TURING 图灵计算机科学丛书

CAMBRIDGE

Introduction to Information Retrieval

信息检索导论

（修订版）

[美] 克里斯托夫·曼宁 [美] 普拉巴卡尔·拉格万 [德] 欣里希·舒策◎著
王斌 李鹏◎译

信息检索领域**知名科学家**扛鼎之作，**斯坦福大学**教材
重点展示**搜索引擎核心技术**以及**机器学习**和**数值计算**方法



中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵计算机科学丛书

信息检索导论

（修订版）

【美】克里斯托夫·曼宁 【美】普拉巴卡尔·拉格万 【德】欣里希·舒策◎著
王斌 李鹏◎译

Introduction to Information Retrieval



人民邮电出版社
北 京

图书在版编目（CIP）数据

信息检索导论 : 修订版 / (美) 克里斯托夫·曼宁 (Christopher Manning) , (美) 普拉巴卡尔·拉格万 (Prabhakar Raghavan) , (德) 欣里希·舒策 (Hinrich Schütze) 著 ; 王斌, 李鹏译. -- 北京 : 人民邮电出版社, 2019.7
（图灵计算机科学丛书）
ISBN 978-7-115-51408-0

Ⅰ. ①信… Ⅱ. ①克… ②普… ③欣… ④王… ⑤李… Ⅲ. ①信息检索－理论－教材 Ⅳ. ①G254.90

中国版本图书馆CIP数据核字(2019)第104282号

内 容 提 要

本书是一本讲授信息检索的经典教材。全书共21章，前8章详述了信息检索的基础知识，包括倒排索引、布尔检索及词项权重计算和评分算法等，后13章介绍了一些高级话题，如基于语言建模的信息检索模型、基于机器学习的检索结果排序方法和Web搜索技术等。另外，本书还着重讨论了文本分类、聚类等信息检索中不可或缺的组成部分。全书语言流畅，由浅入深，一气呵成。

本书适合作为高等院校相关专业高年级本科生和研究生的课程教材，也可供信息检索领域的研究人员和专业人士参考。

◆ 著　[美] 克里斯托夫·曼宁
[美] 普拉巴卡尔·拉格万
[德] 欣里希·舒策
译　王　斌　李　鹏
责任编辑　朱　巍
责任印制　周昇亮

◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
涿州市京南印刷厂印刷

◆ 开本：787×1092　1/16
印张：24.25
字数：620千字　2019年7月第1版
印数：1－3 000册　2019年7月河北第1次印刷
著作权合同登记号　图字：01-2018-7599号

定价：99.00元

读者服务热线：(010)51095183转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

站在巨人的肩上

Standing on Shoulders of Giants



iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants



iTuring.cn

译者简介



王斌 博士，小米公司 AI 实验室 NLP 方向首席科学家，前中国科学院信息工程研究所研究员、博导、中国科学院大学教授。主持国家 973、863、国家自然科学基金、国际合作基金、部委及企业合作等课题 20 余项，在包括 SIGIR、CIKM、ACL、EMNLP、AAAI、IJCAI、TKDE 等在内的会议和刊物上发表学术论文 150 余篇。担任 SIGIR、ACL、CIKM、WWW、IJCAI 等会议的程序委员会委员，同时是中国计算机学会高级会员、中国中文信息学会理事、中文信息学会信息检索专业委员会委员及《中文信息学报》编委。自 2006 年起在中国科学院研究生院讲授“现代信息检索”研究生课程，迄今培养博士、硕士研究生 80 余名。



李鹏 博士，中国科学院信息工程研究所高级工程师，硕士生导师。主持国家重点研发计划、国家自然科学基金青年基金、部委合作项目 10 余项，在包括 SIGIR、CIKM、ECIR 等会议上发表论文 20 余篇，获得省部级科技奖 1 项，并以负责人身份研制了多个实际应用系统。现为中国中文信息学会青年工作委员会委员。

修订版译者序

本书第一版的中文版已经面世近 8 年了，期间经历了 11 次重印。该版本也被多所国内高校作为信息检索课程的教材使用，我在中国科学院大学开设的“信息检索导论”研究生课程也从 2010 年开始采用本书作为教材。在使用的过程中，包括老师、学生、企业从业人员等在内的很多读者都提出了十分宝贵的意见和建议。同仁们和读者们的厚爱，让我感觉欣慰的同时也不断鞭策着我，因此不断对该书进行修正也成了我日常的小小爱好之一。一个典型的场景是：我把电子版带在随身的笔记本电脑中，出差间歇时不时拿出来读读做点修改。本书原书正式出版后，其修订版的电子版也于次年在网上与大家见面，并时有勘误和更新。今年，人民邮电出版社委托我进行修订版的翻译工作，我义不容辞二话没说就撸起袖子干了起来。

