TURING 图灵计算机科学丛书

Springer

信息检索

算法与启发式方法

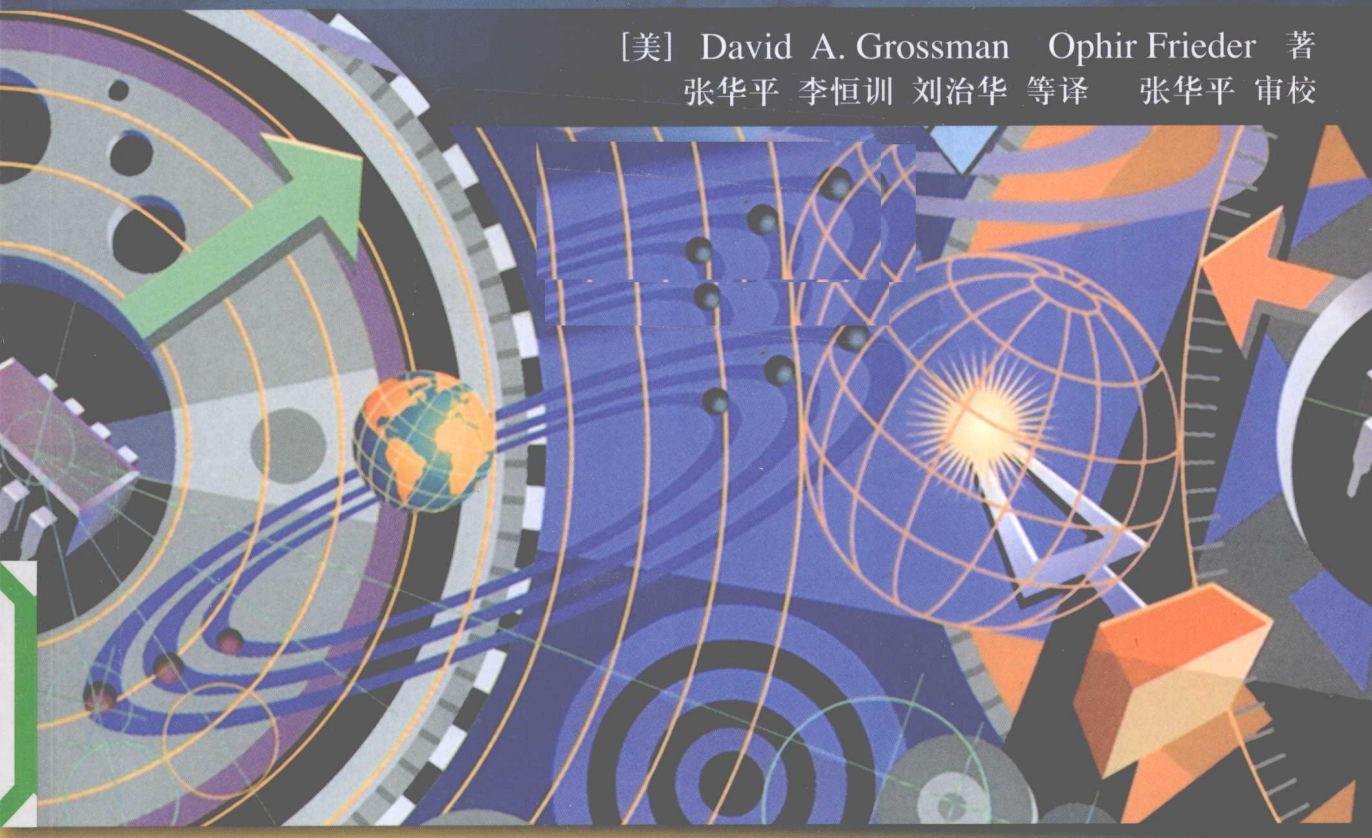
（第2版）

Information Retrieval: Algorithms and Heuristics

Second Edition

[美] David A. Grossman Ophir Frieder 著

张华平 李恒训 刘治华 等译 张华平 审校



人民邮电出版社

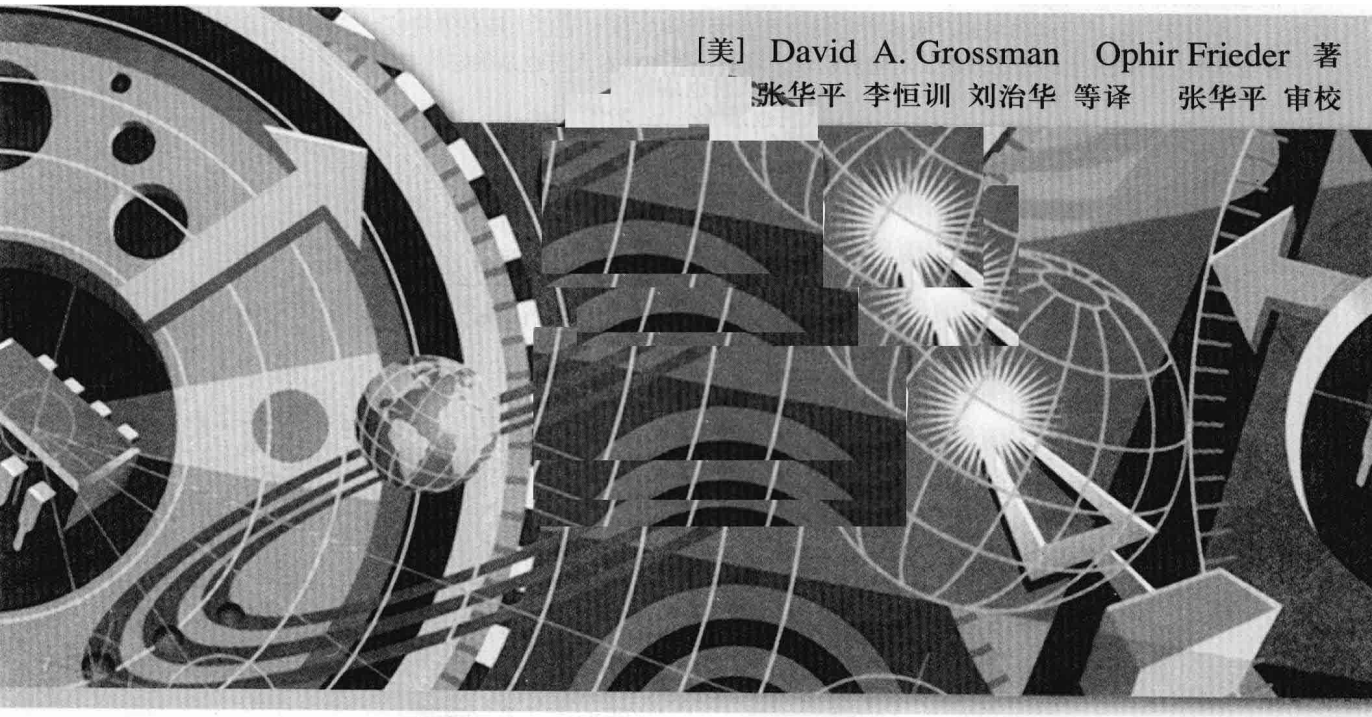
POSTS & TELECOM PRESS

TURING 图灵计算机科学丛书

信息检索
算法与启发式方法
（第2版）

Information Retrieval: Algorithms and Heuristics

Second Edition



[美] David A. Grossman　Ophir Frieder 著
张华平 李恒训 刘治华 等译　张华平 审校

人民邮电出版社
北　京

图书在版编目（CIP）数据

信息检索 : 算法与启发式方法 : 第2版 / (美) 格罗斯曼 (Grossman, D. A.), (美) 弗里德 (Frieder, O.) 著 ; 张华平等译. -- 北京 : 人民邮电出版社, 2010.9
(图灵计算机科学丛书)
书名原文: Information Retrieval:Algorithms and Heuristics, Second Edition
ISBN 978-7-115-23575-6

Ⅰ. ①信… Ⅱ. ①格… ②弗… ③张… Ⅲ. ①情报检索 Ⅳ. ①G252.7

中国版本图书馆CIP数据核字(2010)第142198号

内 容 提 要

