Modern Information Retrieval

the concepts and technology behind search

Second edition

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Modern Information Retrieval

The Concepts and Technology behind Search

Ricardo Baeza-Yates
Berthier Ribeiro-Neto

Second edition

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

Chapter 1

Introduction

1.1 Information Retrieval

Information retrieval (IR) is a broad area of Computer Science focused primarily on providing the users with easy access to information of their interest, as follows.

> Information retrieval deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organization of the information items should be such as to provide the users with easy access to information of their interest.

In terms of scope, the area has grown well beyond its early goals of indexing text and searching for useful documents in a collection. Nowadays, research in IR includes modeling, Web search, text classification, systems architecture, user interfaces, data visualization, filtering, languages.

In terms of research, the area may be studied from two rather distinct and complementary points of view: a computer-centered one and a human-centered one. In the computer-centered view, IR consists mainly of building up efficient indexes, processing user queries with high performance, and developing ranking algorithms to improve the results. In the human-centered view, IR consists mainly of studying the behavior of the user, of understanding their main needs, and of determining how such understanding affects the organization and operation of the retrieval system. In this book, we focus mainly on the computer-centered view of IR, which is dominant in academia and in the market place.

1.1.1 Early Developments

For more than 5,000 years, man has organized information for later retrieval and searching. In its most usual form, this has been done by compiling, storing, organizing,

and indexing clay tablets, hieroglyphics, papyrus rolls, and books. For holding the various items, special purpose buildings called *libraries*, from the Latin word *liber* for book, or *bibliothekes*, from the Greek word *biblion* for papyrus roll, are used.

The oldest known library was created in Elba, in the "Fertile Crescent", currently northern Syria, some time between 3,000 and 2,500 BC. In the seventh century BC, Assyrian king Ashurbanipal created the library of Nineveh, on the Tigris River (today, north of Iraq), which contained more than 30,000 clay tablets at the time of its destruction in 612 BC. By 300 BC, Ptolemy Soter, a Macedonian general, created the Great Library in Alexandria – the Egyptian city at the mouth of the Nile named after the Macedonian king Alexander the Great (356-323 BC). For seven centuries the Great Library, jointly with other major libraries in the city, made Alexandria the intellectual capital of the Western world [1164].

Since then, libraries have expanded and flourished. Nowadays, they are everywhere. They constitute the collective memory of the human race and their popularity is in the rising. In 2008 alone, people in the US visited their libraries some 1.3 billion times and checked out more than 2 billion items – an increase in both yearly figures of more than 10 percent [155].

Since the volume of information in libraries is always growing, it is necessary to build specialized data structures for fast search – *the indexes*. In one form or another, indexes are at the core of every modern information retrieval system. They provide fast access to the data and allow speeding up query processing, as we discuss in Chapter 9.

For centuries indexes have been created manually as sets of categories. Each category in the index is typically composed of labels that identify its associated topics and of pointers to the documents that discuss those topics. While these indexes are usually designed by library and information science researchers, the advent of modern computers has allowed the construction of large indexes automatically, which has accelerated the development of the area of Information Retrieval (IR).

Early developments in IR date back to research efforts conducted in the 50's by pioneers such as Hans Peter Luhn, Eugene Garfield, Philip Bagley, and Calvin Moores, this last one having allegedly coined the term *information retrieval* [1692]. In 1955, Allen Kent and colleagues published a paper describing the precision and recall metrics [903], which was followed by the publication in 1962 of the Cranfield studies by Cyril Cleverdon [394, 395]. In 1963, Joseph Becker and Robert Hayes published the first book on information retrieval [164]. Throughout the 60's, Gerard Salton and Karen Sparck Jones, among others, shaped the field by developing the fundamental concepts that led to the modern technologies of ranking in IR. In 1968, the first IR book authored by Salton was published. In 1971, N. Jardine and C.J. Van Rijsbergen articulated the "cluster hypothesis" [827]. In 1978, the first ACM Conference on IR (ACM SIGIR) was held in Rochester, New York. In 1979, C.J. Van Rijsbergen published *Information Retrieval* [1624], which focused on probabilistic models. In 1983, Salton and McGill published *Introduction to Modern Information Retrieval* [1414], a classic book on IR focused on vector models. Since then, the IR community has grown to include thousands of professors, researchers, students, engineers, and practitioners throughout the world. The main conference in the area, the ACM International Conference on Information Retrieval (ACM SIGIR), now attracts hundreds of attendees and receives hundreds of submitted papers on an yearly basis.

1.1.2 Information Retrieval in Libraries and Digital Libraries

Libraries were among the first institutions to adopt IR systems for retrieving information. Usually, library systems were initially developed by academic institutions and later by commercial vendors. In the first generation, such systems consisted of an automation of existing processes such as card catalogs searching, restricted to author names and titles. In the second generation, increased search functionality was added to include subject headings, keywords, and query operators. In the third generation, which is currently being deployed, the focus has been 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, we distinguish MELVYL developed by the California Digital Library at University of California, and the Cheshire system developed originally at UC Berkeley and lately in cooperation with the University of Liverpool. Further details on these library systems can be found in Chapter 16.

