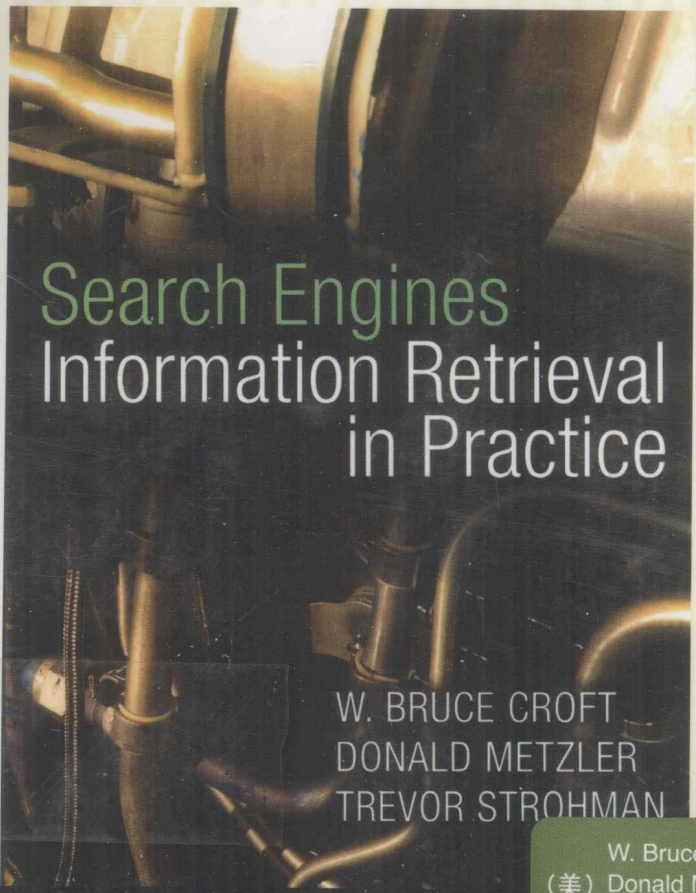
HZ BOOKS
华章教育

经 典 原 版 书 库

PEARSON
Addison Wesley

搜索引擎
信息检索实践

（英文版）



W. Bruce Croft
（美）Donald Metzler 著
Trevor Strohman

机械工业出版社
China Machine Press

经 典 原 版 书 库

搜索引擎

信息检索实践

Search Engines

Information Retrieval in Practice

W. Bruce Croft
(美) Donald Metzler 著
Trevor Strohman

机械工业出版社
China Machine Press

本书版权登记号：图字：01-2009-4966

图书在版编目（CIP）数据

搜索引擎：信息检索实践（英文版）/（美）克罗夫特（Croft, W. B.）等著. —北京：机械工业出版社，2009.10

（经典原版书库）

书名原文：Search Engines: Information Retrieval in Practice

ISBN 978-7-111-28247-1

Ⅰ.搜…　Ⅱ.克…　Ⅲ.互联网络-情报检索-英文　Ⅳ.G354.4

中国版本图书馆CIP数据核字（2009）第161295号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）

责任编辑：迟振春

北京京师印务有限公司印刷

2009年10月第1版第1次印刷

150mm×214mm　·　16.75印张

标准书号：ISBN 978-7-111-28247-1

定价：45.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：（010）68326294

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭示了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短的现状下，美国等发达国家在其计算机科学发展的几十年间积淀和发展的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起到积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章分社较早意识到“出版要为教育服务”。自1998年开始，华章分社就将工作重点放在了遴选、移译国外优秀教材上。经过多年的不懈努力，我们与Pearson，McGraw-Hill，Elsevier，MIT，John Wiley & Sons，Cengage等世界著名出版公司建立了良好的合作关系，从他们现有的数百种教材中甄选出Andrew S. Tanenbaum，Bjarne Stroustrup，Brain W. Kernighan，Dennis Ritchie，Jim Gray，Afred V. Aho，John E. Hopcroft，Jeffrey D. Ullman，Abraham Silberschatz，William Stallings，Donald E. Knuth，John L. Hennessy，Larry L. Peterson等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及珍藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近两百个品种，这些

书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍。 其影印版“经典原版书库”作为姊妹篇也被越来越多实施双语教学的学校所采用。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证。随着计算机科学与技术专业学科建设的不断完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都将步入一个新的阶段，我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。华章分社欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下：

华章网站：www.hzbook.com
电子邮件：hzjsj@hzbook.com
联系电话：(010) 88379604
联系地址：北京市西城区百万庄南街1号
邮政编码：100037

Preface

This book provides an overview of the important issues in information retrieval, and how those issues affect the design and implementation of search engines. Not every topic is covered at the same level of detail. We focus instead on what we consider to be the most important alternatives to implementing search engine components and the information retrieval models underlying them. Web search engines are obviously a major topic, and we base our coverage primarily on the technology we all use on the Web,[1] but search engines are also used in many other applications. That is the reason for the strong emphasis on the information retrieval theories and concepts that underlie all search engines.

