



原书第2版

Ricardo Baeza-Yates Berthier Ribeiro-Neto 著

黄萱菁 张奇 邱锡鹏 译

Modern Information Retrieval

The Concepts and Technology behind Search Second Edition

Modern Information Retrieval

the concepts and technology behind search

[illegible]

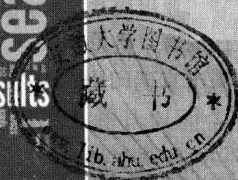
机械工业出版社
China Machine Press

原书第2版

Ricardo Baeza-Yates Berthier Ribeiro-Neto 著

黄萱蓓 张奇 邱锡鹏 译

The Concepts and Technology behind Search Second Edition



本书论述信息检索的概念和技术、这些技术在搜索引擎中的应用,及其对相关领域知识的影响等,主要内容包括:用户界面设计;经典的信息检索模型、结果质量评估和用户相关反馈;文档和查询概念及其相关技术;文档集索引和搜索技术;Web 文档的爬取、检索和排序;结构化文本检索、多媒体检索和企业搜索;图书馆系统和数字图书馆等。

本书内容广泛、细节丰富、深入浅出,可以作为高等院校信息管理与信息系统、计算机科学与技术、图书馆学、情报学、档案学等专业本科生和研究生的教材或参考书,对从事信息检索及系统分析、设计的实际工作者也有较高的参考价值。

Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition (9780321416919).

Copyright © 2011 by Pearson Education Limited.

This translation of Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition (9780321416919) is published by arrangement with Pearson Education Limited.

All rights reserved.

本书中文简体字版由英国 Pearson Education 培生教育出版集团授权出版。

封底无防伪标均为盗版

版权所有,侵权必究

本书法律顾问 北京市展达律师事务所

本书版权登记号:图字:01-2010-6144

图书在版编目(CIP)数据

现代信息检索(原书第2版)/(智)贝泽-耶茨(Baeza-Yates, R.)等著;黄萱菁,张奇,邱锡鹏译. —北京:机械工业出版社,2012.8

(计算机科学丛书)

书名原文:Modern Information Retrieval: The Concepts and Technology behind Search, Second Edition

ISBN 978-7-111-38599-8

I. 现… II. ①贝… ②黄… ③张… ④邱… III. 情报检索 IV. G252.7

中国版本图书馆CIP数据核字(2012)第114931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:盛思源

藁城市京瑞印刷有限公司印刷

2012年10月第1版第1次印刷

185mm×260mm·43.25印张

标准书号:ISBN 978-7-111-38599-8

定价:118.00元

凡购本书,如有缺页、倒页、脱页,由本社发行部调换

客服热线:(010) 88378991; 88361066

购书热线:(010) 68326294; 88379649; 68995259

投稿热线:(010) 88379604

读者信箱:hzsj@hzbook.com

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中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公司较早意识到“出版要为教育服务”。自1998年开始，我们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, Brian W. Kernighan, Dennis Ritchie, Jim Gray, Alfred 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。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方式如下：

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

华章科技图书出版中心

十多年前,我刚刚开始接触信息检索,读了几本经典教材,也看了不少论文,但因为缺乏有关信息检索系统实现的文献,上手很慢。同学从国外回来,带来了 Ricardo Baeza-Yates 撰写的《Information Retrieval: Data Structures and Algorithms》,该书系统地介绍了信息检索领域的重要数据结构和算法,可操作性极强,我简直是如获至宝,也因而记下了 Ricardo 的大名。

几年后, Ricardo 和 Berthier 合著了本书的第 1 版,拜读之后,惊叹于作者不仅具备娴熟的实践技巧,深厚的理论功底,而且还有很强的大局观、洞察力和驾驭素材的能力。该书毫无疑问地成为复旦大学研究生课程“信息检索”的首选教材。

去年春天,好友秦兵教授告诉我,机械工业出版社引进了这本书的第 2 版,打算翻译成中文版,如果我有兴趣,她可以向出版社推荐。虽然此前从未翻译过任何书籍,自己的工作负担也已很重,但出于对本书及作者的推崇,我毫不犹豫地接下了这份任务。

收到出版社寄来的样书后,我发现第 2 版与第 1 版相比可谓截然不同。应该说本书的第 1 版已经足够优秀,被世界上数以百计的大学和学校采纳为教科书,但两位作者仍然大刀阔斧地对许多章节进行了彻头彻尾的修改,并增加了许多新的章节,第 2 版的 60%~70% 由新的素材组成即是印证。

第 2 版的巨大变化来自于以下原因:第一,随着互联网的普及,搜索引擎进入人们的日常生活中,成为获取信息的重要入口,用户需求带动了搜索引擎产业的飞速发展,谷歌、雅虎、必应和百度等企业成长为极有影响力的互联网公司,作者因而在本书中加入了许多和搜索引擎有关的章节,如搜索引擎界面、并行和分布式检索、Web 爬取等;第二,产业界的繁荣吸引了大量的研究人员和从业者,而搜索引擎的普及带来了海量的真实用户数据,这些都极大地促进了信息检索研究水平的提高,本书为此增加了语言模型、排序学习等新的研究内容;第三,撰写第 1 版的时候,作者还是大学教师,在撰写第 2 版之际,他们开创了自己的搜索事业,之后进入了主流搜索引擎公司工作,丰富的经历带来更开阔的视野,对搜索引擎也有了更深入的了解。第 2 版不仅反映了信息检索产业界和学术界的變化,也体现了他们在研究、开发和实现信息检索技术,并将其应用于互联网过程中的心得体会。

