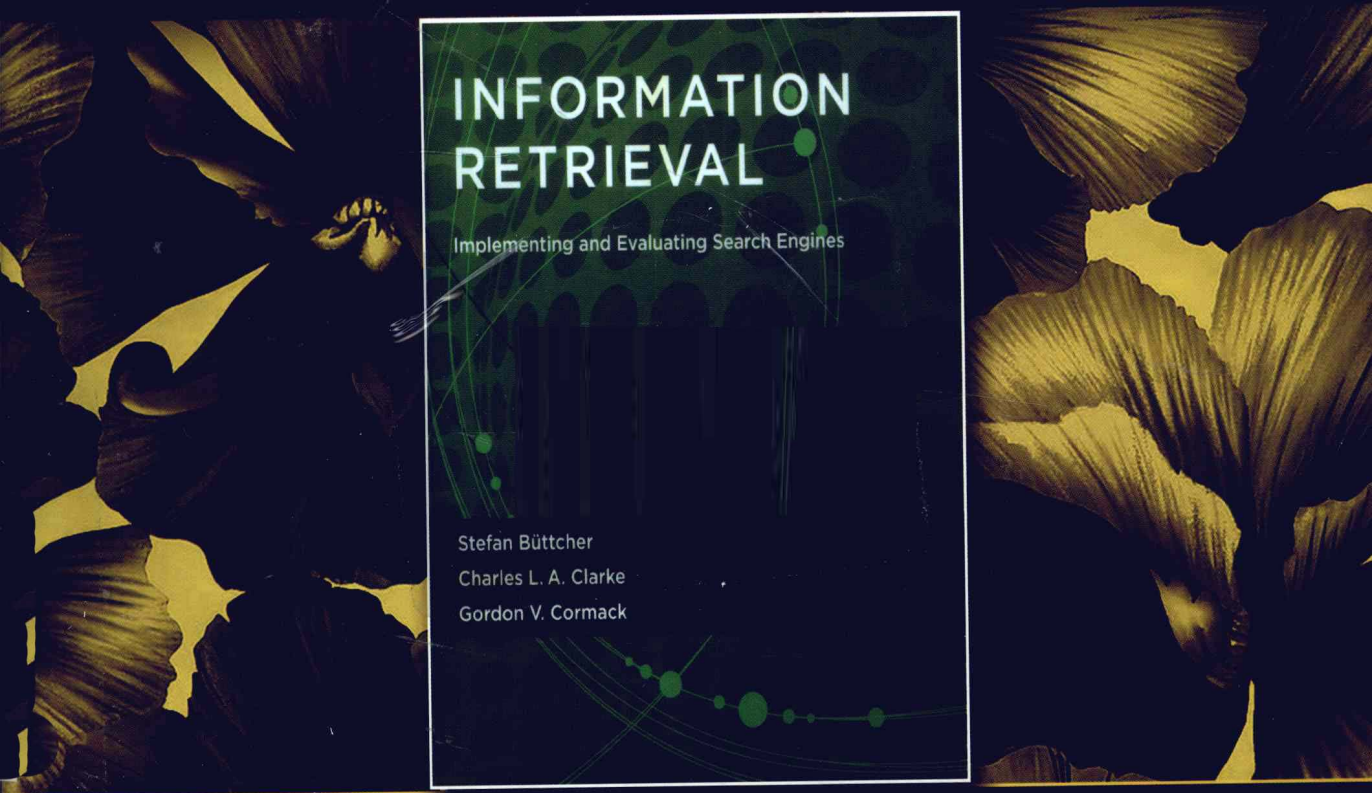
HZ BOOKS
华章教育

计 算 机 科 学 丛 书

信息检索
实现和评价搜索引擎

（美）Stefan Büttcher（加）Charles L. A. Clarke（加）Gordon V. Cormack 著
陈健 黄晋 等译

Information Retrieval
Implementing and Evaluating Search Engines



机械工业出版社
China Machine Press

计 算 机 科 学 丛 书

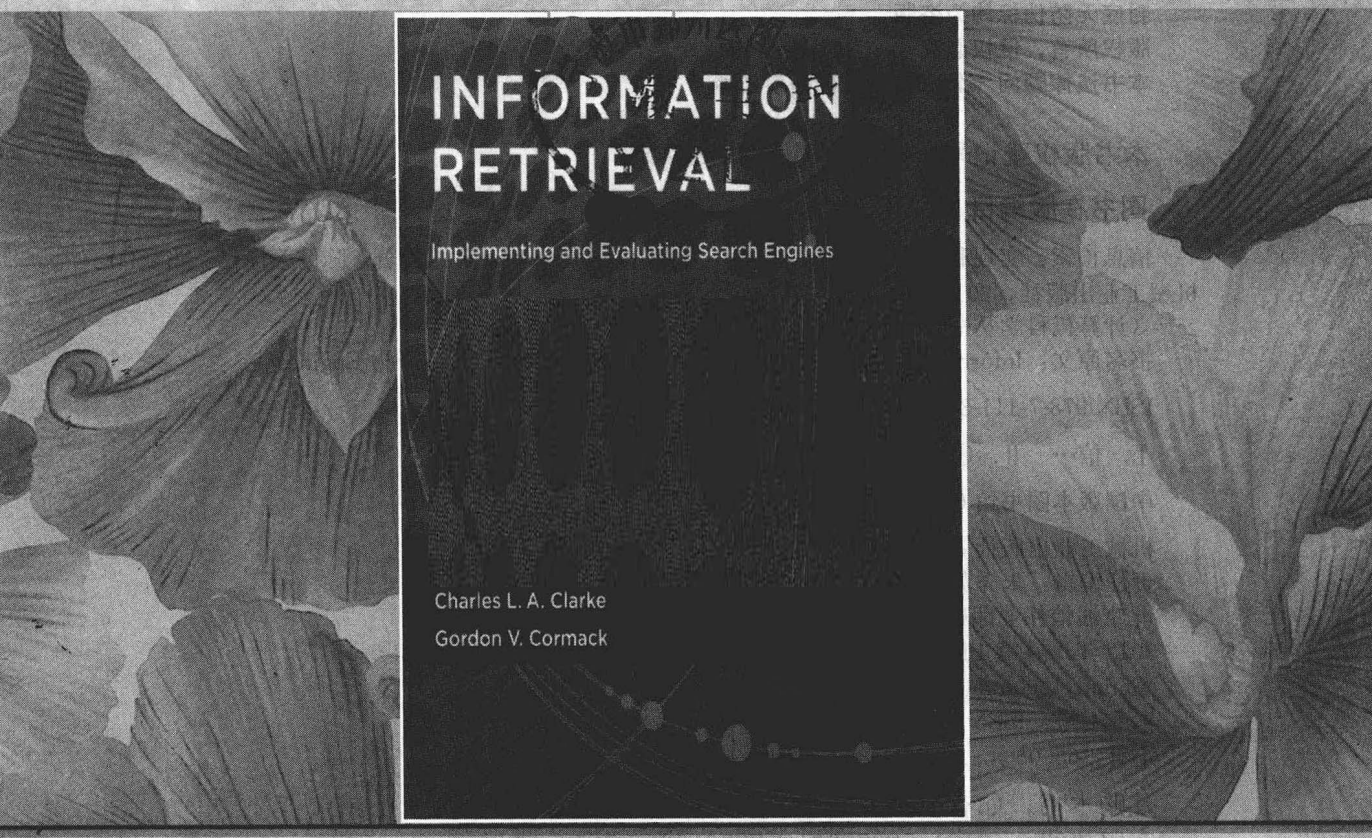
信息检索

实现和评价搜索引擎

(美) Stefan Büttcher (加) Charles L. A. Clarke (加) Gordon V. Cormack 著
陈健 黄晋 等译

Information Retrieval

Implementing and Evaluating Search Engines



机械工业出版社
China Machine Press

本书从多个视角对信息检索技术进行了深入讲解，内容涵盖了信息检索系统的架构、基础技术、词条和词项、静态和动态倒排索引、查询处理、索引压缩技术、概率模型、语言模型、分类和过滤、融合和元学习、评价方法以及并行信息检索、Web检索和XML检索等具体应用。本书以模块化的方式进行组织，理论性强，体系完整，同时强调实践。作者以认真严谨的态度实现了书中绝大部分的主要方法，并详尽地描述了各种方法的适用环境以及取得的效果。

本书可作为高等院校信息管理与信息系统、计算机科学与技术、情报学、图书馆学以及电子商务等专业的高年级本科生和研究生的教材和参考书，对于从事信息检索与网络分析等实际工作的从业人员也具有较高的参考价值。

Stefan Büttcher, Charles L. A. Clarke, Gordon V. Cormack: Information Retrieval: Implementing and Evaluating Search Engines (ISBN 978-0-262-02651-2).