1.1.3 IR at the Center of the Stage

Despite its maturity, until recently, IR was seen as a narrow area of interest restricted 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, invented in 1989 by Tim Berners-Lee, has become a universal repository of human knowledge and culture. Its success is based on the conception of a standard user interface which is always the same, no matter the computational environment used to run the interface, and which allows any user to create their own documents. As a result, millions of users have created billions of documents that compose the largest human repository of knowledge in history. An immediate consequence is that finding useful information on the Web is not always a simple task and usually requires posing a query to a search engine, i.e., running a search. And search is all about IR and its technologies. Thus, almost overnight, IR has gained a place with other technologies at the center of the stage.

1.2 The IR Problem

Users of modern IR systems, such as search engine users, have information needs of varying complexity. In the simplest case, they are looking for the link to the homepage of a company, government, or institution. In the more sophisticated cases, they are looking for information required to execute tasks associated with their jobs or immediate needs. An example of a more complex *information need* is as follows:

> Find all documents that address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).[1]

[1] This is topic 168 of the TREC reference collection, see Chapter 4.

This full description of the user need does not necessarily provide the best formulation for querying the IR system. Instead, the user might want to first translate this information need into a *query*, or sequence of queries, to be posed to the system. In its most common form, this translation yields a set of keywords, or index terms, which summarize the user information need. Given the user query, the key goal of the IR system is to retrieve information that is useful or relevant to the user. The emphasis is on the retrieval of *information* as opposed to the retrieval of *data*.

To be effective in its attempt to satisfy the user information need, the IR system must somehow 'interpret' the contents of the information items i.e., the 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 IR Problem: the primary goal of an IR system is to retrieve all the documents that are relevant to a user query while retrieving as few non-relevant documents as possible.

The difficulty is knowing not only how to extract information from the documents but also knowing how to use it to decide relevance. That is, the notion of *relevance* is of central importance in IR.

One main issue is that relevance is a personal assessment that depends on the task being solved and its context. For example, relevance can change with time (e.g., new information becomes available), with location (e.g., the most relevant answer is the closest one), or even with the device (e.g., the best answer is a short document that is easier to download and visualize). In this sense, no IR system can provide perfect answers to all users all the time.

1.2.1 The User's Task

The user of a retrieval system has to translate their information need into a query in the language provided by the system. With an IR system, such as a search engine, this usually implies specifying a set of words that convey the semantics of the information need. We say that the user is *searching* or *querying* for information of their interest. While searching for information of interest is the main retrieval task on the Web, search can also be used for satisfying other user needs distinct from information access, such as the buying of goods and the placing of reservations, as we discuss in section 1.4.3.

Consider now a user who has an interest that is either poorly defined or inherently broad, such that the query to specify is unclear. To illustrate, the user might be interested in documents about car racing in general and might decide to glance related documents about Formula 1 racing, Formula Indy, and the '24 Hours of Le Mans.' We say that the user is *browsing* or *navigating* the documents in the collection, not searching. It is still a process of retrieving information, but one whose main objectives are less clearly defined in the beginning. The task in this case is more related to exploratory search and resembles a process of quasi-sequential search for information of interest.

In this book, we make a clear distinction between the different tasks the user of the retrieval system might be engaged in. The task might be then of two distinct

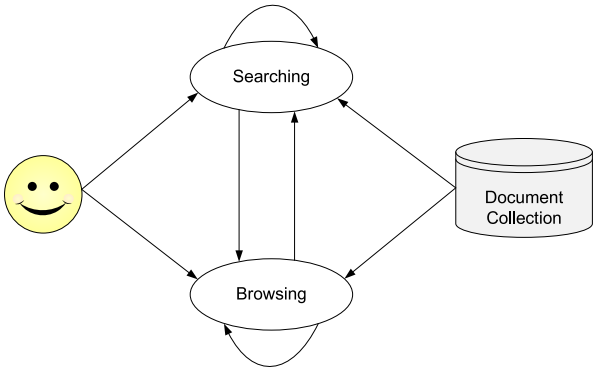

Figure 1.1: The tasks of the user.

types: *searching* and *browsing*, as illustrated in Figure 1.1. This two different tasks are covered in detail in Chapter 2.

1.2.2 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 that satisfies a given query. For instance, a user of an IR system is willing to accept documents that contain synonyms of the query terms in the result set, even when those documents do not contain any query terms. That is, in an IR system the retrieved objects might be inaccurate and small errors are likely to go unnoticed.

In a data retrieval system, on the contrary, a single erroneous object among a thousand retrieved objects means total failure. A data retrieval system, such as a relational database, deals with data that has a well defined structure and semantics, while an IR system deals with natural language text which is not well structured. 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.

1.3 The IR System

In this section we provide a high level view of the software architecture of an IR system. We also introduce the processes of retrieval and ranking of documents in response to a user query.

1.3.1 Software Architecture of the IR System

To describe the IR system, we use a simple and generic software architecture as shown in Figure 1.2. The first step in setting up an IR system is to assemble the

document collection, which can be private or be crawled from the Web. In the second case a crawler module is responsible for collecting the documents, as we discuss in Chapter 12. The document collection is stored in disk storage usually referred to as the *central repository*. The documents in the central repository need to be indexed for fast retrieval and ranking. The most used index structure is an *inverted index* composed of all the distinct words of the collection and, for each word, a list of the documents that contain it. Inverted indexes are discussed in Chapter 9.

