
Information Retrieval, 5, 269–278, 2002

Book Reviews

Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. Richard K. Belew. New York, NY: Cambridge University Press; 2000; 356p. Price: $49.95 (ISBN 0-521-63028-2.)

Over the years a wide variety of textbooks have been written for the field of information retrieval. Many excellent older texts, e.g., van Rijsbergen's *Information Retrieval* (Butterworths 2d. ed. 1979) and Salton and McGill's *Introduction to Modern Information Retrieval* (McGraw-Hill 1983), do not cover developments, such as the Word Wide Web and web search engines, that have come into existence since their publication dates. More recent texts have appeared, including *Information Retrieval: Data Structures & Algorithms* edited by Frakes and Baeza-Yates (Prentice Hall 1992), *Managing Gigabytes: Compressing and Indexing Documents and Images* by Witten, Moffat, and Bell (Morgan Kaufmann, 2d. ed. 1999), and *Modern Information Retrieval* by Baeza-Yates and Ribeiro-Neto (Addison–Wesley 1999) to name a few. In addition a collection of readings, *Readings in Modern Information Retrieval* edited by Sparck Jones and Willet has been published. Despite the variety of textbooks available many instructors of information retrieval courses have, while perhaps assigning a text, also relied heavily on their own selection of readings. The field of information retrieval is rapidly changing. There are a wide variety of perspectives of the field on the part of instructors and wide variety of student backgrounds. All of these factors often make it hard for an instructor to find a single textbook that can be the focus of a class in information retrieval. Belew's book, *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*, or *FOA*, can provide such a focus, especially for courses that seek to provide students with experience in programming a search engine. The reviewer has taught information retrieval over the past 15 years using, either as primary or recommended texts, most of the books mentioned above. Over the past two years the reviewer has used draft versions of *FOA* and its accompanying software and data sets, as well as the published version.

FOA consists of eight chapters. Chapter 1 provides an overview of "... Finding Out About (FOA), the process of actively seeking out information relevant to a topic of interest ..." and the academic discipline of information retrieval that has sought to support the FOA process. Basic concepts, more fully treated in later chapters, are introduced, including: keywords, query syntax, documents, and indexing. Chapter 2, "Extracting Lexical Features", covers automatic tokenization of a document, leading to the creation of an inverted index. Chapter 3, "Weighting and Matching against Indices", describes the weighting of index and query terms and introduces the vector space model of retrieval. The calculation of TF-IDF weights is discussed and there is a section on partial ranking algorithms that ignore the contribution of inconsequential terms to overall document ranking in order to achieve efficiency.

Chapter 4 discusses evaluation of retrieval performance in terms of precision and recall, with an emphasis on the subjective nature of relevance judgments and on relevance feedback. A valuable feature of *FOA* is *RAVE*, or the *Relevance Assessment VEhicle*. Instructors using the text can register students with the author who provides a web site where students view documents from the *AIT* collection (see below) and judge them for relevance. This exercise is useful in several ways. First, it gives students hands on experience with making relevance assessments. Second, all of the assessments entered by students from all classes that have used *FOA* are maintained by the author, so that over time a large set of judgments have been collected for queries and documents from the *AIT* collection. Third, these queries, documents, and judgments, can be used in a final class assignment to evaluate the search engines that the teams of students have developed over the term. Chapter 5, "Mathematical Foundations", is a well-written account of, among other topics, Zipf's law, dimensionality reduction, latent semantic indexing, multidimensional scaling, clustering, and probabilistic information retrieval. Chapter 6, "Inference Beyond the *Index*", develops topics such as citation indexing, hypertext, social relations among authors, and deep interfaces. Chapter 7, "Adaptive Information Retrieval", introduces machine learning approaches to retrieval and text classification, including a discussion of the InfoSpiders algorithm. Chapter 8 summarizes the material covered and suggests directions for future research related to: (a) the Internet, (b) electronic publishing, and (c) education.