本书主要由黄萱菁、张奇和邱锡鹏三人执笔翻译。周雅倩、王秉卿、计峰、丁卓冶、吴龔、周金龙和刘昭等同事和研究生帮助做了许多资料整理、录入、校对等辅助工作,李伟和路红两位同事帮助我们了解了多媒体检索所特有的许多概念,王春华、盛思源两位编辑帮助发现了译稿中的许多不足之处,本书两位原作者帮助澄清了许多问题,复旦大学计算机学院为本书的翻译提供了有力支持,在此一并致谢。

翻译一本书,比我想象的要困难很多。好的译者,不仅要對领域知识有充分的了解和掌握,也需要流畅精彩的文笔。然而,“知易行难”,本书的几位译者都是理工科出身,虽然都是具有一定经历的信息检索研究人员,但第一次从事翻译工作,水平有限,错漏之处在所难免,敬请各位读者谅解并批评指正。

黄萱菁

2012 年春于浦东张江

自从本书第1版出版以来,信息检索(Information Retrieval, IR)领域发生了许多变化,其中许多和Web有关。首先,Web上的海量信息已将搜索引擎转化为寻找和发现用户感兴趣信息的关键工具。其次,由于搜索引擎的本质核心是信息检索系统,这就有力地证明了信息检索技术可以应用于具有巨大查询流量的海量文档集。

紧随这一演变趋势,在本书第1版出现以后的短短几个月内,我们在巴西和智利就开始了搜索引擎的研究。后来,我们进入谷歌和雅虎这两个主要的搜索引擎公司工作,对搜索引擎的一切行为有了更深入的了解。因此,本书第2版不仅反映了信息检索领域的变化,也反映了我们自己正在研究、开发和实现的信息检索技术,以及将其应用于Web的经验。

本书第1版并不是按照标准方式书写的,对于我们觉得没有足够专业知识的领域,我们邀请专家撰写相关章节。所以,从某种意义上说,我们先于Web 2.0的发展趋势进行了团队协作。我们的宗旨是精心协调和监督所有的写作内容,使本书成为有机的整体。在某种程度上,我们的努力颇有成效。事实上,第1版卖得非常好,成为了信息检索领域的畅销书,并已重印多次。该书已被数以百计的大学和学校采纳。它首先被翻译成韩文,其次是中文,还有一个特别低价的版本已在印度出版。因此,第1版出版后仅仅一两年,我们就开始谈论第2版。这个想法一直到2004年我们向出版商提交建议书并获得批准后才得以实现。最终在2005年11月,也就是四年多前,我们开始第2版的工作。今天,我们终于完成了!

在第2版中,我们遵循着和第1版相同的方法,因为它明显行之有效。尽管如此,我们仍然是更多章节的作者或合著者,而且我们采取了更强有力的手段对其他章节的内容进行设计。我们不得不完全修改许多章节,并增加了许多新的章节。因此,第2版的60%~70%是由新素材组成的,和第1版的不同之处主要在以下几个方面:

- 完全重组第1章内容。
- 增加文本分类、Web爬取、结构化文本检索和企业搜索等新章节,以及一个关于开源搜索引擎的新附录。
- 完全重写用户界面、多媒体检索和数字图书馆等章节。
- 扩充章节内容,以包括重要的新进展,例如语言模型、新的评价准则、查询特性、基于集群的信息检索和分布式信息检索、排序学习、搜索引擎界面和个性化等。
- 改进本书网站,其中包括本书所有章节的全套幻灯片和推荐的练习列表,使之成为信息检索的参考教学资源。

最后的成果是,和第1版相比,第2版几乎有两倍的篇幅,并包含两倍以上参考文献。总之,如果你喜欢本书第1版,我们希望你更喜欢这个第2版。万一你不喜欢第1版,我们希望这一次你会改变主意。

Ricardo Baeza-Yates 于西班牙巴塞罗那

Berthier Ribeiro-Neto 于巴西贝洛奥里藏特

2010年12月

第 1 版前言

Modern Information Retrieval: The Concepts and Technology behind Search, 2E

随着 Web 的发展, 以及时尚而廉价的图形用户界面和海量存储设备的问世, 信息检索在过去几年中发生了巨大的变化。传统的信息检索教科书已相当过时, 为此, 最近已经出版了一些新的信息检索书籍。不过, 我们相信, 仍然非常需要这样一本书, 它能够从计算机科学的视角, 而不是从用户为中心的视角, 以严密和完整的方式来介绍这个领域。本书致力于部分地填补这一鸿沟, 它既可以作为信息检索的入门教材, 也可以用于该方向的研究生课程。