和第一版相比，修订版并没有进行大幅度的修改，而是主要修正了第一版中的上百处不当之处。这些不当之处遍布全书，因此这次翻译过程中大量的工作都是对两个版本进行全方位的比对，工作量很大。所以我邀请了李鹏加入。李鹏在信息检索领域有近 10 年的研究积累，也是本书第一版的最早读者之一。在此次翻译过程中，我们一起对两个版本的全文进行了仔细比对，并共同修正了大量译文。此外，我们也对第一版中的一些翻译不当甚至错误之处进行了细致的处理，并更新了原文中的若干错误。

重译过程中，我们得到了一些老师的鼓励、意见和建议，他们是：清华大学的马少平教授，哈尔滨工业大学的刘挺教授、秦兵教授，大连理工大学的林鸿飞教授，中国科学院自动化所的赵军研究员，中国科学院软件研究所的孙乐研究员，复旦大学的黄萱菁教授，中国科学技术大学的俞能海教授，中国科学院大学的何苯副教授。在此深表感谢。

受人之邀，我于 2014 年 9 月加入中国科学院信息工程研究所，重建了信息检索团队。感谢郭莉研究员对我的认可，感谢我的团队对我这些年来的一直支持。我所在的计算所及信工所信息检索团队成员也是本书第一版的忠实读者和修订版的最早读者，他们也对本次重译做出了不少贡献。他们是：孟二利、鲁骁、王书鑫、李文娜、徐安滢、张驭宇、上官建峰、沙灏副研究员、李锐副研究员、王泉副研究员、毛震东副研究员、齐保元博士、邱泳钦博士、周美林、梁棋、余姚霖、郝炜、姜波、殷荣超、李强、付玺、毋建军、江小天、郭舒、王鹏鸣、李铂鑫、黄琪、刘春晓、骆丹、陈建均、马路、郁博文、朱时超、朱茜、林希珣、曹江峡、从鑫、叶邦宇、胡成承、宋旭亮、罗远飞、杨文静、李思旭、郑楷坚、叶永君、万仪方、施振辉、旷志鑫、刘骐、孙悦、游绩榕、丁舶洋、马呈芳、王栋、胡雪丹、卞娅靖、何纯玉、徐雅珺、王梓涵、易传润、陈宇鹏、王飞、管文宇、薛梦鸽、刘陆琛、袁玥、唐恒柱等。感谢的同时，也祝各位前程似锦。

由于我们水平有限，最后的译稿中可能仍然存在不少问题。希望读者能一如既往提出修改的建议和意见。本书的勘误敬请关注图灵社区：www.ituring.com.cn/book/2601。谢谢大家。

王斌

2018 年末写于北京西山脚下

译 者 序

第一次见到这本书的电子版是在2007年的年底，当时北京大学的闫宏飞博士向我推荐了这本书。拿到电子版之后，我便迫不及待地在一周时间内通读了这本书。读完之后便萌发了翻译这本书的冲动，随后我就联系作者、联系剑桥大学出版社并通过朋友寻找获得授权的国内出版社。辗转数月之后，我被告知该书已经交由其他学者翻译，很快便可出版。听到这个消息，虽然我有些遗憾，但也算是心里的一块石头得以落地。所以，当2008年8月人民邮电出版社突然联系并询问我是否有意翻译这本著作时，我心里的惊讶可想而知。当然，惊讶之余我毫不犹豫地接受了这份邀请，并从此开始了长达数月的翻译历程。

