
**AUTOMATIC INDEXING
AND ABSTRACTING
OF DOCUMENT TEXTS**

THE KLUWER INTERNATIONAL SERIES ON INFORMATION RETRIEVAL

Series Editor

W. Bruce Croft

*University of Massachusetts
Amherst, MA 01003*

Also in the Series:

**MULTIMEDIA INFORMATION RETRIEVAL: Content-Based
Information Retrieval from Large Text and Audio Databases**

by Peter Schäuble

ISBN: 0-7923-9899-8

INFORMATION RETRIEVAL SYSTEMS

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
Katholieke Universiteit Leuven, Belgium

KLUWER ACADEMIC PUBLISHERS
New York / Boston / Dordrecht / London / Moscow

eBook ISBN: 0-306-47017-9
Print ISBN: 0-792-37793-1

©2002 Kluwer Academic Publishers
New York, Boston, Dordrecht, London, Moscow

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: <http://www.kluweronline.com>
and Kluwer's eBookstore at: <http://www.ebooks.kluweronline.com>

to Peter, Michael, and Laura

CONTENTS

PREFACE	xi
ACKNOWLEDGEMENTS	xv
PART I	
THE INDEXING AND ABSTRACTING ENVIRONMENT	1
1. THE NEED FOR INDEXING AND ABSTRACTING TEXTS	3
1. Introduction	3
2. Electronic Documents	4
3. Communication through Natural Language Text	5
4. Understanding of Natural Language Text: The Cognitive Process	7
5. Understanding of Natural Language Text: The Automated Process	8
6. Important Concepts in Information Retrieval and Selection	10
7. General Solutions to the Information Retrieval Problem	17
8. The Need for Better Automatic Indexing and Abstracting Techniques	22
2. THE ATTRIBUTES OF TEXT	27
1. Introduction	27
2. The Study of Text	27
3. An Overview of Some Common Text Types	29
4. Text Described at a Micro Level	30
5. Text Described at a Macro Level	38
6. Conclusions	47
3. TEXT REPRESENTATIONS AND THEIR USE	49
1. Introduction	49
2. Definitions	49

3.	Representations that Characterize the Content of Text	50
4.	Intellectual Indexing and Abstracting	55
5.	Use of the Text Representations	60
6.	A Note about the Storage of Text Representations	69
7.	Characteristics of Good Text Representations	70
8.	Conclusions	73
PART II		
METHODS OF AUTOMATIC INDEXING AND ABSTRACTING		75
4.	AUTOMATIC INDEXING: THE SELECTION OF NATURAL LANGUAGE INDEX TERMS	77
1.	Introduction	77
2.	A Note about Evaluation	78
3.	Lexical Analysis	78
4.	Use of a Stoplist	80
5.	Stemming	81
6.	The Selection of Phrases	84
7.	Index Term Weighting	89
8.	Alternative Procedures for Selecting Index Terms	98
9.	Selection of Natural Language Index Terms: Accomplishments and Problems	101
10.	Conclusions	102
5.	AUTOMATIC INDEXING: THE ASSIGNMENT OF CONTROLLED LANGUAGE INDEX TERMS	103
1.	Introduction	103
2.	A Note about Evaluation	104
3.	Thesaurus Terms	106
4.	Subject and Classification Codes	111
5.	Learning Approaches to Text Categorization	115
6.	Assignment of Controlled Language Index Terms: Accomplishments and Problems	131
7.	Conclusions	132
6.	AUTOMATIC ABSTRACTING: THE CREATION OF TEXT SUMMARIES	133
1.	Introduction	133
2.	A Note about Evaluation	134
3.	The Text Analysis Step	136
4.	The Transformation Step	148

