HZ BOOKS
华章教育

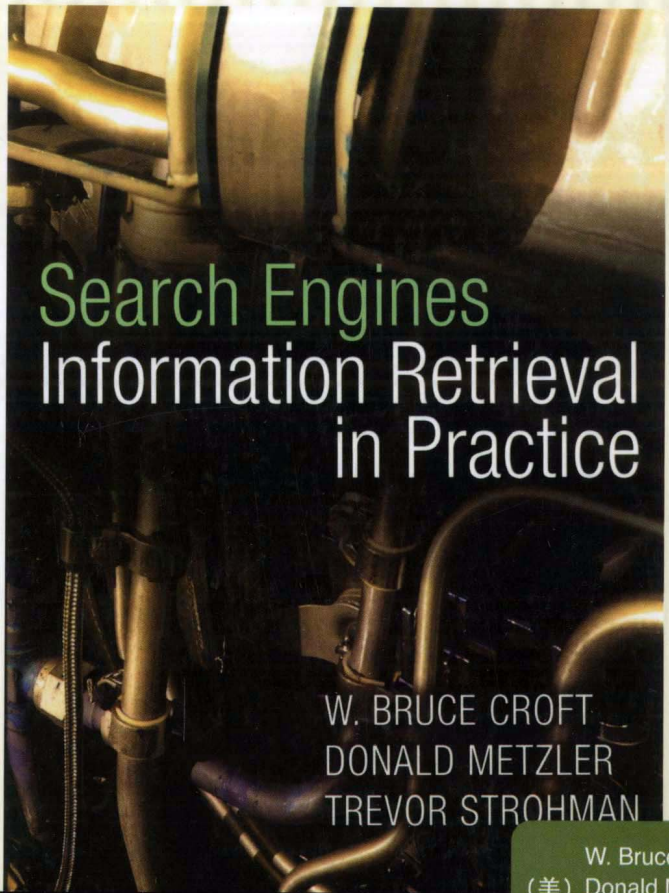
经 典 原 版 书 库

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

2010年12月第1版第2次印刷

150 mm × 214 mm · 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

华章教育

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

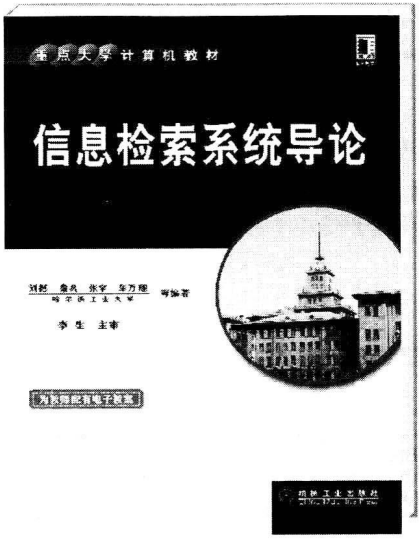
The supplements are available at www.search-engines-book.com, or at www.aw.com.

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

好书推荐

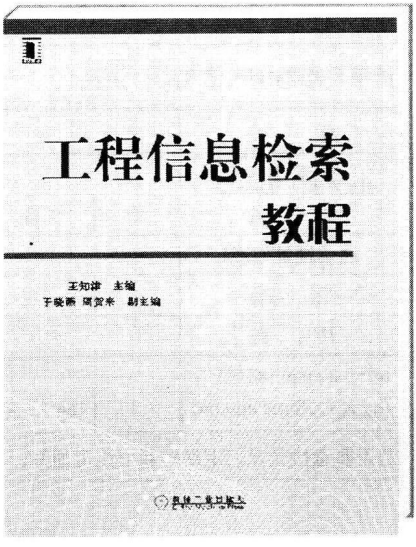


信息检索系统导论

ISBN：978-7-111-24607-7

作者：刘挺、秦兵、张宇、车万翔

定价：35.00



工程信息检索教程

ISBN： 978-7-111-25227-6

作者：王知津

定价：36.00

北京培生信息中心
中国北京海淀区中关村大街甲59号
人大文化大厦1006室

邮政编码：100872
电话：(8610)82504008/9596/9586
传真：(8610)82509915

Beijing Pearson Education
Information Centre
Room1006,CultureSquare No.59 Jia,
Zhongguancun Street
Haidian District, Beijing, China100872
TEL: (8610)82504008/9596/9586
FAX: (8610)82509915

尊敬的老师：

您好！

为了确保您及时有效地申请教辅资源，请您务必完整填写如下教辅申请表，加盖学院的公章后传真给我们，我们将会为您开通属于您个人的唯一账号以供您下载与教材配套的教师资源。

请填写所需教辅的开课信息：

采用教材			□中文版 □英文版 □双语版
作　　者		出版社	
版　　次		ISBN	
课程时间	始于　　年　月　日	学生人数	
	止于　　年　月　日	学生年级	□专科 □本科1/2年级 □研究生 □本科3/4年级

请填写您的个人信息：

学　　校			
院系/专业			
姓　　名		职　　称	□助教 □讲师 □副教授 □教授
通信地址/邮编			
手　　机		电　　话	
传　　真			
official email(必填) (eg:XXX@ruc.edu.cn)		email (eg:XXX@163.com)	

是否愿意接受我们定期的新书讯息通知：　□是　□否

系/院主任：_____（签字）

（系/院办公室章）

_____年_____月_____日

Please send this form to: Service.CN@pearson.com

Website: www.pearsonhighered.com/educator

教师服务登记表

尊敬的老师：

您好！感谢您购买我们出版的 _____ 教材。

机械工业出版社华章公司本着为服务高等教育的出版原则，为进一步加强与高校教师的联系与沟通，更好地为高校教师服务，特制此表，请您填妥后发回给我们，我们将定期向您寄送华章公司最新的图书出版信息。为您的教材、论著或译著的出版提供可能的帮助。欢迎您对我们的教材和服务提出宝贵的意见，感谢您的大力支持与帮助！

个人资料（请用正楷完整填写）

教师姓名		□先生 □女士	出生年月		职务		职称：□教授 □副教授 □讲师 □助教 □其他
学校			学院			系别	
联系电话	办公： 宅电： 移动：			联系地址及邮编			
				E-mail			
学历		毕业院校		国外进修及讲学经历			
研究领域							

主讲课程	现用教材名	作者及出版社	共同授课教师	教材满意度
课程： □专 □本 □研 人数： 学期：□春□秋				□满意 □一般 □不满意 □希望更换
课程： □专 □本 □研 人数： 学期：□春□秋				□满意 □一般 □不满意 □希望更换

样书申请			
已出版著作		已出版译作	
是否愿意从事翻译/著作工作 □是 □否		方向	
意见和建议			

填妥后请选择以下任何一种方式将此表返回：（如方便请赐名片）

地 址：北京市西城区百万庄南街1号 华章公司营销中心 邮编：100037

电 话：(010) 68353079 88378995 传真：(010)68995260

E-mail:hzedu@hzbook.com markerting@hzbook.com 图书详情可登录http://www.hzbook.com网站查询

Contents