之所以愿意翻译这本书不仅仅是由于该书的作者都是学术界甚至业界鼎鼎大名的人物，更主要的是因为本书在内容和组织上都有独到之处。之前也有很多信息检索方面的教材，但是其中很多内容已经过时。信息检索是一门不断发展并和其他领域、技术不断融合的学科。这本书补充了一些近年来受到广泛关注的新内容。比如：基于语言建模的信息检索模型、基于机器学习的检索结果排序方法、检索结果的Snippet生成、聚类标签生成、XML检索、搜索广告、网页作弊，等等。除此之外，本书每章末尾的“参考文献及补充读物”一节也给出了相关技术的最新进展。本书在内容上与传统教材的另一个显著不同之处是加大了文本分类/聚类技术的介绍篇幅，实际上这些技术已经成为当代信息检索不可分割的一部分。另一方面，本书在深度上超过了大部分传统教材。在介绍信息检索技术的同时，本书深入介绍了其背后所依赖的原理。因此，本书不仅可以用作信息检索领域的入门教材，还能满足对该领域进行深入研究的需要。另外，本书给出了很多现实中的运行算法和实施细节，这些内容对于信息检索技术的实际应用有很好的参考价值。最后值得一提的是，本书在结构上也进行了巧妙构思。首先通过一个例子引出基本技术，然后通过基本技术的不断增强来介绍信息检索的其他技术。全书浑然一体，读起来也有一气呵成的感觉。

这么一本优秀的著作在给译者的翻译带来无穷动力的同时，无疑也给翻译带来了无形的压力。为了尽量保证每章译稿的质量并保持译文的前后一致性，整本书的初译工作全部由译者本人独立完成，在翻译过程中我阅读了大量相关的教材和论文，并前后进行了六次自我校对。在校对过程中，有很多学术界同仁提出了很多宝贵的意见和建议。他们包括：中科院研究生院的朱廷劭教授、中科院自动化所的赵军研究员、中科院软件所的孙乐研究员、复旦大学的黄萱菁教授、江西师范大学的王明文教授、江西财经大学的刘德喜博士、北京大学的闫宏飞博士和何靖博士、清华大学的张敏博士、北京语言大学的徐燕博士等。译者所在的中科院计算所信息检索课题组及选修研究生院“现代信息检索”课程的部分学生也提出了大量修改建议，他们是：郎皓、李亚楠、顾智宇、李鹏、李锐、马宏远、张爱华、蒋在帆、沈沉、史亮、卫冰洁、崔雅超、赵琴琴、李恒训、袁平广、邱泳钦、李丹、鲁凯、徐飞、张帅、张启龙、廖凤、钟进文、朱亮、赵娟等。对于他们无私的帮助，我表示由衷的感谢。感谢我所在的前瞻研究实验室主任

李锦涛老师对我的翻译工作给予的支持和肯定。另一个需要感谢的是我的妻子，在前前后后近八个月当中，除上班时间完成自己的科研工作外，我几乎所有的业余时间都用在翻译和校对上，而她却默默地承担起两岁儿子的所有抚育责任。

翻译的过程中，我还有幸与原书的第二作者 Prabhakar Raghavan 教授进行了当面交流，他对我的翻译工作给予了极大鼓励。在与原书作者的邮件交流中，我也澄清了一些理解上的误区，并修正了原书中的多处错误。

虽然得到了众人的帮助，自己也算认真努力，但由于本人专业水平、理解能力和写作功底都十分有限，加上时间上仍显仓促，最后的译稿中一定存在不少理解上的偏差，译文也会有许多生硬之处。希望读者能不吝提出修改的意见和建议，以便对现有译稿不断改进，直至为国内信息检索领域的读者真正造福为止。原书的初稿电子版、相关课件、勘误表、论坛等信息也可以从网站 http://informationretrieval.org 下载。

王斌

2010 年 3 月

前　言

研究表明，直到 20 世纪 90 年代，大多数人还是首选通过其他人而不是使用信息检索系统来获取信息。当然，那时候大多数人也往往通过旅行社来安排自己的行程。然而，在过去的十年中，信息检索效果的不断优化已经使 Web 搜索引擎的质量达到了一个新的水平，大多数用户在大部分情况下都对搜索的结果感到满意。Web 搜索引擎已经成为用户发现和获取信息的常规和首选渠道。以统计数据为证，2004 年美国 Pew 研究中心的一项因特网调查（Fallows 2004）结果表明，有 92%的因特网用户认为因特网是人们获取日常信息的良好渠道。令很多人惊讶的是，信息检索也从一个以学术研究为主的领域，摇身一变而成为人们日常信息获取工具背后所依赖的基础学科。本书主要介绍该学科的核心理论基础，既考虑了研究生科研的需求，也兼顾了高年级本科生学习的需求。

