
INFORMATION STORAGE AND RETRIEVAL SYSTEMS

Theory and Implementation

Second Edition

THE KLUWER INTERNATIONAL SERIES ON INFORMATION RETRIEVAL

Series Editor

W. Brace Croft
University of Massachusetts, Amherst

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: 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 STORAGE AND RETRIEVAL SYSTEMS

Theory and Implementation

Second Edition

by

Gerald J. Kowalski

Central Intelligence Agency

Mark T. Maybury

The MITRE Corporation

KLUWER ACADEMIC PUBLISHERS

NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47031-4
Print ISBN: 0-792-37924-1

©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>

This book is dedicated to my parents who taught me the value of a strong work ethic and my daughters, Kris and Kara, who continue to support my taking on new challenges. (Jerry Kowalski)

CONTENTS

Preface	xi
1 Introduction to Information Retrieval Systems	1
1.1 Definition of Information Retrieval System	2
1.2 Objectives of Information Retrieval Systems	4
1.3 Functional Overview	10
1.3.1 Item Normalization	10
1.3.2 Selective Dissemination of Information	16
1.3.3 Document Database Search	18
1.3.4 Index Database Search	18
1.3.5 Multimedia Database Search	20
1.4 Relationship to Database Management Systems	20
1.5 Digital Libraries and Data Warehouses	21
1.6 Summary	24
2 Information Retrieval System Capabilities	27
2.1 Search Capabilities	28
2.1.1 Boolean Logic	29
2.1.2 Proximity	30
2.1.3 Contiguous Word Phrases	31
2.1.4 Fuzzy Searches	32
2.1.5 Term Masking	32
2.1.6 Numeric and Date Ranges	33
2.1.7 Concept and Thesaurus Expansions	34
2.1.8 Natural Language Queries	36
2.1.9 Multimedia Queries	37
2.2 Browse Capabilities	38
2.2.1 Ranking	38
2.2.2 Zoning	40
2.2.3 Highlighting	40
2.3 Miscellaneous Capabilities	41
2.3.1 Vocabulary Browse	41
2.3.2 Iterative Search and Search History Log	42
2.3.3 Canned Query	43
2.3.4 Multimedia	43
2.4 Z39.50 and WAIS Standards	44
2.5 Summary	47
3. Cataloging and Indexing	51
3.1 History and Objectives of Indexing	52
3.1.1 History	52
3.1.2 Objectives	54

3.2	Indexing Process	56
3.2.1	Scope of Indexing	57
3.2.2	Precoordination and Linkages	58
3.3	Automatic Indexing	58
3.3.1	Indexing by Term	61
3.3.2	Indexing by Concept	63
3.3.3	Multimedia Indexing	64
3.4	Information Extraction	65
3.5	Summary	68
4.	Data Structure	71
4.1	Introduction to Data Structure	72
4.2	Stemming Algorithms	73
4.2.1	Introduction to the Stemming Process	74
4.2.2	Porter Stemming Algorithm	75
4.2.3	Dictionary Look-up Stemmers	77
4.2.4	Successor Stemmers	78
4.2.5	Conclusions	80
4.3	Inverted File Structure	82
4.4	N-Gram Data Structures	85
4.4.1	History	86
4.4.2	N-Gram Data Structure	87
4.5	PAT Data Structure	88
4.6	Signature File Structure	93
4.7	Hypertext and XML Data Structures	94
4.7.1	Definition of Hypertext Structure	95
4.7.2	Hypertext History	97
4.7.3	XML	98
4.8	Hidden Markov Models	99
4.9	Summary	102
5.	Automatic Indexing	105
5.1	Classes of Automatic Indexing	105
5.2	Statistical Indexing	108
5.2.1	Probabilistic Weighting	108
5.2.2	Vector Weighting	111
5.2.2.1	Simple Term Frequency Algorithm	113
5.2.2.2	Inverse Document Frequency	116
5.2.2.3	Signal Weighting	117
5.2.2.4	Discrimination Value	119
5.2.2.5	Problems With Weighting Schemes	120
5.2.2.6	Problems With the Vector Model	121
5.2.3	Bayesian Model	122
5.3	Natural Language	123
5.3.1	Index Phrase Generation	125
5.3.2	Natural Language Processing	128
5.4	Concept Indexing	130
5.5	Hypertext Linkages	132
5.6	Summary	135
6.	Document and Term Clustering	139
6.1	Introduction to Clustering	140
6.2	Thesaurus Generation	143

