Modern Information Retrieval

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

ACM Press
New York

Addison-Wesley

Harlow, England • Reading, Massachusetts
Menlo Park, California • New York
Don Mills, Ontario • Amsterdam • Bonn
Sydney • Singapore • Tokyo • Madrid
San Juan • Milan • Mexico City • Seoul • Taipei

Addison Wesley Longman Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the World.

Typeset in Computer Modern by 56
Printed and bound in the United States of America

First printed 1999

ISBN 0-201-39829-X

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloguing-in-Publication Data
Baeza-Yates, R.(Ricardo)
Modern information retrieval / Ricardo Baeza-Yates, Berthier Ribeiro-Neto.
p. cm.
Includes bibliographical references and index.
ISBN 0-201-39829-X
1. Information storage and retieval systems. I. Ribeiro, Berthier de Araújo Neto, 1960- . II.Title.
Z667.B34 1999
025.04–dc21 99-10033
CIP

Preface

Information retrieval (IR) has changed considerably in recent years with the expansion of the World Wide Web and the advent of modern and inexpensive graphical user interfaces and mass storage devices. As a result, traditional IR textbooks have become quite out of date and this has led to the introduction of new IR books. Nevertheless, we believe that there is still great need for a book that approaches the field in a rigorous and complete way from a computer-science perspective (as opposed to a user-centered perspective). This book is an effort to partially fulfill this gap and should be useful for a first course on information retrieval as well as for a graduate course on the topic.

The book comprises two portions which complement and balance each other. The core portion includes nine chapters authored or coauthored by the designers of the book. The second portion, which is fully integrated with the first, is formed by six state-of-the-art chapters written by leading researchers in their fields. The same notation and glossary are used in all the chapters. Thus, despite the fact that several people have contributed to the text, this book is really much more a textbook than an edited collection of chapters written by separate authors. Furthermore, unlike a collection of chapters, we have carefully designed the contents and organization of the book to present a cohesive view of all the important aspects of modern information retrieval.

From IR models to indexing text, from IR visual tools and interfaces to the Web, from IR multimedia to digital libraries, the book provides both breadth of coverage and richness of detail. It is our hope that, given the now clear relevance and significance of information retrieval to modern society, the book will contribute to further disseminate the study of the discipline at information science, computer science, and library science departments throughout the world.

Ricardo Baeza-Yates, Santiago, Chile
Berthier Ribeiro-Neto, Belo Horizonte, Brazil
January, 1999

To Helena, Rosa, and our children

Amo los libros
exploradores,
libros con bosque o nieve,
profundidad o cielo

de Oda al Libro (I),

Pablo Neruda

I love books
that explore,
books with a forest or snow,
depth or sky

from Ode to the Book (I),

Pablo Neruda

território de homens livres
que será nosso país
e será pátria de todos.
Irmãos, cantai ese mundo
que não verei, mas virá
um dia, dentro de mil anos,
talvez mais... não tenho pressa.

de Cidade Prevista no livro
A Rosa do Povo, 1945

Carlos Drummond de Andrade

territory of free men
that will be our country
and will be the nation of all
Brothers, sing this world
which I'll not see, but which will come
one day, in a thousand years,
maybe more... no hurry.

from Prevised City in the book
The Rose of the People, 1945

Carlos Drummond de Andrade

Acknowledgements

We would like to deeply thank the various people who, during the several months in which this endeavor lasted, provided us with useful and helpful assistance. Without their care and consideration, this book would likely not have matured.

First, we would like to thank all the chapter contributors, for their dedication and interest. To Elisa Bertino, Eric Brown, Barbara Catania, Christos Faloutsos, Elena Ferrari, Ed Fox, Marti Hearst, Gonzalo Navarro, Edie Rasmussen, Ohm Sornil, and Nivio Ziviani, who contributed with writings that reflect expertise we certainly do not fully profess ourselves. And for all their patience throughout an editing and cross-reviewing process which constitutes a rather difficult balancing act.

Second, we would like to thank all the people who demonstrated interest in publishing this book, particularly Scott Delman and Doug Sery.

Third, we would like to commend the interest, encouragement, and great job done by Addison Wesley Longman throughout the overall process, represented by Keith Mansfield, Karen Sutherland, Bridget Allen, David Harrison, Sheila Chatten, Helen Hodge and Lisa Talbot. The reviewers they contacted read an early (and rather preliminary) proposal of this book and provided us with good feedback and invaluable insights. The chapter on Parallel and Distributed IR was moved from the part on Applications of IR (where it did not fit well) to the part on Text IR due to the objective argument of an unknown referee. A separate chapter on Retrieval Evaluation was only included after another zealous referee strongly made the case for the importance of this subject.

Fourth, we would like to thank all the people who discussed this project with us. Doug Oard provided us with an early critique of the proposal. Gary Marchionini was an earlier supporter and provided us with useful contacts during the process. Bruce Croft encouraged our efforts from the beginning. Alberto Mendelzon provided us with an initial proposal and a compilation of references for the chapter on searching the Web. Ed Fox found time in a rather busy schedule to provide us with an insightful review of the introduction (which resulted in a great improvement) and a thorough review of the chapter on Modeling. Marti Hearst demonstrated interest in our proposal early on, provided assistance throughout the editing process, and has been an enthusiastic supporter and partner.

Fifth, we thank the support of our institutions, the Departments of Computer Science of the University of Chile and of the Federal University of Minas Gerais, as well as the funding provided by national research agencies (CNPq in Brazil and CONICYT in Chile) and international collaboration projects, in particular CYTED project VII.13 AMYRI (Environment for Information Managing and Retrieval in the World Wide Web) and Finep project SIAM (Information Systems for Mobile Computers) under the Pronex program.

Most important, to Helena, Rosa, and our children, who put up with a string of trips abroad, lost weekends, and odd working hours.

Contents

Biographies

Biographies of Main Authors

Ricardo Baeza-Yates received a bachelor degree in Computer Science in 1983 from the University of Chile. Later, he received an MSc in Computer Science (1985), a professional title in electrical engineering (1985), and an MEng in EE (1986) from the same university. He received his PhD in Computer Science from the University of Waterloo, Canada, in 1989. He has been the president of the Chilean Computer Science Society (SCCC) from 1992 to 1995 and from 1997 to 1998. During 1993, he received the Organization of the American States award for young researchers in exact sciences. Currently, he is a full professor at the Computer Science Department of the University of Chile, where he was the chairperson in the period 1993 to 1995. He is coauthor of the second edition of the *Handbook of Algorithms and Data Structures*, Addison-Wesley, 1991; and coeditor of *Information Retrieval: Algorithms and Data Structures*, Prentice Hall, 1992. He has also contributed several papers to journals published by professional organizations such as ACM, IEEE, and SIAM.

