

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*

All Rights Reserved  
© 2006 Springer

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

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

## Contents

Preface	xi
Acknowledgements	xiii
<b>1 Information Extraction and Information Technology</b>	<b>1</b>
1.1 Defining Information Extraction	1
1.2 Explaining Information Extraction	4
1.2.1 Unstructured Data	4
1.2.2 Extraction of Semantic Information	5
1.2.3 Extraction of Specific Information	7
1.2.4 Classification and Structuring	8
1.3 Information Extraction and Information Retrieval	10
1.3.1 Information Overload	10
1.3.2 Information Retrieval	12
1.3.3 Searching for the Needle	13
1.4 Information Extraction and Other Information Processing Tasks	16
1.5 The Aims of the Book	17
1.6 Conclusions	20
1.7 Bibliography	21
<b>2 Information Extraction from an Historical Perspective</b>	<b>23</b>
2.1 Introduction	23
2.2 An Historical Overview	23
2.2.1 Early Origins	23
2.2.2 Frame Theory	26
2.2.3 Use of Resources	28
2.2.4 Machine Learning	31
2.2.5 Some Afterthoughts	32
2.3 The Common Extraction Process	36
2.3.1 The Architecture of an Information Extraction System	36
2.3.2 Some Information Extraction Tasks	38
2.4 A Cascade of Tasks	42
2.5 Conclusions	42
2.6 Bibliography	43

---

<b>3 The Symbolic Techniques</b>	<b>47</b>
3.1 Introduction	47
3.2 Conceptual Dependency Theory and Scripts	47
3.3 Frame Theory	54
3.4 Actual Implementations of the Symbolic Techniques	58
3.4.1 Partial Parsing	58
3.4.2 Finite State Automata	58
3.5 Conclusions	63
3.6 Bibliography	63
<b>4 Pattern Recognition</b>	<b>65</b>
4.1 Introduction	65
4.2 What is Pattern Recognition?	66
4.3 The Classification Scheme	70
4.4 The Information Units to Extract	71
4.5 The Features	73
4.5.1 Lexical Features	76
4.5.2 Syntactic Features	80
4.5.3 Semantic Features	83
4.5.4 Discourse Features	84
4.6 Conclusions	85
4.7 Bibliography	86
<b>5 Supervised Classification</b>	<b>89</b>
5.1 Introduction	89
5.2 Support Vector Machines	92
5.3 Maximum Entropy Models	101
5.4 Hidden Markov Models	107
5.5 Conditional Random Fields	114
5.6 Decision Rules and Trees	118
5.7 Relational Learning	121
5.8 Conclusions	122
5.9 Bibliography	122

---

<b>6 Unsupervised Classification Aids</b>	<b>127</b>
6.1 Introduction	127
6.2 Clustering	129
6.2.1 Choice of Features	129
6.2.2 Distance Functions between Two Objects	130
6.2.3 Proximity Functions between Two Clusters	133
6.2.4 Algorithms	133
6.2.5 Number of Clusters	134
6.2.6 Use of Clustering in Information Extraction	136
6.3 Expansion	138
6.4 Self-training	141
6.5 Co-training	144
6.6 Active Learning	145
6.7 Conclusions	147
6.8 Bibliography	148
<b>7 Integration of Information Extraction in Retrieval Models</b>	<b>151</b>
7.1 Introduction	151
7.2 State of the Art of Information Retrieval	152
7.3 Requirements of Retrieval Systems	155
7.4 Motivation of Incorporating Information Extraction	156
7.5 Retrieval Models	160
7.5.1 Vector Space Model	162
7.5.2 Language Model	163
7.5.3 Inference Network Model	167
7.5.4 Logic Based Model	170
7.6 Data Structures	171
7.7 Conclusions	176
7.8 Bibliography	176
<b>8 Evaluation of Information Extraction Technologies</b>	<b>179</b>
8.1 Introduction	179
8.2 Intrinsic Evaluation of Information Extraction	180
8.2.1 Classical Performance Measures	181
8.2.2 Alternative Performance Measures	184
8.2.3 Measuring the Performance of Complex Extractions	187



---

8.3	Extrinsic Evaluation of Information Extraction in Retrieval	191
8.4	Other Evaluation Criteria	193
8.5	Conclusions	195
8.6	Bibliography	196
<b>9</b>	<b>Case Studies</b>	<b>199</b>
9.1	Introduction	199
9.2	Generic versus Domain Specific Character	200
9.3	Information Extraction from News Texts	202
9.4	Information Extraction from Biomedical Texts	204
9.5	Intelligence Gathering	209
9.6	Information Extraction from Business Texts	213
9.7	Information Extraction from Legal Texts	214
9.8	Information Extraction from Informal Texts	216
9.9	Conclusions	218
9.10	Bibliography	219
<b>10</b>	<b>The Future of Information Extraction in a Retrieval Context</b>	<b>225</b>
10.1	Introduction	225
10.2	The Human Needs and the Machine Performances	227
10.3	Most Important Findings	229
10.3.1	Machine Learning	229
10.3.2	The Generic Character of Information Extraction	230
10.3.3	The Classification Schemes	230
10.3.4	The Role of Paraphrasing	231
10.3.5	Flexible Information Needs	232
10.3.6	The Indices	233
10.4	Algorithmic Challenges	233
10.4.1	The Features	234
10.4.2	A Cascaded Model for Information Extraction	234
10.4.3	The Boundaries of Information Units	236
10.4.4	Extracting Sharable Knowledge	237
10.4.5	Expansion	237
10.4.6	Algorithms for Retrieval	238
10.5	The Future of IE in a Retrieval Context	239
10.6	Bibliography	241
	Index	243

## 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.