本书版权登记号：图字：01-2010-8044

图书在版编目（CIP）数据

信息检索：实现和评价搜索引擎/（美）布切尔（Büttcher, S.）等著；陈健等译. —北京：机械工业出版社，2011.12

（计算机科学丛书）

书名原文：Information Retrieval: Implementing and Evaluating Search Engines

ISBN 978-7-111-35990-6

Ⅰ. 信…　Ⅱ. ①布…　②陈…　Ⅲ. 情报检索　Ⅳ. G252.7

中国版本图书馆 CIP 数据核字（2011）第 195664 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：朱秀英

北京诚信伟业印刷有限公司印刷

2012 年 1 月第 1 版第 1 次印刷

185mm×260mm · 26.75 印张

标准书号：ISBN 978-7-111-35990-6

定价：65.00 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：hzjsj@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，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相助，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

华章科技图书出版中心

译 者 序

Information Retrieval: Implementing and Evaluating Search Engines

由于手机、个人电脑、互联网等信息工具的快速发展和进化，个人可获取和管理的信息量呈爆发式增长，如何快速准确地找到所需的信息成为信息处理中的一个难题。信息检索技术是解决该问题的主要方法，其最初来源于图书内容的索引和检索，近些年来由于互联网的发展，以此为基础的搜索引擎技术使其受到了广泛的关注和研究。国内无论是高等院校相关专业方向的研究生，还是对搜索技术感兴趣的研究者和开发人员，都迫切需要一本全面专业的信息检索书籍。

国内引进了多本信息检索领域的书籍，本书是其中较新较有特色的一本。它以模块化的方式进行组织，从多个视角对信息检索技术进行了深入的解析，并补充了相关学科的基本知识，例如通用的符号数据压缩技术、统计分析、机器学习、数据库、Web 结构、XML 等等，使读者免去了查阅大量资料和其他书籍的麻烦。这本书理论性强，体系完整，同时也很强调实践。作者以认真严谨的态度对书中绝大部分的主要方法给出了实现细节和分析，并通过实验对比了这些方法，详尽地描述了各种方法的适用环境以及取得的效果，为信息检索在具体环境下的应用提供了很好的参考。在每一章最后的延伸阅读和参考文献部分，读者还可以了解到该章相关知识点的研究历史、发展和目前最新状况，也可据此对相关内容进行更深入的了解和研究。课后练习也经过了精心的设计，各章习题彼此关联、循序渐进，能够帮助读者更好地理解各章的知识点。

感谢原著作者无私地分享了他们在信息检索领域内的独特见解和研究成果。在过去几个月中，胡清兰、吴灿荣、李仕钊、黄锦捷、李蕾、黄蕉平、黄琎都参与了部分翻译、审校工作。感谢徐亚波老师及其学生给出的宝贵意见。当然，本书的翻译工作得以顺利完成，还要感谢机械工业出版社的王春华编辑和其他所有工作人员在各方面的支持和帮助。最后，对于给予我们无私帮助的那些人致以诚挚的谢意。

由于译者水平有限，书中疏漏在所难免，敬请读者批评指正。

陈健、黄晋

2011 年 6 月 29 日

学术巨匠齐聚一堂编撰了一部信息检索的优秀教材。Stefan Büttcher、Charles Clarke和Gordon Cormack以合计超过五十年的研究经验，组成了横跨三代的信息检索研究泰斗组合。Büttcher是Clarke的博士生，而Clarke是Cormack的博士生。他们三人都以对信息检索的深入洞察和建立实用搜索系统的热情而闻名，这种组合在一个充满世界级的研究专家的领域中是很少见的。