His research interests include algorithms and data structures, text retrieval, graphical interfaces, and visualization applied to databases. He currently coordinates an IberoAmerican project on models and techniques for searching the Web financed by the Spanish agency Cyted. He has been a visiting professor or an invited speaker at several conferences and universities around the world, as well as referee for several journals, conferences, NSF, etc. He is a member of the ACM, AMS, EATCS, IEEE, SCCC, and SIAM.

Berthier Ribeiro-Neto received a bachelor degree in Math, a BS degree in Electrical Engineering, and an MS degree in Computer Science, all from the Federal University of Minas Gerais, Brazil. In 1995, he was awarded a Ph.D. in Computer Science from the University of California at Los Angeles. Since then, he has been with the Computer Science Department of the Federal University of Minas Gerais where he is an Associate Professor.

His main interests are information retrieval systems, digital libraries, interfaces for the Web, and video on demand. He has been involved in a number of research projects financed through Brazilian national agencies such as the Ministry of Science and Technology (MCT) and the National Research Council (CNPq). From the projects currently underway, the two main ones deal with wireless information systems (project SIAM financed within program PRONEX) and video on demand (project ALMADEM financed within program PROTEM III). Dr Ribeiro-Neto is also involved with an IberoAmerican project on information systems for the Web coordinated by Professor Ricardo Baeza-Yates. He was the chair of SPIRE'98 (String Processing and Information Retrieval South American Symposium), is the chair of SBBD'99 (Brazilian Symposium on Databases), and has been on the committee of several conferences in Brazil, in South America and in the USA. He is a member of ACM, ASIS, and IEEE.

Biographies of Contributors

Elisa Bertino is Professor of Computer Science in the Department of Computer Science of the University of Milano where she heads the Database Systems Group. She has been a visiting researcher at the IBM Research Laboratory (now Almaden) in San Jose, at the Microelectronics and Computer Technology Corporation in Austin, Texas, and at Rutgers University in Newark, New Jersey. Her main research interests include object-oriented databases, distributed databases, deductive databases, multimedia databases, interoperability of heterogeneous systems, integration of artificial intelligence and database techniques, and database security. In those areas, Professor Bertino has published several papers in refereed journals, and in proceedings of international conferences and symposia. She is a coauthor of the books *Object-Oriented Database Systems — Concepts and Architectures*, Addison-Wesley 1993; *Indexing Techniques for Advanced Database Systems*, Kluwer 1997; and *Intelligent Database Systems*, Addison-Wesley forthcoming. She is or has been on the editorial boards of the following scientific journals: the *IEEE Transactions on Knowledge and Data Engineering*, the *International Journal of Theory and Practice of Object Systems*, the *Very Large Database Systems (VLDB) Journal*, the *Parallel and Distributed Database Journal*, the *Journal of Computer Security, Data & Knowledge Engineering*, and the *International Journal of Information Technology.*

Eric Brown has been a Research Staff Member at the IBM T.J. Watson Research Center in Yorktown Heights, NY, since 1995. Prior to that he was a Research Assistant at the Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst. He holds a BSc from the University of Vermont and an MS and PhD from the University of Massachusetts, Amherst. Dr. Brown conducts research in large scale information retrieval systems, automatic text categorization, and hypermedia systems for digital libraries and knowledge management. He has published a number of papers in the field of information retrieval.

Barbara Catania is a researcher at the University of Milano, Italy. She received an MS degree in Information Sciences in 1993 from the University of

Genova and a PhD in Computer Science in 1998 from the University of Milano. She has also been a visiting researcher at the European Computer-Industry Research Center, Munich, Germany. Her main research interests include multimedia databases, constraint databases, deductive databases, and indexing techniques in object-oriented and constraint databases. In those areas, Dr Catania has published several papers in refereed journals, and in proceedings of international conferences and symposia. She is also a coauthor of the book *Indexing Techniques for Advanced Database Systems*, Kluwer 1997.

Christos Faloutsos received a BSc in Electrical Engineering (1981) from the National Technical University of Athens, Greece and an MSc and PhD in Computer Science from the University of Toronto, Canada. Professor Faloutsos is currently a faculty member at Carnegie Mellon University. Prior to joining CMU he was on the faculty of the Department of Computer Science at the University of Maryland, College Park. He has spent sabbaticals at IBM-Almaden and AT&T Bell Labs. He received the Presidential Young Investigator Award from the National Science Foundation in 1989, two 'best paper' awards (SIGMOD 94, VLDB 97), and three teaching awards. He has published over 70 refereed articles and one monograph, and has filed for three patents. His research interests include physical database design, searching methods for text, geographic information systems, indexing methods for multimedia databases, and data mining.

Elena Ferrari is an Assistant Professor at the Computer Science Department of the University of Milano, Italy. She received an MS in Information Sciences in 1992 and a PhD in Computer Science in 1998 from the University of Milano. Her main research interests include multimedia databases, temporal object-oriented data models, and database security. In those areas, Dr Ferrari has published several papers in refereed journals, and in proceedings of international conferences and symposia. She has been a visiting researcher at George Mason University in Fairfax, Virginia, and at Rutgers University in Newark, New Jersey.

Dr Edward A. Fox holds a PhD and MS in Computer Science from Cornell University, and a BS from MIT. Since 1983 he has been at Virginia Polytechnic Institute and State University (Virginia Tech), where he serves as Associate Director for Research at the Computing Center, Professor of Computer Science, Director of the Digital Library Research Laboratory, and Director of the Internet Technology Innovation Center. He served as vice chair and chair of ACM SIGIR from 1987 to 1995, helped found the ACM conferences on multimedia and digital libraries, and serves on a number of editorial boards. His research is focused on digital libraries, multimedia, information retrieval, WWW/Internet, educational technologies, and related areas.

Marti Hearst is an Assistant Professor at the University of California Berkeley in the School of Information Management and Systems. From 1994 to 1997 she was a Member of the Research Staff at Xerox PARC. She received her BA, MS, and PhD degrees in Computer Science from the University of California at Berkeley. Professor Hearst's research focuses on user interfaces and robust language analysis for information access systems, and on furthering the understanding of how people use and understand such systems.

Gonzalo Navarro received his first degrees in Computer Science from ESLAI (Latin American Superior School of Informatics) in 1992 and from the University of La Plata (Argentina) in 1993. In 1995 he received his MSc in Computer Science from the University of Chile, obtaining a PhD in 1998. Between 1990 and 1993 he worked at IBM Argentina, on the development of interactive applications and on research on multimedia and hypermedia. Since 1994 he has worked in the Department of Computer Science of the University of Chile, doing research on design and analysis of algorithms, textual databases, and approximate search. He has published a number of papers and also served as referee on different journals (*Algorithmica*, *TOCS*, *TOIS*, etc.) and at conferences (SIGIR, CPM, ESA, etc.).

Edie Rasmussen is an Associate Professor in the School of Information Sciences, University of Pittsburgh. She has also held faculty appointments at institutions in Malaysia, Canada, and Singapore. Dr Rasmussen holds a BSc from the University of British Columbia and an MSc degree from McMaster University, both in Chemistry, an MLS degree from the University of Western Ontario, and a PhD in Information Studies from the University of Sheffield. Her current research interests include indexing and information retrieval in text and multimedia databases.