本书是“信息检索”课程的优秀教材，书中对信息检索的概念、原理和算法进行了详细介绍，内容主要包括检索模型与算法、检索实用策略、跨语言信息检索、查询处理、融合结构化数据和文本、并行信息检索以及分布式信息检索等，并给出了阐述算法的大量实例。

本书有一定的广度和深度，而且所有的内容都用当前的技术阐述，是高等院校计算机及信息管理等专业本科生和研究生的理想教材，对信息检索领域的科研和技术人员也是很好的参考书。

图灵计算机科学丛书

信息检索：算法与启发式方法（第2版）

◆ 著 [美] David A. Grossman Ophir Frieder
译 张华平 李恒训 刘治华 等
审校 张华平
责任编辑 王军花

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子函件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
三河市海波印务有限公司印刷

◆ 开本：787×1092 1/16
印张：15.25
字数：390千字 2010年9月第1版
印数：1-3 000册 2010年9月河北第1次印刷
著作权合同登记号 图字：01-2009-4808号
ISBN 978-7-115-23575-6

定价：49.00元

读者服务热线：(010)51095186 印装质量热线：(010)67129223
反盗版热线：(010)67171154

版 权 声 明

译 者 序

搜索引擎越来越受到普通大众、技术、产业与资本的热力追捧，CNNIC 发布的《第 25 次中国互联网络发展状况统计报告》显示，搜索引擎在网络应用使用率中排名第三，达到了 73.3%。2009 年中国搜索引擎的市场规模达到 69.5 亿元，在经济危机的背景下，搜索引擎营销获得大品牌广告主的认可。

