HZ BOOKS
华章教育

计 算 机 科 学 丛 书

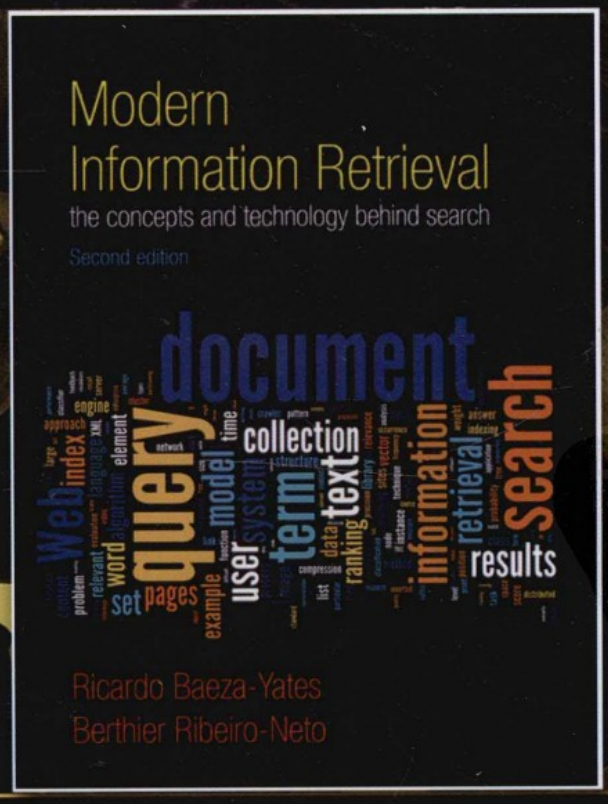
原书第2版

现代信息检索

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



机械工业出版社
China Machine Press

现代信息检索（原书第2版）

Modern Information Retrieval The Concepts and Technology behind Search Second Edition

本书详细介绍了信息检索的所有主要概念和技术，以及有关信息检索方面的所有新变化，使读者既可以对现代信息检索有一个全面的了解，又可以获取现代信息检索所有关键主题的详细知识。本书的主要内容由信息检索领域的代表人物Baeza-Yates和Ribeiro-Neto撰写，对于那些希望深入研究关键领域的读者，书中还提供了由其他主要研究人员撰写的关于特殊主题的发展现状。

与上一版相比，本版在内容和结构上都有大量调整、更新和充实，其中新增内容在60%～70%左右。具体更新情况如下：

- 新增了文本分类、Web爬取、结构化文本检索和企业搜索等章节，以及关于开源搜索的一个附录。
- 全面改写了用户界面、多媒体检索和数字图书馆等内容。
- 拓展了一些章节，介绍了信息检索方面的新的重要进展，如语言模型、新的评价方法、查询的特点、基于集群的信息检索和分布式信息检索等。

作者简介

Ricardo Baeza-Yates 于加拿大滑铁卢大学获得计算机科学博士学位，现为雅虎欧洲和拉丁美洲研究院副总裁，主管雅虎在巴塞罗纳（西班牙）和圣地亚哥（智利）的研究中心，并监管海法研究中心。他曾担任智利计算机科学学会主席、智利大学计算机科学系Web研究中心主任、ICREA教授，并且他还在巴塞罗纳法布拉大学创立了信息与通信技术系Web研究组。现在他仍是智利大学和法布拉大学的兼职教授。他的主要研究方向为算法与数据结构、信息检索、用户界面以及可视化在数据库中的应用等。

Berthier Ribeiro-Neto 于加利福尼亚大学洛杉矶分校获得计算机科学博士学位，现任巴西Minas Gerais联合大学计算机科学系副教授，同时也是ACM、ASIS及IEEE会员。他的主要研究方向是信息检索系统、数字图书馆、Web界面及视频点播。

书号：978-7-111-33174-2
定价：78.00元

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

华章网站 http://www.hzbook.com

网上购书：www.china-pub.com

封面设计：锡 彬

上架指导：计算机/信息检索
ISBN 978-7-111-38599-8

定价：118.00元

计 算 机 科 学 丛 书

原书第2版

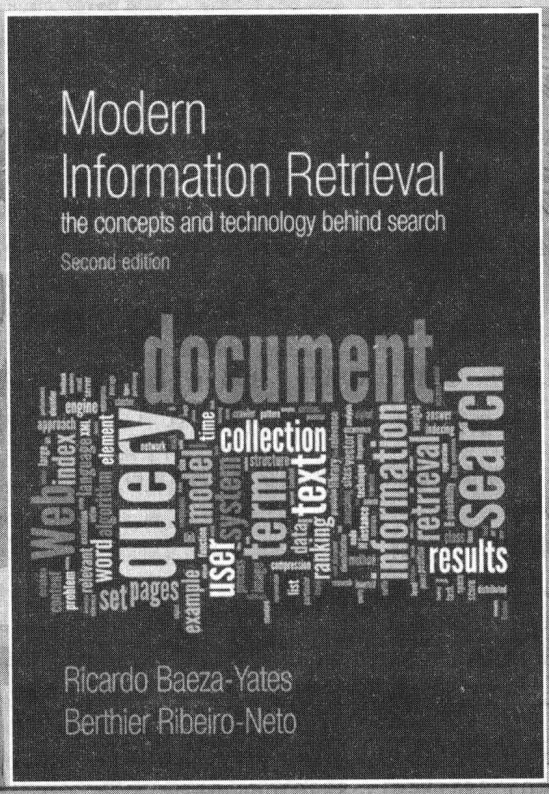
现代信息检索

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



机械工业出版社
China Machine Press

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。

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：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

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

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变化，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、吴龑、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，不仅要对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，并包含两倍以上的参考文献。总之，如果你喜欢本书第1版，我们希望你会更喜欢这个第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和浓厚兴趣表示感谢，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 Diekema、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和浓厚兴趣表示感谢。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.berkerley.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.，

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。

引　言

1.1 信息检索

信息检索（Information Retrieval，IR）是计算机科学的一大领域，主要研究如何为用户访问他们感兴趣的信息提供各种便利的手段，即：

> 信息检索涉及对文档、网页、联机目录、结构化和半结构化记录及多媒体对象等信息项的表示、存储、组织和访问。信息项的表示和组织必须便于用户访问他们感兴趣的信息。

在范围上，信息检索的发展已经远远超出了其早期目标，即对文档集进行索引并从中寻找有用的文档。如今，信息检索的研究包括建模、Web 搜索、文本分类、系统架构、用户界面、数据可视化、过滤和语言处理技术。

在研究方面，信息检索可以从两个相当不同和互补的视角展开研究：以计算机为中心的视角和以人为中心的视角。从以计算机为中心的视角来看，信息检索主要包括建立高效的索引，高性能地处理用户的查询，并开发排序算法以提高检索结果。从以人为中心的视角来看，信息检索主要包括研究用户的行为，理解他们的主要需求，并且相应地确定检索系统的组织和操作。鉴于前者在学术界和市场上的主导地位，本书主要论述以计算机为中心的视角。

1.1.1 信息检索的早期发展

5000 多年来，人类已经知道如何组织信息，为以后的检索和搜索服务。在最通常的形式，它一直是通过编辑、储存、组织和索引泥板、象形文字、纸草卷和书籍实现的。为存放各种物品，人类还使用了特殊用途的建筑物，并称之为图书馆。表示图书馆的英语单词一个是“library”，来自拉丁文的“liber”，表示“书籍”；另一个是“bibliothek”，来自希腊文的“biblion”，表示“纸草卷”。

已知最古老的图书馆在公元前 3000—公元前 2500 年之间成立于厄尔巴。它位于“新月沃地”（Fertile Crescent），即目前的叙利亚北部。在公元前 7 世纪，亚述王亚述巴尼拔在底格里斯河（位于今日的伊拉克北部）建造了尼尼微图书馆，该图书馆在公元前 612 年，也就是被毁灭的那一年，共收藏 30 000 多块泥板。到了公元前 300 年，马其顿将军多利买梭特尔，在尼罗河口以马其顿国王亚历山大大帝（公元前 356—公元前 323 年）命名的亚历山大市，建造了亚历山大图书馆。700 年间，亚历山大图书馆和同城的其他图书馆一道，使得亚历山大成为西方世界的知识之都［1164］。

从那时起，图书馆日渐扩大和繁荣，如今已遍布世界各地。它们构成了人类的集体记忆，并且越来越普遍。仅 2008 年，美国人去图书馆的次数就达到了约 13 亿次，借阅资料超过 20 亿件，并且这个数字每年增加的幅度都在 10%以上［155］。

由于图书馆的信息容量一直在增长，因此有必要建立专门的数据结构——索引，进行快速搜索。不管采用哪种形式，索引都是每一个现代信息检索系统的核心。它们提供快速访问数据的方法以加快查询处理。我们将在第 9 章讨论索引技术。

数百年来，索引的形式都是手动建立的类目集。索引中的每个类目通常由标志相关主题的标签和指向相关文档的指针组成。虽然这些索引通常是由图书馆和信息科学的研究人员设计，但现代计算机的出现使得自动构建大规模索引成为可能，而这也加快了信息检索领域的发展。

信息检索的早期发展可以追溯到20世纪50年代Hans Peter Luhn、Eugene Garfield、Philip Bagley和Calvin Moores等开拓者所进行的研究工作，其中最后一位还发明了信息检索这个术语[1692]。在1955年，Allen Kent和他的同事发表了一篇论文，描述了精度(precision)[⊖]和召回率(recall)[⊖]两项评价指标[903]，1962年Cyril Cleverdon在所进行的Cranfield研究中沿用了它们[394，395]。1963年，Joseph Becker和Robert Hayes出版了关于信息检索的第一部书籍[164]。在20世纪60年代，Gerard Salton和Karen Sparck Jones等人提出了现代信息检索中排序技术的基本概念，从而塑造了这一领域。1968年，Salton出版了他的第一部信息检索书籍。1971年，N. Jardine和C. J. Van Rijsbergen清晰地提出了“聚类假设”(cluster hypothesis)[827]。1978年，第一届ACM信息检索会议(ACM Conference on IR，ACM SIGIR)在纽约州的罗切斯特举行。1979年，C. J. Van Rijsbergen出版了介绍概率检索模型的专著《Information Retrieval》[1624]。1983年，Salton和McGill出版了介绍向量检索模型的经典专著《Introduction to Modern Information Retrieval》[1414]。从那以后，信息检索研究群体日渐扩大，现在已包含来自世界各地成千上万的教授、研究人员、学生、工程师和从业人员。本领域最重要的会议——ACM信息检索国际会议(ACM International Conference on Information Retrieval，ACM SIGIR)，现在每年能吸引数百位参加者和数百篇投稿。

2

1.1.2 图书馆和数字图书馆中的信息检索

图书馆是采用信息检索系统搜寻信息的第一批机构。通常情况下，图书馆系统最初是由学术机构，后来由商业供应商开发。第一代图书馆系统是对现有流程的自动化，例如用作者姓名和书名检索卡片目录。第二代系统则增加了搜索功能，包括主题词和关键字的检索和查询操作。目前正在部署的第三代系统重点则是改进的图形界面、电子表单、超文本功能和开放式系统架构。

传统的图书馆管理系统供应商包括Endeavor信息系统公司、Innovative Interfaces公司和EOS国际公司。在目前正在开发的研究系统中，值得关注的是位于加州大学的加州数字图书馆所开发的MELVYL系统，以及最初由加州大学伯克利分校开发、最近与利物浦大学合作的Cheshire系统。关于这些图书馆系统的进一步详情可参看第16章。

1.1.3 舞台中央的信息检索

虽然已经成熟，但直到最近，信息检索仍被视为只有图书管理员和信息专家感兴趣的狭窄领域。这种偏见已盛行多年，尽管多媒体和超文本的信息检索工具已经在现代个人计算机用户中迅速传播。万维网(World Wide Web，Web)在20世纪90年代初的引入彻底颠覆了所有这些看法。