5. Generation of the Abstract	150
6. Text Abstracting: Accomplishments and Problems	152
7. Conclusions	154
PART III	
APPLICATIONS	155
7. TEXT STRUCTURING AND CATEGORIZATION WHEN SUMMARIZING LEGAL CASES	157
1. Introduction	157
2. Text Corpus and Output of the System	158
3. Methods: The Use of a Text Grammar	161
4. Results and Discussion	165
5. Contributions of the Research	168
6. Conclusions	172
8. CLUSTERING OF PARAGRAPHS WHEN SUMMARIZING LEGAL CASES	173
1. Introduction	173
2. Text Corpus and Output of the System	174
3. Methods: The Clustering Techniques	175
4. Results and Discussion	181
5. Contributions of the Research	188
6. Conclusions	190
9. THE CREATION OF HIGHLIGHT ABSTRACTS OF MAGAZINE ARTICLES	191
1. Introduction	191
2. Text Corpus and Output of the System	192
3. Methods: The Use of a Text Grammar	194
4. Results and Discussion	201
5. Contributions of the Research	204
6. Conclusions	205
10. THE ASSIGNMENT OF SUBJECT DESCRIPTORS TO MAGAZINE ARTICLES	207
1. Introduction	207
2. Text Corpus and Output of the System	208
3. Methods: Supervised Learning of Classification Patterns	210
4. Results and Discussion	217
5. Contributions of the Research	224
6. Conclusions	225

SUMMARY AND FUTURE PROSPECTS	227
1. Summary	227
2. Future Prospects	235
REFERENCES	237
SUBJECT INDEX	261

PREFACE

Currently, we are confronted with a huge quantity of electronic documents that are written in natural language text. We are good at creating the texts, but not as capable at managing their information content. The documents are stored on computer disks or on CD-ROMS to form large collections. Retrieval systems, search engines, browsing tools, and other information management software are at our disposal for selecting relevant documents or information from the collections. When present-day retrieval and information selection tools operate on the content of document texts or make it accessible, they are not sufficiently powerful to identify documents or information that might be relevant to their users.

Text indexing and abstracting are old techniques for organizing the content of natural language text. These processes create a short description or characterization of the original text, which is called a text representation or representative and has a recognized and accepted format. Indexing commonly extracts from or assigns to the text a set of single words or phrases that function as index terms of the text. Words or phrases of the text are commonly called natural language index terms. When the assigned words or phrases come from a fixed vocabulary, they are called controlled language index terms. The index terms, besides reflecting content, can be used as access points or identifiers of the text in the document collection. Abstracting results in a reduced representation of the content of the text. The abstract usually has the form of a continuous, coherent text or of a profile that structures certain information of the original text.

The idea and the first attempts of automating text indexing and abstracting go back to the end of the 1950s. What at that time was a progressive theory has now become an absolute necessity. The manual task of indexing and abstracting is simply not feasible with the ever expanding collections of textual documents (e.g., on the Internet). Automatic indexing and abstracting, besides being efficient, probably produce a more consistent,

objective and more complete final product. The process of automatic indexing and abstracting starts when the text is already electronically stored and can be regarded as a string of characters (including spaces and punctuation marks). As in the case of manual indexing and abstracting, the automated method entails content analysis of the text, selection and generalization of information, and translation into a final form. Current systems that index and abstract texts generate text representations that are similar to those prepared by humans in terms of content and format (e.g., set of index terms, abstract in the form of a fluent text). This is because retrieval and other text management systems support these representations.

Text representations are used in systems that manage document contents. The majority of them are document retrieval systems. The ultimate goal of indexing and abstracting in text retrieval is an effective retrieval operation, so that more relevant and less irrelevant items are found. It is currently assumed that the major problem in current retrieval systems is capturing the meaning that a document may have for its user. Thus, progress can be made by accurately defining a user's need. We do not deny the importance of an accurate representation of the user's need, but accurately defining information needs will only work well with richer semantic representations of the textual content of documents produced by automatic indexing and abstracting. Current text representations that are automatically generated are only crude reflections of the content of document texts. They are often restricted to some terms that frequently occur in the text, to all words from the beginning of the text, or to sentences that contain frequent terms.