作为网络搜索与挖掘十余年的研发者，我在研发与教学工作中发现，目前搜索引擎技术方面的资料往往有两种不足：一种是过于学术化，关注于一个很窄的话题，读起来生涩难懂，缺乏一定学术背景的读者很难真正理解；另外一种是原理性的介绍，往往比较浅显，缺乏一定的深度，读者读起来总觉得意犹未尽。而本书是难得的一本佳作。一方面，它全面综合了信息检索领域的各类研究，融合了最新的研究成果；另一方面，它采用通俗易懂的行文方式，在关键点上采用了详尽的案例来解释抽象的过程，读者并不需要具备专业的数学背景就很容易掌握信息检索技术的精髓所在。

本书主要作为本科生或研究生信息检索课程的教材，也特别适合信息检索领域从事应用开发的研发人员使用。在国外，它已经被多所高校作为研究生与本科生的教材，广受好评。

我非常有幸应人民邮电出版社图灵公司之邀翻译这本信息检索领域的杰作。对我们来说，这是一次学习过程，也是多年研究工作的总结提炼过程。翻译从来就不是一件轻松的事，加之正赶上工作调动，在时间和精力上受到一定影响。但我们依然克服了各种困难，每周固定开会研讨，字斟句酌，抱着严谨的学术态度，尽可能忠实表达作者的原意，同时兼顾语言之美。在“苦行僧”式的翻译学习过程中，我们往往也会从作者敏捷的思维和精巧的话语中领略到一种无穷的智慧。

在此，我要感谢原作者 Grossman 与 Frieder 的卓越工作，感谢我的研究生忍受我的各种严格要求，感谢人民邮电出版社图灵公司给予我们这次学习的机会。同时，我要感谢北京理工大学计算机学院院长黄河燕教授在此工作过程中给我的支持，也要感谢中科院计算所网络重点实验室主任程学旗研究员对我在信息检索领域的指导。最后，我特别感谢我的妻子，感谢她在这段艰难的时光的默默支持。

本书翻译分工如下：由我负责翻译本书的第 1 章、第 7 章与第 9 章，我的研究生李恒训同学负责翻译第 2 章，刘治华同学负责翻译第 3 章，李恒训同学、秦鹏同学和周鹏同学负责翻译第 4 章和第 5 章，张京阳同学负责翻译第 6 章，蒋骅同学负责翻译第 8 章。另外，我对全书进行了多次全面的审校。

限于水平有限，错误在所难免，恳请读者批评指正。

张华平博士

E-mail: kevinzhang@bit.edu.cn

2010 年 1 月 25 日于北京理工大学计算机语言信息处理研究所

序

正如格劳斯曼（Grossman）和弗里德（Frieder）在“前言”中所述，信息检索（IR）在过去 5 年里取得了相当大的进步。对普通人来说，这种进步充分体现在网络商用搜索引擎的日趋成熟；对从事信息检索的人来说，5 年来的进步拓展了网络搜索问题的研究范畴，并突破了诸多局限，随着基础体系架构和信息检索模型的发展，整个检索过程不断地引入了新的视角，同时开展了一系列令人振奋的应用，比如跨语言检索、P2P 搜索和音乐检索，这些都极大地拓宽了信息检索研究的疆域。数据库与信息检索这两个不同领域的学者逐步达成共识：必须整合非结构化与结构化数据的处理技术，我们才能够真正解决未来社会的信息问题。本书阐述了许多这方面的重要进展，也是迄今为止唯一一本这样做的教科书。

本书让我印象最深的两个例子是信息检索语言模型和跨语言检索。语言模型强大而简洁，而且在搭建很多实验和应用时可以使用许多现成的工具，因此很多研究人员都在采用这一模型，并且语言模型已经成为主流信息检索学术会议研究的重要课题。格劳斯曼和弗里德在第 2 章中很好地概述了这一课题，同时也给出了不同平滑技术的例子。跨语言检索是指采用某种语言检索其他多种语言的文本内容，欧洲和美国政府存在广泛的实际需求，这也一直推动着跨语言检索的迅猛发展。充分利用平行语料库和可比语料库，人们研制了一些方法，这些系统的性能现在已接近（在某些方面甚至已超越）单一语言的检索系统。本书专门增加了一章跨语言检索的内容，清晰地阐述了跨语言检索的主要方法，并给出了示例来具体说明如何在真实数据上执行算法。本书覆盖了最新的研究成果，采用准确直接的语言，同时频繁使用了大量实例，可作为研究生或本科生信息检索课程的首选教材。

W. 布鲁斯·克劳福特（W. Bruce Croft）

2004 年 8 月

致　谢

本书的第 1 版于 1998 年出版。从那以后，信息检索领域发生了翻天覆地的变化。

我们要感谢那些帮助我们准备第 1 版大部分材料的人。在此依然要深深地感谢 Paul Kantor 和 Don Kraft，感谢他们在本书编写的早期阶段给出的真知灼见。也要特别感谢 Steven Robertson、K.L. Kwok 和 Jamie Callan，本书的一些章节介绍了他们的相关工作，他们对这些章节做了严格的审查；同时，我们还要感谢 Warren Greiff，感谢他极其耐心地反复教我们推理网络的各种细节问题。