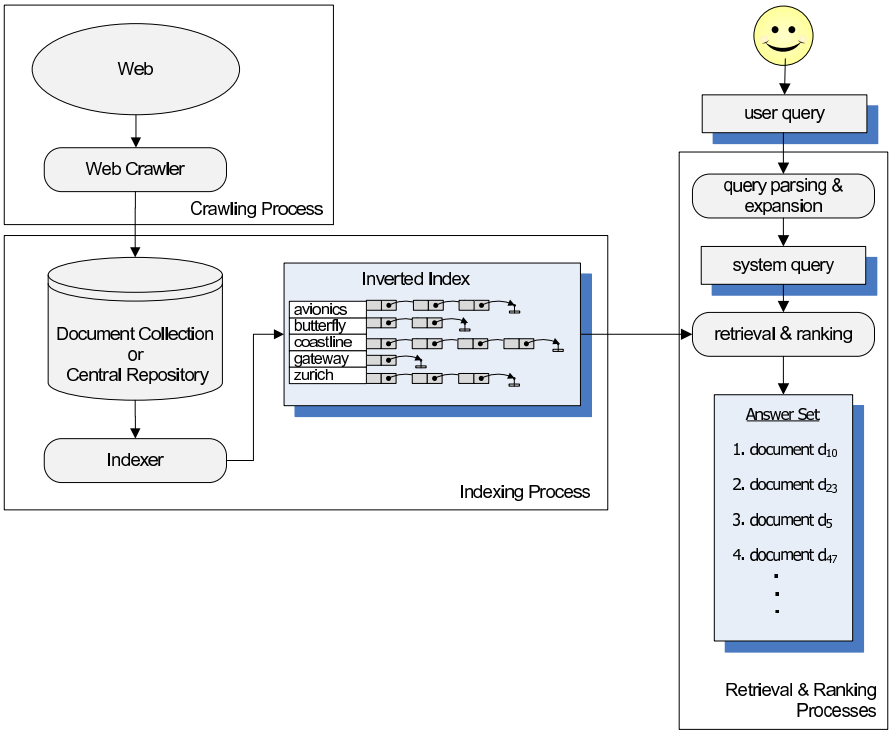

Figure 1.2: High level software architecture of an IR system. Crawling is an additional module required by Web IR systems, such as the search engines.

Given that the document collection is indexed, the retrieval process can be initiated. It consists of retrieving documents that satisfy either a user query or a click in a hyperlink. In the first case, we say that the user is *searching* for information of interest; in the second case, we say that the user is *browsing* for information of interest. In the remaining of this section, we use retrieval as it applies to the searching process. For a more detailed discussion on browsing and how it compares to searching, see Chapter 2.

To search, the user first specifies a *query* that reflects their information need. Next, the user query is parsed and expanded with, for instance, spelling variants of a query word. The expanded query, which we refer to as the *system query*, is then processed against the index to retrieve a subset of all documents. Following, the retrieved documents are ranked and the top documents are returned to the user.

The purpose of ranking is to identify the documents that are most likely to be considered relevant by the user, and constitutes the most critical part of the IR system. Because of this, our coverage of IR models in Chapter 3 is broad and quite detailed.

Given the inherent subjectivity in deciding relevance, evaluating the quality of the answer set is a key step for improving the IR system. A systematic evaluation process allows fine tuning the ranking algorithm and improving the quality of the results, as we discuss in Chapter 4. The most common evaluation procedure consists of comparing the set of results produced by the IR system with results suggested by human specialists.

To improve the ranking, we might collect feedback from the users and use this information to change the results. In the Web, the most abundant form of user feedback are the clicks on the documents in the results set, as we discuss in Chapter 5. Another important source of information for Web ranking are the hyperlinks among pages, which allow identifying sites of high authority, as we discuss in Chapter 11.

There are many other concepts and technologies that bear impact on the design of a full fledged IR system, such as a modern search engine, and most of them are covered in the remaining chapters of the book.

1.3.2 The Retrieval and Ranking Processes

To describe the retrieval and ranking processes, we further elaborate on our description of the modules shown in Figure 1.2, as illustrated in Figure 1.3. Given the documents of the collection, we first apply text operations to them such as eliminating stopwords, stemming, and selecting a subset of all terms for use as indexing terms. The indexing terms are then used to compose document representations, which might be smaller than the documents themselves (depending on the subset of index terms selected).

Given the document representations, it is necessary to build an index of the text. Different index structures might be used, but the most popular one is an *inverted index*, as we discuss in Chapter 9. The steps required to generate the index compose the *indexing process* and must be executed offline, before the system is ready to process any queries. The resources (time and storage space) spent on the indexing process are amortized by querying the retrieval system many times.

Given that the document collection is indexed, the retrieval process can be initiated. The user first specifies a *query* that reflects their information need. This query is then parsed and modified by operations that resemble those applied to the documents. Typical operations at this point consist of spelling corrections and elimination of terms such as stopwords, whenever appropriated. Next, the transformed query is expanded and modified. For instance, the query might be modified using query suggestions made by the system and confirmed by the user. The expanded and modified query is then processed to obtain the set of *retrieved documents*, which is composed of documents that contain the query terms. Fast query processing is made possible by the index structure previously built. The steps required to produce the set of retrieved documents constitute the *retrieval process*.