6.2.1	Manual Clustering	144
6.2.2	Automatic Term Clustering	145
6.2.2.1	Complete Term Relation Method	146
6.2.2.2	Clustering Using Existing Clusters	151
6.2.2.3	One Pass Assignments	153
6.3	Item Clustering	154
6.4	Hierarchy of Clusters	156
6.5	Summary	160
7.	User Search Techniques	165
7.1	Search Statements and Binding	166
7.2	Similarity Measures and Ranking	167
7.2.1	Similarity Measures	168
7.2.2	Hidden Markov Model Techniques	173
7.2.3	Ranking Algorithms	174
7.3	Relevance Feedback	175
7.4	Selective Dissemination of Information Search	179
7.5	Weighted Searches of Boolean Systems	186
7.6	Searching the INTERNET and Hypertext	191
7.7	Summary	194
8.	Information Visualization	199
8.1	Introduction to Information Visualization	200
8.2	Cognition and Perception	203
8.2.1	Background	203
8.2.2	Aspects of Visualization Process	204
8.3	Information Visualization Technologies	208
8.4	Summary	218
9.	Text Search Algorithms	221
9.1	Introduction to Text Search Techniques	221
9.2	Software Text Search Algorithms	225
9.3	Hardware Text Search Systems	233
9.4	Summary	238
10.	Multimedia Information Retrieval	241
10.1	Spoken Language Audio Retrieval	242
10.2	Non-Speech Audio Retrieval	244
10.3	Graph Retrieval	245
10.4	Imagery Retrieval	246
10.5	Video Retrieval	249
10.6	Summary	255

1.1. Information System Evaluation	257
11.1 Introduction to Information System Evaluation	257
11.2 Measures Used in System Evaluations	260
11.3 Measurement Example - TREC Results	267
11.4 Summary	278
References	281
Subject Index	313

PREFACE - Second Edition

The Second Edition incorporates the latest developments in the area of Information Retrieval. The major addition to this text is descriptions of the automated indexing of multimedia documents. Items in information retrieval are now considered to be a combination of text along with graphics, audio, image and video data types. What this means from an Information Retrieval System design and implementation is discussed.

The growth of the Internet and the availability of enormous volumes of data in digital form have necessitated intense interest in techniques to assist the user in locating data of interest. The Internet has over 800 million indexable pages as of February 1999 (Lawrence-99.) Other estimates from International Data Corporation suggest that the number is closer to 1.5 billion pages and the number will grow to 8 billion pages by the Fall 2000 (<http://news.excite.com/news/zd/000510/21/inktomi-chief-gets>, 11 May 2000.) Buried on the Internet are both valuable nuggets to answer questions as well as a large quantity of information the average person does not care about. The Digital Library effort is also progressing, with the goal of migrating from the traditional book environment to a digital library environment.

The challenge to both authors of new publications that will reside on this information domain and developers of systems to locate information is to provide the information and capabilities to sort out the non-relevant items from those desired by the consumer. In effect, as we proceed down this path, it will be the computer that determines what we see versus the human being. The days of going to a library and browsing the new book shelf are being replaced by electronic searching the Internet or the library catalogs. Whatever the search engines return will constrain our knowledge of what information is available. An understanding of Information Retrieval Systems puts this new environment into perspective for both the creator of documents and the consumer trying to locate information.

