Information Extraction: Algorithms and Prospects
in a Retrieval Context

THE INFORMATION RETRIEVAL SERIES

Series Editor:
W. Bruce Croft
University of Massachusetts, Amherst

Also in the Series:

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** *by Gerald Kowalski;* ISBN: 0-7923-9926-9

CROSS-LANGUAGE INFORMATION RETRIEVAL, *edited by Gregory Grefenstette;* ISBN: 0-7923-8122-X

TEXT RETRIEVAL AND FILTERING: ***Analytic Models of Performance***, *by Robert M. Losee;* ISBN: 0-7923-8177-7

INFORMATION RETRIEVAL: UNCERTAINTY AND LOGICS: ***Advanced Models for the Representation and Retrieval of Information*****,** *by Fabio Crestani, Mounia Lalmas, and Cornelis Joost van Rijsbergen;* ISBN: 0-7923-8302-8

DOCUMENT COMPUTING: ***Technologies for Managing Electronic Document Collections***, *by Ross Wilkinson, Timothy Arnold-Moore, Michael Fuller, Ron Sacks-Davis, James Thom, and Justin Zobel;* ISBN: 0-7923-8357-5

AUTOMATIC INDEXING AND ABSTRACTING OF DOCUMENT TEXTS, *by Marie-Francine Moens*; ISBN 0-7923-7793-1

ADVANCES IN INFORMATIONAL RETRIEVAL: ***Recent Research from the Center for Intelligent Information Retrieval***, *by W. Bruce Croft*; ISBN 0-7923-7812-1

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** **Second Edition**, *by Gerald J. Kowalski and Mark T. Maybury;* ISBN: 0-7923-7924-1

PERSPECTIVES ON CONTENT-BASED MULTIMEDIA SYSTEMS, *by Jian Kang Wu; Mohan S. Kankanhalli;Joo-Hwee Lim;Dezhong Hong;* ISBN: 0-7923-7944-6

MINING THE WORLD WIDE WEB: ***An Information Search Approach***, *by George Chang, Marcus J. Healey, James A. M. McHugh, Jason T. L. Wang;* ISBN: 0-7923-7349-9

INTEGRATED REGION-BASED IMAGE RETRIEVAL, *by James Z. Wang;* ISBN: 0-7923-7350-2

TOPIC DETECTION AND TRACKING: Event-based Information Organization, *edited by James Allan;* ISBN: 0-7923-7664-1

LANGUAGE MODELING FOR INFORMATION RETRIEVAL, *edited by W. Bruce Croft; John Lafferty;* ISBN: 1-4020-1216-0

MACHINE LEARNING AND STATISTICAL MODELING APPROACHES TO IMAGE RETRIEVAL, *by Yixin Chen, Jia Li and James Z. Wang;* ISBN: 1-4020-8034-4

INFORMATION RETRIEVAL: ***Algorithms and Heuristics,*** *by David A. Grossman and Ophir Frieder,* 2nd ed.; ISBN: 1-4020-3003-7; PB: ISBN: 1-4020-3004-5

CHARTING A NEW COURSE: Natural Language Processing and Information Retrieval, *edited by John I. Tait;* ISBN: 1-4020-3343-5

INTELLIGENT DOCUMENT RETRIEVAL: Exploiting Markup Structure, *by Udo Kruschwitz;* ISBN: 1-4020-3767-8

THE TURN: Integration of Information Seeking and Retrieval in Context, *by Peter Ingwersen, Kalervo Järvelin;* ISBN: 1-4020-3850-X

NEW DIRECTIONS IN COGNITIVE INFORMATION RETRIEVAL, *edited by Amanda Spink, Charles Cole;* ISBN: 1-4020-4013-X

COMPUTING ATTITUDE AND AFFECT IN TEXT: Theory and Applications, *edited by James G Shanahan, Yan Qu, Janyce Wiebe;* ISBN: 1-4020-4026-1

Information Extraction: Algorithms and Prospects in a Retrieval Context

By

Marie-Francine Moens
Katholieke Universiteit Leuven, Belgium

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 1-4020-4987-0 (HB)
ISBN-13 978-1-4020-4987-3 (HB)
ISBN-10 1-4020-4993-5 (e-book)
ISBN-13 978-1-4020-4993-4 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

To those who search for meaning amidst ambiguous appearances...