1989年蒂姆·伯纳斯-李发明的Web，已成为人类知识和文化的万能信息库。它的成功

⊖ 在信息检索领域，“precision”也译为“查准率”。——译者注

⊖ 在信息检索领域，“recall”也译为“查全率”。——译者注

是基于对标准用户界面的构想——该界面不随计算环境的改变而改变，并允许任何用户创建自己的文件。利用 Web，数百万用户已经创造了数十亿的文档，从而构成了人类历史上最大的知识宝库。一个直接后果是，在 Web 上查找有用的信息并不总是一个简单的任务，通常需要提交查询给搜索引擎，即运行一个搜索任务，这完全就是信息检索技术。因此，几乎在一夜之间，信息检索与其他技术一起，站在了舞台的中央。

1.2 信息检索问题

现代信息检索系统的用户，例如搜索引擎用户，有多种多样的信息需求。在最简单的情况下，他们寻找指向企业、政府或者机构主页的链接；在稍微复杂的情况下，他们寻找完成工作任务或即时需求的信息。更复杂的信息需求（information need）则例如：

寻找所有与联邦政府在全国铁路运输公司（National Railroad Transportation Corporation，AMTRAK）*融资中所扮演角色相关的文档*[⊖]。

用户需求的这种完整表示并不必然就构成提交给信息检索系统的最佳形式。相反地，用户可能首先要将此信息需求转换成查询（query）或者查询序列提交给系统。在其最常见的形式，这种转换总结了用户的信息需求，并产生一组关键字或索引项。给定用户查询，检索系统的主要目标是要获取有用或相关的信息并提交给用户。重点是信息检索，而不是数据检索。

为了有效地满足用户的信息需求，检索系统必须以某种方式“解释”信息项（即库中的文档）的内容，并根据和用户查询相关的程度对文档进行排序。文档内容的“解释”涉及从文档中提取文本的句法和语义信息并利用这些信息来匹配用户的信息需求。

信息检索问题：*信息检索系统的主要目标是检出所有和用户查询相关的文档，并且把检出的不相关文档控制在最低限度。*

信息检索的困难在于不仅需要知道如何从文档中提取信息，而且还要知道如何用它来决定相关性。也就是说，相关性的概念对信息检索至关重要。

一个主要的问题是，对相关性的评估是个性化的，决定于被解决的任务及其上下文。例如，相关性可以随时间而改变（如新信息的出现），随位置而改变（例如，最相关的答案是距离最近的），甚至随设备而改变（例如，最好的答案是一篇简短的、容易下载和可视化的文档）。从这个意义上讲，不存在能在任何时间给任何用户提供完美答案的检索系统。

1.2.1 用户的任务

检索系统的用户必须把他们的信息需求转换成用系统提供的语言所描述的查询。利用信息检索系统，如搜索引擎，通常意味着指定一组词来传达信息的语义。我们称之为用户在*搜索或查询*他们感兴趣的信息。虽然搜索感兴趣的信息是 Web 检索的主要任务，但除了信息获取之外，搜索也可用于满足其他种类的用户需求，如购买商品和订位等，我们将在 1.4.3 节对此加以讨论。

现在考虑这样一种情况，用户的兴趣要么定义不清要么流于泛泛，以至于很难清晰地制定查询。例如，用户可能对关于赛车的一般信息感兴趣，可能会决定浏览与 F1 赛车、印地车赛和勒芒 24 小时耐力赛有关的文档。我们称这种情况为用户在*浏览或者导航*文档集中的

⊖ TREC 参考集的 168 主题。参看第 4 章。

文档，而不是搜索。它仍然是一个信息检索过程，但主要目标起初并不太清楚。这种情况下，任务更多的是探索式搜索，类似于对感兴趣信息的准序列搜索过程。

在这本书中，我们将检索系统的不同用户所进行的任务区分为两种截然不同的类型：*搜索*和*浏览*，如图 1-1 所示。第 2 章将详细介绍
4 这两种不同的任务。

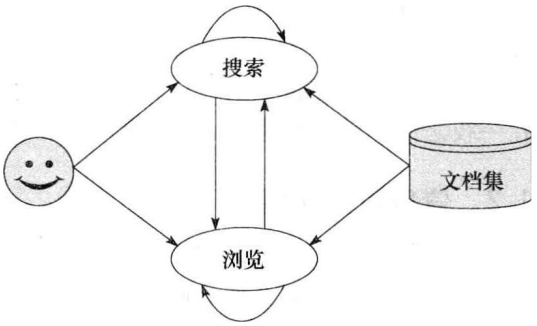


图 1-1 用户的任务

1.2.2 信息检索与数据检索

在信息检索系统的环境中，数据检索通常不足以满足用户的信息需求。数据检索主要包括确定集合中的哪些文档包含用户查询中的关键字。事实上，信息检索系统的用户更注重检出与某个主题相关的信息，而不是检索出符合用户查询的数据。例如，信息检索系统的用户愿意接受结果中包含查询项同义词的文档，即使这些文档没有包含任何查询项。也就是说，在一个信息检索系统中，检出的对象可以是不精确的，小错误可能被忽视。

与此相反，在数据检索系统中，1000 个检索对象中出现一个错误对象就意味着彻底失败。数据检索系统，如关系数据库，其处理对象具有明确定义的结构和语义；而信息检索系统处理的是没有很好结构的自然语言文本。数据检索能够为数据库系统的用户提供解决方案，但却不能解决检索与特定主题相关信息的问题。

1.3 信息检索系统

在本节中，我们提出信息检索系统软件架构的高层视图，并介绍响应用户查询的文档检索和排序过程。

1.3.1 信息检索系统的软件架构

为了描述信息检索系统，我们使用一个简单而通用的软件体系结构，如图 1-2 所示。建立信息检索系统的第一步是建立文档集，它可以是私有的，或者从 Web 上爬取。在第二种情况下，爬虫模块负责收集文档，我们将在第 12 章对此进行讨论。存储在磁盘上的文档集
5 通常称为*中央资源库*（central repository）。中央资源库里的文档需要进行索引，以进行快速检索和排序。最常用的索引结构是*倒排索引*（inverted index），它由文档集中所有不同的词组成，并为每个词建立一个包含这个词的文档列表。倒排索引将在第 9 章讨论。

文档集的索引建立之后，检索过程就可以启动。它既包括检索满足用户查询的文档，也包括点击超链接。在第一种情况下，我们说用户正在*搜索*感兴趣的信息；在第二种情况下，我们说用户在*浏览*感兴趣的信息。本节的其余部分介绍搜索。有关浏览的更详细的讨论，以及两种情况的比较，请参阅第 2 章。

为了进行搜索，用户首先指定一个反映他们信息需求的*查询*。接下来，对用户查询进行分析和扩展，例如加入查询词的拼写变体。扩展的查询，我们称之为*系统查询*，将与倒排索引进行匹配，并检索出一个文档子集。接下来，对文档子集排序并把排在最前面的文档返回给用户。换行排序的目的是找出最有可能被用户认为是相关的文档。这构成了信息检索系统中最关键的部分。正因为如此，第 3 章的信息检索模型介绍将非常详细，
6 且覆盖范围广泛。

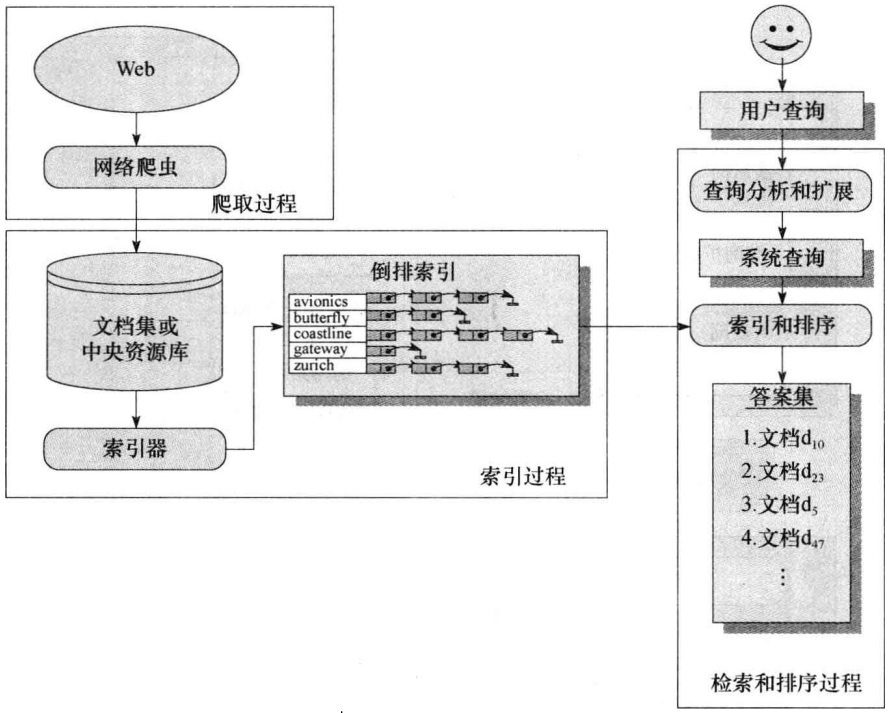


图 1-2　信息检索系统的高层软件架构。其中爬取是 Web 信息检索系统（如搜索引擎）额外要求的一个模块

鉴于判断相关性过程中所固有的主观性因素，评价答案集的质量是提高信息检索系统性能的关键步骤。系统的评价过程允许对排序算法进行微调以提高结果的质量，我们将在第 4 章对此加以讨论。最常见的评价过程是把信息检索系统产生的结果文档集和人类专家建议的结果进行比较。

为了提高排序的性能，我们可以收集用户的反馈，并使用这些信息来对结果重新排序。在 Web 中，最丰富的用户反馈形式是在返回结果上点击链接，我们将在第 5 章讨论。网页排序的另一个重要信息来源是页面间的超链接，可以从中发现权威度较高的页面，我们将在第 11 章讨论。

对一个完全成熟的信息检索系统（例如现代搜索引擎）而言，还有许多其他的概念和技术，其中大多数将在本书的其余章节内介绍。

1.3.2　检索和排序过程

为了描述检索和排序过程，我们对图 1-2 所示的模块进行进一步阐述，如图 1-3 所示。给定文档集中的文档，我们首先进行禁用词消除、词干提取等文本操作，并选择所有项的一个子集作为索引项，然后用索引项来构建文档的表示，这种表示可能比文档本身小（取决于选定的索引项子集）。

给定该文档表示，有必要建立一个文本索引。可以使用不同的索引结构，但最流行的是将在第 9 章讨论的*倒排索引*。生成所需索引的步骤就组成了*索引过程*，该过程必须在系统准备好处理任何查询之前离线执行。在索引过程中所耗费的资源（时间和存储空间）由检索系统在处理多次查询的过程中分摊。