Ohm Sornil is currently a PhD candidate in the Department of Computer Science at Virginia Polytechnic and State University and a scholar of the Royal Thai Government. He received a BEng in Electrical Engineering from Kasetsart University, Thailand, in 1993 and an MS in Computer Science from Syracuse University in 1997. His research interests include information retrieval, digital libraries, communication networks, and hypermedia.

Nivio Ziviani is a Professor of Computer Science at the Federal University of Minas Gerais in Brazil, where he heads the laboratory for Treating Information. He received a BS in Mechanical Engineering from the Federal University of Minas Gerais in 1971, an MSc in Informatics from the Catholic University of Rio in 1976, and a PhD in Computer Science from the University of Waterloo, Canada, in 1982. He has obtained several research funds from the Brazilian Research Council (CNPq), Brazilian Agencies CAPES and FINEP, Spanish Agency CYTED (project AMYRI), and private institutions. He currently coordinates a four year project on Web and wireless information systems (called SIAM) financed by the Brazilian Ministry of Science and Technology. He is cofounder of the Miner Technology Group, owner of the Miner Family of agents to search the Web. He is the author of several papers in journals and conference proceedings covering topics in the areas of algorithms and data structures, information retrieval, text indexing, text searching, text compression, and related areas. Since January of 1998, he is the editor of the 'News from Latin America' section in the Bulletin of the European Association for Theoretical Computer Science. He has been chair and member of the program committee of several conferences and is a member of ACM, EATICS and SBC.

Chapter 1
Introduction

1.1 Motivation

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested. Unfortunately, characterization of the *user information need* is not a simple problem. Consider, for instance, the following hypothetical user information need in the context of the World Wide Web (or just the Web):

> Find all the pages (documents) containing information on college tennis teams which: (1) are maintained by an university in the USA and (2) participate in the NCAA tennis tournament. To be relevant, the page must include information on the national ranking of the team in the last three years and the email or phone number of the team coach.

Clearly, this full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the user must first translate this information need into a *query* which can be processed by the search engine (or IR system).

In its most common form, this translation yields a set of keywords (or index terms) which summarizes the description of the user information need. Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user. The emphasis is on the retrieval of *information* as opposed to the retrieval of *data*.

1.1.1 Information versus Data Retrieval

Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need. In fact, the user of an IR system is concerned more with retrieving *information* about a subject than with retrieving data which satisfies a given query. A data retrieval language aims at retrieving all objects which satisfy clearly defined conditions such as those in a regular expression or in a relational algebra expression. Thus, for a data retrieval system, a single erroneous object among a thousand retrieved objects means total failure. For an information retrieval system, however, the retrieved objects might be inaccurate and small errors are likely to go unnoticed. The main reason for this difference is that information retrieval usually deals with natural language text which is not always well structured and could be semantically ambiguous. On the other hand, a data retrieval system (such as a relational database) deals with data that has a well defined structure and semantics.

Data retrieval, while providing a solution to the user of a database system, does not solve the problem of retrieving information about a subject or topic. To be effective in its attempt to satisfy the user information need, the IR system must somehow 'interpret' the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This 'interpretation' of a document content involves extracting syntactic and semantic information from the document text and using this information to match the user information need. The difficulty is not only knowing how to extract this information but also knowing how to use it to decide relevance. Thus, the notion of *relevance* is at the center of information retrieval. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible.

1.1.2 Information Retrieval at the Center of the Stage

In the past 20 years, the area of information retrieval has grown well beyond its primary goals of indexing text and searching for useful documents in a collection. Nowadays, research in IR includes modeling, document classification and categorization, systems architecture, user interfaces, data visualization, filtering, languages, etc. Despite its maturity, until recently, IR was seen as a narrow area of interest mainly to librarians and information experts. Such a tendentious vision prevailed for many years, despite the rapid dissemination, among users of modern personal computers, of IR tools for multimedia and hypertext applications. In the beginning of the 1990s, a single fact changed once and for all these perceptions — the introduction of the World Wide Web.

The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedent sharing of ideas and information in a scale never seen before. Its success is based on the conception of a standard user interface which is always the same no matter what computational environment is used to run the interface. As a result, the user is shielded from details of communication protocols, machine location, and operating systems. Further, any user can create his own Web documents and make them point to any other Web documents without restrictions. This is a key aspect because it turns the Web into a new publishing medium accessible to everybody. As an immediate consequence, any Web user can push his personal agenda with little effort and almost at no cost. This universe without frontiers has attracted tremendous attention from millions of people everywhere since the very beginning. Furthermore, it is causing a revolution in the way people use computers and perform their daily tasks. For instance, home shopping and home banking are becoming very popular and have generated several hundred million dollars in revenues.

Despite so much success, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. For instance, to satisfy his information need, the user might navigate the space of Web links (i.e., the *hyperspace*) searching for information of interest. However, since the hyperspace is vast and almost unknown, such a navigation task is usually inefficient. For naive users, the problem becomes harder, which might entirely frustrate all their efforts. The main obstacle is the absence of a well defined underlying data model for the Web, which implies that information definition and structure is frequently of low quality. These difficulties have attracted renewed interest in IR and its techniques as promising solutions. As a result, almost overnight, IR has gained a place with other technologies at the center of the stage.

1.1.3 Focus of the Book

Despite the great increase in interest in information retrieval, modern textbooks on IR with a broad (and extensive) coverage of the various topics in the field are still difficult to find. In an attempt to partially fulfill this gap, this book presents an overall view of research in IR from a computer scientist's perspective. This means that the focus of the book is on computer algorithms and techniques used in information retrieval systems. A rather distinct viewpoint is taken by librarians and information science researchers, who adopt a human-centered interpretation of the IR problem. In this interpretation, the focus is on trying to understand how people interpret and use information as opposed to how to structure, store, and retrieve information automatically. While most of this book is dedicated to the computer scientist's viewpoint of the IR problem, the human-centered viewpoint is discussed to some extent in the last two chapters.

We put great emphasis on the integration of the different areas which are closed related to the information retrieval problem and thus, should be treated together. For that reason, besides covering text retrieval, library systems, user interfaces, and the Web, this book also discusses visualization, multimedia retrieval, and digital libraries.

1.2 Basic Concepts

The effective retrieval of relevant information is directly affected both by the *user task* and by the *logical view of the documents* adopted by the retrieval system, as we now discuss.

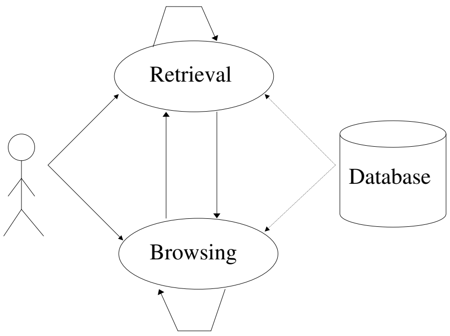

Figure 1.1 Interaction of the user with the retrieval system through distinct tasks.