但是，信息检索并非始于 Web。在应对信息存取的各种挑战的过程中，信息检索逐渐发展成为一门给各种形式的内容搜索提供原理性方法的学科。信息检索起初主要面向科学文献和馆藏记录，但是很快就扩展到其他形式的内容，特别是新闻记者、律师、医生等特定领域专业人士所需的信息内容。信息检索中的很多学术研究都围绕上述内容展开，而其实践方面则主要是为公司或政府部门提供非结构化信息的获取服务，这些领域的研究和实践构成了本书的主要内容。

然而，近年来信息检索革新的主要推动力却来自万维网，因为网络上聚集了数以千万计的网络用户发布的内容。如果这些内容不能及时被发现、标注和分析，并为有需求的人们提供全面的相关性信息，那么它们的存在将毫无意义。到 20 世纪 90 年代末，很多人逐渐意识到，由于 Web 的规模呈指数级增长，继续给整个 Web 建立索引很快会变得毫无可能。但是，卓越的科学创新、一流的工程水平、日益低廉的计算机硬件价格及 Web 搜索商业化基础的壮大等一系列因素，促成了当今主流搜索引擎的产生与成长。这些搜索引擎一天之内能够完成对数十亿网页的数亿次搜索请求，并且每次搜索都能够在亚秒级时间内返回高质量的结果。

本书组织结构及课程设计

本书是我们在斯坦福大学和斯图加特大学所讲授的一系列课程的教学成果总结。这些课程持续的时间从四分之一学期、半学期到一学期不等，主要面向低年级计算机专业的研究生，也曾用于高年级计算机专业的本科生，以及法律、医学信息学、统计、语言学及其他工程学科背景的学生的教学。因此，本书主要的写作原则是提供一个学期的信息检索研究生课程，并尽量覆盖信息检索的学科重点。另一个原则是尽量让每章的内容能在约 75 ~ 90 分钟内完成讲授。

本书前 8 章介绍信息检索的基础知识，特别是搜索引擎的核心理论。这 8 章对于任何信息检索课程来说都是核心内容。第 1 章主要介绍倒排索引，并说明如何通过这种索引实现简单的布

尔查询。第 2 章介绍索引之前的文档预处理过程，并讨论在不同的功能和速度要求下对倒排索引进行改进的方法。第 3 章主要介绍词典搜索的数据结构，并给出查询存在拼写错误或者与被搜索文档中的词汇不能精确匹配时的处理方法。第 4 章主要介绍对文本集合构建倒排索引的几个算法，并着重介绍具有高扩展性的分布式算法，这类算法适用于大规模文档集的索引构建。第 5 章主要介绍倒排索引的压缩技术，包括词典的压缩和倒排记录表的压缩技术，这些技术对于实现大型搜索引擎的亚秒级查询响应十分关键。第 1 ~ 5 章中介绍的索引和查询仅针对**布尔检索**（Boolean retrieval），即一篇文档和查询要么匹配，要么不匹配。那么，如何度量查询和文档的**匹配程度**，或者说如何根据文档和查询的匹配情况对结果打分呢？对这个问题的回答构成了第 6 章和第 7 章词项权重计算和评分算法的主要内容。也就说是，给定查询，我们可以利用这两章介绍的技术，按照文档评分的结果次序输出结果列表。第 8 章主要介绍信息检索系统的评价技术，即根据检索系统返回结果的相关性对不同系统进行评价，从而可以在基准文档集和查询上对不同系统的性能进行比较。

在前 8 章的基础上，本书的第 9 ~ 21 章涵盖了信息检索的一些高级话题。第 9 章介绍了相关反馈和查询扩展技术，其目的在于增加相关文档返回的可能性。第 10 章介绍了采用 XML 和 HTML 等标记语言的结构化文档的检索，这其中我们将结构化文档的检索进行约简，并采用第 6 章所介绍的向量空间模型进行求解。第 11 章和第 12 章介绍基于概率论的信息检索模型。其中，第 11 章介绍传统的概率检索模型，它提供了一个相关度计算框架，在给定一系列查询词项时，能够计算一篇文档与查询相关的概率。这个概率显然可以用于文档的评分和排序。第 12 章给出了另一种方法，即对文档集中的每篇文档建立一个语言模型，然后在每个模型下估计查询生成的概率。这个概率也显然可以用于文档的评分和排序。