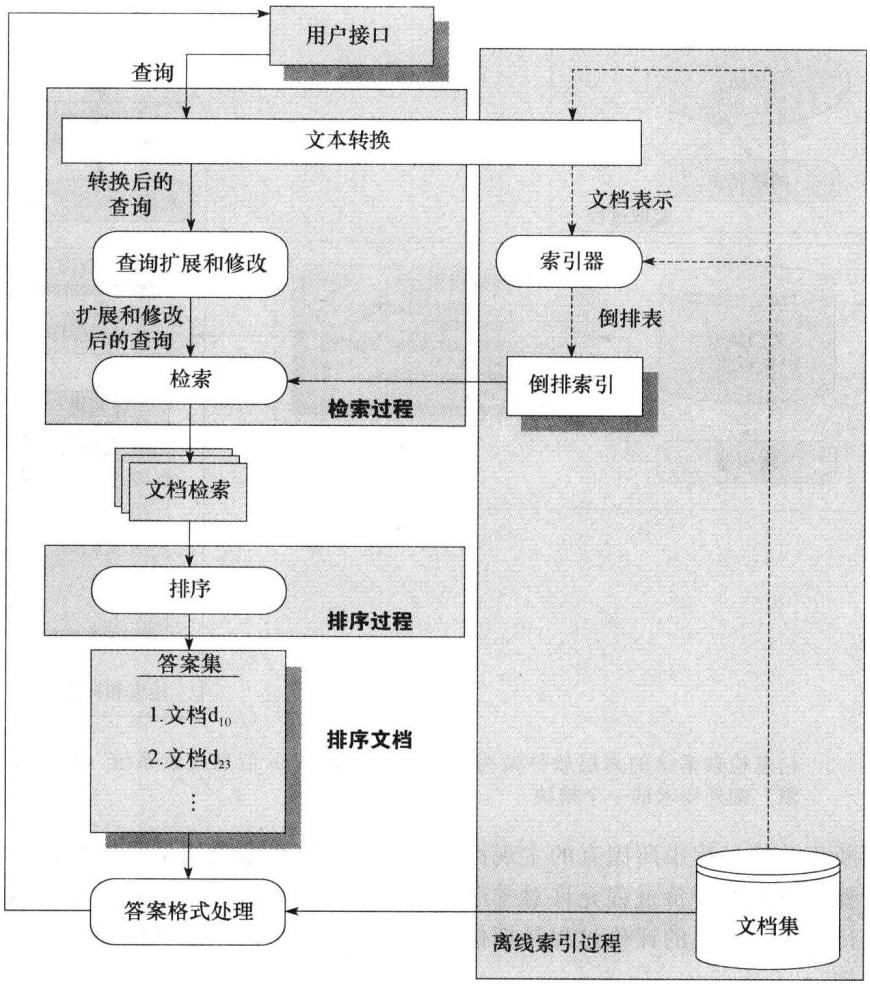


图 1-3 文档的索引、检索和排序过程

检索过程在给定文档集的索引之后启动。用户首先指定一个反映他们信息需求的查询，然后对此查询进行与文档类似的分析和修改操作。这里的典型操作包括适当的拼写校对和禁用词消除等。接下来，对转换后的查询进行扩展和修改。例如，系统可以对查询做出修改建议，并由用户确认。系统对扩展和修改后的查询进行处理，产生检索文档集，即由包含查询项的文档组成的集合。而先前建立的索引结构使得快速的查询处理成为可能。产生检索文档集的必要步骤就构成了检索过程。

接下来，检索出的文档将根据与用户需求的相关性进行似然度排序。由于用户所能感知的检索结果质量完全依赖于排序，因此这是最为关键的步骤。第 3 章我们将对排序过程进行详细介绍。系统将对排在最前面的文档进行格式处理并展现给用户。这些格式处理包括找出
7 文档的标题，根据查询项在文档中出现的上下文生成结果片段等。

1.4 Web

在本节中，我们讨论 Web 及其所揭示的电子出版时代。我们还会讨论 Web 如何改变搜索，也就是说，Web 对搜索任务的主要影响。最后，我们涵盖诸如安全和版权等由百万级的大规模 Web 用户而导致的实际问题。

1.4.1 Web 简史

在第二次世界大战结束时，美国总统富兰克林·罗斯福向后来获得高级政府职位的万尼瓦尔·布什咨询如何将在战争中掌握的技术应用到和平时期。布什首先做了名为“Science, The Endless Frontier”（科学，无尽的前沿）的报告。该报告直接影响了美国国家科学基金会的建立。之后，他写了一篇影响深远的文章“As We May Think”（我们可以想象）[303]，讨论了可能在未来几年发明的新硬件和软件。用布什的话说， 8

> 百科全书将以全新的形式出现，资料之间由网络关联，随时可以放入扩展存储器（memex），并可以不断往里面添加新的信息 [303]。

“As We May Think”影响了许多人，包括 Douglas Engelbart。他在 1968 年 12 月的旧金山秋季联合计算机会议（Fall Joint Computer Conference）上运行了一个演示系统，推出了首个计算机鼠标、视频会议系统、远程会议系统和超文本。它是如此不可思议，以至于成为“所有演示之母”[1690]。演示中最让我们感兴趣的创新之处是超文本（hypertext）。该术语是由 Ted Nelson 在他的项目“世外桃源”（Xanadu）[1691] 中创造的。

超文本允许读者从一个电子文件跳转到另一个，这是蒂姆·伯纳斯-李（Tim Berners-Lee）在 1989 年所面临问题的一个重要属性。当时，伯纳斯-李在日内瓦的欧洲核子研究中心（Conseil Eouropéen pour la Recherche Nucléaire，CERN）工作。那里的研究人员如果想要与他人分享自己的文件就必须重新格式化文件，使其与内部的出版系统兼容 [803]。这很令人厌烦，产生了许多问题，其中许多问题需要由伯纳斯-李去解决。他意识到需要更好的解决方案。

欧洲核子研究中心碰巧是欧洲最大的因特网节点。伯纳斯-李认为，需要把共享的文件分散化，使得研究人员能够自由地分享他们的成果。他认为通过因特网链接的超文本将是一个很好的解决方案，并开始着手实现。1990 年，他写了 HTTP 协议，定义了 HTML 语言，编写了第一个 Web 浏览器——他称之为“万维网”，并搭建了第一个 Web 服务器。1991 年，他在因特网上发布了浏览器和服务器软件。Web 诞生了。

1.4.2 电子出版时代

Web 从一出现就取得了巨大的成功。现在网页的数量已远远超过 200 亿[⊖] [487]，全世界的 Web 用户数也超过 17 亿 [815]。此外，众所周知在 Web 上有超过 1 万亿个不同的 URL[651]，即使其中许多是指向动态页面的指针，而不是静态的 HTML 页面。基于在线广告甚至实现了经济可持续发展的可行模式 [801]。

Web 的出现改变了这个世界，这一点很少有人能预见到。然而，人们想知道 Web 有哪些特性使得它如此成功，或者说，是否存在着某个单一的特性，对 Web 的成功起到决定性的作用？对这个问题的初步答案包括：简单的 HTML 标记语言、低成本的存取、因特网的广泛普及、交互式的浏览器界面，以及搜索引擎。然而，虽然这些技术提供了基本的 Web 基础设施，但不是其流行的根源。那么特性是什么呢？ 9

要强调这里我们提出的观点，让我们观察 200 年前某位作家的一生。

> 她在 1796 年和 1797 年之间完成了她的小说初稿。第一次投稿却被拒绝。因为最终失去了原稿，所以她在 1812 年改写了小说，并终于在 1813 年匿名出版，署名

⊖ 根据 http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html 上的博文，谷歌宣称已经搜集了超过 1 万亿个不同的 URL。

为“一位女士”[400]。

《Pride and Prejudice》（傲慢与偏见）是英国史上最受欢迎的三部书籍之一，另两部是《The Lord of the Rings》（指环王）和《Harry Potter》（哈利·波特）系列丛书。它先后被翻拍为6部电视剧和5部电影[1694]。最近的一部由Keira Knightley和Matthew Macfadyen主演，获得了超过1亿美元的全球票房，Knightley女士并因此获得了奥斯卡奖提名[1693]。

简·奥斯汀一生所有的作品都是匿名发表。在整个20世纪，奥斯汀的小说从未绝版。各种版本的出现对《Pride and Prejudice》的普及功不可没。

Web所带来的人际关系的根本性转变是出版自由。简·奥斯汀没有这种自由，所以她要不必须设法说服出版商相信作品的质量，要不就自己支付出版费用。由于无法支付，因此她不得不耐心等待了15年，直到出版商相信为止。

在Web世界中，这样的情况不会再发生。人们现在可以在Web上发布自己的想法，无须支付任何代价，也无须说服大型出版公司的编委会。而这样的想法经过一夜就会为数百万人所知晓。也就是说，Web几乎完全解除了由大众传媒公司和自然地域壁垒所造成的限制，这导致了自由出版的新时代的诞生。我们称之为电子出版时代。

1.4.3 Web如何改变搜索

Web搜索是信息检索及其相关技术当今最突出的应用。事实上，任何搜索引擎的排序和索引组件在本质上都是信息检索技术。这个事实的一个直接后果就是，Web对信息检索的发展已经产生了重大影响。

Web对搜索的第一个重大影响与文档集自身的特点相关。Web文档集是由分布在数百万个网站上的文档（或网页）组成，并通过超链接将页面上的文字块与其他网页连接。由于Web文档集所固有的分布式性质，因此在建立索引之前，需要收集所有文档的副本并将它们存储在一个中央资源库中。由Web所带来的这个信息检索过程中的新阶段称为网页爬取
10 (crawling)，将在第12章详细讨论。

Web对搜索的第二个重大影响与文档集的大小以及每天提交的用户查询数量有关。Web比以往任何已知的文档集更大，且增长速度更快，以至于现在的搜索引擎所需要处理的文本数量已经远远超过200亿页[487]，远大于以往的任何文档集。此外，虽然对于用户查询的数量有各种规模的估计，但都认为比以往任何时候都多。海量文档集和海量查询流量的结合，使得搜索引擎的性能和可扩展性要求大大超过以往任何信息检索系统[151]。也就是说，性能和可扩展性已成为Web信息检索系统的重要特性，其重要程度远远超过了它们在以往检索系统中的地位。虽然本书不讨论搜索引擎的性能和可扩展性问题，但是读者可以参考第11章关于本主题的文献（见文献讨论章节）。

Web对搜索的第三个重大影响也与海量文档集有关。在非常大的文档集中预测相关性比以前更难。基本上，任何查询都会检索出很多匹配查询项的文档，这意味着检索文档集中有许多噪声。也就是说，检索文档集中的大部分文档似乎与查询相关，但实际上据大多数用户判断却是不相关的。此问题首次出现于早期的Web搜索引擎中，并随着Web的增长变得更严重。幸运的是，Web还提供了标准文档集所没有的、缓解上述问题的新证据来源，如超链接、用户在结果文档中的点击行为等。在第11章中，我们将讨论Web上的相关性预测问题。

Web对搜索的另外两个主要的影响源于这样的事实：Web已不再仅仅是文档和数据库，也是一个商业媒介。直接的含义就是，搜索问题已经超出了对文字资料的寻找，还扩展到其

他用户需求，例如查询一本书的价格、酒店的电话号码、下载软件的链接。对这些类型的信息，提供有效的答案经常需要确定和关注对象相关联的一些结构化数据，如价格、地点，或主要特性描述等。这些新的查询类型将在第7章讨论。

Web对搜索的第五个、也是最后的影响来自于Web广告和其他经济激励。作为大众化互动媒体，Web的持续成功创造了广告和电子商务等形式的经济开发激励机制。这些激励措施也导致了Web垃圾信息的泛滥，也就是把商业信息伪装成纯粹的信息内容。Web上的垃圾信息越来越普遍，有时是如此引人注目，并且与真正的相关内容相混淆，使得寻找相关信息甚至比以前更困难。正因为如此，认为垃圾内容使得相关性变差，也就是说垃圾信息的存在使得现有的排序算法产生的答案比没有垃圾内容的情况差很多，这也不是一点道理都没有。这种困难是如此之大，以至于现在需要谈论敌对Web检索，我们将在第11章对此加以讨论。 11

1.4.4 Web上的实际问题

电子商务是当今Web一个惠及亿万人民的大趋势。在电子交易中，买方通常提交信用资料给供应商进行收费。信用资料最常见的形式就是信用卡号。出于安全原因，这些信息通常是加密的，由机构和公司部署的验证过程自动完成。

除了安全外，另一个引起关注的主要问题是隐私。通常，只要不被公开，人们都愿意交换信息。原因有很多，但最常见的是防止由第三方滥用自己的私人信息。因此，隐私是另一个影响Web的部署却并没有得到妥善解决的问题。

另外两个重要的问题是著作权和专利权。Web数据的广泛分布如何影响各个国家的版权和专利法，目前还很不明朗。这一点很重要，因为它影响了建立和部署大型数字图书馆的业务。举例来说，网站是否要像出版商一样监督发布的所有信息？如果是的话，如果发布的信息被滥用，它是否需要负责（即使它不是信息源）？

此外，其他值得关注的实际问题包括扫描、光学字符识别（Optical Character Recognition，OCR），以及跨语言检索（用一种语言提交查询，但检索出的文档是另一种语言）。但是，本书将不会对这些实际问题进行详细介绍，因为它不是我们的主要关注点。有兴趣的读者可以参考Lesk的著作[1005]。