本书是由相互补充和平衡的两部分组成。核心部分包括由本书设计者撰写或合著的 9 章。第二部分和第一部分紧密相连, 共分为 6 章。这部分由相关领域的领先研究人员撰写, 介绍最新的研究进展。所有章节采用相同的符号和术语。因此, 尽管事实上邀请了多位撰稿人, 但这本书并不是由不同作者撰写的章节汇编成的合著, 而是一本教科书。此外, 与合著相比, 本书的主要作者精心设计了全书的内容和结构, 以便展示现代信息检索中所有重要方面的内在联系。

从信息检索模型到文本索引, 从信息检索可视化工具和界面到 Web, 从多媒体信息检索到数字图书馆, 本书都广泛涵盖, 而且细节丰富。考虑到信息检索对现代社会显而易见的相关性和重要性, 我们希望本书对世界各地的信息科学、计算机科学与图书馆学等学科研究的进一步传播起到促进作用。

Ricardo Baeza-Yates 于智利圣地亚哥

Berthier Ribeiro-Neto 于巴西贝洛奥里藏特

1998 年 10 月

我们对在过去几年间向我们提供了有用和有益的意见、评论和建议的人们致以衷心的感谢。本书内容和素材组织的改进，很大程度上归功于他们。如果没有他们的帮助，第2版的质量将大大下降。仍然存在的任何错误——希望只有少量，完全是我们的责任。

第一，我们对所有撰稿人所体现出的奉献精神 and 浓厚兴趣表示感谢，他们是 Eric Brown、Carlos Castillo、Marcos Gonçalves、David Hawking、Marti Hearst、Mounia Lalmas、Yoelle Maarek、Christian Middleton、Gonzalo Navarro、Dulce Ponceleón、Edie Rasmussen、Malcolm Slaney 和 Nivio Ziviani。他们所体现的专业知识是我们所欠缺的。

第二，我们感谢对第2版的新内容提供直接或者间接贡献或影响的人们，他们是 Omar Alonso（他指出我们偏离了众包的重要趋势）、Paolo Boldi（Web 图压缩）、Pavel Calado（文本分类）、Marco Cristo（他对于文本分类章节的意见导致了对素材的整体重组）、Christos Faloutsos（多维索引）、Winston Hsu（多媒体）、Flavio Junqueira（分布式检索）、Edleno Moura（检索评价）、Vanessa Murdock（查询困难性）、Martin Porter（词干提取算法）、Mark Sanderson（他的尖锐意见导致检索评价章节的重大改进）、Fabrizio Silvestri（URL 排序）和 Gleb Skobeltsyn（对等网络信息检索）。另外，我们还感谢巴西米纳斯吉拉斯州联邦大学 Marcos Gonçalves 的多位研究生的贡献，他们评阅了文本分类章节并书写了大量意见。

第三，我们需要感谢所有提供第1版勘误信息、提出改进建议和对第2版草稿提出修改意见的人们。对于勘误表，我们只提及发现错误的第一人，否则名单将太长。他们是：Omar Alonso、Jose Hilario Canos、Berkant Barla Cambazoglu、Ernie Davis、Anne Dieckema、Bill Dimm、Joaquim Gabarro、Jamie Geddes、Eduardo Graells、Kyoung-Soo Han、Claudia Hauff、Shoujie He、Ben Houston、Puay-Leng Lee、Songwook Lee、Shian-Hua Lin、Mildrid Ljosland、Chang-Tien Lu、Mari Carmen Marcos、Peter Mika、Vanessa Murdock、Joanna Plattner、Luz Rello、Hee-Cheol Seo、Ben Shneiderman、Helge Grenager Solheim、Ellen Spertus、Markus Stocker、Kazunari Sugiyama、Satoru Takabayashi、Juha Takkinen、Luong Minh Thang、Yannis Tzitzikas、Fredrik Wallenberg、Theo van der Weide、John Westbrook、Judith Winter、Sui Xi、Peng Yong、Hugo Zaragoza 和 Yonghui Zhang。上述名单可能不全。

第四，我们特别感谢 David Fernandes，本书网站上有他制作的教学幻灯片。他也耐心指出了许多小错误和不一致的地方。我们也需要提及我们的雇主雅虎和谷歌，他们为我们完成撰写本书的艰巨任务提供了隐性支持。

第五，我们感谢 Pearson Education 公司的编辑。他们是 Kate Brewin、Simon Plumtree、Owen Knight 和 Rufus Curnow。在最重要的出版过程中，他们给予了支持。Anita Atkinson 和 Jenny Oates 分别是本书的文字编辑和校对，我们感谢她们的帮助。

最后也是最重要的，感谢 Helena、Rosa 和我们的孩子，他们再次忍受了我们一连串的国际旅行、周末加班和不规律的工作时间。在过去的4年里，他们总是在问：你们什么时候完成这本书？

