# Book Review

## Introduction to Information Retrieval

**Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze**
(Stanford University, Yahoo! Research, and University of Stuttgart)

Cambridge: Cambridge University Press, 2008, xxi+482 pp; hardbound,
ISBN 978-0-521-86571-5, $60.00

*Reviewed by*
*Olga Vechtomova*
*University of Waterloo*

*Introduction to Information Retrieval* by Manning, Raghavan, and Schütze is an up-to-date, thorough, and systematic introduction to information retrieval (IR) from a computer science perspective. Written as a textbook, its main audience is graduate and senior undergraduate students taking IR courses. The book will also be valuable to researchers in other computer science fields, such as computational linguistics, as well as to professional practitioners wishing to delve into the IR field.

The book is structured into 21 chapters, which gradually unfold the subject of information retrieval, starting with the fundamentals (such as Boolean retrieval, document indexing, vector-space model, and evaluation in IR) and moving on to more advanced topics (such as probabilistic models, XML retrieval, text classification, machine learning for IR, document clustering, and Web retrieval). Pedagogical features of the book include short exercises at the end of each section and brief overviews of related research literature at the end of each chapter. Some of the major strengths of the book are its accessibility, clarity, and good balance between theory and practice. There are many concrete examples throughout the book that facilitate understanding of complex topics.

Although the book covers a broad selection of the major established and emerging topics in IR, it largely bypasses two important subjects, in my opinion: natural language processing techniques in IR and interactive information retrieval. Although the authors refer to some research done in these areas in various chapters, they do not give them the same thorough treatment given to other topics in the book. To compensate, in the preface the authors provide references to the detailed coverage of these and some other topics in other textbooks. It also might have been useful if the authors introduced some specialized IR tasks, such as opinion retrieval or enterprise search, which might benefit from more advanced NLP techniques.

Chapter 1 gives a succinct and focused introduction to the main concepts in IR, such as term, index, document, query, recall, precision, and so on. It outlines the main principles of Boolean retrieval, briefly criticizes it, and compares it to ranked retrieval. The authors also present a good real-world example of a commercial Boolean retrieval system.

Chapter 2 provides a detailed discussion of the initial stages of the document indexing process that include tokenization, stemming and lemmatization, stopwords removal, and approaches to dealing with phrases at the indexing stage, namely bigram indexing and the use of positional indexes. In this chapter the authors discuss some linguistic aspects of these processes. For example, when examining tokenization, they discuss various morphological and other aspects of languages that complicate this process (e.g., hyphenation in English, compound nouns in German, and word

sequences in East Asian languages). A brief outline of word segmentation approaches is provided. The authors also give a very good summary of the key relevant research works in these areas at the end of the chapter.

In Chapters 3, 4, and 5 the authors discuss data structures for indexes, index construction, and compression. These three chapters will most likely be of least interest to the computational linguistics community. However, two topics that might be of interest are the use of wildcard queries by users and spelling correction of queries, discussed in Chapter 3. In their discussion of wildcard queries, the authors examine only the use of wildcards in queries to represent different morphological variants of a word (e.g., American vs. British), the user's uncertainty in the correct spelling of a word, and stemmed words (e.g., *judicia** to represent both *judicial* and *judiciary*). It would have been interesting if the authors also discussed the use of wildcards to represent entire words, since some commercial search engines started to provide this functionality, allowing users to search for a phrase with a user-specified number of words in the middle (e.g., the use of *fine * me* to represent *fine by me*, *fine with me*, and *fine for me*).

Chapters 6 and 7 introduce the fundamentals of ranked document retrieval, term frequency and inverse document frequency, and the vector-space model. The authors briefly touch upon phrase queries, and how they can be handled by the vector-space model. Phrase or proximity-based retrieval is an important problem in IR, but unfortunately, the authors do not discuss in detail different approaches to proximity-based term weighting.