本书涵盖了搜索引擎的各个重要组成部分，从爬虫到索引到查询过程。大部分章节用于介绍索引、检索方法和评价的核心主题。重点放在实现和实验上，以让读者了解到信息检索系统的底层细节，包括索引压缩和索引更新策略，同时让读者理解在实际中哪一种方法效果更好。关于评价的两章提供了评价搜索引擎的方法论和统计学基础，使得读者能够知道：例如改变搜索引擎的排名公式是否对检索结果的质量有一个正面的影响。关于分类的一章介绍了对高级搜索操作非常有用的机器学习技术，例如如何将查询限制在某种特定语言书写的文档中，或者如何过滤搜索结果中的不良信息。关于并行信息检索和Web搜索的章节描述了从一个基本的信息检索系统变为一个涵盖数十亿文档并同时为成千上万的用户服务的大规模检索服务系统时所必须做出的改变。

通过引用数以百计的研究文献，作者对当今信息检索研究状况给出了指导性的概述，这个概述的高度远远超过了那些一般的综述。通过使用一个运行样例集和一个通用框架，他们具体描述了在每个环节中的重要方法——为什么这些方法行得通，它们是如何实现的，以及它们是如何工作的。为了写这本书，作者几乎实现和测试了每一个重要的方法，进行了数百次实验，并增加了对实验结果的阐述。每一章最后的练习题鼓励读者自己动手去建立系统并进行探索。

这本书是所有信息检索研究者和从业人员的必读教材！

Amit Singhal，Google Fellow

前　言

Information Retrieval: Implementing and Evaluating Search Engines

信息检索奠定了现代搜索引擎的基石。在这本教材中，我们针对计算机科学、计算机工程和软件工程的研究生以及专业人员介绍了信息检索。选择的主题引起了大部分读者的兴趣，涵盖了算法、数据结构、索引、检索和评价的核心主题，为读者今后的学习提供广博的基础。同时考虑 Web 搜索引擎、并行系统和 XML 检索在已有和新的应用场景的特性。

我们的目的是在理论与实践之间取得平衡，稍微偏向于实践，强调实现和实验。只要有可能，本书中的方法都通过实验进行了对比和验证。每一章都包含了练习和学生项目。本书其中一位作者开发的一个多用户开源信息检索系统 Wumpus，提供了模型实现，可作为学生练习的基础。可以通过 www.wumpus-search.org 获取 Wumpus。

本书组织

本书以模块化结构组织，可分为 5 个部分。第一部分提供了介绍性的材料。第二至第四部分，每部分专注于一个重要主题领域：索引、检索和评价。阅读完第一部分后，第二至第四部分都可以分别单独阅读。第五部分主要基于前面部分的内容来介绍具体的应用领域。

第一部分涵盖了信息检索的基础知识。第 1 章讨论基本概念，包括信息检索系统的架构、术语、文本特征、文档格式、词项分布、语言模型和测试集。第 2 章介绍 3 个重要主题（索引、检索和评价）的基础。这 3 个主题稍后在各自所属的部分（第二至第四部分）有详细介绍。这一章也为读者可以独立阅读每个主题或多或少地提供了基础。第一部分的最后一章，即第 3 章，继续介绍了在第 1 章中引入、在第 2 章中结束的部分主题。它涉及的问题与具体的自然（即人类）语言相关，特别是**分词**（tokenization）——为了进行索引和检索而将一个文档转化成一个词项序列的过程。一个信息检索系统必须能够处理由多种自然语言混合的文档，而这一章就是从这方面讨论几种主要语言的重要特性。

第二部分主要讨论倒排索引的创建、访问和维护。第 4 章讨论建立和访问**静态**（static）索引的算法，这种索引适用于不常变动的文档集，即当文档发生变动时，有足够的时间来重新从头建立索引。第 5 章讨论索引访问和查询过程，这一章介绍一种轻量级的方法来处理文档结构，并使用这种方法来支持布尔约束。第 6 章介绍索引压缩。第 7 章提出用于维护**动态**（dynamic）文档集的算法，也就是文档的更新相对于查询次数是频繁的，同时要求更新必须迅速。