An intuitive solution to generating rich semantic representations of the natural language texts is to analyze them and to interpret their words and phrases based on complete linguistic, domain world, and contextual knowledge. Given the current state of natural language processing, this is not possible, nor is it always desirable. Linguistic knowledge refers to the lexical, syntactic and semantic properties of the texts' language and the typical properties of the discourse. Domain knowledge describes the concepts and subconcepts of the subject domain and their relationships. The contextual knowledge concerns communicative knowledge, which deals with the preferences and needs of those who use information in the texts. A working hypothesis in the domain of information retrieval is that valid text representations can be made without subjecting text to a complete and complex language-dependent processing. This is a valid hypothesis to start with. In the course of this book we will develop and defend a few lesser hypotheses. First, it is stated that knowledge of discourse structures – whether inherent or not to the text type or genre – and of surface linguistic cues that signal them is very useful for automatically indexing and abstracting a text's content. This knowledge also allows us to focus upon certain information in texts that is relevant for specific communication

needs. It is also possible to learn discourse structures from texts with statistical techniques. Finally, domain knowledge is important to identify topical concepts in texts. Knowledge of concepts and their variant textual patterns can be learned from example texts.

The book has ambitious objectives: to study automatic indexing and abstracting in all its facets and to describe the latest novel techniques in automatic indexing and abstracting. In addition, it confronts the many problems that automatic indexing and abstracting of text pose. Although, the book focuses upon indexing and abstracting of written text, many findings are also important for spoken textual documents, which are increasingly used for communication and storage of information.

This book is organized as follows:

The first part, “*The Indexing and Abstracting Environment*”, places the problem in a broad context and defines important concepts of the book. The first chapter, “*The Need for Indexing and Abstracting Texts*”, justifies the urgency for better methods for automatic indexing and abstracting of text content. From a broad viewpoint, some pertinent problems in information retrieval and text management in general are discussed. The current solutions to these problems are outlined. In the course of this chapter, the real need for better automatic indexing and abstracting techniques becomes clear. The second chapter of this part, “*The Attributes of Text*”, elaborates on the features of text. It gives an overview of the different components and structures that make up a text. The last chapter of this part, “*Text Representations and their Use*”, discusses the properties and use of different text representations for document and information retrieval.

The second part of the book, “*Methods of Automatic Indexing and Abstracting*”, gives an overview of existing techniques of automatic indexing and abstracting. Currently, such a detailed overview is lacking in the literature. The different chapters deal with the major forms of text representations: “*Automatic Indexing: The Selection of Natural Language Index Terms*”, “*Automatic Indexing: The Assignment of Controlled Language Index Terms*”, and “*Automatic Abstracting: The Creation of Text Summaries*”. The content of this part provides the context for the applications discussed in the third part and justifies the choice of certain techniques in the applications.

The third part of the book considers “*Applications*”. Four important problems are described for two collections of texts, written in Dutch. The problems mainly regard indexing with controlled language index terms, text classification, and abstracting. One corpus contains the texts of legal cases, while the other is composed of magazine articles. Solutions are proposed and tested with the help of software for indexing and abstracting, which the author designed and implemented. The applications elaborate on novel techniques and improve existing ones for automatic indexing and

abstracting. The first chapter “*Text Structuring and Categorization when Summarizing Legal Cases*”, deals with a successful initial categorization and structuring of the criminal cases. A text grammar is employed to represent knowledge of case structures, of concepts typical for the criminal law domain, and of the information focus. In the next chapter, “*Clustering of Paragraphs when Summarizing Legal Cases*”, a number of lengthy passages of the legal cases are summarized by extracting representative paragraphs and key terms. The techniques for identifying the representative textual units rely upon the distribution of lexical items in the legal texts and demonstrate the usefulness of clustering based on the selection of representative objects. In the third chapter entitled “*The Creation of Highlight Abstracts of Magazine Articles*”, the portability of the text grammar approach for text abstracting is demonstrated, in the process of creating highlight abstracts of magazine articles. Here, the typical discourse patterns of news stories are taken advantage of. In the last chapter of this part, “*The Assignment of Subject Descriptors to Magazine Articles*”, the technique learns the typical text patterns of the broad subject classes of the articles from a limited set of example texts and applies this knowledge for assigning subject descriptors to new, previously unseen articles.

The book concludes with a summary, an overview of the contributions of the research, and directions for future research.