第 13 ~ 18 章介绍了信息检索中各种形式的机器学习和数值方法。其中，第 13 ~ 15 章主要关注文档分类的问题，即在给定一系列文档及其归属类别的前提下，将新文档分配到某个或者某几个类别中去。第 13 章首先指出统计分类是一个成功的搜索引擎所必需的关键技术之一，接着介绍了概念虽然简单但分类效率却很高的朴素贝叶斯算法，最后给出了文本分类的评价技术。第 14 章将第 6 章所讲述的向量空间模型应用于文本分类，介绍了几种基于向量空间模型的分类方法，主要包括 Rocchio 和 kNN（k nearest neighbor）两种分类算法。本章最后给出了用于选择分类方法的偏差-方差折中准则，而偏差-方差折中也是学习问题的一个重要特点。第 15 章介绍了支持向量机，这是目前公认的效果最好的文本分类算法之一。另外，本章还将分类问题和一些看上去与文本分类无关的问题（比如如何从给定的训练集合中推导出检索的评分函数）联系起来。

第 16 ~ 18 章主要介绍文档的聚类技术。第 16 章在概述信息检索中的一些重要聚类应用的基础上，主要介绍了两个扁平聚类算法：K-均值算法和 EM 算法。前者是一个效率很高并被广泛应用的算法；后者虽然计算复杂度高一些，但是灵活性更好。第 17 章介绍信息检索对层次聚类（而非扁平聚类）的应用需求，并介绍了一些能产生层次簇结构的聚类算法。这一章还探讨了自动生成聚类标签的难题。第 18 章介绍了一些线性代数的方法，它们是对聚类方法的扩展，并且为线性代数方法在信息检索的应用提供了极具吸引力的前景，其中最具代表性的方法是隐性语义索引。

第 19 ~ 21 章主要介绍 Web 搜索这个具体的应用。第 19 章概述了 Web 搜索所面临的基本挑战，并给出了 Web 信息检索中的一些普遍使用的技术。第 20 章介绍了一个基本网络采集器的

体系结构和必要需求。最后，第 21 章讨论了链接分析在 Web 搜索中的作用，其中用到了线性代数和高级概率论中的方法。

本书并没有囊括信息检索的所有主题，因为有些主题超出了信息检索入门课程的范围。当然，感兴趣的读者可以参见如下参考书籍。

- *Cross-language IR*（跨语言检索）：Grossman 和 Frieder，2004，第 4 章；Oard 和 Dorr，1996。
- *Image and multimedia IR*（图像和多媒体信息检索）：Grossman 和 Frieder，2004，第 4 章；Baeza-Yates 和 Ribeiro-Neto，1999；第 6、11 和 12 章；del Bimbo，1999；Lew，2001；Smeulders 等人，2000。
- *Speech retrieval*（语音检索）：Coden 等人，2002。
- *Music retrieval*（音乐检索）：Downie，2006。
- *User interfaces for IR*（信息检索中的用户界面）：Baeza-Yates 和 Ribeiro-Neto，1999，第 10 章。
- *Parallel and peer-to-peer IR*（并行和 P2P 检索）：Grossman 和 Frieder，2004，第 7 章；Baeza-Yates 和 Ribeiro-Neto，1999，第 9 章；Aberer，2001。
- *Digital libraries*（数字图书馆）：Baeza-Yates 和 Ribeiro-Neto，1999，第 15 章；Lesk，2004。
- *Information science perspective*（基于信息科学视角的信息检索）：Korfhage，1997；Meadow 等人，1999；Ingwersen 和 Järvelin，2005。
- *Logic-based approaches to IR*（基于逻辑的信息检索）：van Rijsbergen，1989。
- *Natural language processing techniques*（自然语言处理技术）：Manning 和 Schütze，1999；Jurafsky 和 Martin，2008；Lewis 和 Jones，1996。

预备知识

所有 21 章都需要数据结构和算法、线性代数以及概率论的基本知识。为方便读者和教师使用本书，各章更具体的要求如下。

第 1 ~ 5 章需要数据结构和算法的基本知识。第 6 章和第 7 章还另外需要线性代数的知识，包括向量和内积的基本概念。第 8 ~ 10 章不需要其他的预备知识。第 11 章需要概率论的基础知识，其中，11.1 节简单介绍了第 11 ~ 13 章所需要的基本概念。第 15 章假定读者熟悉非线性优化的基本概念，当然如果读者对非线性优化没有深入的了解，在阅读上也不会有太多的问题。第 18 章需要线性代数的基本知识，包括矩阵的秩、特征向量等概念，18.1 节对这些概念有一个简单的介绍。第 21 章还需要了解特征值和特征向量的概念。