1.5 本书的组织结构

1.5.1 本书的重点

虽然信息检索越来越引起人们的兴趣，但广泛覆盖本领域众多主题的现代信息检索教科书仍很难找到。本书从计算机科学家的视角出发，介绍信息检索领域的整体研究现状，试图部分地填补这一鸿沟。这意味着本书的关注点是信息检索系统所使用的计算机算法和技术。图书馆专家和信息科学研究人员的视角则截然不同，他们从以用户为中心的角度解释信息检索系统，其关注点不是如何自动地结构化、存储和检索信息，而是试图理解人们如何解释和使用信息。虽然本书的大部分章节专注于从计算机科学家的视角研究信息检索系统，但在本书的用户界面部分和最后两章的部分章节依然讨论了以人为中心的视角。

本书着重强调与信息检索紧密相关的不同领域需要整合在一起。因此，除了覆盖文本检索、图书馆系统、用户界面和Web之外，本书也介绍了可视化、多媒体信息检索和数字图书馆。

虽然有多位专家撰写了部分章节，本书依然是一本教科书，其内容和结构由两个主要作者进行了精心设计，他们也撰写或合写了全书17章中的12章。此外，所有其他作者撰写的 12

章节都已审慎修改、编辑，并被整合进入统一的框架。该框架规定了结构一致性、统一风格、共同词汇表、共同书目，以及适当的交叉引用。在每章的结尾讨论了研究问题、趋势和参考文献。这种讨论对研究生以及研究人员应该是有价值的。

1.5.2 本书的内容

由于信息检索是有五十多年历史的学科，一本书只能涵盖本领域全部知识的有限部分。尽管如此，为了获得对信息检索技术广泛深入的理解，仍然需要了解一些核心的关键概念、方法和技术。为了尽量覆盖这些概念和技术，我们撰写了17章内容，构成了本书第2版。由于第1版是十多年前出版的，因此本书第2版的所有章节和第1版截然不同。事实上，一半以上的素材是新的或已重写，目的或者是为了全面覆盖最新的研究结果，或者是为了简化符号，或者是为了介绍第1版尚未涉及的相关主题。为了说明这一点，本书增加了文本分类、结构化文本检索、Web爬取，以及企业搜索等章节。此外，对相关反馈、多媒体、Web、图书馆系统、数字图书馆、检索评价和建模的章节已进行了大量修改和更新。

图1-4说明了本书的组织结构。本章介绍信息检索问题、Web的简史，并分析其对信息检索的影响。因为搜索已经成为信息检索技术的主要应用领域，所以第2章论述用户搜索界面的设计。第2章是全新的，和第1版的用户界面章节截然不同，旨在为读者理解信息检索问题提供一个自顶向下的视角。

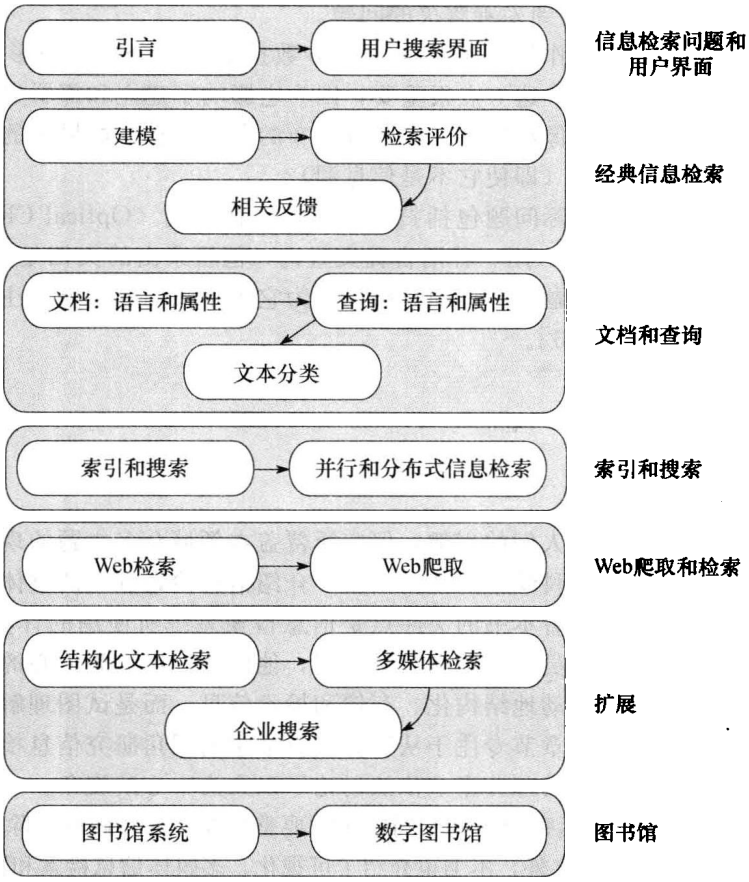


图1-4 本书的组织结构

接下来的三章涉及经典的信息检索模型，包括排序模型、结果质量评价和用户相关反馈。这三章超过一半的素材是新的，证明了信息检索在过去 10 年的快速发展。我们的讨论广泛而深入。第 3 章讨论 14 个不同的信息检索模型，这些模型旨在对答案集中的每个文档打分并按照分数排序。我们从经典的布尔、向量和概率模型开始，然后对这 3 个经典模型中的每一个引入 3 个变种，包括基于集合的模型、广义向量模型、BM25 和语言模型。第 4 章就信息检索系统的结果质量评价，探讨了许多技术细节。我们首先进行了简单的历史回顾，包括 Cleverdon 针对索引系统评价的开创性工作，及其如何发展成为 Cranfield 范式。接下来介绍精度-召回率关系图，针对相关性分级的折扣累积增益（Discounted Cumulative Gain，DCG）指标，以及针对不完全相关性评价的二元偏好（Bpref）指标。我们同时也讨论了斯皮尔曼系数和肯德尔等级相关系数等排序关联度指标，并详尽介绍了 TREC 文档集和其他众多的小测试文档集。最后，讨论 Web 特有的评价方法，例如并排面板（side-by-side panels），并讨论如何将点击数据解释为相关性指标的方法。第 5 章讨论用户进行相关反馈的隐式和显式方法，以及如何使用它们来改变最终的排序。这些方法和查询扩展技术直 13
接交织在一起。这三章涵盖了所有经典的信息检索基本概念，即解决信息检索问题并评价结果的技术和方法。

接下来的三章讨论文档和查询的概念和相关技术，以及如何通过文本分类组织文档和查询。第 6 章讨论文本属性，例如词汇在文档的分布、自然语言模型、SGML、HTML 和 XML 等标注语言、文本处理与分析，以及压缩方法。第 7 章讨论各种查询属性，包括查询关键词的分布、Web 查询的特点，以及基于关键词的查询语言、结构化形式和查询协议等。第 8 章讨论组织文档和查询的算法和方法。我们讨论的重点是文档分类，因为这是最常见的情况。我们区分无监督和监督的文本分类算法。对于无监督的方法，我们介绍文本聚类算法，如 K 均值算法及其变种。对于监督方法，我们讨论 6 种不同类型的文本分类算法，即 14
决策树、最近邻、Rocchio、朴素贝叶斯、支持向量机和集成分类器。我们还会详细讨论如何评价分类结果。鉴于文本分类如今是信息检索的一项关键技术，所以本章是全新的，也是本书和第 1 版的重要区别之一。

再接下来的两章讨论索引和搜索文档集使用的技术。第 9 章讨论各种索引和搜索技术，包括序列搜索、倒排索引和后缀数组。我们还介绍索引压缩技术，以及如何使用它们来提高检索速度。第 10 章讨论并行和分布式索引以及（查询）搜索过程的体系结构和算法。提交给搜索引擎的海量查询只能由分布式计算机集群处理，这是现代 Web 的主要趋势。

之后的两章覆盖 Web 文档的爬取、检索和排序。第 11 章讨论 Web 检索，介绍 Web 的属性，搜索引擎的体系结构，HITS 和 Page rank 等链接分析算法，以及 Web 文档排序。虽然本章没有包括该领域的全部研究——当然任何一章都不可能，但它的确说明了搜索引擎如何获益于信息检索算法和技术。第 12 章讨论 Web 爬取。我们首先简要回顾了 Web 爬取技术的发展历史，然后讨论 Web 爬取的架构和实施问题。接下来是调度算法，这是任何爬取算法的核心部分，它确定下一步应该爬取哪些网页。最后，我们讨论 Web 爬取的评价过程。

Web 搜索的扩展包括结构化文本检索和多媒体检索，这两个是与 Web 日益相关的主要领域，另外还包括企业搜索。第 13 章讨论结构化文本检索，这是全新的章节，反映了自本书第 1 版出版以来该领域的迅速发展。其中包括早期的文本检索模型，XML 的索引和排序模型，XML 检索的评价方法和 XML 查询语言。第 14 章的多媒体信息检索也是全新的内容，从信息检索视角出发，从自顶向下的角度讨论多媒体检索。涵盖的内容包括基于内容的图像检索、音频和音乐检索，以及视频检索。将基于内容的图像检索、音频、音乐和视频检

索组合成一个单一的搜索机制需要融合模型，我们也会对此进行讨论。最后介绍 MPEG 标准。第 15 章讨论在机构和企业内部检索信息的企业搜索系统，包括它们和 Web 搜索系统的区别，以及在设计和实现方面的挑战。

本书的最后两章包括图书馆系统和数字图书馆。第 16 章讨论商业文档数据库、集成图书馆系统（Integrated Library System，ILS）和联机公共检索目录。商业文档数据库仍是当今最大的信息检索系统。例如 LEXIS-NEXIS 有一个由超过十亿的文档组成的数据库，每年提供数百万次查询。第 17 章已全面改写，提供对数字图书馆最新技术和趋势的详尽描述。首先是历史概述，随后讨论基本概念、社会和经济问题，以及 7 个独特的数字图书馆系统。最后，我们还讨论了数字图书馆中的重要个案，例如学位论文网络数字图书馆（networked digital library of theses and dissertations）、国家科学数字图书馆（national science digital library）和 ETANA 考古数字图书馆（ETANA archaeological digital library）。

15

本书还包括两个附录。附录 A 评论 27 个开源搜索系统，包括 HtDig、Indri、Lucene、MG4J、Omega、Omnifind、SwishE、Swish++、Terrier 和 Zettair。附录 A 对这些搜索系统从索引构建时间、查询处理性能和存储需求等方面进行比较分析。附录 B 是对本书有贡献的所有作者的简介。最后是全书所使用的 1800 多篇参考文献。

虽然本书第 2 版的大部分资料还是纯粹的教科书风格，但我们还是为对研究有兴趣的读者增加了更多的内嵌引用。虽然我们试图平衡内容的广度和深度，但由于我们自己的专长和研究兴趣，有一些题目论述得更详细些。如果我们错过了一些主题或重要的细节或引用，我们在此提前道歉。

从经典的信息检索到 Web，从信息组织算法到现代数字图书馆，从搜索引擎所使用的索引和搜索技术到结构化文本搜索、多媒体搜索等需要扩充的新技术，本书第 2 版旨在从一个广泛而深入的视角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，以及对相关领域（如信息科学、多媒体、数据库和数字图书馆）知识的影响。

1.6 本书的教学资源网站

本书的网站是 http://www.mir2ed.org，其中包含全书所有章节的幻灯片，可以作为教学资源使用。除幻灯片之外，也包括了词汇表、练习题和对面向不同听众的不同课程的详细教学建议，例如：

- 信息检索，计算机科学专业，本科生水平；
- 高级信息检索，计算机科学专业，研究生水平；
- 多媒体检索，计算机科学专业，本科生水平；
- 信息检索，信息系统专业，本科生水平；
- 信息检索，图书馆学专业，本科生水平；
- Web 检索，通识教育，本科或研究生水平；
- 数字图书馆，通识教育，本科或研究生水平。

此外，网站提供一个参考文档集供实验之用，包含 1239 篇来自 Cystic Fibrosis 参考集的文档，100 个信息需求和详尽的相关性评价数据 [1454]。而且，网站包括连接不同大学的信息检索课程、研究组、出版机构以及与信息检索及本书相关的其他资源的链接。