本书增加了语言模型和跨语言信息检索相关的内容，为此，我们要特别感谢翟成祥、Doug Oard 和 James Mayfield，感谢他们富有洞察力的点评与建议。

同时，本书还得到了很多人一些非常重要的反馈。通过讲授本科“信息检索”以及研究生“高级信息检索”等课程，我们从学生中得到了诸多反馈，最终提高了本书的易读性，使其更加适合课堂教学。John Juilfs 为我们花了无数的时间，仔细检查了这一版中所有的新增示例，并复查了所有数学方面的内容（当然，我们对剩余的任何错误仍然要负全责，而他帮我们避免了大部分错误）。Abdur Chowdhury 为 5.4 节提供了素材，并进行了仔细的检查。Wei Gen Yee 对第 8 章的修改提供了重要的材料。Steven Beitzel 和 Eric Jensen 为我们更新第 5 章和第 7 章的内容提供了重要的素材。Rebecca Cathey 为第 4 章提供了示例。Michael Saelee 为我们制作并细化了各种图表。最后，管伟反复仔细地检查了所有的示例和符号。

也要感谢 Wendy Grossman、Ben Goldfarb、David Roberts 以及 Jordan Wilberding，他们给出了详尽的点评意见。同样要感谢所有其他读过本书并给出了精彩反馈的人，感谢给我们提供了精神支持的朋友和同事。

如果没有家人和亲朋精神上持续不断的支持，我们也很难完成本书。为此，特别感谢 Mary Catherine McCabe 和 Nazli Goharian，她们为我们牺牲了无数时间，给了我们坚持不懈的鼓励和支持！

我们俩对所有这些人，道以最诚挚的谢意！

前　言

我们在1998年写本书第1版时，Web还是比较新鲜的事物。实际上，信息检索是一个比较老的研究领域，只不过还没有引起广泛关注罢了。如今，Google已成为流行词汇，Google索引了网页40多亿页。1998年，只有几所学校为研究生开设了信息检索课程；而如今，信息检索在本科阶段都已普及。文献[Goharian等人，2004]总结了我们讲授本科生信息检索课程的经验，详细分析了在课堂上讨论的每个专题内容，并介绍了课程教学的效果。

信息检索指的是搜索任何形式的信息，包括结构化数据、文本、视频、图像、声音、乐谱、DNA序列等。事实是，多年来，数据库系统用来搜索结构化数据，信息检索则用来搜索文档。本书的作者原本就从事结构化搜索领域的研究，但是在过去十年的大部分时间里，都在研究文档的检索。对我们来说，客观世界的数据类型本身就是不可知的，因此，我们没有必要特别区分结构化数据与非结构化数据。1998年，我们在本书第1版中有一章内容专门讲述数据整合，书评人则认为，收录该部分内容的唯一原因就是它涉及我们最新的一些研究罢了。而现在，这种评述或者辩解已然没有任何意义了，因为我们已经引入了信息的中间表示结构（mediator），可以对结构化和非结构化数据同时进行操作。而且，XML（eXtensible Markup Language，可扩展标记语言）已经广泛地应用于数据库与信息检索领域。

我们主要关注ad hoc信息检索问题[①]。简单来说，ad hoc信息检索指的是针对用户提交的各种不同查询，搜索出相关的文档集合。像Google这样的系统可能已经解决了这个问题，但是，Google的性能评测并没有公布。一些经典系统的准确率最高也只能达到40%[TREC, 2003]。在对现有算法深入理解的基础上，我们仍有很大的改进空间。

市面上信息检索教材的内容相对散乱，并不适合我们的日常教学。这些教材在许多关键检索模型的细节上往往避而不谈。推理网络是许多系统都要用到的核心模型，但是，几乎没有教材详细介绍推理网络。另外，许多教材都没有详细描述系统的效率，即单一查询的执行速度。或许，对于那些只关注检索效果的人来说，检索效率能够引起的潜在兴趣特别有限；但是对于从业者来说，对效率的关注可以超越其他所有的指标。

另外，针对每种方法，我们都给出了详细可行的实例。当介绍具体方法时，我们很容易在细节方面轻描淡写，不过，实例可以让我们更忠实地阐述方法的本质所在。我们发现贯穿于整本书的一个实例能让学生们从中受益。值得一提的是，本书每个描述核心检索算法的章节要么经过了算法原创者的评审（我们感谢他们的慷慨奉献，更多的感谢详见致谢），要么经过了精通该算法的专家审校。因此，就我们所知，本书所述检索算法的诸多细节目前还很难在其他出版物中找到。

---

① ad hoc信息检索是信息检索中一个专门的任务形式，与routing相对应，前者的数据集相对稳定，而查询多变；后者反之。——译者注

我们的目标是写一本特别专注于 ad hoc 信息检索的书。为达到这个目标，我们基于模型建立了本领域一套完整的分类体系，主要包括文档和查询比较的算法模型，以及一些可以内嵌到所有算法模型中对性能进行优化的实用策略。本书中介绍了所有的基本方法，还有一系列的工具集。我们提供了足够详尽的说明，阅读本书的学生或者其他读者都可以方便地实现其中的方法或者工具。*Managing Gigabytes* [Witten 等人，1999]一书非常出色地阐述了倒排索引压缩的策略。我们引用了其中最新而且最有效的研究成果，但还是推荐读者将 *Managing Gigabytes* 作为本教材优秀的辅助读物。

