batch retrieval phase of a Web retrieval interaction or any other batch retrieval task which is designed using hard techniques like Concept Formal Analysis and soft techniques like Rough-Set Theory.

The fourth part reports a selection of *Web Applications* developed by means of SC techniques. The chapter by Cristo, Ribeiro-Neto, Golgher, and de Moura presents an analysis on key concepts and variables related to search advertising, both in the commercial and in the technology fronts. It studies some SC approaches to the Web topic as the case of content-targeted advertising based on Bayesian Networks. The chapter by Domingo-Ferrer, Mateo-Sanz, and Sebé provides a fast method for generating numerical hybrid microdata in a way that preserves attribute means, variances and covariances, as well as (to some extent) record similarity and subdomain analyses. Finally, the chapter by Sobrino-Cerdeiriña, Fernández-Lanza, and Graña-Gil proposes a general model for implementing large dictionaries in natural language processing applications which is able to store a considerable amount of data relating to the words contained in these dictionaries. Additionally, it shows how this model can be applied to implement and transform a Spanish dictionary of synonyms into a computational framework able to represent relations of synonymy between words.

Ultimately, the goal of this book is to show that Web IR can be a stimulating area of research where SC technologies can be applied satisfactorily. This book is a proof of this and we think that it will not be the last one.

Granada
October, 2005

Enrique Herrera-Viedma
Gabriella Pasi
Fabio Crestani

Acknowledgments

We would like to thank the authors of the papers, that with their effort showed that it is possible to improve the performance of the Web technologies through SC techniques and made possible the apparition of this book.

Our gratitude also goes to Ricardo Baeza-Yates for his foreword, and to the reviewers (Miyamoto, Kraft, Sobrino, Losada, Huete, Domingo-Ferrer, Damiani, Martín-Bautista, Dominich, Ribeiro-Neto, Bordogna, Olsina, Mich, Fan, Olivas, Peis, Loia, Sicilia). Without their help and collaboration we could not have assured the high quality of this book (we received 22 contributions and each paper was reviewed by at least three referees).

Finally, thanks to Janusz Kacprzyk, the series editor of Studies in Fuzziness and Soft Computing, for accepting our proposal of this volume.

Foreword

The Web currently is the largest repository of data available, comprising a maremagnum of different media over more than ten billion interconnected pages. However, volume is not necessarily the main challenge, as content and link spamming makes information retrieval even harder. Indeed, searching for information in the Web has been called “adversarial Web retrieval”. Hence, the main challenges are:

- to keep an up-to-date index of all the pages of the Web by crawling it,
- to assess how much we can trust the content of a given page (this is a key issue also for the semantic Web),
- to compute the relevance of the page with respect to the user query, and
- to give a personalized answer.

These challenges imply several sub-challenges and several related problems such as what advertising can be shown in the answer page or how to use different sources of information to rank a page (content, links, usage, etc.)

Soft-computing can help in many of the challenges above, specially in off-line tasks where we can preprocess data to build additional data structures that are fast enough for on-line use. Important examples are new retrieval models, categorization of documents, link analysis, trust models, creation of pseudo-semantic resources, fuzzy search, linguistic processing, adaptive interfaces, etc. This book contains several of the problems and applications mentioned above and it is one step forward on the fascinating research path that lies in front of us.

Barcelona, Spain
October 2005

Ricardo Baeza-Yates

Contents

Part I Document Classification

A Dynamic Hierarchical Fuzzy Clustering Algorithm for Information Filtering

Gloria Bordogna, Marco Pagani, and Gabriella Pasi 3

A Theoretical Framework for Web Categorization in Hierarchical Directories using Bayesian Networks

Luis M. de Campos, Juan M. Fernández-Luna, and Juan F. Huete 25

Personalized Knowledge Models Using RDF-Based Classification

Vincenzo Loia and Sabrina Senatore 45

A Genetic Programming Approach for Combining Structural and Citation-Based Evidence for Text Classification in Web Digital Libraries

*Baoping Zhang, Weiguo Fan, Yuxin Chen, Edward A. Fox,
Marcos André Gonçalves, Marco Cristo, and Pável Calado* 65

Part II Semantic Web

Adding a Trust Layer to Semantic Web Metadata

Paolo Ceravolo, Ernesto Damiani, and Marco Viviani 87

A Fuzzy Linguistic Multi-agent Model Based on Semantic Web Technologies and User Profiles

*Enrique Herrera-Viedma, Eduardo Peis,
and José M. Morales-del-Castillo* 105

Fuzzy Concept-Based Models in Information Browsers

Elena García Barriocanal and Miguel-Ángel Sicilia 121

**Evaluation of Term-based Queries
using Possibilistic Ontologies**

Yannick Loiseau, Mohand Boughanem, and Henri Prade 135

Part III Web Information Retrieval

Formal Theory of Connectionist Web Retrieval

Sándor Dominich, Adrienn Skrop, and Zsolt Tuza 163

Semi-fuzzy Quantifiers for Information Retrieval

David E. Losada, Félix Díaz-Hermida, and Alberto Bugarín 195

**Helping Users in Web Information Retrieval Via Fuzzy
Association Rules**

M.J. Martín-Bautista, D. Sánchez, J.M. Serrano, and M.A. Vila 221

**Combining Soft and Hard Techniques for the Analysis
of Batch Retrieval Tasks**

Francisco J. Valverde-Albacete 239

Part IV Web Application

Search Advertising

*Marco Cristo, Berthier Ribeiro-Neto, Paulo B. Golgher, and Edleno
Silva de Moura* 259

**Information Loss in Continuous Hybrid Microdata:
Subdomain-Level Probabilistic Measures**

Josep Domingo-Ferrer, Josep Maria Mateo-Sanz, and Francesc Sebé 287

**Access to a Large Dictionary of Spanish Synonyms: A Tool
for Fuzzy Information Retrieval**

*Alejandro Sobrino-Cerdeiríña, Santiago Fernández-Lanza, and Jorge
Graña-Gil* 299