16

最后，本书网站还将公开发布与本书相关的重要新成果和补充信息，以及勘误表。

1.7 文献讨论

现在市面上已经有许多关于信息检索的其他书籍，由于目前对该主题的广泛兴趣，最近

也出现了一些新书。下面我们将本书与这些之前出版的书进行简单的比较。

信息检索领域的经典参考书是 van Rijsbergen 的《Information Retrieval》[1624]（网上可以找到），以及 Salton 和 McGill 的《Introduction to Modern Information Retrieval》[1414]。本书对于数据和信息检索的区别借鉴了前者，对于信息检索过程的定义则受到了后者的影响。然而，25 年过去了，这两本书现已过时，不能涵盖信息检索的新进展。

其他三本众所周知的信息检索著作是 Frakes 和 Baeza-Yates 编辑的《Information Retrieval: Data Structures & Algorithms》[582]，Witten、Moffat 和 Bell 撰写的《Managing Gigabytes-Compressing and Indexing Documents and Images》[1709]，以及 Lesk 的《Practical Digital Libraries; Books, Bytes, & Bucks》[1005]。这三本书都和本书互为补充。第一本偏重信息检索的数据结构和算法，有助于迅速实现已知算法的原型。第二本偏重索引和压缩技术，同时除了文本之外，也覆盖图像。本书由此借鉴了文本化图像的概念。第三本偏重数字图书馆及其实际问题，例如历史、分布、可用性、经济意义和知识产权。关于经典信息检索较新的著作包括 Hersh 的 [749]，Chowdhury 的《Introduction to Modern Information Retrieval》第 3 版 [382]。这两本著作的视角都比本书窄。Meadow、Boyce、Kraft 和 Barry 的《Text Information Retrieval Systems》第 3 版 [1112] 着重介绍信息及其表示。Allen 的《Information Tasks: Toward a User-Centered Approach to Information Systems》[32] 是关于信息系统的一般性著作，它采用以用户为中心，而不是以计算机为中心的视角阐述检索。从信息搜寻的视角，则需要提及 Marchionini 的《Information Seeking in Electronic Environments》[1082]，以及 Tedd 和 Hartley 的《Information Seeking in The Online Age: Principles and Practice》[977]。

某些章节有补充书籍。例如，许多书籍讨论信息检索和超文本，包括 Agosti 和 Smeaton 编辑的《Information retrieval and hypertext》[20]。多媒体检索也是如此，例如 Steinmetz 和 Nahrstedt 的《Multimedia-Computing, Communications and Applications》[1534]，以及 Alessi 和 Trollip 的《Multimedia for learning: methods and development》[25]。Hersh 的《Information Retrieval-A Health and Biomedical Perspective》[749] 是一本有趣的书，从健康和生物医药角度讨论信息检索。虽然标题中没有信息检索，但 Rosenfeld 和 Morville 的《Information Architecture for the World Wide Web: Designing Large-Scale Web Sites》第 3 版 [1157] 介绍了 Web 上的信息架构，是本书第 11 章的有益补充。Menasce 和 Almeida 的《Capacity Planning for Web Performance: Metrics, Models, and Methods》[1118] 阐述了如何利用排队论来预测 Web 服务器的行为。Chakrabarti 的《Mining the Web: Discovering Knowledge from Hypertext Data》[349] 介绍了 Web 知识挖掘的方法。此外，还有许多书籍说明如何从 Web 发现信息，如何使用搜索引擎。

Sparck Jones 和 Willet 编辑的《Readings in Information Retrieval》[1510]，与其说是一本合著，不如说是论文集。本书具有连贯性和广泛性，是更合适的学科教材。不过，该论文集仍然是有价值的研究工具书。Grefenstette 编辑的《Cross-Language Information Retrieval》是一本与跨语言信息检索有关的论文集 [674]。读者如对这个特定主题感兴趣，那么这本论文集就是本书很好的补充。此外，Maybury 编辑的《Intelligent Multimedia Information Retrieval》是一本偏重智能多媒体检索的论文集 [1101]，而 Strzalkowski 编辑的《Natural Language Information Retrieval》则关注自然语言信息检索 [1538]。为了纪念 Karen Sparck Jones，Tait 编辑的《Charting a New Course: Natural Language Processing and Information Retrieval》讨论自然语言处理与信息检索的关系 [1554]。其他一些合著探讨了 [17]

信息检索和不确定性与逻辑的关系 [444]、语言模型 [453]、认知检索 [1515] 和 TREC 评价 [1654]。

Korfhage 的《Information Storage and Retrieval》[931] 覆盖的材料比本书少很多，且不够具体。例如，该书没有详细讨论数字图书馆、Web、多媒体，以及并行处理。类似地，Kowalski 和 Maybury 的《Information Storage and Retrieval Systems：Theory and Implementation》第 2 版 [937]，还有 Shapiro 等人的《Automated Information Retrieval：Theory and Text-Only Methods》[1453] 都没有详细介绍这些内容，且定位也不同。Grossman 和 Frieder 的《Information Retrieval：Algorithms and Heuristics》[682] 没有讨论 Web、数字图书馆和可视化界面。Berry 和 Browne 的《Understanding Search Engines-Mathematical Modeling and Text Retrieval》[194] 是一本在搜索引擎语境下讨论经典信息检索的著作。其他一些专著则分别偏重于信息检索的数学基础 [505]、检索的几何解释 [1625]，以及标记结构在信息检索中的智能应用等 [942]。

近期 Ingwersen 和 Jarvelin 关于信息搜寻的著作《The Turn：Integration of Information Seeking and Retrieval in Context》[810] 力图从延伸的认知角度，而非基于 Cranfield 范式的实验模型，来解释信息检索。这直接影响了系统的评价方法。采用认知角度阐述信息检索，但却专注于搜索引擎的另一本专著是 Belew 的《Finding Out About：A Cognitive Perspective on Search Engine Technology and the WWW》[170]。最近的一本探索性搜索著作是 White 和 Roth 的《Exploratory Search：Beyond the Query-Response Paradigm》[1686]。

最近，Manning、Raghavan 和 Schutze 撰写了一部介绍经典信息检索和 Web 搜索的著作《Introduction to Information Retrieval》[1081]，该书的视角相当连贯而优雅，着重于基本概念，因此没有像本书一样探讨搜索界面等问题，也没有详尽介绍各种信息检索模型。此外，该书对于检索质量的评价和 Web 爬取介绍得很少，并且没有介绍结构化文本检索、多媒体检索、图书馆搜索系统和数字图书馆等内容。

Croft、Metzler 和 Strohman 的《Search Engines-Information Retrieval in Practice》是一本更新的著作 [449]，着重介绍搜索引擎，可作为本科生教材。该书提供的材料可用于讲述如何应用信息检索技术搭建搜索引擎的介绍性课程，因此其覆盖范围比本书窄，且没有包含搜索界面、相关反馈、查询扩展、多媒体和图书馆等材料。另外，对于建模、检索评价、文本分类和结构化文本检索的材料也不如本书详尽深入。

最后，几乎和本书同时出版的是 Büttcher、Clarke 和 Cormack 的《Information Retrieval：Implementing and Evaluating Search Engines》[304]。该书偏重信息检索系统的实现和评价，内容包括 XML 检索、并行搜索引擎和 Web 搜索。

对关注研究结果的读者而言，讨论信息检索及相关主题的学术期刊主要包括：

- 《Journal of the American Society of Information Sciences and Technology》（JASIST，Wiley and Sons）
- 《ACM Transactions on Information Systems》（TOIS）
- 《Information Retrieval》（Kluwer）
- 《Information Processing and Management》（IP&M，Elsevier）
- 《ACM Transactions on the Web》
- 《IEEE Transactions on Knowledge and Data Engineering》（TKDE）
- 《Information Systems》（Elsevier）
- 《Knowledge and Information Systems》（KAIS，Springer）

- 《Data and Knowledge Engineering》(DKE, Springer)
- 《D-Lib Magazine》
- 《International Journal on Digital Libraries》(Springer)

与信息检索最相关的会议包括:

- ACM SIGIR International Conference on Information Retrieval (ACM SIGIR 信息检索国际会议)
- ACM International Conference on Web Search and Data Mining (WSDM) (ACM Web 搜索和数据挖掘国际会议)
- World Wide Web Conference (WWW), search track (万维网会议搜索分会)
- ACM Conference on Information Knowledge and Management (CIKM) (ACM 信息与知识管理会议)
- European Conference on IR (ECIR) (欧洲信息检索会议)
- String Processing and Information Retrieval Symposium (SPIRE) (国际字符串处理和信息检索会议)
- Text REtrieval Conference (TREC) (文本检索会议)
- INitiative for the Evaluation of XML retrieval (INEX) (INEX XML 检索评测)
- Cross Language Evaluation Forum (CLEF) (跨语言评价论坛)
- International Conference on Multimedia Retrieval (ICMR) (国际多媒体检索会议), 该会议新近由 ACM MIR 和 CIVR 合并而成
- Joint ACM-IEEE Conference on Digital Libraries (JCDL) (ACM-IEEE 数字图书馆联合会议)
- European Conference on Digital Libraries (ECDL) (欧洲数字图书馆会议) 19

第2章

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

用户搜索界面

——Marti Hearst 著

2.1 介绍

本书大部分的内容描述搜索引擎和信息检索系统的算法。而本章关注的是搜索系统的用户和搜索系统所显示的视窗：*用户搜索界面*。用户搜索界面的作用是帮助用户理解和表达他们的信息需求，并帮助用户制定他们的查询，在可用的信息源中进行选择，理解搜索结果，以及跟踪他们的搜索进程。

本书的第1版很少提到有关如何建立有效搜索界面的问题。在这些年间，人们开始逐步认识到哪些想法是有效的，哪些是无效的。本章简要总结了一些在学术研究以及商业系统部署方面最先进的搜索界面设计方法，讨论人们是如何搜索的、现今的搜索界面、搜索界面的可视化以及对用户搜索界面的评价。

2.2 人们如何搜索

搜索任务的范围可以从相对简单的（例如，查找有争议的事实或者查找天气信息）到丰富而复杂的（例如，求职和规划假期）。搜索界面应该支持一定范围内的任务，同时也要考虑到人们希望如何寻找到他们想要的信息。本节总结了与在线信息搜寻过程相关的理论模型和经验观察。

2.2.1 信息查找与探索式搜索

与搜索界面交互的不同方式取决于任务的类型、搜索过程中投入的时间和精力，以及信息搜寻者的专业知识。Web搜索引擎中所使用的简单的交互式对话最适合寻找问题的答案、搜索网站，或者作为搜索的起点寻找其他资源。但是，正如Marchionini [1085]指出的那样，搜索引擎“依次接收信息”的界面本身有局限性，在许多情况下正在被专业搜索引擎所取代——如对于旅游和健康信息的搜索，专业引擎能够提供更丰富的互动模式。

Marchionini [1085]给出了*信息查找*（information lookup）和*探索式搜索*（exploratory search）的区别。信息查找任务类似于事实检索或问题回答，只需要简短而离散的信息即可：数字、日期、名称或文件和网站的名称。标准的Web搜索（以及标准数据库管理系统查询）在这些方面可以做得很好。

Marchionini将信息搜寻任务中的探索式搜索的类别划分为*学习*和*调查*两类。学习搜索需要多个查询响应对，并需要用户花费时间扫描和读取多个信息项，并综合这些内容来形成新的理解。调查指的是一个更长期的过程，意指“在相对较长的一段时间内进行多次迭代，返回的结果可能要在整合进个人和专业知识库之前，进行严格的评估。”[1085]调查搜索可能被用做辅助计划安排、发现知识鸿沟，或者监视一个持续性的话题。有些种类的调查搜索关注于发现全体或大部分的可用相关信息（高召回率），比如诉讼研究或学术研究等。

其他人的一些工作支持了这种观点，O'Day和Jeffries[1219]在研究了那些反复出现的深度复杂信息需求之后（他们主要关注于商业智能领域），发现信息搜寻过程是由一系列相

互关联但又不完全相同的搜索所组成的。他们还发现，一个搜索目标的结果往往会引发新的目标，从而引发新的搜索方向，但问题的背景和先前的搜索会从搜索的前一个阶段延续到下一阶段。他们还发现，搜索所带来的主要价值体现在搜索过程中持续的学习和所获得的信息，而不只是最后的搜索结果。