Next, the retrieved documents are ranked according to a *likelihood* of relevance to the user. This is a most critical step because the quality of the results, as perceived by the users, is fundamentally dependent on the ranking. In Chapter 3 we discuss

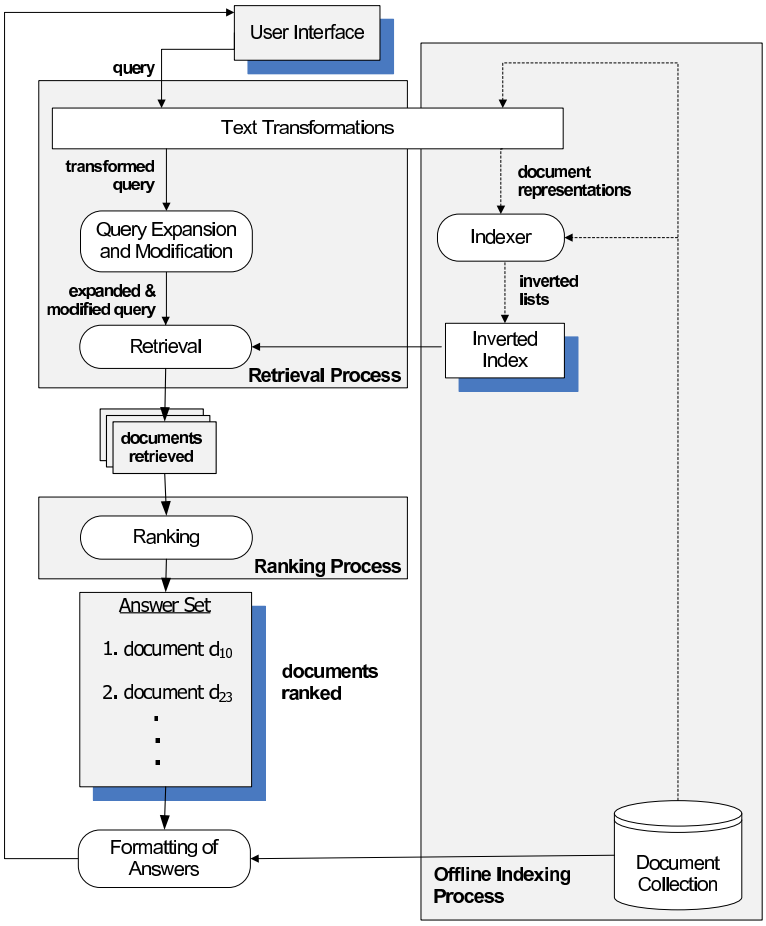

Figure 1.3: The processes of indexing, retrieval, and ranking of documents.

the ranking process in great detail. The top ranked documents are then formatted for presentation to the user. The formatting consists of retrieving the title of the documents and generating snippets for them, i.e., text excerpts that contain the query terms, which are then displayed to the user.

1.4 The Web

In this section we discuss the creation of the Web and its major implication – the advent of the e-publishing age. We also discuss how the Web changed search, i.e., the major impacts of the Web on the search task. At the end, we cover practical issues, such as security and copyright, which derive directly from the massive presence of millions of users on the Web.

1.4.1 A Brief History

At the end of World War II, US President Franklin Roosevelt asked Vannevar Bush, then occupying very high level government positions, for recommendations on applications of technologies learnt during the war to peace times. Bush first produced a report entitled "Science, The Endless Frontier" which directly influenced the creation of the National Science Foundation. Following, he wrote "As We May Think" [303], a remarkable paper that discussed new hardware and software gadgets that could be invented in the upcoming years. In Bush's words,

> Whole new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified [303].

"As We May Think" influenced people like Douglas Engelbart who, at the Fall Joint Computer Conference in San Francisco in December of 1968, ran a demonstration in which he introduced the first ever computer mouse, video conferencing, teleconferencing, and hypertext. It was so incredible that it became known as "the mother of all demos" [1690]. Of the innovations displayed, the one that interests us the most here is *hypertext*. The term was coined by Ted Nelson in his Project Xanadu [1691].

Hypertext allows the reader to jump from one electronic document to another, which was one important property regarding the problem Tim Berners-Lee faced in 1989. At the time, Berners-Lee worked in Geneva at the CERN – *Conseil Européen pour la Recherche Nucléaire*. There, researchers who wanted to share documentation with others had to reformat their documents to make them compatible with an internal publishing system [803]. It was annoying and generated many questions, many of which ended up been directed towards Berners-Lee. He understood that a better solution was required.

It just so happened that CERN was the largest Internet node in Europe. Berners-Lee reasoned that it would be nice if the solution to the problem of sharing documents were decentralized, such that the researchers could share their contributions freely. He saw that a networked hypertext, through the Internet, would be a good solution and started working on its implementation. In 1990, he wrote the HTTP protocol, defined the HTML language, wrote the first browser, which he called "World Wide Web", and the first Web server. In 1991, he made his browser and server software available in the Internet. The Web was born.

1.4.2 The e-Publishing Era

Since its inception, the Web became a huge success. The number of Web pages now far exceeds 20 billion[2] [487] and the number of Web users in the world exceeds 1.7 billion [815]. Further, it is known that there are more than one trillion distinct URLs on the Web [651], even if many of them are pointers to dynamic pages, not static HTML pages. Further, a viable model of economic sustainability based on online advertising was developed [801].

[2] In its blog at http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html Google announced to have collected over 1 trillion distinct URLs.

The advent of the Web changed the world in a way that few people could have anticipated. Yet, one has to wonder on the characteristics of the Web that have made it so successful. Is there a single characteristic of the Web that was most decisive for its success? Tentative answers to this question include the simple HTML markup language, the low access costs, the wide spread reach of the Internet, the interactive browser interface, the search engines. However, while providing the fundamental infrastructure for the Web, these technologies were not the root cause of its popularity. What was it then?