1.2.1 The User Task

The user of a retrieval system has to translate his information need into a query in the language provided by the system. With an information retrieval system, this normally implies specifying a set of words which convey the semantics of the information need. With a data retrieval system, a query expression (such as, for instance, a regular expression) is used to convey the constraints that must be satisfied by objects in the answer set. In both cases, we say that the user searches for useful information executing a *retrieval* task.

Consider now a user who has an interest which is either poorly defined or which is inherently broad. For instance, the user might be interested in documents about car racing in general. In this situation, the user might use an interactive interface to simply look around in the collection for documents related to car racing. For instance, he might find interesting documents about Formula 1 racing, about car manufacturers, or about the '24 Hours of Le Mans.' Furthermore, while reading about the '24 Hours of Le Mans', he might turn his attention to a document which provides directions to Le Mans and, from there, to documents which cover tourism in France. In this situation, we say that the user is *browsing* the documents in the collection, not searching. It is still a process of retrieving information, but one whose main objectives are not clearly defined in the beginning and whose purpose might change during the interaction with the system.

In this book, we make a clear distinction between the different tasks the user of the retrieval system might be engaged in. His task might be of two distinct types: information or data *retrieval* and *browsing*. Classic information retrieval systems normally allow information or data retrieval. Hypertext systems are usually tuned for providing quick browsing. Modern digital library and Web interfaces might attempt to combine these tasks to provide improved retrieval capabilities. However, combination of retrieval and browsing is not yet a well established approach and is not the dominant paradigm.

Figure 1.1 illustrates the interaction of the user through the different tasks we identify. Information and data retrieval are usually provided by most modern information retrieval systems (such as Web interfaces). Further, such systems might also provide some (still limited) form of browsing. While combining information and data retrieval with browsing is not yet a common practice, it might become so in the future.

Both retrieval and browsing are, in the language of the World Wide Web, 'pulling' actions. That is, the user requests the information in an interactive manner. An alternative is to do retrieval in an automatic and permanent fashion using software agents which *push* the information towards the user. For instance, information useful to a user could be extracted periodically from a news service. In this case, we say that the IR system is executing a particular retrieval task which consists of *filtering* relevant information for later inspection by the user. We briefly discuss filtering in Chapter 2.

1.2.2 Logical View of the Documents

Due to historical reasons, documents in a collection are frequently represented through a set of index terms or keywords. Such keywords might be extracted directly from the text of the document or might be specified by a human subject (as frequently done in the information sciences arena). No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a *logical view of the document.* For a precise definition of the concept of a document and its characteristics, see Chapter 6.

Modern computers are making it possible to represent a document by its full set of words. In this case, we say that the retrieval system adopts a *full text* logical view (or representation) of the documents. With very large collections, however, even modern computers might have to reduce the set of representative keywords. This can be accomplished through the elimination of *stopwords* (such as articles and connectives), the use of *stemming* (which reduces distinct words to their common grammatical root), and the identification of noun groups (which eliminates adjectives, adverbs, and verbs). Further, compression might be employed. These operations are called *text operations* (or transformations) and are covered in detail in Chapter 7. Text operations reduce the complexity of the document representation and allow moving the logical view from that of a full text to that of a set of *index terms.*

The full text is clearly the most complete logical view of a document but its usage usually implies higher computational costs. A small set of categories (generated by a human specialist) provides the most concise logical view of a document but its usage might lead to retrieval of poor quality. Several intermediate logical views (of a document) might be adopted by an information retrieval system as illustrated in Figure 1.2. Besides adopting any of the intermediate representations, the retrieval system might also recognize the internal structure normally present in a document (e.g., chapters, sections, subsections, etc.). This

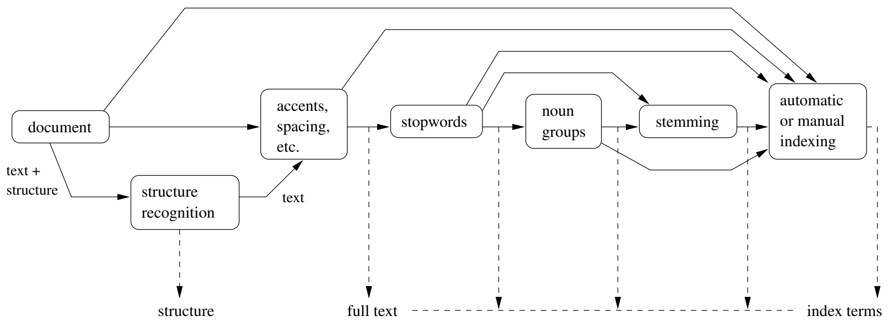

Figure 1.2 Logical view of a document: from full text to a set of index terms.

information on the structure of the document might be quite useful and is required by structured text retrieval models such as those discussed in Chapter 2.

As illustrated in Figure 1.2, we view the issue of logically representing a document as a continuum in which the logical view of a document might shift (smoothly) from a full text representation to a higher level representation specified by a human subject.

1.3 Past, Present, and Future

1.3.1 Early Developments

For approximately 4000 years, man has organized information for later retrieval and usage. A typical example is the table of contents of a book. Since the volume of information eventually grew beyond a few books, it became necessary to build specialized data structures to ensure faster access to the stored information. An old and popular data structure for faster information retrieval is a collection of selected words or concepts with which are associated pointers to the related information (or documents) — the *index*. In one form or another, indexes are at the core of every modern information retrieval system. They provide faster access to the data and allow the query processing task to be speeded up. A detailed coverage of indexes and their usage for searching can be found in Chapter 8.

For centuries, indexes were created manually as categorization hierarchies. In fact, most libraries still use some form of categorical hierarchy to classify their volumes (or documents), as discussed in Chapter 14. Such hierarchies have usually been conceived by human subjects from the library sciences field. More recently, the advent of modern computers has made possible the construction of large indexes automatically. Automatic indexes provide a view of the retrieval problem which is much more related to the system itself than to the user need.

In this respect, it is important to distinguish between two different views of the IR problem: a computer-centered one and a human-centered one.

In the computer-centered view, the IR problem consists mainly of building up efficient indexes, processing user queries with high performance, and developing ranking algorithms which improve the 'quality' of the answer set. In the human-centered view, the IR problem consists mainly of studying the behavior of the user, of understanding his main needs, and of determining how such understanding affects the organization and operation of the retrieval system. According to this view, keyword based query processing might be seen as a strategy which is unlikely to yield a good solution to the information retrieval problem in the long run.

In this book, we focus mainly on the computer-centered view of the IR problem because it continues to be dominant in the market place.

1.3.2 Information Retrieval in the Library

Libraries were among the first institutions to adopt IR systems for retrieving information. Usually, systems to be used in libraries were initially developed by academic institutions and later by commercial vendors. In the first generation, such systems consisted basically of an automation of previous technologies (such as card catalogs) and basically allowed searches based on author name and title. In the second generation, increased search functionality was added which allowed searching by subject headings, by keywords, and some more complex query facilities. In the third generation, which is currently being deployed, the focus is on improved graphical interfaces, electronic forms, hypertext features, and open system architectures.

