
Information Retrieval, 4, 79–83, 2001

Book Reviews

Managing Gigabytes—Compressing and Indexing Documents and Images (Second Edition). Ian H. Witten, Alistair Moffat and Timothy C. Bell, San Francisco, CA: Morgan Kaufmann; 1999; 576 pp. Price: $54.95 (ISBN: 1-55860-570-3.)

Books on general information retrieval (IR) have a variety of foci. Some cover the conceptual aspects of IR, (Baeza-Yates and Ribeiro-Neto 1999, Lancaster and Warner 1993, Meadow 1992) while others focus on algorithms and code for implementing systems (Frakes and Baeza-Yates 1992) and other still focus on specific domains. (Hersh 1996) This book by Witten, Moffat, and Bell is algorithm-oriented, but provides a nice conceptual overview of IR as well. Along with the freely available MG ("managing gigabytes") system and source code, the work of these authors provides an entry point to learning about IR, being able to implement functioning systems, and carrying out research.

The title of the book implies that its focus is on compression, and that is certainly a major theme that runs through the various chapters. However, the book also covers general IR, describing automated indexing, querying, basic principles of evaluation, and even a modestly visionary view of the information explosion.

The book begins with an introductory overview chapter. This is followed by a chapter on text compression which presents a comprehensive discussion of different basic approaches. The subsequent chapter covers indexing, beginning with a general discussion on inverted indexes and then details on the compression of such indexes. After this is a chapter on querying which discusses general retrieval principles and efficient means for carrying them out, including though the use of the compressed indexes described in the previous chapter.

The next four chapters cover a variety of compression methods for different information types—inverted indexes, pictures, text images, and mixed picture and text images. The discussion covers not only the major standards but also research work in their own labs at the cutting edge. These are followed by two summary chapters. The first brings all of the indexing, compression, and query methods together to describe overall system implementation. The second covers the general information explosion, with a particular focus on "managing gigabytes" on the World Wide Web and in digital libraries. The book ends with two appendices, one describing the MG system which implements many of the techniques described in the book (with a pointer to source code on an FTP site) and the other describing the various collections publicly available in the New Zealand Digital Library Project.

The book is generally well-written and produced. Most technical discussions are easy to read and comprehend. Both the text font and the images are visually pleasing. I readily admit to being a fan of this New Zealand- and Australian-based group, as I have used the MG system in my own IR research for several years. With the source code freely available, and this book describing the rationale behind it, the bar is lowered for learning about IR and performing research.

What are the uses of this book? Those who design and implement IR systems and wish to know more about algorithms and coding will find it valuable. And certainly those interested in compression of text and images will find it desirable to have as well. Those who teach IR courses will also find the book useful, especially those who teach about IR algorithms and coding.

In all, this is a well-written book that describes how to build IR systems, with a strong focus on compression methods. But its coverage of general IR should also lead others to take a look at it as well.

References

Baeza-Yates R and Ribeiro-Neto B (1999), Eds. Modern Information Retrieval, McGraw-Hill, New York.
Frakes W and Baeza-Yates R (1992), Eds. Information Retrieval: Data Structures and Algorithms, Prentice-Hall, Englewood Cliffs, NJ.
Hersh W (1996) Information Retrieval: A Health Care Perspective, Springer-Verlag, New York.
Lancaster F and Warner A (1993) Information Retrieval Today, Information Resources Press, Arlington, VA.
Meadow C (1992) Text Information Retrieval Systems, Academic Press, San Diego.

William Hersh
Associate Professor & Chief
Division of Medical Informatics and Outcomes Research
Oregon Health Sciences University
Portland, OR, USA

Foundations of Statistical Natural Language Processing. Christopher D. Manning and Hinrich Schütze. Cambridge, MA: MIT Press; 1999; 620 pp. Price: $60.00 (ISBN: 0-262-13360-1.)

Natural language processing has long seemed to be the magic bullet that will bring information retrieval much closer to human capabilities. It is rather frustrating that 20 years of work along these lines has not produced the best information retrieval systems. Generally speaking, systems based entirely on natural language concepts are not at all competitive with systems based on statistical analysis of texts. In addition, although adding natural language features appears to improve the performance of poor systems, no one has yet shown a way to add these features to the best systems and generate any further improvement. There is a growing suspicion that in fact the river is currently flowing the other way, and that ways of thinking about text that have been developed for the purposes of information retrieval have more to contribute to the problem of natural language processing than vice-versa.

The existence of the large and rapidly growing body of material on application of methods of numerical computation to the analysis of natural language texts is the motivation for this excellent book. The book is intended to serve as a reference manual for researchers, supplemented by a 54-page bibliography, and as a textbook for advanced students in computer science.

In spite of the direction of influence mentioned above, the authors are open minded and point out, for example, that a non-quantitative tagger developed at the University of