第 1 版致谢

Modern Information Retrieval: The Concepts and Technology behind Search, 2E

我们对在过去几个月的写作过程中向我们提供了有用和有益帮助的各位人士致以衷心的感谢。如果没有他们的关心，本书很可能无法完成。

第一，我们对所有撰稿人所体现出的奉献精神 and 浓厚兴趣表示感谢。他们是 Elisa Bertino、Eric Brown、Barbara Catania、Christos Faloutsos、Elena Ferrari、Ed Fox、Marti Hearst、Gonzalo Navarro、Edie Rasmussen、Ohm Sornil 和 Nivio Ziviani。他们所体现的专业知识是我们所欠缺的。我们也感谢他们在编辑和交叉审阅过程中给予的耐心，这是一种相当难以平衡的工作。

第二，我们要感谢对出版本书感兴趣的所有人士，特别是 Scott Delman 和 Doug Sery。

第三，对于 Addison Wesley Longman 出版社对我们的兴趣和给予的鼓励，以及在整个过程中所做的优秀工作，我们在此深表感谢。他们的代表是 Keith Mansfield、Karen Sutherland、Bridget Allen、David Harrison、Sheila Chatten、Helen Hodge 和 Lisa Talbot。他们联系的评阅人阅读了本书的早期（也是非常原始的）方案，并提供了很好的反馈意见，显示了深刻的洞察力。鉴于一位匿名评阅人的客观评论，“并行和分布式检索”章节从不很合适的“信息检索应用”部分移到了“文本信息检索”部分。鉴于检索评价的重要性，另一位热心的评阅人强烈建议我们将它单列为一章。

第四，我们要感谢和我们讨论过本书撰写计划的所有人士。Doug Oard 很早就评阅了本书的草案。Gary Marchionini 是本书的早期支持者，并在我们写书的过程中保持联系。Bruce Croft 从一开始就鼓励我们。Alberto Mendelzon 提供了 Web 搜索章节的初始方案和参考文献列表。Ed Fox 在百忙之中对第 1 章“引言”提出了富有洞察力的评阅意见，使我们极大地改进了这一章。他也认真评阅了信息检索建模的内容。Marti Hearst 很早就对我们的方案深表兴趣，在整个编辑过程中提供了帮助，并且是一个热情的支持者和伙伴。

第五，我们感谢我们所在的机构，智利大学和巴西米纳斯吉拉斯州联邦大学计算机科学系的支持，以及来自国家研究机构——巴西科技发展委员会（CNPq）、智利国家科技研究委员会（CONICYT）和国际合作项目的经费资助，特别是拉美科技发展项目（CYTED）项目“Web 信息管理与检索环境（Environment for Information Managing and Retrieval in the World Wide Web, AMYRI, 编号 VII.13）”和巴西科学研究与发展项目资助署（Finep）项目“移动计算机的信息系统（Information Systems for Mobile Computers, SIAM）”。

最重要的是，感谢 Helena、Rosa 和我们的孩子，他们忍受了我们一连串的国际旅行、周末加班和不规律的工作时间。

我们感谢以下复制版权材料的许可：

图

图 2-1 和图 2-12 来自 Yelp!, <http://www.yelp.co.uk/>, Yelp! Inc.; 图 2-3 来自 NextBio.com; 图 2-5、图 4-13b、图 11-10c、图 11-11a 和图 11-13 来自 www.google.co.uk 提供的谷歌系统截图; 图 2-6 来自 <http://biosearch.berkeley.edu>, M. A. Hearst 版权所有; 图 2-7 来自 Microsoft Corporation 的产品截图重印许可; 图 2-13 来自 Findex、FindEx.com, Inc. 及其许可者版权所有©2010; 图 2-15 来自 “Graphical query specification and dynamic result previews for a digital library, Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology (UIST’98) pp.143-151 (Jones, S.1998)”, <http://doi.acm.org/10.1145/288392.288595>, Association for Computing Machinery, Inc. 版权所有©1998, 重印经许可; 图 2-16 来自 “Research: TileBars”, <http://people.ischool.berkeley.edu/~hearst/research/tilebars.html>, M. A. Hearst 版权所有; 图 2-17a 来自 “Search User Interfaces, Cambridge University Press (Hearst, M. A.2009)” 的图 10-17a, M. A. Hearst 版权所有; 图 2-17b 来自 “INSYDER: a content-based visual-information-seeking system for the web, International Journal on Digital Libraries, pp.25-41 (Reiterer, H., Tullius, G. and Mann, T. M.2005)”, 许可来自 Springer Science + Business Media and CCC 及 H. Reiterer 教授; 图 2-18 来自 “Using thumbnails to search the Web, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI’01), pp.198-205 (Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J. and Pirolli, P.2001)”, <http://doi.acm.org/10.1145/365024.365098>, Association for Computing Machinery, Inc. 版权所有©2001, 重印经许可; 图 2-20a 来自 “Evaluating a system for interactive exploration of large, hierarchically structured document repositories, Proceedings of the IEEE Symposium on Information Visualization (INFOVIS’04), pp.127-134 (Granitzer, M., Kienreich, W., Sabol, V., Andrews, K. and Klieber, W.2004)”, IEEE 版权所有©2004; 图 2-20b 来自 “Search result visualisation with xFIND, Proceedings of User Interfaces to Data Intensive Systems (UIDIS 2001), pp.50-58 (Andrews, K., Gutl, C., Moser, J., Sabol, V. and Lackner, W.2001)”, IEEE 版权所有©2001; 图 2-21 来自 <http://kylescholz.com/projects/wordnet/>, Kyle Scholz; 图 2-22 来自 “The Word tree, an interactive visual concordance, IEEE Transactions on Visualization and Computer Graphics, 14 (6), pp.1221-1228 (Wattenberg, M. and Fernanda, B.2008)”, IEEE 版权所有©2008; 图 2-23 来自婴儿名字流行度图 NameVoyager, <http://www.babynamewizard.com>; 图 2-24 来自 “Avian flu case study with nSpace and GeoTime, Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST’06) pp.27-34 (Proulx, P. et al.2006)”, IEEE 版权所有©2006; 图 5-4 仿自 “Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search, ACM Transactions on Information Systems, 25 (2) (Joachims, T., Granka, L., Pan, B., Hembrooke, H.,