This book provides a theoretical and practical explanation of the latest advancements in information retrieval and their application to existing systems. It takes a system approach, discussing all aspects of an Information Retrieval System. The importance of the Internet and its associated hypertext linked structure are put into perspective as a new type of information retrieval data structure. The total system approach also includes discussion of the human interface and the importance of information visualization for identification of relevant information. With the availability of large quantities of multi-media on the Internet (audio, video, images), Information Retrieval Systems need to address multi-modal retrieval. The Second Edition has been expanded to address how Information Retrieval Systems are

expanded to include search and retrieval on multi-modal sources. The theoretical metrics used to describe information systems are expanded to discuss their practical application in the uncontrolled environment of real world systems.

The primary goal of writing this book is to provide a college text on Information Retrieval Systems. But in addition to the theoretical aspects, the book maintains a theme of practicality that puts into perspective the importance and utilization of the theory in systems that are being used by anyone on the Internet. The student will gain an understanding of what is achievable using existing technologies and the deficient areas that warrant additional research. The text provides coverage of all of the major aspects of information retrieval and has sufficient detail to allow students to implement a simple Information Retrieval System. The comparison algorithms from Chapter 11 can be used to compare how well each of the student's systems work.

The first three chapters define the scope of an Information Retrieval System. The theme, that the primary goal of an Information Retrieval System is to minimize the overhead associated in locating needed information, is carried throughout the book. Chapter 1 provides a functional overview of an Information Retrieval System and differentiates between an information system and a Database Management System (DBMS). Chapter 2 focuses on the functions available in an information retrieval system. An understanding of the functions and why they are needed help the reader gain an intuitive feeling for the application of the technical algorithms presented later. Chapter 3 provides the background on indexing and cataloging that formed the basis for early information systems and updates it with respect to the new digital data environment.

Chapter 4 provides a discussion on word stemming and its use in modern systems. It also introduces the underlying data structures used in Information Retrieval Systems and their possible applications. This is the first introduction of hypertext data structures and their applicability to information retrieval. Chapters 5, 6 and 7 go into depth on the basis for search in Information Retrieval Systems. Chapter 5 looks at the different approaches to information systems search and the extraction of information from documents that will be used during the query process. Chapter 6 describes the techniques that can be used to cluster both terms from documents for statistical thesauri and the documents themselves. Thesauri can assist searches by query term expansion while document clustering can expand the initial set of found documents to similar documents. Chapter 7 focuses on the search process as a mapping between the user's search need and the documents in the system. It introduces the importance of relevance feedback in expanding the user's query and discusses the difference between search techniques against an existing database versus algorithms that are used to disseminate newly received items to user's mail boxes.

Chapter 8 introduces the importance of information visualization and its impact on the user's ability to locate items of interest in large systems. It provides the background on cognition and perception in human beings and then how that knowledge is applied to organizing information displays to help the user locate

needed information. Chapter 9 describes text-scanning techniques as a special search application within information retrieval systems. It describes the hardware and software approaches to text search.

Chapter 10 discusses how information retrieval is applied to multimedia sources. Information retrieval techniques that apply to audio, imagery, graphic and video data types are described along with likely future advances in these areas. The impacts of including these data types on information retrieval systems are discussed throughout the book.

Chapter 11 describes how to evaluate Information Retrieval Systems focusing on the theoretical and standard metrics used in research to evaluate information systems. Problems with the measurement's techniques in evaluating operational systems are discussed along with possible required modifications. Existing system capabilities are highlighted by reviewing the results from the Text Retrieval Conferences (TREC).

Although this book covers the majority of the technologies associated with Information retrieval Systems, the one area omitted is search and retrieval of different languages. This area would encompass discussions in search modifications caused by different languages such as Chinese and Arabic that introduce new problems in interpretation of word boundaries and "assumed" contextual interpretation of word meanings, cross language searches (mapping queries from one language to another language, and machine translation of results. Most of the search algorithms discussed in Information retrieval are applicable across languages. Status of search algorithms in these areas can be found in non-U.S. journals and TREC results.