The target audience for the book is primarily undergraduates in computer science or computer engineering, but graduate students should also find this useful. We also consider the book to be suitable for most students in information science programs. Finally, practicing search engineers should benefit from the book, whatever their background. There is mathematics in the book, but nothing too esoteric. There are also code and programming exercises in the book, but nothing beyond the capabilities of someone who has taken some basic computer science and programming classes.

The exercises at the end of each chapter make extensive use of a Java™-based open source search engine called Galago. Galago was designed both for this book and to incorporate lessons learned from experience with the Lemur and Indri projects. In other words, this is a fully functional search engine that can be used to support real applications. Many of the programming exercises require the use, modification, and extension of Galago components.

[1] In keeping with common usage, most uses of the word "web" in this book are not capitalized, except when we refer to the World Wide Web as a separate entity.

Contents

In the first chapter, we provide a high-level review of the field of information retrieval and its relationship to search engines. In the second chapter, we describe the architecture of a search engine. This is done to introduce the entire range of search engine components without getting stuck in the details of any particular aspect. In Chapter 3, we focus on crawling, document feeds, and other techniques for acquiring the information that will be searched. Chapter 4 describes the statistical nature of text and the techniques that are used to process it, recognize important features, and prepare it for indexing. Chapter 5 describes how to create indexes for efficient search and how those indexes are used to process queries. In Chapter 6, we describe the techniques that are used to process queries and transform them into better representations of the user's information need.

Ranking algorithms and the retrieval models they are based on are covered in Chapter 7. This chapter also includes an overview of machine learning techniques and how they relate to information retrieval and search engines. Chapter 8 describes the evaluation and performance metrics that are used to compare and tune search engines. Chapter 9 covers the important classes of techniques used for classification, filtering, clustering, and dealing with spam. Social search is a term used to describe search applications that involve communities of people in tagging content or answering questions. Search techniques for these applications and peer-to-peer search are described in Chapter 10. Finally, in Chapter 11, we give an overview of advanced techniques that capture more of the content of documents than simple word-based approaches. This includes techniques that use linguistic features, the document structure, and the content of nontextual media, such as images or music.

Information retrieval theory and the design, implementation, evaluation, and use of search engines cover too many topics to describe them all in depth in one book. We have tried to focus on the most important topics while giving some coverage to all aspects of this challenging and rewarding subject.

Supplements

A range of supplementary material is provided for the book. This material is designed both for those taking a course based on the book and for those giving the course. Specifically, this includes:

- Extensive lecture slides (in PDF and PPT format)

- Solutions to selected end–of–chapter problems (instructors only)
- Test collections for exercises
- Galago search engine

The supplements are available at www.search-engines-book.com, or at www.aw.com.

Acknowledgments

First and foremost, this book would not have happened without the tremendous support and encouragement from our wives, Pam Aselton, Anne-Marie Strohman, and Shelley Wang. The University of Massachusetts Amherst provided material support for the preparation of the book and awarded a Conti Faculty Fellowship to Croft, which sped up our progress significantly. The staff at the Center for Intelligent Information Retrieval (Jean Joyce, Kate Moruzzi, Glenn Stowell, and Andre Gauthier) made our lives easier in many ways, and our colleagues and students in the Center provided the stimulating environment that makes working in this area so rewarding. A number of people reviewed parts of the book and we appreciated their comments. Finally, we have to mention our children, Doug, Eric, Evan, and Natalie, or they would never forgive us.

Bruce Croft
Don Metzler
Trevor Strohman

Contents

1

Search Engines and Information Retrieval

> "Mr. Helpmann, I'm keen to get into Information Retrieval."
>
> Sam Lowry, *Brazil*

1.1 What Is Information Retrieval?

This book is designed to help people understand search engines, evaluate and compare them, and modify them for specific applications. Searching for information on the Web is, for most people, a daily activity. Search and communication are by far the most popular uses of the computer. Not surprisingly, many people in companies and universities are trying to improve search by coming up with easier and faster ways to find the right information. These people, whether they call themselves computer scientists, software engineers, information scientists, search engine optimizers, or something else, are working in the field of *Information Retrieval*.[1] So, before we launch into a detailed journey through the internals of search engines, we will take a few pages to provide a context for the rest of the book.