Traditional library management system vendors include Endeavor Information Systems Inc., Innovative Interfaces Inc., and EOS International. Among systems developed with a research focus and used in academic libraries, we distinguish Okapi (at City University, London), MELVYL (at University of California), and Cheshire II (at UC Berkeley). Further details on these library systems can be found in Chapter 14.

1.3.3 The Web and Digital Libraries

If we consider the search engines on the Web today, we conclude that they continue to use indexes which are very similar to those used by librarians a century ago. What has changed then?

Three dramatic and fundamental changes have occurred due to the advances in modern computer technology and the boom of the Web. First, it became a lot cheaper to have access to various sources of information. This allows reaching a wider audience than ever possible before. Second, the advances in all kinds of digital communication provided greater access to networks. This implies that the information source is available even if distantly located and that the access can be done quickly (frequently, in a few seconds). Third, the freedom to post whatever information someone judges useful has greatly contributed to the popularity of the Web. For the first time in history, many people have free access to a large publishing medium.

Fundamentally, low cost, greater access, and publishing freedom have allowed people to use the Web (and modern digital libraries) as a highly *interactive* medium. Such interactivity allows people to exchange messages, photos, documents, software, videos, and to 'chat' in a convenient and low cost fashion. Further, people can do it at the time of their preference (for instance, you can buy a book late at night) which further improves the convenience of the service. Thus, high interactivity is the fundamental and current shift in the communication paradigm. Searching the Web is covered in Chapter 13, while digital libraries are covered in Chapter 15.

In the future, three main questions need to be addressed. First, despite the high interactivity, people still find it difficult (if not impossible) to retrieve information relevant to their information needs. Thus, in the dynamic world of the Web and of large digital libraries, which techniques will allow retrieval of higher quality? Second, with the ever increasing demand for access, quick response is becoming more and more a pressing factor. Thus, which techniques will yield faster indexes and smaller query response times? Third, the quality of the retrieval task is greatly affected by the user interaction with the system. Thus, how will a better understanding of the user behavior affect the design and deployment of new information retrieval strategies?

1.3.4 Practical Issues

Electronic commerce is a major trend on the Web nowadays and one which has benefited millions of people. In an electronic transaction, the buyer usually has to submit to the vendor some form of credit information which can be used for charging for the product or service. In its most common form, such information consists of a credit card number. However, since transmitting credit card numbers over the Internet is not a safe procedure, such data is usually transmitted over a fax line. This implies that, at least in the beginning, the transaction between a new user and a vendor requires executing an off-line procedure of several steps before the actual transaction can take place. This situation can be improved if the data is encrypted for security. In fact, some institutions and companies already provide some form of encryption or automatic authentication for security reasons.

However, security is not the only concern. Another issue of major interest is privacy. Frequently, people are willing to exchange information as long as it does not become public. The reasons are many but the most common one is to protect oneself against misuse of private information by third parties. Thus, privacy is another issue which affects the deployment of the Web and which has not been properly addressed yet.

Two other very important issues are copyright and patent rights. It is far from clear how the wide spread of data on the Web affects copyright and patent laws in the various countries. This is important because it affects the business of building up and deploying large digital libraries. For instance, is a site which supervises all the information it posts acting as a publisher? And if so, is it responsible for a misuse of the information it posts (even if it is not the source)?

Additionally, other practical issues of interest include scanning, optical character recognition (OCR), and cross-language retrieval (in which the query is in one language but the documents retrieved are in another language). In this book, however, we do not cover practical issues in detail because it is not our main focus. The reader interested in details of practical issues is referred to the interesting book by Lesk [8].

1.4 The Retrieval Process

At this point, we are ready to detail our view of the retrieval process. Such a process is interpreted in terms of component subprocesses whose study yields many of the chapters in this book.

To describe the retrieval process, we use a simple and generic software architecture as shown in Figure 1.3. First of all, before the retrieval process can even be initiated, it is necessary to define the text database. This is usually done by the manager of the database, which specifies the following: (a) the documents to be used, (b) the operations to be performed on the text, and (c) the text model (i.e., the text structure and what elements can be retrieved). The text operations transform the original documents and generate a logical view of them.

Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the *inverted file* as indicated in Figure 1.3. The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times.

Given that the document database is indexed, the retrieval process can be initiated. The user first specifies a *user need* which is then parsed and transformed by the same text operations applied to the text. Then, *query operations* might be applied before the actual *query*, which provides a system representation for the user need, is generated. The query is then processed to obtain the *retrieved documents*. Fast query processing is made possible by the index structure previously built.

Before been sent to the user, the retrieved documents are ranked according to a *likelihood* of relevance. The user then examines the set of ranked documents in the search for useful information. At this point, he might pinpoint a subset of the documents seen as definitely of interest and initiate a *user feedback* cycle. In such a cycle, the system uses the documents selected by the user to change the query formulation. Hopefully, this modified query is a better representation

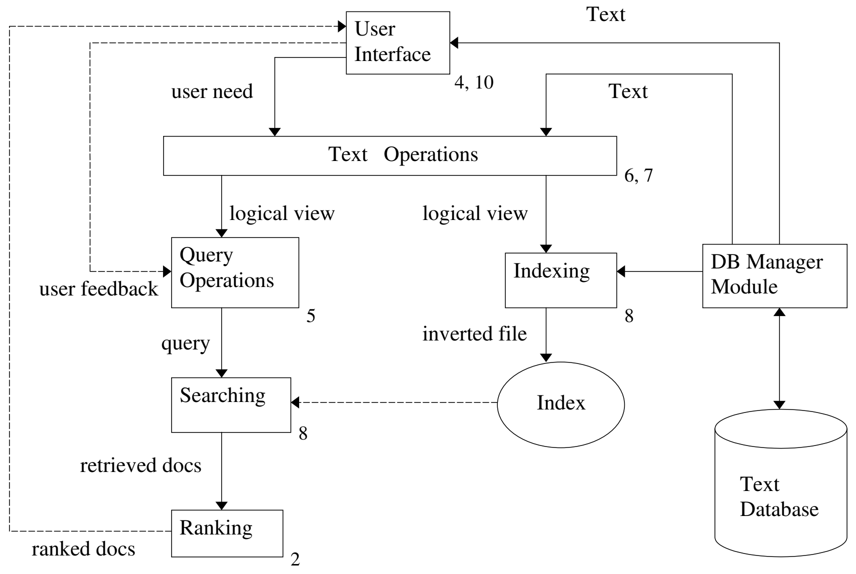

Figure 1.3 The process of retrieving information (the numbers beside each box indicate the chapters that cover the corresponding topic).

of the real user need.

The small numbers outside the lower right corner of various boxes in Figure 1.3 indicate the chapters in this book which discuss the respective subprocesses in detail. A brief introduction to each of these chapters can be found in section 1.5.