To emphasize the point we make here, let us wander through the life of a writer who lived two hundred years ago.

> She finished the first draft of her novel between 1796 and 1797. The first attempt of publication was refused. The originals were eventually lost, so she rewrote the novel in 1812 and it was finally published in 1813. Her authorship was made anonymous under the reference *"By a Lady"* [400].
>
> *Pride and Prejudice* is likely one of the three best loved books in the UK, jointly with *The Lord of the Rings* and the *Harry Potter* series of books. It has been the subject of six TV productions and five film versions [1694]. The last of these, starring Keira Knightley and Matthew Macfadyen, grossed over 100 million dollars worldwide and provided Ms. Knightley with an academy award nomination [1693].
>
> Jane Austen published anonymously her entire life. Throughout the twentieth century, Austen's novels have never been out of print. A variety of print editions have appeared as a tribute to the popularity of *Pride and Prejudice*.

The fundamental shift in human relationships, introduced by the Web, was *freedom to publish*. Jane Austen did not have that freedom, so she had to either convince a publisher of the quality of her work or pay for the publication of an edition of it herself. Since she could not pay for it, she had to be patient and wait for the publisher to become convinced. It took 15 years.

In the world of the Web, this is no longer the case. People can now publish their ideas on the Web and reach millions of people over night, without paying anything for it and without having to convince the editorial board of a large publishing company. That is, restrictions imposed by mass communication media companies and by natural geographical barriers were almost entirely removed by the invention of the Web, which has led to a freedom to publish that marks the birth of a new era. One which we refer to as *The e-Publishing Era*.

1.4.3 How the Web Changed Search

Web search is today the most prominent application of IR and its techniques. Indeed, the ranking and indexing components of any search engine are fundamentally IR pieces of technology. An immediate consequence is that the Web has had a major impact in the development of IR, as we now discuss.

The first major impact of the Web on search is related to the characteristics of the document collection itself. The Web collection is composed of documents (or pages) distributed over millions of sites and connected through hyperlinks , i.e., links that

associate a piece of text of a page with other Web pages. The inherent distributed nature of the Web collection requires collecting all documents and storing copies of them in a central repository, prior to indexing. This new phase in the IR process, introduced by the Web, is called *crawling* and is extensively discussed in Chapter 12.

The second major impact of the Web on search is related to the size of the collection and the volume of user queries submitted on a daily basis. Given that the Web grew larger and faster than any previous known text collection, the search engines have now to handle a volume of text that far exceeds 20 billion pages [487], i.e., a volume of text much larger than any previous text collection. Further, the volume of user queries is also much larger than ever before, even if estimates vary widely. The combination of a very large text collection with a very high query traffic has pushed the performance and scalability of search engines to limits that largely exceed those of any previous IR system [151]. That is, performance and scalability have become critical characteristics of the IR system, much more than they used to be prior to the Web. While we do not discuss performance and scalability of search engines in this book, the reader is referred to Chapter 11 for references on the topic (see section on bibliography).

The third major impact of the Web on search is also related to the vast size of the document collection. In a very large collection, predicting relevance is much harder than before. Basically, any query retrieves a large number of documents that match its terms, which means that there are many *noisy* documents in the set of retrieved documents. That is, documents that seem related to the query but are actually not relevant to it according to the judgement of a large fraction of the users are retrieved. This problem first showed up in the early Web search engines and became more severe as the Web grew. Fortunately, the Web also includes new sources of evidence not present in standard document collections that can be used to alleviate the problem, such as hyperlinks and user clicks in documents in the answer set. In Chapter 11 we discuss the issue of predicting relevance on the Web.

Two other major impacts of the Web on search derive from the fact that the Web is not just a repository of documents and data, but also a medium to do business. One immediate implication is that the search problem has been extended beyond the seeking of text information to also encompass other user needs such as the price of a book, the phone number of a hotel, the link for downloading a software. Providing effective answers to these types of information needs frequently requires identifying structured data associated with the object of interest such as price, location, or descriptions of some of its key characteristics. These new classes of queries are discussed in Chapter 7.

The fifth and final impact of the Web on search derives from Web advertising and other economic incentives. The continued success of the Web as an interactive media for the masses created incentives for its economic exploration in the form of, for instance, advertising and electronic commerce. These incentives led also to the abusive availability of commercial information disguised in the form of purely informational content, which is usually referred to as *Web spam*. The increasingly pervasive presence of spam on the Web has made the quest for relevance even more difficult than before, i.e., spam content is sometimes so compelling that it is confused with truly relevant content. Because of that, it is not unreasonable to think that spam makes relevance negative, i.e., the presence of spam makes the current ranking algorithms produce answers sets that are worst than they would be if the Web were

spam free. This difficulty is so large that today we talk of Adversarial Web Retrieval, as we discuss in the section on spam in Chapter 11.

1.4.4 Practical Issues on the Web

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 submits to the vendor credit information to be used for charging purposes. In its most common form, such information consists of a credit card number. For security reasons, this information is usually encrypted, as done by institutions and companies that deploy automatic authentication processes.

Besides security, 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 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 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 interested reader is referred to the book by Lesk [1005].