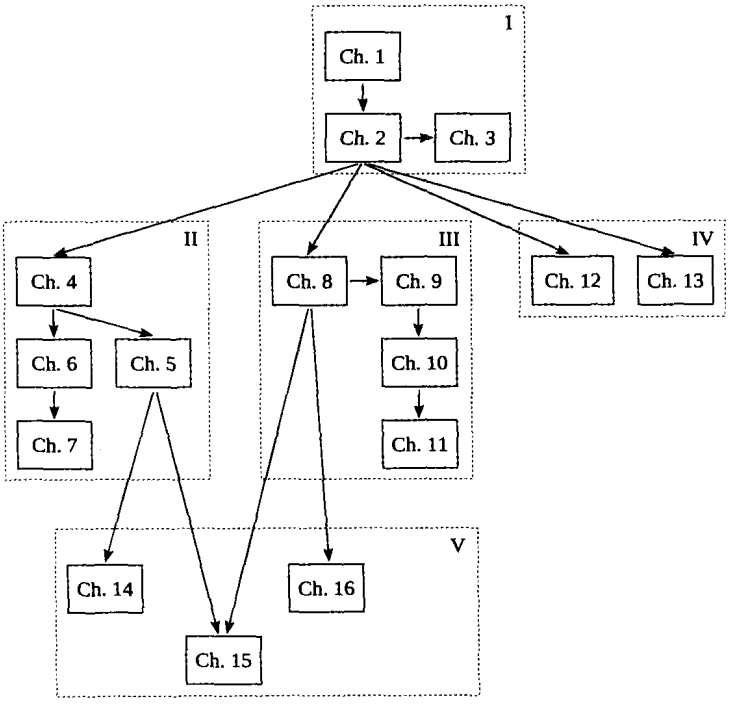
第三部分介绍了检索方法和算法。第 8 章和第 9 章介绍并比较两种基于文档内容的重要排名检索方法：概率模型和语言模型。通过使用文档结构、反馈和查询扩展，可考虑利用一些显式的相关信息来提高这些方法的有效性。我们讨论了每种方法的细节。第 10 章介绍用于文档分类和过滤的技术，包括用于分类的基本的机器学习算法。第 11 章介绍将证据和参数调整进行整合的技术，以及元学习算法及其在排名中的应用。

信息检索评价是第四部分的主题，用独立的章节分别介绍了有效性和效率。第 12 章给出了基本的有效性度量指标，探讨了用于评价有效性的统计基础，并讨论了一些在最近 10 年里提出的度量指标，它们已经超出了传统信息检索评价方法的范围。第 13 章介绍了从响应时间和吞吐量来评价信息检索系统性能的方法。

第五部分是全书的最后一部分，内容涉及一些具体的应用领域，借用并扩展了来自前四个部分的一些基本内容。第 14 章介绍了并行搜索引擎的架构和操作。第 15 章讨论了关于 Web 搜索引擎的一些主题，包括链接分析、抓取和重复检查。第 16 章介绍了 XML 文档集上的信息检索。

书中的每一章都包含了一个小节为深入阅读提供了参考文献，还提供了一组练习题。练习题一般偏向于考查和扩展相应章节介绍的概念。有些练习只需用铅笔和纸花上几分钟就能做好；有些则是需要大量编程的项目。这些参考文献和练习题同时也为我们提供了机会来学习一些在该章的正文部分没有涵盖的重要概念和主题。

下面的示意图展示了本书的各章和各部分之间的关系。箭头表示各章之间的依赖关系。本书的组织使得读者可以关注主题的不同方面。从数据库系统实现的观点来教授的课程可以包括第 1～2、4～7 和 13～14 章。专注于理论的传统信息检索课程可以包括第 1～3、8～12 和 16 章。关于 Web 检索基础的课程可以包括第 1～2、4～5、8 和 13～15 章。每一种涵盖的章节数约占全书的 1/2～2/3，可以在一个 3～4 个月的研究生课程中完成。



本书的组织。各章之间的箭头表示它们之间的依赖关系

背景

我们假设读者拥有计算机科学、计算机工程、软件工程或相关学科的本科相当的基本背景知识，包括：(1) 基本数据结构的概念，例如链表数据结构、B-树和哈希函数；(2) 算法和时间复杂度分析；(3) 操作系统、磁盘设备、内存管理和文件系统。另外，我们假设一些读者熟悉初等概率论和统计学，包括如随机变量、分布和概率群分布函数等概念。

致谢

我们的很多同事花费了大量的时间帮助我们审阅了与其专业领域相关的章节的草稿。我们在这里特别感谢 Eugene Agichtein，Alina Alt，Lauren Griffith，Don Metzler，Tor Myklebust，Fabrizio Silvestri，Mark Smucker，Torsten Suel，Andrew Trotman，Olga Vechtomova，William Webber 和 Justin Zobel 为我们提出了很多宝贵的意见。同时感谢匿名审稿人为我们提供了积极的意见和反馈。