Consider now the user interfaces available with current information retrieval systems (including Web search engines and Web browsers). We first notice that the user almost never declares his information need. Instead, he is required to provide a direct representation for the query that the system will execute. Since most users have no knowledge of text and query operations, the query they provide is frequently inadequate. Therefore, it is not surprising to observe that poorly formulated queries lead to poor retrieval (as happens so often on the Web).

1.5 Organization of the Book

For ease of comprehension, this book has a straightforward structure in which four main parts are distinguished: text IR, human-computer interaction (HCI) for IR, multimedia IR, and applications of IR. *Text IR* discusses the classic problem of searching a collection of documents for useful information. *HCI for IR* discusses current trends in IR towards improved user interfaces and better data visualization tools. *Multimedia IR* discusses how to index document images and other binary data by extracting features from their content and how to search them efficiently. On the other hand, document images that are predominantly text (rather than pictures) are called *textual images* and are amenable to automatic extraction of keywords through metadescriptors, and can be retrieved using text IR techniques. *Applications of IR* covers modern applications of IR such as the Web, bibliographic systems, and digital libraries. Each part is divided into topics which we now discuss.

1.5.1 Book Topics

The four parts which compose this book are subdivided into eight topics as illustrated in Figure 1.4. These eight topics are as follows.

The topic *Retrieval Models & Evaluation* discusses the traditional models of searching text for useful information and the procedures for evaluating an information retrieval system. The topic *Improvements on Retrieval* discusses techniques for transforming the query and the text of the documents with the aim of improving retrieval. The topic *Efficient Processing* discusses indexing and searching approaches for speeding up the retrieval. These three topics compose the first part on Text IR.

The topic *Interfaces & Visualization* covers the interaction of the user with the information retrieval system. The focus is on interfaces which facilitate the process of specifying a query and provide a good visualization of the results.

The topic *Multimedia Modeling & Searching* discusses the utilization of multimedia data with information retrieval systems. The focus is on modeling, indexing, and searching multimedia data such as voice, images, and other binary data.

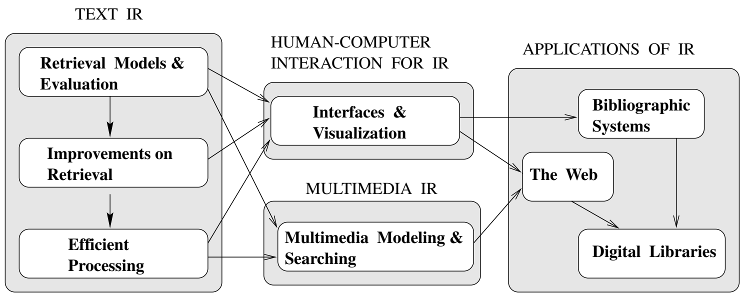

Figure 1.4 Topics which compose the book and their relationships.

The part on applications of IR is composed of three interrelated topics: *The Web*, *Bibliographic Systems*, and *Digital Libraries*. Techniques developed for the first two applications support the deployment of the latter.

The eight topics distinguished above generate the 14 chapters, besides this introduction, which compose this book and which we now briefly introduce.

1.5.2 Book Chapters

Figure 1.5 illustrates the overall structure of this book. The reasoning which yielded the chapters from 2 to 15 is as follows.

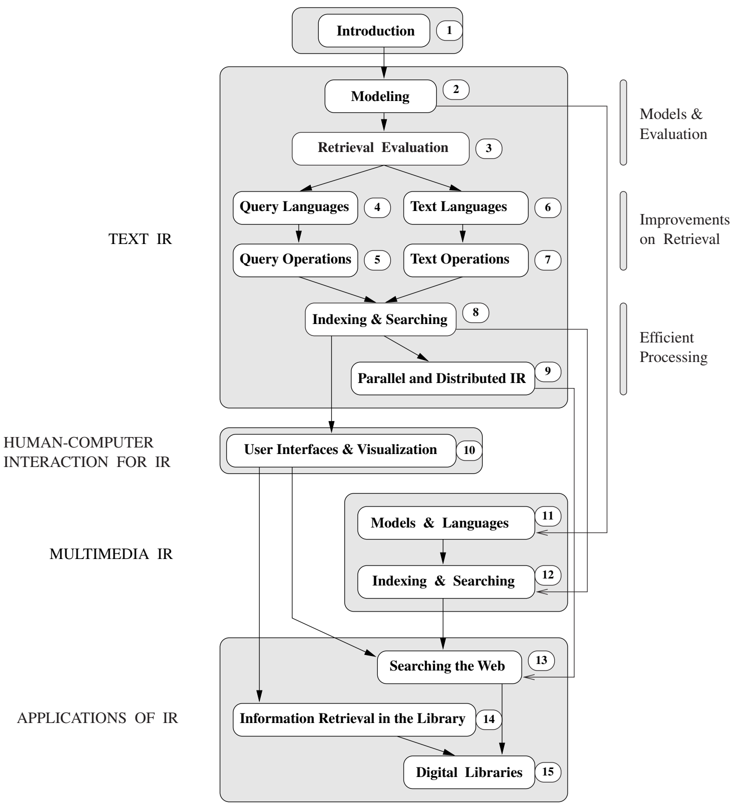

Figure 1.5 Structure of the book.

In the traditional keyword-based approach, the user specifies his information need by providing sets of keywords and the information system retrieves the documents which best approximate the user query. Also, the information system might attempt to rank the retrieved documents using some measure of relevance. This ranking task is critical in the process of attempting to satisfy the user information need and is the main goal of *modeling* in IR. Thus, information retrieval models are discussed early in Chapter 2. The discussion introduces many of the fundamental concepts in information retrieval and lays down much of the foundation for the subsequent chapters. Our coverage is detailed and broad. Classic models (Boolean, vector, and probabilistic), modern probabilistic variants (belief network models), alternative paradigms (extended Boolean, generalized vector, latent semantic indexing, neural networks, and fuzzy retrieval), structured text retrieval, and models for browsing (hypertext) are all carefully introduced and explained.