1.5 Organization of the Book

1.5.1 Focus of the Book

Despite the increased 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 IR 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 in the user interfaces chapter and to some extent in the last two chapters.

We put great emphasis on the integration of the different areas which are closely related to IR 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.

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 have also authored or co-authored 12 of the 17 chapters in the book. Further, all the contributed chapters have been judiciously revised, 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.

1.5.2 Book Contents

Given that IR is a fifty-year-old discipline, a single book can cover only a limited portion of all knowledge in the area. Even though, there are key concepts, methods, and technologies that are central to gaining a broad and deep understanding of IR. Our best effort to cover these concepts and technologies resulted in the 17 chapters that compose this second edition of Modern Information Retrieval. All chapters in this new edition are quite distinct from those in the first edition, given that more than ten years have passed since the publication of the first edition. In fact, more than half of the material is new or has been rewritten to cover recent research and results, to simplify notation, or to cover related topics that had not been covered in the first edition. To illustrate, we added chapters on text classification, structured text retrieval, Web crawling, and enterprise search. Further, the chapters on relevance feedback, multimedia, the Web, library systems, digital libraries, retrieval evaluation, and modeling have been considerably modified and updated.

Figure 1.4 illustrates the organization of the book. The first chapter, this one, introduces the *information retrieval problem*, provides a brief history of the Web, and analyses its impact on IR. Chapter 2 discusses the design of user interfaces for search, given that search has become a main application domain of IR technologies. This chapter is entirely new, quite distinct from the chapter on user interfaces found in the first edition, and aims at providing the reader with a top-down view of the IR problem.

The next three chapters cover classic IR, from ranking models to evaluation of the quality of the results and to user relevance feedback. More than half of the material in these three chapters is new, a testimony to the rapid development of IR in the last ten years. Our discussion is broad and deep. In Chapter 3 we discuss fourteen distinct IR models aimed at associating a rank with each document in the answer set and ordering them according to this rank. We start with the classic Boolean, vector, and probabilistic models and proceed with coverage of three variants of each of the classic models, including among them the set-based, generalized vector, BM25, and language models. In Chapter 4 we provide a detail discussion of techniques for evaluating the quality of results in IR systems. We start with a brief historical review of Cleverdon's seminal work on evaluation of indexing systems and on how that evolved into the Cranfield paradigm. We then cover precision-recall figures, the DCG metric for dealing with graded relevance scores, the Bpref metric for dealing with incomplete

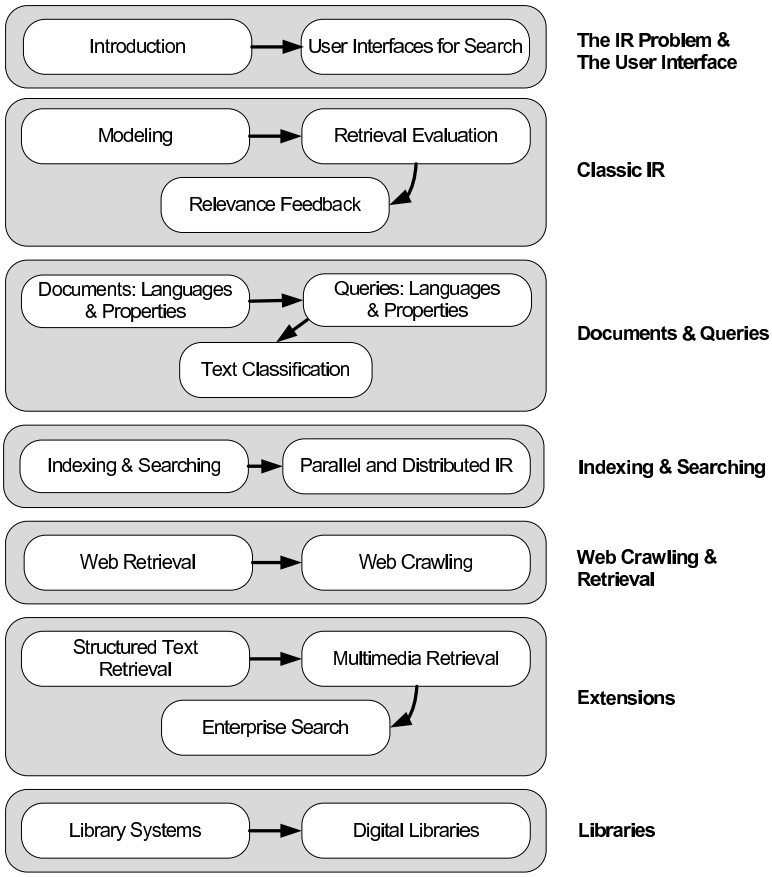

Figure 1.4: Organization of the chapters in the book.

relevance assessments. We also discuss rank correlation metrics such as the Spearman and Kendall Tau coefficients. The TREC reference collections, as well as various small test collections, are discussed at length. At the end, we cover Web specific methods such as side-by-side panels and discuss how to interpret clickthrough data as an indicative of relevance. In Chapter 5 we discuss implicit and explicit methods of compiling relevance feedback from the users and then using them to change the final ranking. These methods are directly intertwined with query expansion techniques. These three chapters cover all the fundamental concepts involved in classic IR, i.e., the techniques and methods for solving the IR problem and evaluating the results.