在第 2 版中有什么新内容呢？许多核心的检索方法仍没有改变。自 1998 年以来，在信息检索领域引入语言模型的论文不计其数。因此，我们专门增加了语言模型的章节。跨语言信息检索（使用一种语言提交查询，而搜索出另一种语言的文档）在本书第 1 版刚问世时尚处于萌芽期，而如今它已经取得了长足的进步，我们在参考了最近 100 多篇相关文献之后，特别增加了一整章内容来介绍跨语言检索的最新研究进展。

自然而然，我们还讨论了许多当前的热点话题，比如 XML、P2P 信息检索、文本查重、文档并行聚类、不同检索策略的融合以及信息中间表示等。

最后，一些细心的本科生和研究生发现了上一版的一些错误，我们一一作了修正。这里，我们要感谢他们的努力。

本书主要作为本科生或研究生信息检索课程的教材。本书已经在我们的研究生课程中实际使用过，我们结合了同学们的反馈制作了一套与教材配套的幻灯片，可以在课堂教学过程中使用。这些资源可以从 www.ir.iit.edu 上获取。

另外，对于要搭建信息检索系统或相关应用程序的读者来说，如何选择恰当的检索方法和工具集用于产品开发，本书将会非常有用。我们曾收到几个读者来信，反映本书第 1 版对他们有帮助，我们将他们的意见和建议都吸收进了本书新版之中。

虽然我们强调本书的重点是算法而不是商用产品，但是据我们所知，本书中包含了大多数商用产品所采用的方法。我们相信读者或许会发现某些商用产品正在使用本书给出的信息检索方法，还能够将本书作为参考来更多地了解这些产品中采用的技术。

最后，我们注意到信息检索领域每天都在发生新的变化。有关本领域更多最新的研究成果，最好的资源有《ACM 信息系统杂志》（*the ACM Transactions on Information Systems*）、《美国信息科学与技术学报》（*the Journal of the American Society for Information Science and Technology*）、《信息处理与管理》（*Infornmtion Processing and Management*）和《信息检索》（*Information Retrieval*）等杂志。其他相关论文可以查询各种信息检索会议，比如 ACM SGIR（www.sigir.org）、NIST TREC（trec.nist.gov）、ACM CIKM （www.cikm.org）。

目　录

第1章 引 言

社会文明伊始，人类就致力于书面形式的沟通。从洞穴石刻到卷轴书写，从印刷出版物到电子图书馆，沟通一直都是人类有史以来最为关注的事情。如今，随着数字图书馆和电子信息通信的涌现，人们的的确确需要不断提高有效管理海量信息的技术。各行各业都需要特别关注信息的溯源、处理、存储和检索等领域理论与实践的研究进展。在本书中，我们主要介绍电子文档查找和检索方面最新的研究成果。

我们介绍的重点就是如何检索出满足用户查询的信息。也就是说，我们讨论的是 ad hoc 信息检索算法与方法，简言之就是信息检索。图 1-1 显示了 ad hoc 信息检索的基本处理流程。在用户查询前，我们需要对静态或相对静态的文档集建立索引。用户提交一个查询请求后，系统会自动给出与查询相关的文档集合，同时按照各篇文档与查询计算出的相关度进行排序，最后将排序后的结果集推送给用户。目前，有许多用来解决文档排序问题的方法（即效果），这是本书讨论的一个重点。另外，还有一些技术用来实现文档的快速排序（即效率），这些也将在本书中给予讨论。

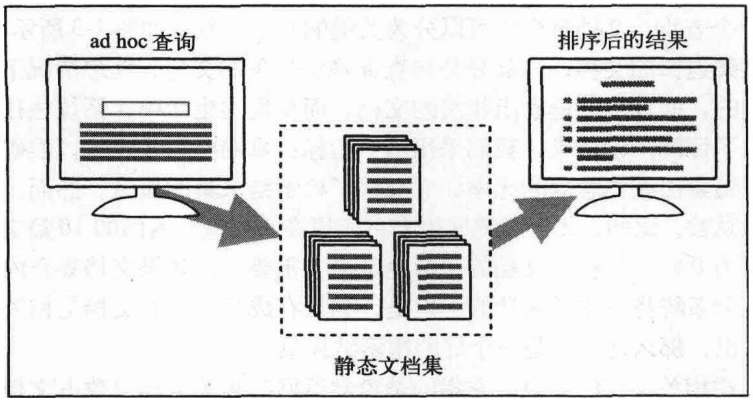


图 1-1 文档检索

信息检索（Information Retrieval，IR）致力于找到相关文档，而不是简单的模式匹配。然而，在信息检索系统的实际评测中，大家往往发现系统会遗漏许多相关文档[Blair 和 Maron，1985]。此外，用户对于信息检索系统精准率的期望也越来越高[Gordon，1997]。