有几个班的研究生起草了早期的一些材料。我们感谢他们的耐心和忍耐。4 个学生——Mohamad Hasan Ahmadi，John Akinyemi，Chandra Prakash Jethani 和 Andrew Kane——非常严谨地审阅了草稿，帮助我们找出和解决了很多问题。另外 3 个学生——Azin Ashkan，Maheedhar Kolla 和 Ian Mackinnon——志愿帮助我们在 2007 年秋季学期进行了一次课内评价，对第一部分中的很多练习有很大的贡献。Jack Wang 校对了第 3 章中关于 CJK 语言的材料。Kelly Itakura 提供了日文输入。

Web 站点

本书的作者维护了一个关于本书材料的 Web 站点，包括勘误以及引用文章的链接，参见 ir.uwaterloo.ca/book。

为了便于参考，以下列表总结了本书中常见的符号。其他必要的符号也在表中列出了。

C	文本集
d	文档
$E[X]$	随机变量 X 的期望值
$f_{t,d}$	词项 t 出现在文档 d 中的次数
l_{avg}	文档集中所有文档的平均长度
l_C	文档集C的大小，以词条数来衡量
l_d	文档 d 的长度，以词条数来衡量
l_t	t 的位置信息列表的长度（即出现次数）
$\mathcal{M}$	概率分布；通常是一个语言模型或一个压缩模型
N	文档集中的文档数
N_t	包含词项 t 的文档数
n_r	相关文档数
$n_{t,r}$	包含词项 t 的相关文档数
$\Pr[x]$	事件 x 的概率
$\Pr[x\|y]$	给定 y，事件 x 的条件概率
q	查询
q_t	词项 t 在查询 q 中出现的次数
t	词项
V	文本集的词汇表
$\vec{x}$	向量
$[\vec{x}]$	向量 $\vec{x}$ 的长度
$[\chi]$	集合 χ 的势（集合 χ 的大小——译者注）

目　录

Information Retrieval: Implementing and Evaluating Search Engines

第一部分 基础知识

第1章

Information Retrieval: Implementing and Evaluating Search Engines

绪 论

1.1 什么是信息检索

信息检索（Information Retrieval，IR）被认为是对大规模电子文本和其他人类语言数据进行表示、搜索和处理的技术。信息检索系统和服务现在已经非常普遍了，成千上万的人每天都使用它们来方便地进行商务、教育和娱乐。Google、Bing 等 Web 搜索引擎，是目前为止最普遍和大量使用信息检索服务的形式，提供获取最新技术信息、搜索人和组织、总结新闻和事件以及简化比较购物的途径。电子图书馆系统帮助医学界和学术界的研究人员了解他们研究领域内最新的期刊文章和会议报告。消费者使用本地搜索服务来找到提供所需产品和服务的零售商。在大型公司中，企业搜索系统作为电子邮件、备忘录、技术报告和其他业务文档的存储库，通过保存这些文档和提供相应的手段获得文档蕴涵的知识来提供企业记忆。桌面搜索系统则允许用户搜索他们的个人电子邮件、文档和文件。

1.1.1 Web 搜索

对 Web 搜索引擎的一般用户而言，通常希望只要在一个文本框里输入一个简短的查询——几个简单的词，然后点击一下搜索按钮，马上就可以得到问题的精确答案。在这简单直观的界面后面是一组计算机集群，包括成千上万台协同工作的机器，用来产生最有可能满足查询中所包含信息的网页排名列表。这些机器要识别包含查询词的网页集合，计算每个网页的得分，消除重复和多余的页面，生成余下页面的摘要，最后将摘要和链接返回给用户以便浏览。

为了达到期望的亚秒级响应时间目标，Web 搜索引擎结合多层缓存和复制机制、利用最常见的查询，并开发并行处理机制，使它们能够随着快速增长的网页和用户数量同步扩展。为了得到精确的查询结果，它们存储 Web 的一个**“快照”**（snapshot）。这个快照必须由运行在成千上万台机器构成的集群上的 Web **爬虫**（crawler）不断采集和更新，定期下载每个页面的最新副本，周期也许是每周一次。包含快速变化信息的高品质网页，例如新闻服务，则可能每天或每小时都会被更新。