Once a new retrieval algorithm (maybe based on a new retrieval model) is conceived, it is necessary to evaluate its performance. Traditional evaluation strategies usually attempt to estimate the costs of the new algorithm in terms of time and space. With an information retrieval system, however, there is the additional issue of evaluating the relevance of the documents retrieved. For this purpose, text reference collections and evaluation procedures based on variables other than time and space are used. Chapter 3 is dedicated to the discussion of *retrieval evaluation.*

In traditional IR, queries are normally expressed as a set of keywords which is quite convenient because the approach is simple and easy to implement. However, the simplicity of the approach prevents the formulation of more elaborate querying tasks. For instance, queries which refer to both the structure and the content of the text cannot be formulated. To overcome this deficiency, more sophisticated query languages are required. Chapter 4 discusses various types of *query languages.* Since now the user might refer to the structure of a document in his query, this structure has to be defined. This is done by embedding the description of a document content and of its structure in a text language such as the Standard Generalized Markup Language (SGML). As illustrated in Figure 1.5, Chapter 6 is dedicated to the discussion of *text languages.*

Retrieval based on keywords might be of fairly low quality. Two possible reasons are as follows. First, the user query might be composed of too few terms which usually implies that the query context is poorly characterized. This is frequently the case, for instance, in the Web. This problem is dealt with through transformations in the query such as query expansion and user relevance feedback. Such *query operations* are covered in Chapter 5. Second, the set of keywords generated for a given document might fail to summarize its semantic content properly. This problem is dealt with through transformations in the text such as identification of noun groups to be used as keywords, stemming, and the use of a thesaurus. Additionally, for reasons of efficiency, text compression can be employed. Chapter 7 is dedicated to *text operations.*

Given the user query, the information system has to retrieve the documents which are related to that query. The potentially large size of the document collection (e.g., the Web is composed of millions of documents) implies that specialized indexing techniques must be used if efficient retrieval is to be achieved. Thus, to speed up the task of matching documents to queries, proper *indexing* and *search-*

ing techniques are used as discussed in Chapter 8. Additionally, query processing can be further accelerated through the adoption of *parallel and distributed IR* techniques as discussed in Chapter 9.

As illustrated in Figure 1.5, all the key issues regarding Text IR, from modeling to fast query processing, are covered in this book.

Modern user interfaces implement strategies which assist the user to form a query. The main objective is to allow him to define more precisely the context associated to his information need. The importance of query contextualization is a consequence of the difficulty normally faced by users during the querying process. Consider, for instance, the problem of quickly finding useful information in the Web. Navigation in hyperspace is not a good solution due to the absence of a logical and semantically well defined structure (the Web has no underlying logical model). A popular approach for specifying a user query in the Web consists of providing a set of keywords which are searched for. Unfortunately, the number of terms provided by a common user is small (typically, fewer than four) which usually implies that the query is vague. This means that new user interface paradigms which assist the user with the query formation process are required. Further, since a vague user query usually retrieves hundreds of documents, the conventional approach of displaying these documents as items of a scrolling list is clearly inadequate. To deal with this problem, new data visualization paradigms have been proposed in recent years. The main trend is towards visualization of a large subset of the retrieved documents at once and direct manipulation of the whole subset. *User interfaces* for assisting the user to form his query and current approaches for *visualization* of large data sets are covered in Chapter 10.

Following this, we discuss the application of IR techniques to multimedia data. The key issue is how to model, index, and search structured documents which contain multimedia objects such as digitized voice, images, and other binary data. *Models and query languages* for office and medical information retrieval systems are covered in Chapter 11. *Efficient indexing and searching* of multimedia objects is covered in Chapter 12. Some readers may argue that the models and techniques for multimedia retrieval are rather different from those for classic text retrieval. However, we take into account that images and text are usually together and that with the Web, other media types (such as video and audio) can also be mixed in. Therefore, we believe that in the future, all the above will be treated in a unified and consistent manner. Our book is a first step in that direction.

The final three chapters of the book are dedicated to applications of modern information retrieval: the Web, bibliographic systems, and digital libraries. As illustrated in Figure 1.5, Chapter 13 presents the Web and discusses the main problems related to the issue of searching the Web for useful information. Also, our discussion covers briefly the most popular search engines in the Web presenting particularities of their organization. Chapter 14 covers commercial document databases and online public access catalogs. Commercial document databases are still the largest information retrieval systems nowadays. LEXIS-NEXIS, for instance, has a database with 1.3 billion documents and attends to over 120 million query requests annually. Finally, Chapter 15 discusses modern digital libraries. Architectural issues, models, prototypes, and standards are all covered. The discussion also introduces the '5S' model (streams, structures, spaces, scenarios and societies) as a framework for providing theoretical and practical unification of digital libraries.

1.6 How to Use this Book

Although several people have contributed chapters for this book, it is really a textbook. The contents and the structure of the book have been carefully designed by the two main authors who also authored or coauthored nine of the 15 chapters in the book. Further, all the contributed chapters have been judiciously edited and integrated into a unifying framework that provides uniformity in structure and style, a common glossary, a common bibliography, and appropriate cross-references. At the end of each chapter, a discussion on research issues, trends, and selected bibliography is included. This discussion should be useful for graduate students as well as for researchers. Furthermore, the book is complemented by a Web page with additional information and resources.

1.6.1 Teaching Suggestions

This textbook can be used in many different areas including computer science (CS), information systems, and library science. The following list gives suggested contents for different courses at the undergraduate and graduate level, based on syllabuses of many universities around the world:

- **Information Retrieval** (Computer Science, undergraduate): this is the standard course for many CS programs. The minimum content should include Chapters 1 to 8 and Chapter 10, that is, most of the part on Text IR complemented with the chapter on user interfaces. Some specific topics of those chapters, such as more advanced models for IR and sophisticated algorithms for indexing and searching, can be omitted to fit a one term course. The chapters on Applications of IR can be mentioned briefly at the end.
- **Advanced Information Retrieval** (Computer Science, graduate): similar to the previous course but with more detailed coverage of the various chapters particularly modeling and searching (assuming the previous course as a requirement). In addition, Chapter 9 and Chapters 13 to 15 should be covered completely. Emphasis on research problems and new results is a must.
- **Information Retrieval** (Information Systems, undergraduate): this course is similar to the CS course, but with a different emphasis. It should include Chapters 1 to 7 and Chapter 10. Some notions from Chapter 8 are

useful but not crucial. At the end, the system-oriented parts of the chapters on Applications of IR, in particular those on Bibliographic Systems and Digital Libraries, must be covered (this material can be complemented with topics from [8]).

- **Information Retrieval** (Library Science, undergraduate): similar to the previous course, but removing the more technical and advanced material of Chapters 2, 5, and 7. Also, greater emphasis should be put on the chapters on Bibliographic Systems and Digital Libraries. The course should be complemented with a thorough discussion of the user-centered view of the IR problem (for example, using the book by Allen [1]).
- **Multimedia Retrieval** (Computer Science, undergraduate or graduate): this course should include Chapters 1 to 3, 6, and 11 to 15. The emphasis could be on multimedia itself or on the integration of classical IR with multimedia. The course can be complemented with one of the many books on this topic, which are usually more broad and technical.
- **Topics in IR** (Computer Science, graduate): many chapters of the book can be used for this course. It can emphasize modeling and evaluation or user interfaces and visualization. It can also be focused on algorithms and data structures (in that case, [2] and [17] are good complements). A multimedia focus is also possible, starting with Chapters 11 and 12 and using more specific books later on.
- **Topics in IR** (Information Systems or Library Science, graduate) similar to the above but with emphasis on non-technical parts. For example, the course could cover modeling and evaluation, query languages, user interfaces, and visualization. The chapters on applications can also be considered.
- **Web Retrieval and Information Access** (generic, undergraduate or graduate): this course should emphasize hypertext, concepts coming from networks and distributed systems and multimedia. The kernel should be the basic models of Chapter 2 followed by Chapters 3, 4, and 6. Also, Chapters 11 and 13 to 15 should be discussed.
- **Digital Libraries** (generic, undergraduate or graduate): This course could start with part of Chapters 2 to 4 and 6, followed by Chapters 10, 14, and 15. The kernel of the course could be based on the book by Lesk [8].