The following three chapters discuss concepts and technologies related to documents, queries, and how to organize them through text classification. In Chapter 6 we discuss text properties such as the distribution of words in documents and models of natural language, markup languages such as SGML, HTML, and XML, document processing and parsing, and compression methods. In Chapter 7 we discuss query properties such as the distribution of keywords in queries and characteristics of Web queries, as well as query languages based on keywords, on structural forms, and on query protocols. In Chapter 8 we discuss algorithms and methods for organizing documents and queries. Our discussion is focused on classification of documents because this is the most common case. We distinguish between unsupervised and supervised algorithms for text classification. Regarding unsupervised methods, we cover text clustering algorithms such as K-means and its variants. Regarding supervised methods, we discuss six distinct types of text classification algorithms namely decision trees, nearest neighbors, Rocchio, naive Bayes, support vector machines, and ensemble classifiers. We also discuss in detail how to evaluate their results. This chapter is entirely new and covers an important gap in the first edition of Modern Information Retrieval, given that text classification is a key technology in IR nowadays.

The following two chapters discuss technologies used for indexing and searching text collections. In Chapter 9 we discuss various indexing and searching techniques, from sequential searching to inverted indexes and suffix arrays. We also cover compression index techniques and how to use them to speed up retrieval. Chapter 10 discusses architectures and algorithms for parallelizing and distributing indexing and searching (query) processes. This is a key trend in the modern Web, given that the massive volumes of queries submitted to the search engines can only be processed by distributed clusters of machines.

The next two chapters cover crawling, retrieval, and ranking of Web documents. Our discussion of Web retrieval in Chapter 11 covers properties of the Web, architecture of search engines, link analysis algorithms such as HITS and Page rank, and ranking on the Web. While the chapter does not cover all the research in the area, which would not be possible in any single chapter, it does illustrate how search engines take advantage of IR algorithms and techniques. Our discussion of Web crawling in Chapter 12 starts with a historical account of how Web crawling developed. We then discuss Web crawling architectures and implementation issues. Following, we cover scheduling algorithms for determining which pages should be collected next, a central part of any crawling algorithm. At the end, we discuss procedures for evaluating the crawler.

Our extensions to the Web cover structured text retrieval and multimedia retrieval, two main areas which have become increasingly more associated with the Web, fol-

lowed by enterprise search. Chapter 13, on structured text retrieval, is entirely new and reflects the rapid evolution of the area since the publication of the first edition of Modern Information Retrieval. It covers early text retrieval models, XML retrieval from indexing to ranking, methods of evaluating XML retrieval, and XML query languages. Chapter 14, on multimedia retrieval, is also entirely new, we changed the focus to provide a top-down view of multimedia from an IR perspective. It covers content-based image retrieval, followed by audio and music retrieval, and then video retrieval. The combination of content-based, audio, music, and video retrieval into a single search mechanism requires fusion models, which are also discussed. At the end, the MPEG standards are presented. Chapter 15 discusses enterprise search systems, i.e., systems aimed at retrieving information within organizations and corporations, how they differ from Web search systems, and challenges during their design and implementation.

The last two chapters in the book cover library systems and digital libraries. Chapter 16 covers commercial document databases, integrated library systems (ILS), and online public access catalogs. Commercial document databases are still the largest information retrieval systems nowadays. LEXIS-NEXIS, for instance, has a database with more than one billion documents and attends hundreds of million queries annually. Chapter 17, which has been fully rewritten, provides detailed coverage of the latest technologies and trends in digital libraries, starting with a historical overview, followed by a discussion of fundamentals, social-economical issues, and 7 distinct digital library systems. At the end, we also discuss important case studies in digital libraries, such as the networked digital library of theses and dissertations, the national science digital library, and the ETANA archaeological digital library.

The book also includes two appendices. The first reviews 27 publicly available search systems, including HtDig, Indri, Lucene, MG4J, Omega, Omnifind, SwishE, Swish++, Terrier, and Zettair. It includes a comparative analysis of these search systems in terms of time to index, query processing performance, and storage requirements. The second appendix presents the biographies of all the authors that contributed to this book. We end with exactly 1,800 references that are used throughout the book.

In this second edition we added more inline references for people interested in research, although there are still material covered in a pure textbook style. Although we tried to balance the breadth and the depth of the content, clearly there are some topics covered in more detail due to our own expertise and research interests. We apologize in advance if we missed some topic or some important detail or reference.

From classic IR to the Web, from algorithms for organizing the information to the modern digital libraries, from the IR technologies of indexing and searching used by the search engines to the new technologies required to deploy extensions such as structured text and multimedia retrieval, this second edition of Modern Information Retrieval aims at providing a broad but also deep view of IR, its concepts and technologies, the applications of these technologies to search engines, as well as their impact on neighbor fields of knowledge such as information sciences, multimedia, databases, and digital libraries.

1.6 The Book Web Site: A Teaching Resource

The book Web site, which contains slides covering all chapters of the book to be used as teaching material, is located at

http://www.mir2ed.org

Besides the slides, it also includes a glossary, exercises and detailed teaching suggestions for different courses targeted at distinct audiences, such as:

- Information Retrieval, Computer Science, Undergraduate Level;
- Advanced Information Retrieval, Computer Science, Graduate Level;
- Multimedia Retrieval, Computer Science, Undergraduate Level;
- Information Retrieval, Information Systems, Undergraduate Level;
- Information Retrieval, Library Sciences, Undergraduate Level;
- Web Retrieval, Generic, Undergraduate or Graduate Level;
- Digital Libraries, Generic, Undergraduate or Graduate Level.