The book's accompanying CD-ROM contains not only a hyper-linked version of *FOA*, but also a hyper-linked version of van Rijsbergen's *Information Retrieval*, now out of print. FOA is also hyper-linked to van Rijsbergen's text. Other content of the CD-ROM includes four Java software packages that constitute a search engine. These packages can be used as starting points for programming assignment extensions by teams of students. The four packages provide: creation of an inverted file, query processing, retrieval evaluation, and general utilities. The software is also available in C, although Java is the preferred language.

The CD-ROM includes four data sets to be used with the search engine: the "Artificial Intelligence Thesis" (AIT) corpus, consisting of approximately 5,000 Ph.D. and Masters dissertation abstracts on the topic of artificial intelligence; AI Genealogy data, providing additional data related to the AIT collection; ancillary data derived from AIT; and an Encyclopedia Britannica (EB5) data set, corresponding to all encyclopedia entries classified under Section V of the EB Propaedia on the topic of "Human Society." Finally the CD-ROM includes version 0.95 of McCallum's *Rainbow* text categorization software, as well as distribution 1.0 of the Reuters-21578 test collection for text categorization.

Beyond the book and CD-ROM, one of the useful features of *FOA* is the web-based support made available for instructors and students. The *RAVE* environment has already been mentioned. In addition the author maintains a discussion board for students and instructors. The author also provides links from sections in *FOA* to articles in Sparck Jones and Willett's *Readings in Information Retrieval*, which he recommends being included as an additional text for classes using *FOA*.

FOA is a comprehensive resource for teaching a programming-based information retrieval class. The text provides an integrated introduction to many topics in information retrieval with a strong emphasis on mathematical and machine learning models, as well as giving a clear account of implementation details for the programming assignments. The links to

van Rijsbergen's text on the CD-ROM and to *Readings in Information Retrieval* connect the reader to the extensive literature of the field. Finally, the Web focus of the book enhances its value as a teaching tool.

Paul Thompson
Senior Research Engineer
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire 03755 USA
E-mail: Paul.Thompson@dartmouth.edu

Information Retrieval: Algorithms and Heuristics. David A. Grossman and Ophir Frieder. Norwell: Kluwer Academic Publishers, 1998.

This book presents a comprehensive overview of the most important subjects of Information Retrieval (IR) today, describing the main theories, algorithms and heuristics. It presents a large number of techniques in great detail, supporting the explanations with detailed examples and critic reviews of work published in journals and conferences. Besides classical IR material, an important number of novel topics in IR are introduced.

"Information Retrieval: Algorithms and Heuristics" is composed of 9 chapters. The main chapters of the book can be divided into two parts: first, an in-depth explanation of today's most important retrieval strategies and techniques (chapters 2 to 4) and second, an exploration of some of the current topics in information retrieval (chapters 5 to 7). The book opens with a short introduction (9 pages) and ends with two very short chapters on the Text Retrieval Conference (TREC) (chapter 8, four pages) and the future directions of information retrieval (chapter 9, three pages). Each chapter ends with a summary and exercises.

Chapter 2, Retrieval Strategies (80 pages), describes in a detailed manner today's main ad-hoc retrieval strategies: the Vector Space Model, Probabilistic Retrieval, Inference Networks, Extended Boolean Retrieval, Latent Semantic Indexing, Neural Networks, Genetic Algorithms, and Fuzzy Set Retrieval. An in-depth description is given of each one of the different approaches, as well as explicit very thorough examples using a demonstration corpus of four documents and a five word vocabulary. The necessary background material is provided, so that anyone with a basic knowledge of mathematics and probability can understand all strategies and follow the examples.

Probabilistic Retrieval is treated more in depth than any other strategy, due to its importance and the large number of variants in existence. An attempt is made to present in a consistent manner several distribution models (although only the most basic ones are described) as well as term component strategies. The strengths and weaknesses of the different models are detailed, and a good selection of references to the literature is given. A comparison of published results is often provided (although not consistently: some articles on weighting and distribution models are cited without a discussion on their results).

The practitioner will miss perhaps a clear comparison of the different models, their strengths and weaknesses, as well as some heuristics for choosing a particular model given