Radlinski, F. and Gay, G. 2007)”, <http://doi.acm.org/10.1145/1229179.1229181>, Association for Computing Machinery, Inc. 版权所有©2007, 重印经许可; 图 7-4 和图 7-5 来自 “The impact of caching on search engines, Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGR’07) (Baeza-Yates, R. et al. 2007)”, <http://doi.acm.org/10.1145/1277741.1277775>, Association for Computing Machinery, Inc. 版权所有©2007, 重印经许可; 图 7-6 来自 “Query usage mining in search engines, Web Mining Applications and Techniques (Baeza-Yates, R. (Scime, A. ed.) 2004)”, Idea Group, 重印经出版商 IGI Global 许可; 图 10-1 改编自 “Load balancing for term-distributed parallel retrieval, Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 348-355 (Moffat, A., Webber, W. and Zobel, J. 2006)”, <http://doi.acm.org/10.1145/1148170.1148232>, Association for Computing Machinery, Inc. 版权所有©2006, 重印经许可; 图 10-12 和图 10-13 来自 “Challenges on distributed web retrieval, Proceedings of ICDE 2007, pp. 6-20 (2007)”, IEEE 版权所有©2007; 图 10-14 来自 “A pipelined architecture for distributed text query evaluation, Information Retrieval, 10 (3), pp. 205-231 (Webber, W., Moffat, A., Zobel, J. and Baeza-Yates, R. 2007)”, 许可来自 Springer Science + Business Media; 图 11-1 来自 “Graph structure in the web: experiments and models, Proceedings of the North Conference on World Wide Web, pp. 309-320 (Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. 2000)”, Elsevier 版权所有 (2000); 图 11-3a 来自 M. Crovella, 1998; 图 11-3b 来自 “Self-similarity in World Wide Web traffic: evidence and possible causes, SIGMETRICS’96: Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modelling of Computer Systems, 24, pp. 160-169 (Crovella, M. E. and Bestavros, A. 1996)”, <http://doi.acm.org/10.1109/90.650143>, Association for Computing Machinery, Inc. 版权所有©1996, 重印经许可; 图 11-4 和图 11-5 来自 “Generic damping functions for propagating importance in linkbased ranking algorithms, Internet Mathematics, 3 (4), pp. 445-478 (Baeza-Yates, R., Boldi, P. and Castillo, C. 2006)”, A. K. Peters, Ltd. 版权所有 2006; 图 11-7 仿自 “Challenges in building large-scale information retrieval systems: invited talk presentation”, <http://research.google.com/people/jeff/WSDM09-keynote.pdf>, Jeffrey Dean; 图 11-8 来自 “Design trade-offs for search engine caching, TWEB, 2 (4) (Baeza-Yates, R. A., Gionis, A., Juncqueira, F., Murdock, V., Plachouras, V. and Silvestri, F. 2008)”, <http://doi.acm.org/10.1145/1409220.1409223>, Association for Computing Machinery, Inc. 版权所有©2008, 重印经许可; 图 11-10a 来自 Ask 系统截图, IAC Search & Media, Inc. 保留所有权利©2010. ASK.COM、ASK JEEVES、ASK 商标、ASK JEEVES 商标及其他出现在 Ask.com 和 Ask Jeeves 网站上的商标属于 IAC Search & Media, Inc. 及其授权者; 图 11-10b 及图 11-15 来自 Bing 系统截图, 重印经 Microsoft Corporation 许可; 图 12-8 来自 “Synchronizing a database to improve freshness, Proceedings of ACM International Conference on Management of Data (SIGMOD), pp. 117-128 (Cho, J. and Garcia-Molina, H. 2000)”, <http://doi.acm.org/10.1145/342009.335391>, Association for Computing Machinery, Inc. 版权所有©2000, 重印经许可; 图 13-9 来自 INEX 2006 评估界面, 由 Mounia Lalmas 教授提供; 图 14-4 来自