In addition, a reference collection (containing 1239 documents on Cystic Fibrosis and 100 information requests with extensive relevance evaluation [1454]) 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 [1624] (available on the Web) and Salton and McGill [1414]. 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, more than 25 years later, both books are now outdated and do not cover many of the new developments in information retrieval.

Three other well known references in information retrieval are the book edited by Frakes and Baeza-Yates [582], the book by Witten, Moffat and Bell [1709], and the book by Lesk [1005]. All these three books complement 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. Later books on classic IR are the ones by Hersh [751], in its second edition, and the one by Chowdhury [382], in its third edition. Both books have a much narrower view of the field than the view we take in this book. A book focused on information and its representation is the one by Meadow, Boyce, Kraft, and Barry [1112], also in its third edition. 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 [32]. On the information seeking view we have to mention Marchionini's [1082] and Tedd & Hartley's [977] books.

There are other complementary books for specific chapters. For example, there are many books on IR and hypertext, like [20]. The same is true for books on multimedia, such as the book by Steinmetz and Nahrstedt [1534] and the book by Alessi and Trollip [25]. An interesting IR book with a perspective on health and biomedical information is the one by Hersh [749]. Although not an information retrieval title, the book by Rosenfeld and Morville [1157] on information architecture of the Web, is a good complement to our chapter on searching the Web. The book by Menasce and Almeida [1118] demonstrates how to use queueing theory for predicting Web server performance. The book by Chakrabarti [349] discusses methods for knowledge mining on the Web. 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 [1510] 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 edited by Grefenstette [674]. This book is a good complement to ours for people interested in this particular topic. Additionally, a collection of papers focusing on intelligent IR was edited by Maybury [1101], and a collection of papers focusing on natural language IR was edited by Strzalkowski [1538]. A collection on the relation of natural language processing and information retrieval, in honour of Karen Sparck Jones, was edited by Tait [1554]. Other edited collections explore the relations of IR with uncertainty and logic [444], language modeling [453], cognitive retrieval [1515], and TREC evaluation [1654].

The book by Korfhage [931] 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 and Maybury [937] (second edition) and Shapiro *et al.* [1453] do not cover these topics in detail, and have a different orientation. The book by Grossman and Frieder [682] does not discuss the Web, digital libraries, or visual interfaces. A book that discusses classic IR in the context of search engines is the one by Berry and Browne [194]. Other specific books look at the mathematical foundations of IR [505], the geometry of retrieval [1625] or the intelligent use of the markup structure in IR [942].

The recent book by Ingwersen and Jarvelin on information seeking proposes an extended cognitive view of IR, beyond the laboratory model based on the Cranfield paradigm [810]. This has direct implication on how the system is evaluated, for instance. Another book that takes a cognitive view of the IR process, but focused on search engines, is the one by Belew [170]. A more recent book on exploratory search is one by White and Roth [1686].

A recent work is the book by Manning, Raghavan and Schütze [1081], which provides a coherent and elegant view of classic IR and Web retrieval. The book is focused on basic concepts and, as a result, does not cover search interfaces and does not provide an extensive discussion of IR models, as we do here. Additionally, the discussion on retrieval quality evaluation and Web crawling is minimalistic and there are no chapters on structured text retrieval, multimedia, libraries search systems, and digital libraries.

An even more recent book is the one by Croft, Metzler and Strohman [449], which focuses on search engines and is aimed at undergraduate courses. The book provides material for an introductory course on the use of IR technologies to build a search engine. The coverage is on purpose more superficial than the one provided here and does not include material on search interfaces, relevance feedback and query expansion, multimedia, and libraries. Further, the material on modeling, retrieval evaluation, text classification, and structured text retrieval in this book is more extensive and deep.

Finally, almost together with this book, Büttcher, Clarke and Cormack [304] published a book that focuses in the implementation and evaluation of information retrieval systems including XML retrieval, parallel search engines, and Web search.
For people interested in research results, the main journals on IR and related topics are:

- *Journal of the American Society of Information Sciences and Technology* (*JASIST*, Wiley and Sons),
- *ACM Transactions on Information Systems* (TOIS),
- *Information Retrieval* (Kluwer),
- *Information Processing and Management* (IP&M, Elsevier),
- *ACM Transactions on the Web*,
- *IEEE Transactions on Knowledge and Data Engineering* (TKDE),
- *Information Systems* (Elsevier),
- *Knowledge and Information Systems* (KAIS, Springer),
- *Data and Knowledge Engineering* (DKE, Springer),
- *D-Lib Magazine*, and
- *International Journal on Digital Libraries* (Springer).

Regarding conferences on IR, the most relevant ones include:

- ACM SIGIR International Conference on Information Retrieval,
- ACM International Conference on Web Search and Data Mining (WSDM),
- World Wide Web Conference (WWW), search track,
- ACM Conference on Information Knowledge and Management (CIKM),

- European Conference on IR (ECIR),
- String Processing and Information Retrieval Symposium (SPIRE).
- Text REtrieval Conference (TREC),
- INitiative for the Evaluation of XML retrieval (INEX),
- Cross Language Evaluation Forum (CLEF),
- International Conference on Multimedia Retrieval (ICMR), the fusion of ACM MIR and CIVR,
- Joint ACM-IEEE Conference on Digital Libraries (JCDL), and
- European Conference on Digital Libraries (ECDL).