More bibliography useful for many of the courses above is discussed in the last section of this chapter.

1.6.2 The Book's Web Page

As IR is a very dynamic area nowadays, a book by itself is not enough. For that reason (and many others), the book has a Web home page located and mirrored in the following places (mirrors in USA and Europe are also planned):

- Brazil: `http://www.dcc.ufmg.br/irbook`
- Chile: `http://sunsite.dcc.uchile.cl/irbook`

Comments, suggestions, contributions, or mistakes found are welcome through email to the contact authors given on the Web page.

The Web page contains the Table of Contents, Preface, Acknowledgements, Introduction, Glossary, and other appendices to the book. It also includes exercises and teaching materials that will be increasing in volume and changing with time. In addition, a reference collection (containing 1239 documents on Cystic Fibrosis and 100 information requests with extensive relevance evaluation [14]) is available for experimental purposes. Furthermore, the page includes useful pointers to IR syllabuses in different universities, IR research groups, IR publications, and other resources related to IR and this book. Finally, any new important results or additions to the book as well as an errata will be made publicly available there.

1.7 Bibliographic Discussion

Many other books have been written on information retrieval, and due to the current widespread interest in the subject, new books have appeared recently. In the following, we briefly compare our book with these previously published works.

Classic references in the field of information retrieval are the books by van Rijsbergen [16] and Salton and McGill [12]. Our distinction between data and information retrieval is borrowed from the former. Our definition of the information retrieval process is influenced by the latter. However, almost 20 years later, both books are now outdated and do not cover many of the new developments in information retrieval.

Three more recent and also well known references in information retrieval are the book edited by Frakes and Baeza-Yates [2], the book by Witten, Moffat, and Bell [17], and the book by Lesk [8]. All these three books are complementary to this book. The first is focused on data structures and algorithms for information retrieval and is useful whenever quick prototyping of a known algorithm is desired. The second is focused on indexing and compression, and covers images besides text. For instance, our definition of a textual image is borrowed from it. The third is focused on digital libraries and practical issues such as history, distribution, usability, economics, and property rights. On the issue of computer-centered and user-centered retrieval, a generic book on information systems that takes the latter view is due to Allen [1].

There are other complementary books for specific chapters. For example, there are many books on IR and hypertext. The same is true for generic or specific multimedia retrieval, as images, audio or video. Although not an information retrieval title, the book by Rosenfeld and Morville [11] on information architecture of the Web, is a good complement to our chapter on searching the

Web. The book by Menasce and Almeida [10] demonstrates how to use queueing theory for predicting Web server performance. In addition, there are many books that explain how to find information on the Web and how to use search engines.

The reference edited by Sparck Jones and Willet [5], which was long awaited, is really a collection of papers rather than an edited book. The coherence and breadth of coverage in our book makes it more appropriate as a textbook in a formal discipline. Nevertheless, this collection is a valuable research tool. A collection of papers on cross-language information retrieval was recently edited by Grefenstette [3]. This book is a good complement to ours for people interested in this particular topic. Additionally, a collection focused on intelligent IR was edited recently by Maybury [9], and another collection on natural language IR edited by Strzalkowski will appear soon [15].

The book by Korfhage [6] covers a lot less material and its coverage is not as detailed as ours. For instance, it includes no detailed discussion of digital libraries, the Web, multimedia, or parallel processing. Similarly, the books by Kowalski [7] and Shapiro *et al.* [13] do not cover these topics in detail, and have a different orientation. Finally, the recent book by Grossman and Frieder [4] does not discuss the Web, digital libraries, or visual interfaces.

For people interested in research results, the main journals on IR are: *Journal of the American Society of Information Sciences* (*JASIS*) published by Wiley and Sons, *ACM Transactions on Information Systems, Information Processing & Management* (IP&M, Elsevier), *Information Systems* (Elsevier), *Information Retrieval* (Kluwer), and *Knowledge and Information Systems* (Springer). The main conferences are: ACM SIGIR International Conference on Information Retrieval, ACM International Conference on Digital Libraries (ACM DL), ACM Conference on Information Knowledge and Management (CIKM), and Text REtrieval Conference (TREC). Regarding events of regional influence, we would like to acknowledge the SPIRE (South American Symposium on String Processing and Information Retrieval) symposium.

References

[1] Bryce L. Allen. *Information Tasks: Toward a User-Centered Approach to Information Systems.* Academic Press, San Diego, CA, 1996.

[2] W.B. Frakes and R. Baeza-Yates. *Information Retrieval: Data Structures & Algorithms.* Prentice Hall, Englewood Cliffs, NJ, USA, 1992.

[3] Gregory Grefenstette. *Cross-Language Information Retrieval.* Kluwer Academic Publishers, Boston, USA, 1998.

[4] David A. Grossman and Ophir Frieder. *Information Retrieval: Algorithms and Heuristics.* Kluwer Academic Publishers, 1998.

[5] K. Sparck Jones and P. Willet. *Readings in Information Retrieval.* Morgan Kaufmann Publishers, Inc., 1997.

[6] Robert Korfhage. *Information Storage and Retrieval.* John Wiley & Sons, Inc., 1997.

[7] Gerald Kowalski. *Information Retrieval Systems, Theory and Implementation.* Kluwer Academic Publishers, Boston, USA, 1997.

[8] Michael Lesk. *Practical Digital Libraries; Books, Bytes, & Bucks.* Morgan Kaufmann, 1997.

[9] Mark T. Maybury. *Intelligent Multimedia Information Retrieval.* MIT Press, 1997.

[10] Daniel A. Menasce and Virgilio A.F. Almeida. *Capacity Planning for Web Performance: Metrics, Models, and Methods.* Prentice Hall, 1998.

[11] Louis Rosenfeld and Peter Morville. *Information Architecture for the World Wide Web.* O'Reilly & Associates, 1998.

[12] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Co., New York, 1983.

[13] Jacob Shapiro, Vladimir G. Voiskunskii, and Valery J. Frants. *Automated Information Retrieval : Theory and Text-Only Methods.* Academic Press, 1997.

[14] W.M. Shaw, J.B. Wood, R.E. Wood, and H.R. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13:347–366, 1991.

[15] Tomek Strzalkowski, editor. *Natural Language Information Retrieval.* Kluwer Academic Publishers, 1999. To appear.

[16] C.J. van Rijsbergen. *Information Retrieval.* Butterworths, 1979.

[17] I.H. Witten, A. Moffat, and T.C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Van Nostrand Reinhold, New York, 1994.