书中的标记符号

本书正文中用铅笔符号（✎）标记例子，用剪刀符号（✂）标记高级或者难度较大的章节，用问号（?）标记习题，其中分别用[★]、[★★]、[★★★]符号标识“容易”“难度适中”“难度大”等难易程度。

致谢

首先感谢剑桥大学出版社允许本书的电子样稿在网上公布，各种反馈促进了本书的写作过程。感谢 Lauren Cowles，她是一名出色的编辑，在本书样式、组织、覆盖面等方面提出了非常好的建议。如果说这本书最终达到了我们的写作预期的话，那么很大程度上应归功于她。

我们对那些在写作过程中对样稿提出建议和错误纠正意见的人们表示衷心的感谢。他们是 Cheryl Aasheim、Josh Attenberg、Daniel Beck、Luc Bélanger、Tom Breuel、Daniel Burckhardt、Georg Buscher、Fazli Can、Alex Fraser、Dinquan Chen、Ernest Davis、Pedro Domingos、Rodrigo Panchiniak Fernandes、Paolo Ferragina、Norbert Fuhr、Vignesh Ganapathy、Elmer Garduno、Xiubo Geng、David Gondek、Sergio Govoni、Corinna Habets、Ben Handy、Donna Harman、Benjamin Haskell、Thomas Hühn、Deepak Jain、Ralf Jankowitsch、Dinakar Jayarajan、Vinay Kakade、Mei Kobayashi、Wessel Kraaij、Rick Lafleur、Florian Laws、Hang Li、David Losada、David Mann、Ennio Masi、Frank McCown、Paul McNamee、Christoph Müller、Sven Meyer zu Eissen、Alexander Murzaku、Gonzalo Navarro、Scott Olsson、Daniel Paiva、Tao Qin、Megha Raghava、Ghulam Raza、Michal Rosen-Zvi、Klaus Rothenhäusler、Kenyu L. Runner、Alexander Salamanca、Grigory Sapunov、Tobias Scheffer、Nico Schlaefer、Evgeny Shadchnev、Ian Soboroff、Benno Stein、Marcin Sydow、Andrew Turner、Jason Utt、Huey Vo、Travis Wade、Mike Walsh、Changliang Wang、Renjing Wang 及 Thomas Zeume。

很多人主动或应我们的要求对每个章节提出了非常细致的意见。就这点，我们要特别感谢如下这些人：James Allan、Omar Alonso、Ismail Sengor Altingovde、Vo NgocAnh、Roi Blanco、Eric Breck、Eric Brown、Mark Carman、Carlos Castillo、Junghoo Cho、Aron Culotta、Doug Cutting、Meghana Deodhar、Susan Dumais、Johannes Fürnkranz、Andreas Heß、Djoerd Hiemstra、David Hull、Thorsten Joachims、Siddharth Jonathan J. B.、Jaap Kamps、Mounia Lalmas、Amy Langville、Nicholas Lester、Dave Lewis、Stephen Liu、Daniel Lowd、Yosi Mass、Jeff Michels、Alessandro Moschitti、Amir Najmi、Marc Najork、Giorgio Maria Di Nunzio、Paul Ogilvie、Priyank Patel、Jan Pedersen、Kathryn Pedings、Vassilis Plachouras、Daniel Ramage、Stefan Riezler、Michael Schiehlen、Helmut Schmid、Falk Nicolas Scholer、Sabine Schulte im Walde、Fabrizio Sebastiani、Sarabjeet Singh、Alexander Strehl、John Tait、Shivakumar Vaithyanathan、Ellen Voorhees、Gerhard Weikum、Dawid Weiss、Yiming Yang、Yisong Yue、Jian Zhang 及 Justin Zobel。

最后，我们还要感谢书稿的审阅人员，他们为本书提出了大量高质量的建议。感谢他们为本书的内容和结构提出了重要建议，我们对他们表示深深的谢意，他们是：Pavel Berkhin、Stefan Büttcher、Jamie Callan、Byron Dom、Torsten Suel 及 Andrew Trotman。

第 13、14 和 15 章的部分初稿基于 Ray Mooney 慷慨提供的报告幻灯片。尽管后来的内容做了大量的修改，但是我们对 Ray Mooney 为这 3 章所做的贡献表示由衷的感谢，特别在各种分类器算法的时间复杂度分析方面，本书基本沿用了 Ray Mooney 的工作。