更广泛地说，信息搜寻可以被看做是更大过程当中的一部分，正如文献 [1272，1401，1400] 提到的意义建构（sensemaking）那样。意义建构是一个迭代的过程，它从一个大的信息集合中制订出一套概念表示方法。Russell 等人 [1401] 观察到，在意义建构中，主要的工作都致力于如何把好的表示方法、思考形式，以及面临的问题结合起来。他们描述了为给定的任务制定和明确其中的重要概念的过程。搜索只是这一过程中的一个部分；有些意义建构过程可能自始至终都需要搜索的参与，而另一些则是先进行一组搜索，然后再进行一系列的分析和综合。那些深层的分析任务需要进行意义构建，同时伴随着搜索，例如法律发现过程、流行病学（疾病跟踪）、通过研究顾客投诉来改善服务并获取商业智能等。 22

2.2.2　信息搜寻的经典模型与动态模型

研究人员已经构造出很多关于人们如何搜索的理论模型。Sutcliffe 和 Ennis[1547] 提出的信息搜寻过程的经典模型将其定义为由 4 个主要活动所构成的周期性过程：

- 明确问题
- 表达信息需求
- 构造查询
- 评价结果

信息搜寻过程的标准模型包含一个潜在的假设，即用户的信息需求是静态的，信息搜寻的过程是一个对于查询项进行连续提炼的过程，直到所有且仅有那些与原始信息有关的文档被检索出来为止。最近的模型强调了搜索过程的动态特性，并指出用户在搜索的同时也在学习，当他们看到检索结果或者其他文档代理时，其信息需求会进行相应的调整。这种动态过程有时称为搜索的采摘模型（berry picking model）[157]。

如今的 Web 搜索引擎的快速响应时间，使得搜索用户能够采用一个较为普遍的查询来“试水”，在看到返回结果后，以显示的文字为基础，重构他们的查询方式，试图更“接近”所需的目标 [158，755，1082]。例如，一个复杂的查询“一个 1/2 英寸的燃气烧烤炉软管连接器，用于 3/8 英寸的家用插座”，这个查询很可能是失败的，典型的搜索用户会选择一个更为普遍的查询，如“燃气炉软管连接器”，甚至“燃气软管”，查看搜索引擎的返回结果，然后重构查询，或者访问相应的网站，在其中浏览网页，寻找所需要的产品。

这样的做法在采摘方法中是常见的策略，有时也称为定向（orienteering）[1219，1569]。进行定向信息搜寻的用户会给出一个快速、不精确的查询，希望近似地得到信息空间的一部分内容，然后再进行一系列的本地导航操作，从而获得更贴近用户兴趣的信息 [158，1082]。可用性研究和 Web 搜索日志表明这种方法是常见的。用户很可能会重构他们的查询，一份对搜索日志的分析说明 52%的用户重构了查询 [820]。

有些信息搜寻模型关注于搜索过程中使用的策略，以及用户在下一个步骤如何做出选择。在某些情况下，这些模型是为了反映专业搜索用户自觉的规划行为。在其他情况下，这些模型是为了捕捉缺少计划性的一般搜索用户的潜在反应。Bates[156] 建议，搜索用户的行为可以被搜索策略所刻画，搜索策略反过来由搜索战术（tactic）的序列所组成。Bates [156] 也讨论了监测当前搜索进度、衡量延续当前策略及改变策略的成本和收益的重要性。

Russell 等人 [1401] 也关注于监测搜索策略的进度，并以*成本结构分析*或者收益递减分析
[23] 作为整个过程的目标或是子目标。这种成本结构分析方法，后来被 Pirolli 和 Card 扩展为*信息搜寻理论*（information foraging theory）[1271，1269]，使用进化生物学立场的理论框架，对人们在信息结构内的导航策略进行了建模与预测。

2.2.3 导航与搜索

并非所有的搜索都开始于在搜索框中输入关键词查询。许多网站和一些搜索引擎允许用户通过仔细阅读某种信息结构（information structure）来选择搜索的起点。*导航*（navigation）和*浏览*（browsing）这两个词在这里可以交换使用，它们表示相同的含义——搜索用户通过一系列浏览和选择操作，对信息结构进行查看，并能够在可用信息的多个视图当中进行切换。当信息结构（如在一个网站上的超链接）非常符合用户的信息需求时，用户往往更喜欢浏览而不是关键词搜索。Hearst 等人的研究 [737] 发现，在多次使用了精心设计的分面分类系统后，自我描述的搜索用户往往会逐渐转变为通过浏览获取信息。

浏览往往是首选，因为*识别*（recognize）出一部分信息要比*召回*（recall）或记住它更为容易。但是，如果花费了过长的时间来寻找感兴趣的标签，或者找不到所需要的信息，那么浏览链接所获得的收益就会递减。也就是说，浏览只有在合适的链接时可用，并对潜在信息具有有意义的提示内容时（有时称为*信息线索* [1269]）才能够有良好的效果。

使用合适的导航结构，某个交互界面可能需要数次点击来引领搜索用户寻找他们的目标，但这并不一定是坏事。Spool [1523] 声称，一般来说，搜索用户对于跟踪多个链接并不十分反感，不过他们反感于跟踪那些与他们的目标无关的链接。因此，只要搜索用户在搜寻目标信息的过程中，没有丢失信息的“线索”，交互界面就算表现良好。Spool 讨论了一个用户要寻找某个特定的激光打印机软件驱动程序的例子。假设用户首先点击“打印机”，然后是“激光打印机”，然后按如下的链接顺序：

惠普激光打印机
惠普激光打印机型号 9750
惠普激光打印机型号 9750 的软件
惠普激光打印机型号 9750 的软件的驱动程序
惠普激光打印机型号 9750 在 Win98 操作系统下的软件驱动程序

这样的交互是可以接受的，因为每次细化对于当前的任务都是有道理的，没有一个地方需要后退来尝试另一种选择：即搜索踪迹永远不会变“冷”（即偏离用户需求）。但如果在中途某个时候，搜索用户通过点击没有看到更接近目标的链接，那么这样的经验就会非常令人沮丧，而交互界面从可用性的角度来说就是失败的。

2.2.4 对搜索过程的观察

人们对于搜索过程的研究已经有很多了，获得的成果可以帮助指导搜索界面的设计。这
[24] 些研究提到的一个共同的观察是用户经常会微调他们的查询，因为这会比第一次就试图给定准确的查询要容易。另一个原因是，搜索用户经常搜索他们先前已访问过的信息 [853，1130]，而在看到以前搜索过的材料之后，用户的搜索策略也会相应地有所不同 [150，853]。研究人员已经开发出了这样的搜索界面，其中特别考虑了搜索用户重新访问信息的可能性 [466，518]，同时支持查询历史和对以前访问过的信息条目进行重新访问。

研究表明，人们难以确定文档是否与主题相关 [451，1402，1687]，而人们对一个主题

了解得越少，就越难判断搜索结果是否与主题相关 [1516，1620]。对于 Web 搜索引擎，搜索用户往往只关注排名靠前的搜索结果，而偏颇地认为排名第一或第二的文档要好于那些排名较低的文档 [663，844]。

研究还表明，人们很难估计在搜索结果中有多少是相关的，他们对一个主题越不了解，也就越有可能自信地认为相关信息都已经访问过了 [1551]。此外，人们往往在找到几个结果后就终止搜索过程，即使文档集中可能还会有更好的结果 [1547]。

有些搜索可用性的研究评估了搜索过程本身的影响，并对专家和新手进行了对比，虽然这种划分形式还没有达成共识的分类标准 [81]。研究指出，专家会使用与新手不同的搜索策略 [771，990，1687]，但也许更说明问题的是，其他研究发现了搜索知识和领域经验之间的交互效应 [771，832]。在一项研究中，能够找到高质量文档的杰出分析师的总体特点是分析的持续性，那些阅读更多文档、花费更多时间的人比其他人完成得更好 [1247]。在另一项研究中，搜索专家比新手更耐心，并有积极的态度，这往往会带来更好的搜索结果 [1551]。

2.3　现今的搜索界面

典型搜索会话的核心过程是由查询描述、搜索结果检查和查询重构组成的。随着搜索过程的进行，搜索用户会更加了解他们想要的主题，以及可用的信息来源。

本节将要介绍几种用户界面的组件，它们已经成为了搜索界面中的标准，并表现出了很高的可用性。在描述这些组件的同时，我们也将介绍它们所支持的设计特点。在理想的情况下，这些组件被集成在一起，以支持搜索进程的不同部分，但分开讨论会更有助于我们对它们的了解。

2.3.1　启动搜寻

信息搜寻的过程是如何开始的？在今天，网络已经在很大程度上取代了传统的物理信息源，如电话簿和百科全书等。对于网上信息系统的用户，开始搜索会话的最常用的方法是访问 Web 浏览器，并使用 Web 搜索引擎。

另一种开始搜索的方法，是从以前访问过的网站收藏中选择一个网站，这些收藏通常存储在浏览器中的书签中。这种方法曾经被大量地使用，然而随着搜索引擎服务变得更快捷也更准确，这种方法就不再那么流行了 [1569]。在其他一些书签系统中，用户将偏爱的网站链接存储在一个网站上（因此从任何连接的计算机都可以访问），其中还可以看到其他人都选择保存了什么网址，这种书签系统已经在一小部分用户中深受欢迎。这些网站（delicious. com 和 furl. net，即现在的 diigo. com，就是这方面的例子）允许用户设定内容的标签（label 或 tag），按主题搜索或浏览，以及按网站标题进行文本搜索。

网站目录曾经也是一个常见的出发点。在较早的时候，Yahoo. com 的目录在当时是最流行的导航起点，但现在网络目录已基本上被搜索引擎所取代，一方面因为网络规模变得太大，没办法手动构造目录，另一方面也因为 Web 搜索的精度不断提高 [1267]。不过，有一些学者认为，搜索用户应该对信息的来源有更多的认识，并认为在搜索结果列表中，这些信息应更加突出地显示 [1355]。如果想了解更多关于网站目录的信息，请见 11.8.2 节。

2.3.2　查询描述

一旦选定搜索起点，用户表达自己信息需求的主要方法就是在搜索框中输入一些词语或

25

者从目录以及其他信息组织中选择链接。对于 Web 搜索引擎来说，查询是通过文本形式指定的。如今这通常是通过在键盘上输入文字的方式来实现的，但在未来，伴随着我们逐渐开始以移动设备作为输入媒体，通过语音命令进行查询的方式有可能会越来越普遍。

在如今的 Web 查询中，输入的文字通常很短，一般由 1～3 个词语组成 [820，819]。多词查询往往视为一个短语，但查询也可能是由多个主题所组成的。短查询反映了标准的使用场景，用户查看搜索引擎返回的结果。如果结果是不相关的，用户会重构他们的查询；如果结果是令人满意的，用户就会定位到最相关的网站，在该网站上继续微调查询 [158，539，755，1082]。这种先用普遍的查询来寻找信息空间中有用的部分，然后跟随相关网站超链接的搜索行为，是 Web 搜索当中应用定向策略的一个示范 [1219，1569]。有证据表明，在许多情况下，用户虽然倾向于更详尽地表示他们的信息需求，但过去的搜索引擎使用经验告诉他们，这种方法不能很好地工作，而关键词查询与定向相结合会表现得更好 [201，1288]。

在 Web 搜索出现之前，商业文本搜索系统通常支持布尔运算和基于命令的语法，而实际上并没有支持关键词查询。但是，布尔运算符和命令行语法已经被一再地证实难以让大多数用户理解，试图使用它们的人经常会犯一些错误 [499，672，699，755，763]。

虽然大多数 Web 搜索引擎支持一些布尔形式的语法，但最近一项针对 Web 查询日志的
26 研究表明，在超过 150 万的查询中，仅发现 2.1%含有布尔运算，7.6%含有其他查询语法，主要是双引号短语 [819]。另一项研究考察了近 60 万用户在 2006 年期间，总计时间超过 13 周、数百万的交互日志。他们发现，1.1%的查询包含 4 个主要的 Web 运算符（双引号、+、－和 site:）中至少一个运算符，只有 8.7%的用户始终使用运算符 [1685]。7.2.1 节将介绍更多关于 Web 查询的内容。