IBM Almaden 研究中心；图 14-6 和图 14-8 来自 IBM Almaden 研究中心 QBIC 系统，Jim Hafner 的许可；图 14-9 来自 “A bipartite graph model for associating images and text, IJ-CAI-2007 Workshop on Multimodal Information Retrieval (Srinivasan, S. H. and Slaney, M. 2007)”；图 14-10 来自 “Image retrieval on large-scale image databases, Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR 07), pp. 17-24 (Horster, E., Lienhart, R. and Slaney, M. 2007)”, <http://doi.acm.org/10.1145/1282280.1282283>, Association for Computing Machinery, Inc. 版权所有©2007, 重印经许可；图 14-13 和图 14-14 来自 Kyogu Lee；图 14-16 来自 Carnegie Mellon 大学计算机学院技术报告 “Video skimming for quick browsing based on audio and image characterization, Technical Report CMU-CS-95-186 (Smith, M. A. and Kanade, T. 1995)”；图 14-17 来自 “Video manga: generating semantically meaningful video summaries, MULTIMEDIA ’99: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1), pp. 383-392 (Uchihashi, S. et al. 1999)”, <http://doi.acm.org/10.1145/319463.319654>, Association for Computing Machinery, Inc. 版权所有©1999, 重印经许可；图 14-18 来自 Sarnoff Corporation 的 Harpreet Sawhney；图 14-19 来自 “Salient stills, ACM Transactions on Multimedia Computing, Communications and Applications, 1 (1), pp. 16-36 (Teodosio, L. and Bender, W. 2005)”, <http://doi.acm.org/10.1145/1047936.1047940>, Association for Computing Machinery, Inc. 版权所有©2005, 重印经许可；图 14-20 来自 “PanoramaExcerpts: Extracting and packing panoramas for video browsing, MULTIMEDIA ’97: Proceedings of the Fifth ACM International Conference on Multimedia, pp. 427-436 (Taniguchi, Y., Akutsu, A. and Tonomura, Y. 1997)”, <http://doi.acm.org/10.1145/266180.266396>, Association for Computing Machinery, Inc. 版权所有©1997, 重印经许可；图 14-21 来自 “Hierarchical brushing in a collection of video data, Proceedings of Hawaii International Conference on System Science (HICSS) (2001)”, IEEE 版权所有©2001；图 14-26 来自 “Automatic recognition of audiovisual speech: recent progress and challenges, Proceedings of the IEEE (Potamianos, G., Neti, C., Gravier, G., Garg, A. and Senior, A. W. 2003)”, IEEE 版权所有©2003；图 14-28 来自 “Multimedia edges: finding hierarchy in all dimensions, Proceedings of 9th ACM International Conference on Multimedia (Slaney, M., Ponceleon, D. and Kaufman, J. 2001)”, <http://doi.acm.org/10.1145/500141.500149>, Association for Computing Machinery, Inc. 版权所有©2001, 重印经许可；图 14-29 来自 “Comparison of automatic shot boundary detection algorithms, SPIE Image and Video Processing VII, 3656, 290-301 (Lienhart, R. 1999)”, SPIE；图 15-3 来自 Oxfam Australia；图 15-5 来自 “Evaluation by comparing result sets in context, Proceedings of the 15th ACM International Conference on Information and Knowledge Management pp. 94-101 (Thomas, P. and Hawking, D. 2006)”, <http://doi.acm.org/10.1145/1183614.1183632>, ACM 版权所有©2006；图 16-1 来自 Edie Rasmussen, 许可来自 The Network Development and MARC Standards Office；图 16-2 来自 “Find... books or journals”, <http://www.library.ubc.ca/home/research.html>, 不列颠哥伦比亚大学网站 (2010), 许可后使用；图 16-4、图 16-5、图 16-6 和图 16-7 来自 DIALOG, Dialog® 的界面及截屏, 经 Dialog LLC. 许可后改编, Dialog 产品名是 Dialog LLC. 的注册商标；图 16-4 来自 EBSCO Publishing, Inc. 的许可。

表

表 4-2 改编自 “Overview of the sixth text retrieval conference (TREC-6), Proceedings of the Sixth Text REtrieval Conference (TREC-6) (Voorhees, E. and Harman, D. 1997)”；表 7-3 来自 “From e- sex to e- commerce: Web search changes, Computer, 35 (3), pp. 107-109 (Spink, A. , Jansen, B. J. , Wolfram, D. and Saracevic, T. 2002)”，IEEE 版权所有© 2002。