Gerard Salton, a pioneer in information retrieval and one of the leading figures from the 1960s to the 1990s, proposed the following definition in his classic 1968 textbook (Salton, 1968):

> Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.

Despite the huge advances in the understanding and technology of search in the past 40 years, this definition is still appropriate and accurate. The term "informa-

[1] Information retrieval is often abbreviated as IR. In this book, we mostly use the full term. This has nothing to do with the fact that many people think IR means "infrared" or something else.

tion" is very general, and information retrieval includes work on a wide range of types of information and a variety of applications related to search.

The primary focus of the field since the 1950s has been on text and text *documents*. Web pages, email, scholarly papers, books, and news stories are just a few of the many examples of documents. All of these documents have some amount of structure, such as the title, author, date, and abstract information associated with the content of papers that appear in scientific journals. The elements of this structure are called attributes, or fields, when referring to database records. The important distinction between a document and a typical database record, such as a bank account record or a flight reservation, is that most of the information in the document is in the form of text, which is relatively unstructured.

To illustrate this difference, consider the information contained in two typical attributes of an account record, the account number and current balance. Both are very well defined, both in terms of their format (for example, a six-digit integer for an account number and a real number with two decimal places for balance) and their meaning. It is very easy to compare values of these attributes, and consequently it is straightforward to implement algorithms to identify the records that satisfy queries such as "Find account number 321456" or "Find accounts with balances greater than $50,000.00".

Now consider a news story about the merger of two banks. The story will have some attributes, such as the headline and source of the story, but the primary content is the story itself. In a database system, this critical piece of information would typically be stored as a single large attribute with no internal structure. Most of the queries submitted to a web search engine such as Google[2] that relate to this story will be of the form "bank merger" or "bank takeover". To do this search, we must design algorithms that can compare the text of the queries with the text of the story and decide whether the story contains the information that is being sought. Defining the meaning of a word, a sentence, a paragraph, or a whole news story is much more difficult than defining an account number, and consequently comparing text is not easy. Understanding and modeling how people compare texts, and designing computer algorithms to accurately perform this comparison, is at the core of information retrieval.

Increasingly, applications of information retrieval involve multimedia documents with structure, significant text content, and other media. Popular information media include pictures, video, and audio, including music and speech. In

[2] http://www.google.com

some applications, such as in legal support, scanned document images are also important. These media have content that, like text, is difficult to describe and compare. The current technology for searching non-text documents relies on text descriptions of their content rather than the contents themselves, but progress is being made on techniques for direct comparison of images, for example.

In addition to a range of media, information retrieval involves a range of tasks and applications. The usual search scenario involves someone typing in a query to a search engine and receiving answers in the form of a list of documents in ranked order. Although searching the World Wide Web (*web search*) is by far the most common application involving information retrieval, search is also a crucial part of applications in corporations, government, and many other domains. *Vertical search* is a specialized form of web search where the domain of the search is restricted to a particular topic. *Enterprise search* involves finding the required information in the huge variety of computer files scattered across a corporate intranet. Web pages are certainly a part of that distributed information store, but most information will be found in sources such as email, reports, presentations, spreadsheets, and structured data in corporate databases. *Desktop search* is the personal version of enterprise search, where the information sources are the files stored on an individual computer, including email messages and web pages that have recently been browsed. *Peer-to-peer search* involves finding information in networks of nodes or computers without any centralized control. This type of search began as a file sharing tool for music but can be used in any community based on shared interests, or even shared locality in the case of mobile devices. Search and related information retrieval techniques are used for advertising, for intelligence analysis, for scientific discovery, for health care, for customer support, for real estate, and so on. Any application that involves a *collection*[3] of text or other unstructured information will need to organize and search that information.

Search based on a user query (sometimes called *ad hoc search* because the range of possible queries is huge and not prespecified) is not the only text-based task that is studied in information retrieval. Other tasks include *filtering*, *classification*, and *question answering*. Filtering or tracking involves detecting stories of interest based on a person's interests and providing an alert using email or some other mechanism. Classification or categorization uses a defined set of labels or classes

[3] The term *database* is often used to refer to a collection of either structured or unstructured data. To avoid confusion, we mostly use the term *document collection* (or just *collection*) for text. However, the terms *web database* and *search engine database* are so common that we occasionally use them in this book.