Contents

Preface

Information extraction (IE) is usually defined as the process of selectively structuring and combining data that are explicitly stated or implied in one or more natural language documents. This process involves a semantic classification of certain pieces of information and is considered as a light form of text understanding. IE has a history going back at least three decades and different approaches have been developed. Currently, there is a considerable interest in using these technologies for information retrieval, since there is an increasing need to localize precise information in documents, for instance, as the answer to a question, rather than retrieving the entire document or a list of documents. Advanced retrieval models such as language modeling answer that need by building a probabilistic model of the content of a document. Question answering systems are trying to take the next step by inferring answers to a natural language question from a document collection. In these and other information retrieval models a semantic classification of entities, relations between entities, and of semantically relevant portions of texts (phrases, sentences, maybe passages) is very valuable to advance the state of the art of text searching. When talking about a semantic Web, semantic classification becomes of primordial importance, but also in other tasks that involve information selection and filtering, such as text summarization and information synthesis from different documents, IE is an indispensable preprocessing step.

The book gives an overview and explanation of the most successful and efficient algorithms for information extraction, and how they could be integrated in an information retrieval system. Special focus is on approaches that are fairly generic, i.e., that can be applied for processing heterogeneous document collections rather than a specific domain or text type and that are as much language independent as possible. The book contains a wealth of information on past and current milestones in information extraction, on necessary knowledge and resources involved in the extraction processes, and on the final aims of an extraction system. Additionally, a very important focus is on current statistical and machine learning techniques for information detection and classification. In an information retrieval context, these techniques can be used to learn and fine tune traditional knowledge engineered rules and patterns.

The book has grown from the results of a project on *Generic Technology for Information Extraction from Texts* (researched at the Katholieke Universiteit Leuven, Belgium) from 2000-2004 and sponsored by the Institute for the Promotion of Innovation by Science and Technology in Flanders) and from a graduate course on *Text Based Information Retrieval* taught at the same university to students in Artificial Intelligence, Informatics, and Electrical Engineering. This book is meant to give a comprehensive overview of the field of information extraction, especially as it is used in an information retrieval context. It is aimed at researchers in information extraction or related disciplines, but the many illustrations and real world examples make it also suitable as a handbook for students.

Acknowledgements

First, I would like to thank Rik De Busser, who is currently a Ph.D. student in Linguistics at La Trobe University in Melbourne, Australia, and who helped with the redaction of the first three chapters of this book. Secondly, I thank Prof. Jos Dumortier, the director of the *Interdisciplinary Centre for Law and Information Technology* at the K.U.Leuven for the opportunities given to our research group *Legal Informatics and Information Retrieval*. Many thanks go the staff of this group and especially to Roxana Angheluta, Jan De Beer, Koen Deschacht and Wim De Smet for participating in weekly project discussions. I am very grateful to Prof. Danny De Schreye, Head of the Informatics Department in the Faculty of Engineering for the many encouragements to pursue research in the domain of artificial intelligence. I sincerely thank Prof. Paul Van Orshoven, dean of our faculty, Prof. Yves Willems, former dean of the Faculty of Engineering, and Prof. Marc Vervenne, Rector of the K.U.Leuven, who gave me a marvelous chance to continue and perpetuate my research and teaching in the domain of information retrieval. Information extraction from written texts by a machine is a first step towards their automatic understanding. The task compares to decoding the symbols of an old language and gradually learning the meaning of the inscriptions. I am very grateful to the late Prof. Jan Quaegebeur (K.U.Leuven) and Prof. John Callender (University of California Los Angeles, USA). A long time ago they arouse in me the profound interest in content extraction from texts. I surely must thank Dr. Donna Harman (NIST, USA), Prof. Ed Hovy (University of Southern California, USA) and Prof. Karen Sparck Jones (University of Cambridge, UK) for creating influential and valuable ideas in the fields of information retrieval and text analysis. The final thank you goes to my family for their patience on Sunday afternoons.