文字

原书 159 页的引文来自 <http://trec.nist.gov>, NIST。

在某些情况下，我们已经无法追溯版权材料的所有者，读者如能提供任何帮助信息，我们将不胜感激。

出版者的话	2.2.3 导航与搜索	18
译者序	2.2.4 对搜索过程的观察	18
第2版前言	2.3 现今的搜索界面	19
第1版前言	2.3.1 启动搜寻	19
第2版致谢	2.3.2 查询描述	19
第1版致谢	2.3.3 查询描述界面	20
出版商致谢	2.3.4 检索结果显示	22
	2.3.5 查询重构	24
	2.3.6 组织搜索结果	26
第1章 引言	2.4 搜索界面的可视化	32
1.1 信息检索	2.4.1 可视化布尔语法	32
1.1.1 信息检索的早期发展	2.4.2 可视化查询结果中的	33
1.1.2 图书馆和数字图书馆中的	查询项	
信息检索	2.4.3 可视化词语和文档间	36
1.1.3 舞台中央的信息检索	的关系	
1.2 信息检索问题	2.4.4 文本挖掘的可视化	38
1.2.1 用户的任务	2.5 搜索界面的设计和评价	40
1.2.2 信息检索与数据检索	2.6 趋势和研究问题	42
1.3 信息检索系统	2.7 文献讨论	42
1.3.1 信息检索系统的软件架构	第3章 信息检索建模	44
1.3.2 检索和排序过程	3.1 信息检索模型	44
1.4 Web	3.1.1 建模和排序	44
1.4.1 Web 简史	3.1.2 信息检索模型描述	44
1.4.2 电子出版时代	3.1.3 信息检索模型的分类体系	45
1.4.3 Web 如何改变搜索	3.2 经典信息检索	47
1.4.4 Web 上的实际问题	3.2.1 基本概念	47
1.5 本书的组织结构	3.2.2 布尔模型	49
1.5.1 本书的重点	3.2.3 项权重	50
1.5.2 本书的内容	3.2.4 TF-IDF 权重	52
1.6 本书的教学资源网站	3.2.5 文档长度归一化	56
1.7 文献讨论	3.2.6 向量模型	57
第2章 用户搜索界面	3.2.7 概率模型	59
2.1 介绍	3.2.8 经典模型之间的简单比较	64
2.2 人们如何搜索	3.3 其他集合论模型	64
2.2.1 信息查找与探索式搜索	3.3.1 基于集合的模型	64
2.2.2 信息搜寻的经典模型与	3.3.2 扩展布尔模型	68
动态模型		

3.3.3 模糊集模型	70	4.5.4 众包	124
3.4 其他代数模型	72	4.5.5 使用点击数据的评价	125
3.4.1 广义向量空间模型	72	4.6 实践说明	126
3.4.2 潜在语义索引模型	74	4.7 趋势和研究问题	127
3.4.3 神经网络模型	75	4.8 文献讨论	127
3.5 其他概率模型	76	第5章 相关反馈与查询扩展	129
3.5.1 BM25 模型	77	5.1 介绍	129
3.5.2 语言模型	78	5.2 反馈方法的框架	129
3.5.3 随机差异模型	83	5.3 显式相关反馈	131
3.5.4 贝叶斯网模型	85	5.3.1 向量模型的相关反馈: Rocchio 方法	131
3.6 其他模型	90	5.3.2 概率模型的相关反馈	133
3.6.1 超文本模型	90	5.3.3 相关反馈的评价	134
3.6.2 基于 Web 的模型	91	5.4 基于点击的显式反馈	134
3.6.3 结构化文本检索	91	5.4.1 眼动追踪和相关性评价	134
3.6.4 多媒体检索	92	5.4.2 用户行为	135
3.6.5 企业和垂直搜索	92	5.4.3 点击作为用户偏好的指标	136
3.7 趋势和研究问题	92	5.5 通过局部分析的隐式反馈	138
3.8 文献讨论	93	5.5.1 通过局部聚类的隐式反馈	138
第4章 检索评价	96	5.5.2 通过局部上下文分析的 隐式反馈	140
4.1 介绍	96	5.6 通过全局分析的隐式反馈	141
4.2 Cranfield 范式	97	5.6.1 基于相似度同义词典的 查询扩展	141
4.2.1 历史简述	97	5.6.2 基于统计同义词典的 查询扩展	143
4.2.2 参考集	98	5.7 趋势和研究问题	145
4.3 检索指标	98	5.8 文献讨论	145
4.3.1 精度和召回率	98	第6章 文档: 语言及属性	147
4.3.2 单值总结: $P@n$, MAP, MRR, F	102	6.1 介绍	147
4.3.3 面向用户的指标	105	6.2 元数据	148
4.3.4 折扣累积增益	106	6.3 文档格式	149
4.3.5 二元偏好	109	6.3.1 文本	149
4.3.6 排序相关性测度	111	6.3.2 多媒体	149
4.4 参考文档集	115	6.3.3 图形和虚拟现实	150
4.4.1 TREC 参考集	115	6.4 标记语言	151
4.4.2 其他参考集	121	6.4.1 SGML	151
4.4.3 其他小规模测试文档集	121	6.4.2 HTML	153
4.5 基于用户的评价	122	6.4.3 XML	155
4.5.1 实验室中的人工实验	122		
4.5.2 并排面板	122		
4.5.3 A/B 测试	123		