另外一个与 ad hoc 信息检索相关的问题是文档分发（document routing）或文档过滤（document filtering）。两者的区别在于文档分发的查询是稳定不变的，而文档集则不断变化。比如企业往往会根据预定义的查询条件将公司邮件分门别类，分发到不同部门中（也就是关于销售的电子邮件分发给销售部，与市场相关的电子邮件被分发给市场部，等等）。图 1-2 给出了文档分发的示意图。文档分发算法与方法许多文献均有介绍，本书中不予详述。

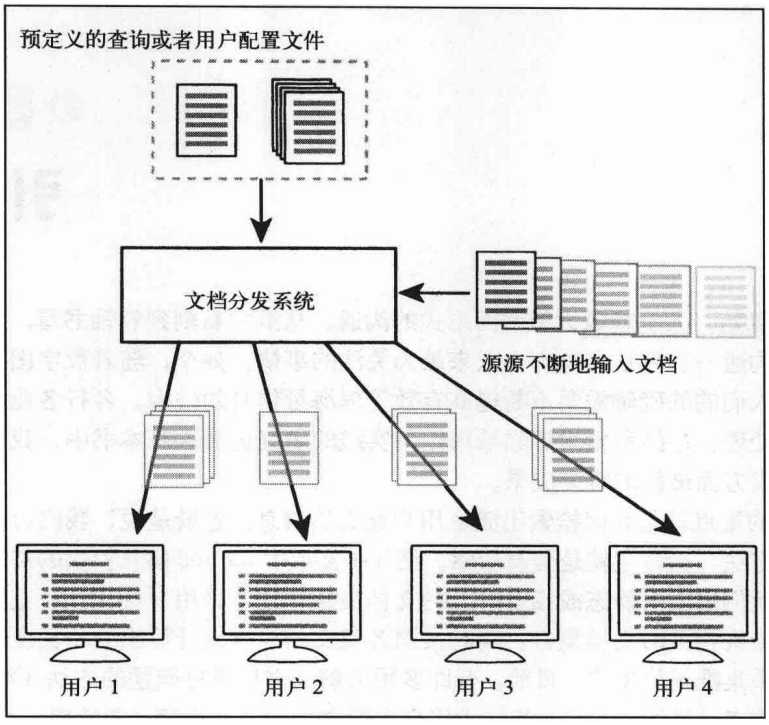


图 1-2　文档分发

针对任何一个查询，文档集合均可以分为关键的几个类别，如图 1-3 所示。在文档集合中，一部分是检索系统返回的文档，一部分是与查询确实相关的文档。理想情况下，这两个文档集应该是完全等同的，即我们只检索出相关的文档。而在现实生活中，系统往往会检索到许多不相关的文档。为了评测检索效果，我们采用两个指标：*准确率*和*召回率*。准确率是检索返回的相关文档数目占检索出文档总数的比率，它反映了检索结果集的质量。然而，准确率并没有考虑到相关文档的总数。比如，某个系统准确率可能很高，如检索返回的 10 篇文档中有 9 篇文档相关（即准确率为 0.9），但相关文档的总数也是至关重要的。如果文档集合内确实仅有 9 个相关文档，那么这个系统将是非常成功的。但是，如果有成千上万的文档是相关的，应该都被检索到而没有被检出，那么这可不是一个好的搜索结果集。

召回率会考虑相关文档的总数，它指的是检索返回的相关文档总数占文档集中相关文档总数的比率。计算所有相关文档的总数不是那么容易的，唯一可信的方法就是人工通读整个文档集并作出判断，这显然是不可行的。人们往往会采用一些近似估计的方法获取该值（参见第 9 章）。[Kantor，1994]综述了各类信息检索性能的评测方法，同时也对信息检索做了一个较好的概述。

2

在不同的召回率指标下，我们往往需要计算其对应的准确率。以某个查询 q 当做例子来考虑，如图 1-3 所示。对于这个查询，我们估计有两篇相关的文档。现在，假设当用户提交查询 q 时，返回 10 篇文档，其中包含这两篇相关文档。本例中，假定文档 2 和文档 5 是相关的，图 1-4 中的斜线表明在检索返回两篇文档后，我们找到了其中的一篇相关文档，从而达到 50%的*召回率*。此时，我们检索到了两篇文档并且其中一篇是相关的，因此*准确率*也是 50%。