上述感谢的名单并不完整，我们仍然在不断整理来自各方面的反馈。当然，和其他作者一样，我们不一定留意到每一条建议，这点还请大家谅解。另外需要指出的是，本书的出版版本文责自负。

作者还要感谢斯坦福大学和斯图加特大学提供的优良的学术环境。在此环境中，大家可以自由交流思想，并通过授课来促进本书的写作和完善。C. D. Manning 感谢他的家人给他时间投入到本书的写作上，并希望明年能利用周末的时间更多地陪伴家人。P. Raghavan 感谢他的家人默默提供的支持，并感谢 Yahoo!公司提供了一个良好的写作环境。H. Schütze 要感谢他的父母、家人和朋友在他写作期间给予他的支持。

联系方式

欢迎大家将更多的反馈意见和修改建议发邮件至informationretrieval@yahoogroups.com。

符号对照表

符　号	边栏页码	含　义
γ	90	γ编码
γ	237	$\gamma(d)$表示分类或者聚类函数：$\gamma(d)$是d所属的类或者簇
Γ	237	第13、14章中的有监督学习方法：$\Gamma(\mathbb{D})$是从训练集$\mathbb{D}$上学到的分类函数γ
λ	370	特征值
$\vec{\mu}(.)$	269	类质心（在Rocchio分类中）或簇质心（在K-均值和质心聚类中）
Φ	105	训练样本
σ	374	奇异值
$\Theta(\cdot)$	10	算法复杂度的紧上界
ω, ω_k	328	聚类结果中的一个簇
Ω	328	聚类结果或簇集合$\{\omega_1,\cdots,\omega_K\}$
$\text{argmax}_x f(x)$	164	使函数f取最大值的x的值
$\text{argmin}_x f(x)$	164	使函数f取最小值的x的值
c, c_j	237	分类中的一个类别
cf_t	82	词项t的文档集频率（该词项在整个文档集中出现的总次数）
$\mathbb{C}$	237	类别集合$\{c_1,\cdots,c_J\}$
C	248	取值为类别集合$\mathbb{C}$中元素的随机变量
$\boldsymbol{C}$	369	词项-文档矩阵
d	4	文档集D中的第d篇文档的索引号
d	65	一篇文档
$\vec{d},\vec{q}$	163	文档向量及查询向量
D	326	所有文档的集合$\{d_1,\cdots,d_N\}$
D_c	269	类别c中的文档集
$\mathbb{D}$	237	第13～15章中的已标记文档集$\{<d_1,c_1>,\cdots,<d_N,c_N>\}$，即训练集
df_t	108	词项t的文档频率（文档集中出现t的文档数目）
H	91	熵
H_M	93	第M个调和数
$I(X;Y)$	252	随机变量X和Y的互信息
idf_t	108	词项t的逆文档频率
J	237	类别数目
k	267	集合中排名前k的元素，如kNN中的前k个邻居、检索文档的前k个结果以及词汇表V中选出的前k个特征
k	50	k个字符组成的序列
K	326	簇的个数
L_d	214	文档d的长度（以词条为单位计数）
L_a	242	测试文档或应用文档的长度（以词条为单位计数）
L_{ave}	64	文档的平均长度（以词条为单位计数）

