Book Reviews

Information Storage and Retrieval Systems

Tutorials - 6.5
Mid-Summer 2000

Author: Robert Korfhage
Published by: John Wiley and Sons, 1997.
ISBN: 0471143383

Reviewed by *Adrian P. O'Riordan*

The volume of information accessible over networks has exceeded the user's capability to sift through and find relevant information. Sophisticated information retrieval (IR) software has been developed to counter the information overload problem. The 1990s and the Internet revolution has seen an explosion in automatic IR research, which dates back to the 1950s. While the main focus of the field has been the retrieval of text, recent efforts addressing content-based retrieval of other media is beginning to show promise.

Information Storage and Retrieval Systems covers a broad range of topics constituting the burgeoning field of IR. Enough theory is presented for an in-depth discussion on querying, retrieval functions, and text analysis. However, there is not enough information to implement the techniques directly. Consequently, this book is an appropriate introduction to the field for graduate and senior undergraduate students. System implementors will have to look elsewhere for an in-depth discussion on the implementation of information retrieval systems.

Rather than focusing on particular IR systems, the book takes a more conceptual approach. IR systems are described as being composed of a number of sub-parts such as querying, query-document matching, and text analysis. In the first section of the book, chapters are dedicated to each of these identified sub-parts. The focus is on the essential algorithms, rather than the actual implementation technologies. For example, the chapter on query-document matching presents a comprehensive set of techniques for the important task of comparing queries and documents for assessing similarity. The measures are mostly based on simple word frequency statistics. The presentation is easy to follow even if you are not familiar with probability and statistics. Implementation topics, including file indexing, hashing, and string matching algorithms, are contained in two appendices of the book.

Popular retrieval models such as the Boolean model, the Vector Space Model, and less known techniques such as fuzzy set theoretic models and various probabilistic models are discussed in each of the above mentioned chapters. The diversity of methods makes it possible to elucidate similarities, where they exist, between the models. For example, the fuzzy set theoretic models can be viewed as extensions to the Boolean model, and the Vector Space and probabilistic models share the crucial idea of index term weighting. The wealth of different probabilistic models are supplemented by numerous references.

The second section of the book is more interesting by being less conventional. There is a chapter about the retrieval of more structured textual information, such as web-pages and the retrieval of other media besides text. Citation processing, information filtering, image processing, and sound processing are all discussed. The emphasis is placed on how each of these types of retrieval differs from the standard paradigm of unstructured text retrieval. None of these topics are presented in any great depth, but numerous references for further investigation are provided.

Two chapter are dedicated to document user access and other policy issues such as service, funding, copyright access, and privacy. While it is typical to see such topics discussed in books geared toward a library science readership, these topics are a welcome addition to a book that is more specifically about the technology of text retrieval systems.

A criticism of the book is that there is little discussion of actual retrieval systems, either operational commercial systems or research test-beds and prototypes. Even the widely used SMART system, which has acted as a test-bed for decades of research is only given a brief mention [2]. A related criticism is that too few practical exercises are provided. Finally, the rapidly expanding sub-field of text mining (related to database mining) is not discussed.

These criticisms aside, overall this is a very good textbook. A welcome addition to a field which supplements, rather than replaces classic, but older texts on IR, such as Salton and McGill's *Introduction to Modern Information Retrieval* [1], and Van Rijsbergen's *Information Retrieval* [3]. Anybody who has taken a course in database systems or data structures will be able to extend their knowledge by reading this book.

References

1

Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval* McGraw-Hill Companies, March 1984.

2

SMART system ftp://ftp.cs.cornell.edu/pub/smart/.

3

Van Rijsbergen, C. J. *Information Retrieval*, 2nd edition, Butterworth-Heinemann 1979.