6.4.4	RDF	157	7.3	趋势和研究问题	203
6.4.5	HyTime	158	7.4	文献讨论	204
6.5	文本属性	159	第8章	文本分类	205
6.5.1	信息论	159	8.1	介绍	205
6.5.2	自然语言建模	159	8.2	文本分类的特性描述	206
6.5.3	文本相似度	162	8.2.1	机器学习	206
6.6	文档预处理	163	8.2.2	文本分类问题	206
6.6.1	文本的词汇分析	163	8.2.3	文本分类算法	207
6.6.2	去除禁用词	164	8.3	无监督算法	208
6.6.3	词干提取	165	8.3.1	聚类	208
6.6.4	关键词选择	166	8.3.2	朴素文本分类	212
6.6.5	同义词典	166	8.4	监督算法	212
6.7	组织文档	168	8.4.1	决策树	214
6.7.1	分类体系法	168	8.4.2	k 近邻分类器	218
6.7.2	分众分类法	169	8.4.3	Rocchio 分类器	219
6.8	文本压缩	170	8.4.4	概率朴素贝叶斯文档分类	221
6.8.1	基本概念	170	8.4.5	支持向量机分类器	224
6.8.2	统计方法	171	8.4.6	集成分类器	231
6.8.3	统计方法: 建模	171	8.4.7	关于监督算法的结束语	234
6.8.4	统计方法: 编码	173	8.5	特征选择或降维	234
6.8.5	字典方法	179	8.5.1	项-类别出现列联表	235
6.8.6	压缩预处理	180	8.5.2	索引项文档频率	236
6.8.7	文本压缩技术的比较	181	8.5.3	TF-IDF 权重	236
6.8.8	结构化文本压缩	182	8.5.4	互信息	236
6.9	趋势和研究问题	183	8.5.5	信息增益	237
6.10	文献讨论	185	8.5.6	卡方检验	237
第7章	查询: 语言及属性	187	8.5.7	特征选择的作用	238
7.1	查询语言	187	8.6	评价指标	238
7.1.1	基于关键词的查询	188	8.6.1	列联表	238
7.1.2	非关键词查询	190	8.6.2	准确率和错误率	239
7.1.3	结构化查询	192	8.6.3	精度和召回率	239
7.1.4	查询协议	194	8.6.4	F 测度和 F_1	240
7.2	查询属性	195	8.6.5	交叉检验	241
7.2.1	Web 查询的特征	195	8.6.6	标准文档集	241
7.2.2	用户搜索行为	197	8.7	类别组织——构建分类体系	242
7.2.3	查询意图	197	8.8	趋势和研究问题	244
7.2.4	查询主题	199	8.9	文献讨论	244
7.2.5	查询会话与任务	200	第9章	索引和搜索	247
7.2.6	查询难度	200	9.1	介绍	247

9.2 倒排索引	249	10.4.3 在 SIMD 架构上的并行 信息检索	306
9.2.1 基本概念	249	10.5 基于集群的信息检索	310
9.2.2 完全倒排索引	250	10.6 分布式信息检索	310
9.2.3 搜索	252	10.6.1 介绍	310
9.2.4 排序	256	10.6.2 索引	313
9.2.5 构建	257	10.6.3 查询处理	315
9.2.6 压缩的倒排索引	260	10.6.4 Web 问题	320
9.2.7 结构化查询	261	10.7 联合搜索	320
9.3 签名文件	262	10.8 在对等网络中的检索	322
9.4 后缀树和后缀数组	264	10.9 趋势和研究问题	325
9.4.1 结构: trie 树和后缀树 ..	265	10.10 文献讨论	326
9.4.2 简单字符串搜索	266	第 11 章 Web 检索	327
9.4.3 复杂模式的搜索	267	11.1 介绍	327
9.4.4 构建	268	11.2 一个有挑战性的问题	328
9.4.5 压缩的后缀数组	270	11.3 Web	329
9.5 序列搜索	273	11.3.1 特性	329
9.5.1 简单字符串: Horspool	274	11.3.2 Web 图的结构	331
9.5.2 复杂模式: 自动机和位 并行	276	11.3.3 对 Web 建模	332
9.5.3 更快的位并行算法	279	11.3.4 链接分析	334
9.5.4 正则表达式	281	11.4 搜索引擎架构	335
9.5.5 多重模式	282	11.4.1 基本架构	335
9.5.6 近似搜索	283	11.4.2 基于集群的架构	336
9.5.7 搜索压缩文本	285	11.4.3 缓存	337
9.6 多维索引	287	11.4.4 多级索引	339
9.7 趋势和研究问题	288	11.4.5 分布式架构	340
9.8 文献讨论	289	11.5 搜索引擎排序	342
第 10 章 并行与分布式信息检索	293	11.5.1 排序信号	342
10.1 介绍	293	11.5.2 基于链接的排序	343
10.2 分布式信息检索系统的分类 ..	294	11.5.3 简单的排序函数	345
10.3 数据划分	296	11.5.4 排序学习	345
10.3.1 文档集划分	297	11.5.5 学习排序函数	346
10.3.2 文档集选择	298	11.5.6 质量评价	347
10.3.3 倒排索引划分	299	11.5.7 Web 垃圾	348
10.3.4 划分其他索引	302	11.6 管理 Web 数据	348
10.4 并行信息检索	303	11.6.1 为文档分配标识符	348
10.4.1 介绍	303	11.6.2 元数据	349
10.4.2 在 MIMD 架构上的并行 信息检索	305	11.6.3 压缩 Web 图	349
		11.6.4 处理重复数据	349