（续）

符　号	边栏页码	含　义
M	4	词汇表大小（即$\|V\|$）
M_a	242	测试文档或应用文档的词汇量
M_{ave}	71	文档集中每篇文档的平均词汇量
M_d	218	文档d的模型
N	4	检索或训练文档集中的文档数目
N_c	240	类别c中的文档数目
$N(\omega)$	275	事件ω发生的次数
$O(\cdot)$	10	算法复杂度的界
$O(\cdot)$	203	事件的优势率
P	142	正确率
$P(\cdot)$	202	概率
P	425	转移概率矩阵
q	55	查询
R	143	召回率
s_i	53	字符串
s_i	103	域评分布尔值
$\mathrm{sim}(d_1,d_2)$	111	文档d_1和d_2的相似度
T	40	文档集中所有词条的数目
T_{ct}	240	词t在c类文档中的出现次数
t	4	词汇表V中第t个词项的索引号
t	56	词汇表中的一个词项
$\mathrm{tf}_{t,d}$	107	词项t在文档d中的出现频率（即t在d中的出现次数）
U_t	246	表示词项t存在与否的随机变量，当t存在时，值为1，否则为0
V	190	文档中的所有词项$\{t_1,\cdots,t_M\}$组成的词汇表（也称为词典lexicon）
$\vec{v}(d)$	111	文档d经长度归一化后的文档向量
$\vec{V}(d)$	110	文档d未经长度归一化的文档向量
$\mathrm{wf}_{t,d}$	115	词项t在文档d中的权重
w	103	权重，比如域的权重或者词项的权重
$\vec{w}^{\mathrm{T}}\vec{x}=b$	269	超平面方程：$\vec{w}$是超平面的法向量，w_i是$\vec{w}$的第i个分量
$\vec{x}$	204	基于词项表示的文档向量$\vec{x}=(x_1,\cdots,x_M)$，更一般地说，为文档的特征表示
X	246	取值为词汇表V中元素的随机变量（比如，某个文档位置k上的词）
$\mathbb{X}$	237	文本分类中的文档空间
$\|A\|$	56	集合A的势：集合A中的元素个数
$\|S\|$	570	方阵$\boldsymbol{S}$的行列式
$\|s_i\|$	53	s_i的长度（以字符计）
$\|\vec{x}\|$	128	向量$\vec{x}$的大小
$\|\vec{x}-\vec{y}\|$	121	向量$\vec{x}$、$\vec{y}$的欧氏距离，也即向量$(\vec{x}-\vec{y})$的大小

目　录

第 1 章　布尔检索 …… 1
1.1　一个信息检索的例子 …… 2
1.2　构建倒排索引的初体验 …… 5
1.3　布尔查询的处理 …… 8
1.4　对基本布尔操作的扩展及有序检索 …… 11
1.5　参考文献及补充读物 …… 13
第 2 章　词项词典及倒排记录表 …… 14
2.1　文档分析及编码转换 …… 14
2.1.1　字符序列的生成 …… 14
2.1.2　文档单位的选择 …… 16
2.2　词项集合的确定 …… 16
2.2.1　词条化 …… 16
2.2.2　去除停用词 …… 19
2.2.3　词项归一化 …… 20
2.2.4　词干还原和词形归并 …… 23
2.3　基于跳表的倒排记录表快速合并算法 …… 26
2.4　含位置信息的倒排记录表及短语查询 …… 28
2.4.1　二元词索引 …… 28
2.4.2　位置信息索引 …… 29
2.4.3　混合索引机制 …… 31
2.5　参考文献及补充读物 …… 32
第 3 章　词典及容错式检索 …… 34
3.1　词典搜索的数据结构 …… 34
3.2　通配符查询 …… 36
3.2.1　一般的通配符查询 …… 37
3.2.2　支持通配符查询的k-gram索引 …… 38
3.3　拼写校正 …… 39
3.3.1　拼写校正的实现 …… 39
3.3.2　拼写校正的方法 …… 40
3.3.3　编辑距离 …… 40
3.3.4　拼写校正中的 k-gram索引 …… 42
3.3.5　上下文敏感的拼写校正 …… 43
3.4　基于发音的校正技术 …… 44
3.5　参考文献及补充读物 …… 45
第 4 章　索引构建 …… 46
4.1　硬件基础 …… 46
4.2　基于块的排序索引方法 …… 47
4.3　内存式单遍扫描索引构建方法 …… 50
4.4　分布式索引构建方法 …… 51
4.5　动态索引构建方法 …… 54
4.6　其他索引类型 …… 56
4.7　参考文献及补充读物 …… 57
第 5 章　索引压缩 …… 59
5.1　信息检索中词项的统计特性 …… 59
5.1.1　Heaps定律：词项数目的估计 …… 61
5.1.2　Zipf定律：对词项的分布建模 …… 62
5.2　词典压缩 …… 63
5.2.1　将词典看成单一字符串的压缩方法 …… 63
5.2.2　按块存储 …… 64
5.3　倒排记录表的压缩 …… 66
5.3.1　可变字节码 …… 67
5.3.2　γ编码 …… 68
5.4　参考文献及补充读物 …… 74
第 6 章　文档评分、词项权重计算及向量空间模型 …… 76
6.1　参数化索引及域索引 …… 76
6.1.1　域加权评分 …… 78
6.1.2　权重学习 …… 79
6.1.3　最优权重g的计算 …… 80
6.2　词项频率及权重计算 …… 81
6.2.1　逆文档频率 …… 81
6.2.2　tf-idf权重计算 …… 82
6.3　向量空间模型 …… 83
6.3.1　内积 …… 83
6.3.2　查询向量 …… 86