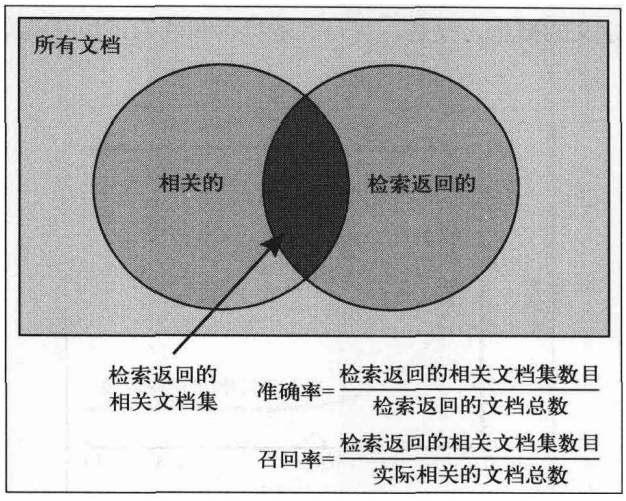


图 1-3 结果集：检索返回的相关文档集、相关文档集以及检索返回的文档集

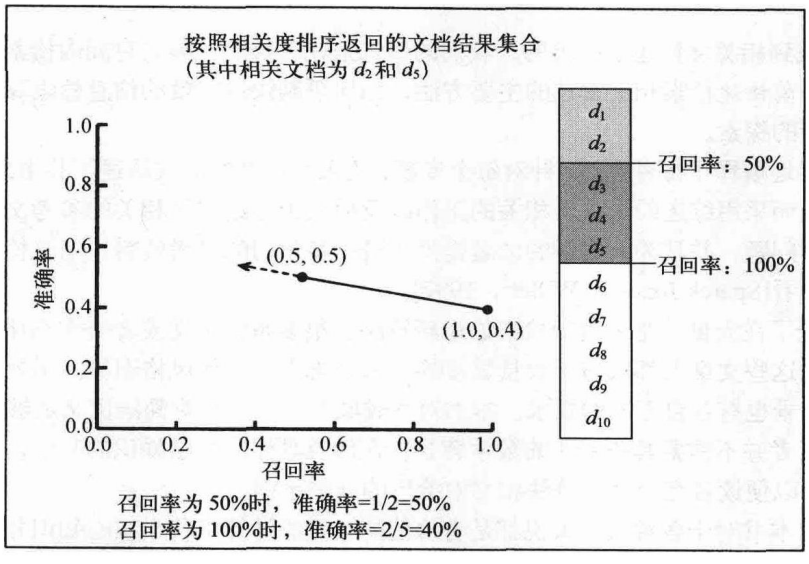


图 1-4 两种召回率以及相应的准确率

为了达到 100%的召回率，我们必须继续检索文档直到所有的相关文档都被返回。在这个例子中，必须返回 5 篇文档才能检索到两个相关文档。此时，准确率为 40%，因为在 5 篇返回的文档中有两篇是相关的。因此，在召回率达到期望值的时候，我们都需要计算相应的准确率。在不同的召回率上计算出准确率，最终形成变化的图形，我们称之为*准确率/召回率曲线*。

图 1-5 给出了一条典型的准确率/召回率曲线。通常情况下，要想取得更高的召回率，我们必须返回更多的文档才能达到期望的召回率。在一个完美的系统中，只检索输出相关的文档。这意味着在任意的召回率条件下，其准确率都为 1.0。最优准确率/召回率曲线如图 1-5 所示。

平均准确率指的是在不同召回率条件下准确率的平均值。在一个标准的文档集上进行测试，当前大部分系统的平均准确率在 0.2 到 0.3 之间。当然，这里有一些比较模糊的因素，因为相关

性还是一个不能清晰定义的概念，但是，有一点是特别明确的：信息检索领域的效果还有很可观的提升空间。

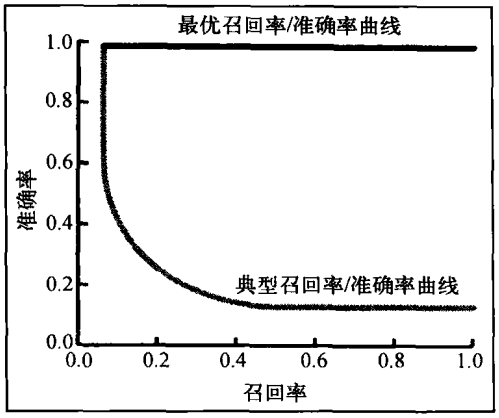


图 1-5　召回率/准确率曲线：典型曲线与最优曲线

仅仅查找到相关文档还是不够的，我们的目标是在可接受的响应时间内检索到相关文档。本书介绍了当前快速检索相关文档的主要方法。如何研制快速高效的信息检索算法？这还需要我们做进一步的探索。

我们详尽地解释了每种算法，针对每个专题，还采用示例的形式阐述了其中最重要的算法。然后，我们转而采用综述的形式对相关的工作以及后续工作给出了相关的参考文献。我们阐述
4 了算法的关键问题，并且为有兴趣的读者提供了进一步学习的参考资料。信息检索研究的主要论文集可以参看[Sparck Jones 和 Willett，1997]。

本书讨论了在大量信息中进行检索的最新算法。很多研究论文或者著作都详尽地论述了这
5 些算法，但是这些文章大都散布于数量繁多的学术杂志上，写作风格迥异。另外，这些文章对读者的知识背景也有各自不同的要求。本书对本领域做了一个相对简洁但又足够细致的综述。