来看一个简单的例子，假设你身边有一台可以连接到 Internet 的计算机，请你用一分钟的时间打开一个浏览器并尝试在一个主流商业 Web 搜索引擎上输入“information retrieval”进行查询，搜索引擎通常在一秒钟之内就会响应。花一些时间来看看前 10 个结果，每一个结果都列出一个网页的 URL，通常还提供了一个标题和一个从网页正文提取的简短的文字片段。总的来说，结果来自多个不同的网站，包括与主要教科书、期刊、会议和其他研究者

相关的网站。类似这样常见的**信息性**（informational）查询，Wikipedia[⊖]的文章可能出现在结果集中。前 10 个结果是否包含了不合适的内容？它们的顺序是否还可以再改进呢？让我们看一下紧接的后 10 个结果，看看其中的一个是否可以替代前 10 个中的一个。

现在，我们考虑数以百万计的包含了“information”和“retrieval”这两个词的网页集合。这个集合包含了很多与信息检索主题相关的网页，但这些网页不如排在最前面的 10 个结果那么一般化，例如学生主页和个别研究论文。另外，这个集合还包括很多恰好包含了这两个词但却与这个主题没有任何直接联系的网页。搜索引擎的排名算法基于一系列的特征，从这上百万个可能的网页中选择出排在最前面的网页，这些特征包括网页的内容和结构（例如标题）、网页之间的关系（例如它们之间的链接关系）以及整个 Web 的结构和内容。对于某些查询，用户的某些特征，例如她所在的地理位置或以往的搜索行为，也可能会起到一定的作用。权衡这些特征，以便以用户所期望的查询相关性来对网页进行排名，是**相关性排名**（relevance ranking）的一个例子。在不同的背景和要求下有效实现和评价相关性排名算法是信息检索的核心问题，也是本书的中心议题。

1.1.2 其他搜索应用

桌面和文件系统搜索是信息检索广泛应用的另一个实例。桌面搜索引擎提供对存储在本地硬盘，甚至内部网络上的其他硬盘内的文件进行搜索和浏览的手段。与 Web 搜索引擎相比，桌面搜索引擎系统需要对文件格式和创建时间有更多的了解。例如，用户可能希望只搜索他们的电子邮件，或只知道文件创建或下载的大致时间。因为文件可能变化很快，所以这些系统必须与操作系统的文件系统层直接交互，并且要设计得能够处理高负荷的更新。

介于桌面和普通的 Web 搜索之间，企业级信息检索系统为企业和其他组织提供文档管理和搜索服务。这些系统的实现细节千差万别。有一些基本上就是将 Web 搜索引擎运用到企业内部网，仅抓取组织内部可见的网页并提供类似标准 Web 搜索引擎那样的搜索界面。另一些则提供更一般化的文档和内容管理服务，可以方便地进行显式更新、版本控制和访问控制。在许多行业，这些系统能满足有关电子邮件和保持其他商务联系的监管要求。

数字图书馆和其他专业信息检索系统支持对高品质资料集的访问，这些资料集常常是专有的，可包括报纸文章、医学期刊、地图、图书等，由于版权的问题不能放在一个普通的网站里。考虑到这些资料的编辑质量和有限范围，通常可利用结构化的特征——作者、标题、日期和其他出版数据——来缩小查询需求和提高检索效率。此外，电子图书馆可能包含由**光学字符识别**（Optical Character Recognition，OCR）系统从印刷资料中提取的电子文本，由 OCR 输出带来的字符识别错误会给检索过程带来更复杂的情况。

1.1.3 其他信息检索应用

搜索是信息检索领域的中心任务，但信息检索还涵盖了存储、处理和检索人类语言数据等各种相互关联的问题：

- **文档路由、过滤和选择性传播**（routing，filtering，and selective dissemination）是典型信息检索过程的反过程。一个典型的搜索应用是在一组给定的文档集合上评价一个输入查询，而一个路由、过滤或传播系统则将最新创建或发现的文档与一组预先由用户提供的固定查询进行比较，最符合给定查询的文档被确定为用户最有可能感兴

⊖ en.wikipedia.org/wiki/Information_retrieval