Web 排序已经经历了 3 个主要阶段。第一阶段大约从 1994—2000 年，大多数的搜索引擎使用统计排序，但是没有使用网页内查询项的（位置）邻近信息和网页相对重要性的信息。那时，整个 Web 的规模还比较小，不太可能有相关的信息源为那些较为复杂的查询提供答案。并且有可能会检索出那些缺失查询中关键词的网页，许多用户无法理解这样的行为方式。（例如，AltaVista 引入了强制运算符，用加号表示，即允许用户可以在一个词前增添一个加号，表示这个词必须出现在查询中，但只有那些极富经验的用户才会利用这种查询运算符。）

在 1997 年左右，谷歌转向了只采用合取查询的方式，这意味着只有所有查询项都出现在网页中时，网页才会被检索到。他们还增加了查询项的邻近信息和网页的重要性打分（见 11.5.2 节的 PageRank 算法），这大大提高了许多查询的相关性，特别是导航查询；比如，以“丰田”（Toyota）作为查询，会检索到丰田公司的主页，而不是那些“丰田”出现次数最多的页面。其他的 Web 搜索引擎也紧跟着这种趋势，合取排序成了常态。

随着网络上可用信息数量的增加，老练的搜索用户发现，把较长的查询看做短语往往会找到高度相关的结果。过去，如果搜索用户有复杂的信息需求，并试图充分地表达给 Web 搜索引擎时，这样的尝试往往都会失败。例如，如果一个搜索用户想知道“我在哪里可以找到 1985 年的卡罗拉的轮毂?”以这种形式编写的查询由于合取约束，将无法返回任何结果。现在，Web 搜索引擎已经变得越来越精细，能够去掉一些无意义的项，而只匹配重要的查询项，在排名较高的文档中，这些查询项彼此相邻。另外，可以使用其他在 Web 搜索中已经证明有效的方法进行排序。有关查询语言的更多细节见 7.1 节。

2.3.3 查询描述界面

文本查询的标准界面是一个搜索框，用户输入查询时，通过按键盘上的回车键或点击与

表单相关的按钮进行查询。研究表明，查询长度与输入框宽度之间有一定的关系；小的输入框会阻碍长查询，而宽形式的输入框则会鼓励长查询 [171，585]。

有些输入框被分为多个组件，允许用户更自由地输入查询文本，并跟随着一些查询过滤的输入框。例如，在 yelp. com 上，用户在第一个输入框中输入一个普遍的查询，通过在第二个输入框中输入位置信息，对搜索进行改进（见图 2-1）。表单允许选择以前用到的信息，这些信息有时是结构化的，并允许设置为未来使用的参数。例如，yelp. com 的表单会显示用户的本地 27
位置（如果过去曾经指定过）以及其他近期指定过的位置，并可以选择添加额外的位置。

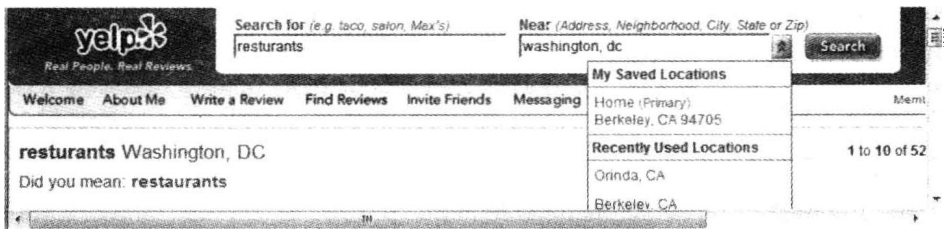


图 2-1　yelp. com 的查询表单，显示了对结构化查询以及存储之前查询信息的支持

一个在搜索框中使用得越来越普遍的策略是，通过灰色文字来暗示什么类型的信息应该输入到搜索框中。例如，在 zvents. com 搜索中（见图 2-2），第一个搜索框上标有“你要买什么?”，而第二个框标有“什么时候（今晚、本周末、……）”。当用户将光标放在搜索输入框上时，灰色的文字消失，用户可以输入自己的查询项。

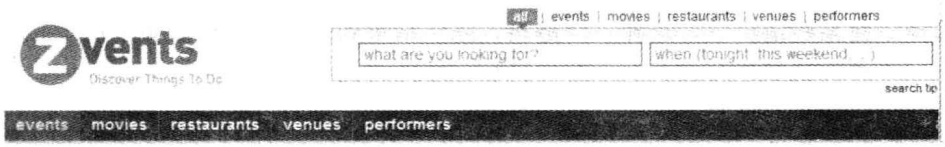


图 2-2　zvents. com 的查询表单，表单中的灰色文字说明什么类型的信息可以被输入

这个例子也说明了现在的搜索引擎支持专用的输入方式。例如，网站 zvents. com 会识别诸如“明天”一类对时间敏感的查询词，并以事先设定好的方式来进行处理。它还能够更加灵活地处理更正式的日期格式，因而搜索“星期三”（wed）的“喜剧”（comedy）时会自动计算最近的星期三的具体日期。这是一个很好的例子，说明我们应该通过设计界面来反映人们是如何思考的，而不是要求用户遵循那些不可靠且流于形式的标准程序来思考。（这种放宽查询描述的方法，更适用于那些“非正式的”（casual）界面，在这些系统当中，日期并不是最关键的要素。非正式的日期格式在填写税表时是不能接受的，因为发生错误的代价太大了。）

一个已经显著改善了查询描述的创新是动态生成的查询建议列表，当用户输入查询时，表单实时显示查询建议 [1684]。这种方法也称为*自动填充*（auto-complete）、*自动建议*（auto-suggest），或*动态查询建议*（dynamic query suggestion）。通过对大规模的日志进行研究发现，用户在大约 1/3 的时间里，点击了雅虎搜索助手提供的动态查询建议 [61]。这一主题将在 11. 7. 2 节介绍 Web 搜索引擎时进行详细解释。 28

通常显示的查询建议是那些前缀字符与之前输入的字符匹配的词语，但在某些情况下，显示的是只有中间字符匹配的词语。如果用户输入多个词的查询，那么显示的查询建议可能是之前输入内容的同义词，但在词法上并不匹配。举例来说，Netflix. com 用灰色字体显示可能需要的词，然后通过一个下拉列表框显示可以点击的词语。

在动态查询建议界面中，匹配的显示也有着不同的方式。有些界面根据类别信息对建议进行着色。在大多数情况下，用户必须移动鼠标到所需的查询建议上以选择它并用来填充查询框。在某些情况下，查询可以立即进行；在另一些情况下，用户必须输入回车键或点击“搜索”按钮才能进行查询。

查询建议可能来自多种资源。在某些情况下，列表是根据用户自己的查询历史获得的，在其他情况下，它基于其他用户的热门查询。这个列表也可以来自于网站设计人员认为重要的一组元数据，例如在药理文献搜索时显示的一组已知疾病或基因的名字（见图 2-3），在电子商务网站搜索时显示的产品列表，或者在电影网站上搜索时显示的热门电影列表。这些建议也可以来自网站内部的所有文本。

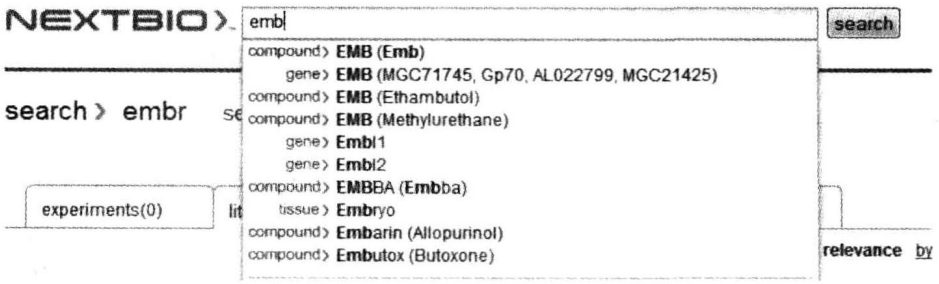


图 2-3 按类型分组的动态查询建议，源自 NextBio. com

查询描述的另一个形式包括从一些信息显示中进行选择，最典型的是在超链接或保存的书签中选择。在某些情况下，选择一个链接，除了结果列表，还会产生更多的链接来进行下一步导航。这种查询描述会在 2. 3. 6 节中进行详细的讨论。

2. 3. 4 检索结果显示

当显示搜索结果时，或者是显示全部文档，或者将文档的一些有代表性的内容提交给搜索用户。这种文档代理（surrogate）指的是文档的摘要，这是一个成功的搜索界面的重要组成部分。而文档代理的设计和检索结果显示是目前研究和实验比较活跃的领域。

文档代理的质量极大地影响对搜索结果列表相关性的感知。在 Web 搜索中，页面标题
29 通常与 URL 一起加亮显示，有时也会与其他元数据一起显示。在对信息集合进行搜索时，出版日期和作者等元数据往往会显示（但这类元数据较少应用于网页）。文本摘要（summary）（也称为摘要（abstract）、提取（extract）、摘录（excerpt），或片段（snippet））包含了从文档中提取的文本，它们对检索结果的评估是至关重要的。

一项研究评价了搜索结果中的哪个属性会获得更多的点击，并从中找到了许多能带来正面效果的因素，包括更长的文本摘要、包含查询关键词的标题、标题组合、包含作为短语匹配的查询的摘要和网址（URL）、更短的 URL，以及域名中包含查询项的 URL [390]。

目前，标准的结果显示是一个文本摘要的垂直列表，有时也称为搜索引擎结果页（Search Engine Results Page，SERP）。在某些情况下，摘要是对包含查询项的文档的摘录。在其他情况下，通过混合（blended）结果（也称为全能搜索，universal search）技术，有些特殊的元数据与标准的文本结果一起显示给用户。例如，以“彩虹”作为查询，返回的搜索结果可能包含一行彩虹的示例图像（见图 2-4），或者查询运动队的名称可能检索出最近的比赛得分和一个购买门票或浏览比赛直播时间表的链接（见图 2-5）。Nielsen [1206] 指出，

在某些情况下，搜索结果列表可以直接满足信息需求，从而使搜索引擎变成“答案引擎”。

图 2-4　雅虎搜索中对于查询“彩虹”的搜索结果页面。结果从上到下分别包括：查询改善建议，彩虹图片的链接，关于彩虹的百科文章和一些说明的图片，以及名为“彩虹”的摇滚乐团的百科文章

图 2-5　谷歌搜索引擎对于查询“火箭”的搜索结果界面。根据查询词的不同意义，显示了不同种类的信息。第一个是篮球队主页的链接，以及一些该网站内部的“深度链接”（deep links）。接下来，是与球队有关的其他链接，关于火箭的百科页面，以及火箭队的视频链接和如何制造火箭的链接

对排序的研究发现，邻近性信息可以很有效地提高搜索的精度 [391，733，1563]。可
30 用性研究建议，将查询项出现在文档时的上下文显示出来，可以有助于用户评估结果的相关性 [1593，1682]。这有时称为*上下文关键字*（Keywords in Context，KWIC）、*查询偏置摘要*（query-biased summary）、*面向查询的摘要*（query-oriented summary），或*用户主导摘要*（user-directed summary）。

查询项的加亮显示（highlighting）可以在视觉上改善搜索结果列表的可用性，这个观点已经有几十年了 [974，1005，1063，1082]。加亮是指在视觉把某个部分和其他部分进行区别，可以通过粗体文字、改变文字或背景的颜色、改变文字的大小，以及其他方法来实现。加亮显示既可以用于搜索结果中的文档代理，也可以用于检索出的文档本身。有些界面通过可视化的方法给文档中加亮显示的部分做一个概述 [160，305，661，766]。

确定哪些文字可以用做摘要和多少文字应当显示，是一个具有挑战性的问题。通常，最相关的文章是那些包含所有查询项，并且查询项是彼此紧密相连的，但对于不那么匹配的结
31 果，在显示连续的句子以增加结果的连贯性与显示包含查询项的句子中，要进行一个权衡与取舍。有些研究结果表明，完整地显示整个句子比将句子切分开有更好的效果 [80，1380，1683]，但在另一方面，很长的句子通常也不是我们想要的结果。