本书的读者并不需要具备专业的数学背景。在涉及具体的数学知识的时候，我们都会快速复习关键点，以便读者充分理解算法和书中给出的详细示例。

我们相信本书对于各种读者来说都是有价值的。熟悉计算机科学核心知识并希望能学习更多信息检索算法的读者应该能从本书中受益匪浅。我们将解释信息检索领域存在的一些基本问题，并具体介绍人们以前是如何处理这些问题的。

正在使用和维护 RDBMS（关系数据库管理系统）的读者也可以看一看这本书。第 6 章具体介绍了一些算法，这类算法将文本检索看做是 RDBMS 的一种应用。因此，我们可以将结构化数据和文本综合起来处理。另外，我们还单独用一节介绍如何利用关系数据库来处理半结构化文档，如采用 XML 标记的文档。

为了引导读者全面理解 ad hoc 信息检索中的关键问题，我们将全书分为几个独立的，但又存在有机联系的部分。第一部分包括第 2 章和第 3 章，全面综述了信息检索领域主要的算法模型与实用策略。所有的模型和策略都紧紧围绕一个关键的问题，即如何提高检索的准确性。第 2 章介绍了 9 种信息检索模型，这些模型要么是专门为信息检索而研制的，要么是为信息检索改造而成的，其目的都是针对特定用户查询，用来提高相关文档排序或评测结果的性能。第 3

章介绍了一些检索相关的工具，这些工具能用来提升第2章中任何一个算法的性能。

在第3章中，我们重点介绍一些信息检索的技术，这些技术均可以应用于大部分甚至所有的检索模型当中。这里介绍的几个技术与语言密切相关，如语法解析和主题知识库，而其他的很多技术则具有语言无关性，如 n 元处理过程。我们在第3章中明确说明了，这里介绍的有些技术实际上就是一个独立的处理系统。在实际应用过程中，这些技术有机地组合在一起可以取得最佳效果。如何精准地判定出这些技术的最优组合策略，采用什么样的执行顺序，实际运行过程中采用什么样的底层模型，才能确保最佳的性能，至今仍然是未知的课题。

介绍完这些提高信息检索准确性的模型和策略之后，我们将注意力转向效率方面。在第4章中，我们介绍了多种文档读取访问方法。既介绍了倒排索引的构建与使用，也介绍了其他类型的表现形式，如签名文件。每一种访问模式都有各自的优点和缺点。选择任何一种方案，我们都必须在存储代价、可维护性与查询检索处理速度之间寻求一种综合平衡。介绍完存储访问 6
模式后，我们将介绍几种数据压缩方法。

第2章和第3章集中阐述了传统信息检索模型的基本原理、常用技术以及处理策略。第4章简要介绍跨语言信息检索。第5章分析了信息检索中的效率问题。在第6章、第7章与第8章中，我们把重点放在信息检索领域内的3个专题上。这3个专题分别是数据整合、并行以及分布式信息检索系统。之所以选择讨论这几个专题，是因为它们如今也是业界热切关注的问题。

一般来说，结构化数据和半结构化数据之间有一个清晰的界限。结构化数据大都通过关系数据库管理系统来存储与读取，而文本等半结构化数据一般都是通过信息检索系统来存储与读取的。每一个处理系统都只支持自己的数据存储格式以及相应的处理方法。然而，如今结构化数据和半结构化数据之间的差别正在逐步消失。实际上，我们不再仅仅关注结构化数据和半结构化数据，而是更关注如何采用相同的存储策略来处理文本以及经常出现的非结构化数据，如图像。

为了解决结构化数据和非结构化数据之间的整合问题，Oracle、IBM、Microsoft 等商业公司都在各自的关系数据库系统中加入了信息检索功能，而文本检索厂商（如 Convera、Verity）也在自己的系统中加入了关系处理模块。然而，在所有这些情况下，额外功能的增加都需要另外新增独立的处理模块。在第6章中，我们将讨论如何增加处理模块，并给出一种可行的替代方法，该方法将信息检索处理能力作为关系数据库的一种应用。使用该方法，无需额外的软件开发，就能获得关系数据库系统本身具备的一些传统优势（即可移植性、并发性、数据的可恢复性等）。所有的关系数据库厂商在数据库系统中都提供了并行处理功能，因此，如果信息检索系统作为关系数据库系统的应用程序，那么信息检索系统也将具备并行处理能力。

认识到并行能力对信息检索系统的重要性之后，我们将进一步介绍该领域的最新进展。在第7章中，我们首先介绍信息检索系统早期的并行方法。这些方法主要利用 SIMD（单指令多数据流）技术，采用多个处理器来高效地并行扫描文本。然而，随着对并行处理技术理解的不断加深，人们研究出了基于倒排索引的方法，倒排索引方法能有效地减少文本扫描过程中不必要的输入/输出开销。我们将讨论各种不同的方法。最后，第7章将总结并行信息检索领域最新 7
的研究成果，其中主要介绍文档聚类并行算法。

在当今的信息处理领域，任何一种方法都必须考虑到最常见的应用场景——万维网。因此，在第8章中，我们将讨论 Web 检索中必须考虑的一个问题——分布式信息检索系统。我们将对一些早期的基本理论进行综述，最终详细地讨论了 P2P 信息检索。