The book is interdisciplinary. Its subject, “*Automatic Indexing and Abstracting of Document Texts*”, is an essential element of information retrieval research. Information retrieval is a discipline that has its foundations in information science, computer science, and statistics. The research especially studies text and its automatic analysis. This is the research domain of computational linguistics, a subdiscipline of computer science. Because of the nature of the two text corpora used in the research, legal texts and magazine articles, the research encounters the disciplines of law and communication science. The field of cognitive science is touched upon when the cognitive process of indexing and abstracting yields models for the automatic processes.

ACKNOWLEDGEMENTS

This publication is a slightly shortened version of my doctoral dissertation defended on June 28, 1999 in the Faculty of Sciences at the Katholieke Universiteit Leuven, Belgium. Though it is impossible to acknowledge the contributions of those who have helped me, I would like to mention those whose assistance was direct and vital to the completion of this work.

The far origins of this book lie in my work on ancient Egyptian language guided by Professor J. Quaegebeur (Katholieke Universiteit Leuven, Belgium) and Professor J. Callender (University of California Los Angeles, California, U.S.A.) who have awoken in me a profound interest in the analysis of language and texts.

I am very grateful to Professor J. Dumortier, my supervisor, who gave me the magnificent chance to explore the subject of this book. He gave me the opportunity to work at the Interdisciplinary Centre for Law and Information Technology (ICRI) (Katholieke Universiteit Leuven, Belgium), which is a very stimulating environment for creative research. It was under his supervision that the research contained in this volume started some five years ago.

I must express my gratitude to the advisors of my doctoral dissertation at the Katholieke Universiteit Leuven, Belgium: Professor H. Olivié, Professor L. Verstraelen, and Professor J. Dumortier. Their continuous encouragement have greatly facilitated its preparation. I thank Professor H. Olivié for his helpful advice.

I also thank the members of the examination jury, Professor D. De Schreye (Katholieke Universiteit Leuven, Belgium), Professor J. Leysen (Koninklijke Militaire School, Belgium), and Professor J. Hobbs (Stanford Research Institute, California, U.S.A.), who by their remarks and suggestions allowed me to achieve the final goals of this publication.

It is with deep respect that I thank Professor A. Oosterlinck, Rector of the Katholieke Universiteit Leuven, Belgium, and Professor J. Herbots, Dean of

the Faculty of Law for having given me the opportunity to work at the Katholieke Universiteit Leuven, Belgium. I must also thank Professor J. Berlamont, Dean of the Faculty of Applied Sciences, who has given me the opportunity of pursuing a doctoral training in Computer Science at the Katholieke Universiteit Leuven, Belgium, and Professor L. Vanquickenbome, Dean of the Faculty of Sciences, who allowed me to defend my doctoral degree. I thank Professor S. Vandewalle for taking care of my dossier regarding the doctoral training.

I am most indebted to my colleague Drs. C. Uyttendaele who provided invaluable help in one of the projects described in the book and who translated most of the legal texts from Dutch to English. I am also grateful to Mrs. T. Bouwen for the verification of some of the results contained in this publication. I thank Dr. W. Wetterstrom (Harvard University, Massachusetts, U.S.A.) who helped me correcting my English in the preface and summary. I wish also to thank Professor J. Zeleznikow (La Trobe University, Australia) for his helpful comments. I thank the anonymous reviewers of my research papers that are integrated in this book.

Additionally, I am grateful to Dr. C. Belmans and Ir. J. Huens (Katholieke Universiteit Leuven, Belgium) and Mr. L. Misseeuw and Mr. P. Huyghe (Roularta Media Group) for their technical assistance in making the text corpora available. I am grateful to Mrs. N. Verbiest for the administrative support. I wish to thank my family and colleagues for their continuous encouragement.

Finally, I would like to express my gratitude to the organizations that provided me with grant support during my studies and research: the Belgian American Educational Foundation (BAEF), the Ministerie van Onderwijs Bestuur van het Hoger Onderwijs en het Wetenschappelijk Onderzoek, the Onderzoeksfonds K.U.Leuven, the Nationaal Fonds voor Wetenschappelijk Onderzoek (NFWO), the Vlaams Instituut voor de bevordering van het Wetenschappelijk-Technologisch onderzoek in de industrie (IWT), the Vlaamse Leergangen Leuven, and the Vlaamse Wetenschappelijke Stichting.