Chapter 8 is devoted to evaluation in IR, and presents major evaluation frameworks, such as the Text Retrieval Conference (TREC) and classical evaluation measures, such as mean average precision, precision at different cutoff points, as well as the more recently developed measure NDCG (normalized discounted cumulative gain) for evaluation with graded (non-binary) relevance judgments. The penultimate section in the chapter discusses various approaches to presenting retrieved documents in the ranked list shown to the user, such as snippets, and query-independent and query-biased document summaries.

Chapter 9 reviews relevance feedback and query expansion. Query expansion (QE) following relevance feedback is one of the most effective techniques in IR. The authors provide an overview of the main types of QE: **local**, whereby the query is modified on the basis of retrieved documents, and **global**, which is query independent. Among the local methods, they introduce here the classic Rocchio algorithm. Probabilistic approaches to query expansion following relevance feedback are discussed in detail in Chapter 11 after the authors introduce probabilistic models of IR. Query expansion following relevance feedback can be either automatic (AQE), whereby the system selects terms and adds them to the query, or interactive (IQE), whereby the selected terms are shown to the user for further selection. The authors only discuss AQE in the context of relevance feedback. Among the global QE methods, they mention the use of manual and automatically generated thesauri, as well as approaches to QE on the Web, such as suggestion of related queries.

Chapter 10 discusses XML retrieval, including such topics as a vector-space model for XML retrieval and INEX (Initiative for the Evaluation of XML retrieval), the main evaluation framework for XML retrieval.

Chapters 11 and 12 focus on probabilistic information retrieval and language modeling, respectively. Chapter 11 introduces the theoretical underpinnings of probabilistic IR models, and describes the Robertson and Spärck Jones probabilistic model and the BM25 term weighting function. Chapter 12 starts by describing the basic approach to language modeling in IR and then reviews some of its extensions.

Chapters 13, 14, and 15 discuss approaches to text classification, starting with naive Bayes classification, and then moving on to vector-space classification and support vector machines. All topics are presented in sufficient detail, supplemented with references to the key papers in these areas.

In Chapters 16 and 17 the authors introduce document clustering. Flat clustering methods (*K*-means and expectation maximization) and clustering evaluation methods are discussed in Chapter 16, and Chapter 17 is devoted to hierarchical clustering. Here, the authors present different agglomerative clustering algorithms, such as single-link, complete-link, group-average, and centroid similarity, as well as top–down (divisive) hierarchical clustering. An important problem in clustering is the labeling of clusters. The authors discuss and compare two approaches to labeling: differential cluster labeling, where label terms are selected on the basis of their distribution in one cluster compared to the others, and cluster-internal labeling, where a label is selected only on the basis of the cluster being labeled.

Chapter 18 introduces latent semantic indexing. This is a rather theoretical chapter, and readers might have benefited from a more extensive discussion of the use and practical applications of LSI.

The remaining three chapters are devoted to Web-based IR. Among the topics discussed are spam, types of user information needs, Web crawling and indexing, link-based approaches to document ranking such as PageRank, Markov chains, and hubs and authorities.

To sum up, *Introduction to Information Retrieval* is a comprehensive, authoritative, and well-written overview of the main topics in IR. The book offers a good balance of theory and practice, and is an excellent self-contained introductory text for those new to IR. Although the book does not cover advanced NLP techniques for IR, it is recommended for experts in computational linguistics who wish to learn about IR. Although many computational linguists are familiar with the material covered in the chapters on text classification, they will most certainly find chapters on different IR models and methods very useful.

*Olga Vechtomova* is an Associate Professor in the Department of Management Sciences, cross-appointed in the David R. Cheriton School of Computer Science, University of Waterloo, Canada. She received her PhD in Information Science from City University, London in 2001. Her interests are in query expansion, relevance feedback, lexical cohesion, applications of natural language processing techniques to IR, and user interaction with IR systems. Vechtomova's address is: Department of Management Sciences, Faculty of Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada; e-mail: ovechtom@engmail.uwaterloo.ca.