还有证据表明，在搜索结果摘要中显示的这类信息应根据查询意图和搜索会话目标的不同而相应地变化。有些研究显示，在某些信息需求下较长的答案会比较短的答案表现得更好 [861，1034，1237]。而当搜索用户决定直接进入到一个知名网站的主页时，简短的结果列表会比长的详细信息更好。在一般情况下，用户对*已知项进行搜索*时往往倾向于可以指示所需信息的较短的代理。主页搜索本质上是对地址的搜索；用户知道网站的名字，希望找到它的网址（URL）。同样，能够简要说明的事实性信息需求可以被简短的结果满足。相反地，如果用户有一个复杂的信息需求，更深层次的文档摘要可以带来更好的搜索体验。这一点对于那些更丰富的任务来说也是正确的，如建议搜寻或获取相似的主题。

其他种类的文档信息可以有效地显示在搜索结果页面中。图 2-5 和图 2-7 显示了*站内链接*（sitelink）和*深度链接*（deep link）的应用，它们在网站主页的下方显示了网站内部较受欢迎的网页。在另一个例子中，生物科学文献检索研究发现，大多数参与者强烈主张在搜索结果的旁边显示从期刊文章中提取的图片 [736]。在图 2-6 中，对 BioText 系统的截图显示了这种思想，也说明了加亮或粗体显示查询项的作用，以及用户可以或多或少看到查询项的上下文环境的机制。

2.3.5 查询重构

在指定了查询和产生了结果之后，有一些工具可以用来帮助用户重构他们的查询，或将信息搜寻过程引领到一个新的方向。对搜索引擎日志的分析表明，查询重构是一种常见的活动；一项研究发现，在一次会话期间，超过 50％的搜索用户至少进行了一次查询修改，有接近 1/3 的人进行了 3 次或更多次查询修改 [820]。

在最重要的查询重构技术中，有一种是显示与查询或检出的文档相关的索引项。其中的一个特殊情况是拼写校对或建议；据估计，10％～15％的查询会出现错字 [461]。在 Web 搜索出现之前，拼写建议主要是基于字典的 [944]。在 Web 搜索出现后，查询日志已应用于开发检测和纠正拼写错误的高精度算法中 [461，1018]。在搜索界面中，通常只有一个更改建议显示；点击更改建议就可以重新执行查询。多年以前，搜索结果中会显示那些据推测不正确的拼写，今天一些搜索引擎已经可以交错显示原始查询的结果和拼写校对后的结果，或将原始查询结果与拼写校对后的结果分别显示。

除了拼写建议外，搜索界面越来越多地采用相关项建议技术，通常称为查询项扩展(term expansion)。对日志所进行的研究发现，如果能提出较好的查询建议，那么在 Web 搜索中它会是一个频繁使用的功能。对日志的研究发现，大约 8%的查询都是由查询项建议产 32
生的 [819]（但它没有显示有多少比例的查询会显示这样的建议），而另一个发现是大约 6%的用户选择点击查询项建议 [61]。

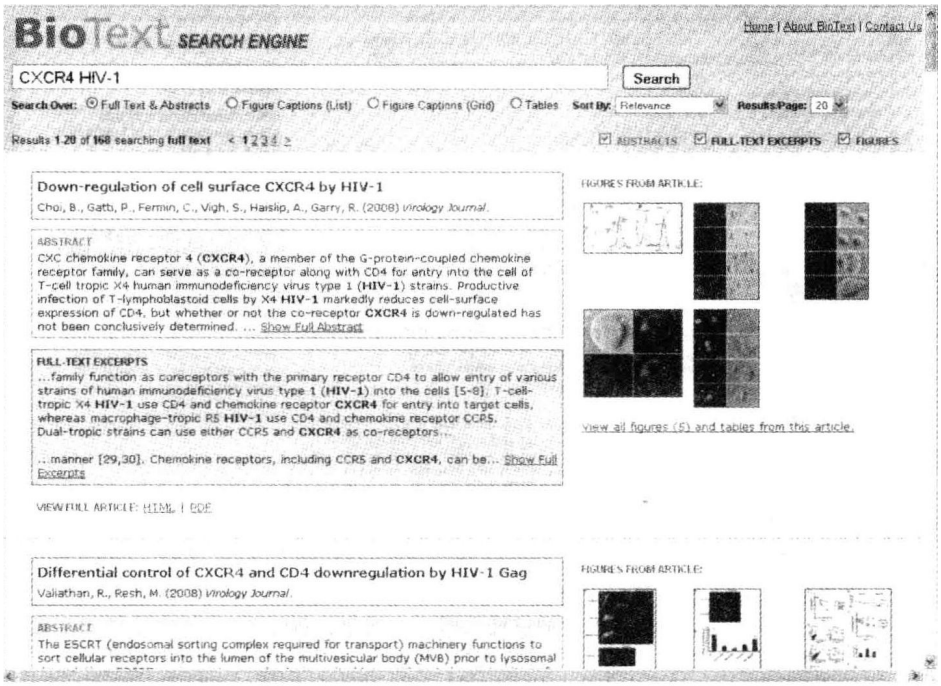


图 2-6　BioText 系统的搜索结果，其中显示了丰富的文档代理信息，包括文章中所抽取的图片、加亮或黑体显示的查询项，以及扩展或缩短文档摘要的选项，来自 http://biosearch.berkeley.edu

较早使用查询项扩展或建议的工作建立在十几个同义词库上，通常是在显示搜索结果前强迫用户从中选择 [285，492]。最近的研究表明，提供较少的建议，只需要用户进行一次点击，或将相关的查询项组合起来进行一次点击选择，是一种更为可取的方法 [59，501，1681]。图 2-4 和图 2-7 显示了只需一次点击的查询项扩展的例子。

有些查询项建议是基于某一特定用户的整个搜索会话，然而有的则是基于之前提出相同或相似查询的其他用户的行为。一种策略是显示其他用户提供的类似查询；另一种方法是从之前提出相同查询的用户所点击的文档中提取查询项。在某些情况下，相同的算法被用做实时查询自动建议。

相关反馈是另一种方法，其目标是帮助我们进行查询重构，将在第 5 章详细讨论它。其主要思想是让用户指出，对于查询哪些文档是相关的（也可以是不相关的）。在另一些搜索 33
系统中，也可能让用户指出从文章中抽取的哪些索引项是相关的 [918]。系统通过这个信息，可以计算出一个新的查询，并使用某种算法，显示一个新的检索集合 [1402]。

相关反馈已被证明在非交互式或人工设置情况下，都可以大大改进排名顺序 [31，895]。然而，这种方法从可用性角度并不认为是成功的，也没有出现在标准的用户界面中 [451，1402]。这源于几个因素：用户不善于评价特定文档的相关性，特别是对他们不熟悉的主题 [1620，1687]；另外，相关反馈的益处是不一致的，这在可用性方面是有问题的；此外，相关反馈的优势大多体现在需要大量相关文档的任务中，但这在 Web 搜索中并不常

见；事实上有一些证据表明，相关反馈的优势在搜索整个 Web 时就会消失了 [1570]（大多数相关反馈的优势只是在应用于小规模文档集时才可以显现出来）。

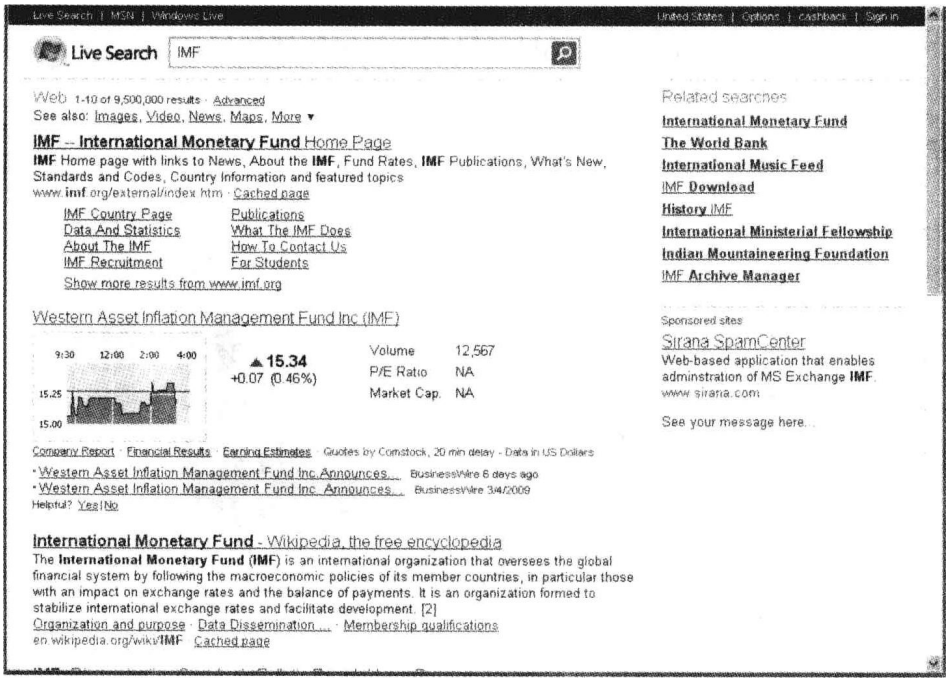


图 2-7 微软的 Live 搜索引擎对于查询“IMF”的结果页面，包括相关的查询建议（在右侧），可选择的“垂直”搜索的链接（图片、视频、新闻等），站内链接如金融统计和一些百科文章

关于相关反馈的变体（自动计算“相关文档”功能）已证明在某些情况下有着积极的影响。在生物医学文献检索系统 PubMed 中，在给定期刊文章的旁边显示一部分相关的文章，
34 这项功能在生物学家中是广受欢迎的。一项研究表明，在显示相关文档的会话中，18.5%的情况下，用户会点击所建议的文档 [1033]。

2.3.6 组织搜索结果

搜索人员通常表示他们希望将搜索结果分成若干有意义的组，以方便理解搜索结果和决定下一步如何做。在一项提取搜索结果分组的纵向研究表明，在应用分组机制的情况下，用户的搜索习惯发生了改变 [862]。现在，有两种搜索结果分组的常用方法：*分类系统*（category system），特别是*分面分类*（faceted category）和*聚类*（clustering）。在本节中，对这两种方法进行详细介绍，对它们的可用性进行比较。

分类系统将一组有意义的标签组织在一起来反映某个领域的相关概念。它们通常是手动构造的，尽管为文档自动设定类别已经可以达到一定的准确率了 [1446]（见第 8 章）。好的分类系统有连贯和（相对）完整的特点，它们的结构也是可预测的，在同一个信息集合内的搜索结果是一致的。

在用于组织搜索结果和表达信息集合结构的分类结构中，最常见的是*扁平的*（flat）、*层次的*（hierarchical）和*分面的*（faceted）分类结构。扁平分类是话题或对象的一个列表。它们可以用做分组、过滤（缩小），或者对搜索界面中的文档集进行排序。大多数网站将信息分类组织，选择相应的类别可以缩小显示的信息集合。在一些实验中，Web 搜索引擎自动

地按扁平分类组织信息；研究表明用户对这种设计给予了积极的回应［519，945］。在庞大的 Web 内容中找出适用的类别子集是非常困难的，相应地，分类系统对于内容更为集中的信息集合似乎有更好的效果。

在线层次组织（hierarchical organization）在桌面文件系统浏览器中是最常见的。在早期的 Web 中，像雅虎所使用的那种层次化目录系统能够将流行的网站组织成可浏览的结构。然而，当信息的集合变得很大，而且结构之间存在相互的链接时，保持严格的层次化结构就会变得很困难。另外，Web 的大小远远超过了在这个系统中可管理的浏览内容，而搜索引擎的应用极大地替代了目录结构的浏览。

层次化对于目录形式的结果会非常有效，如一本书或较小的文档集。Superbook 系统［527，528，974］是一个早期的搜索界面，它用大规模文档的结构来显示查询项命中的情况。在用户指定对一本书的查询后，搜索结果会显示在层次目录中（见图 2-8）。当用户从目录视图中选择一页时，页面会自动显示在右侧，此页内的查询项会被加亮反转显示。最近，有些科研项目应用这个思想来组织企业网［363，1173，1711］。

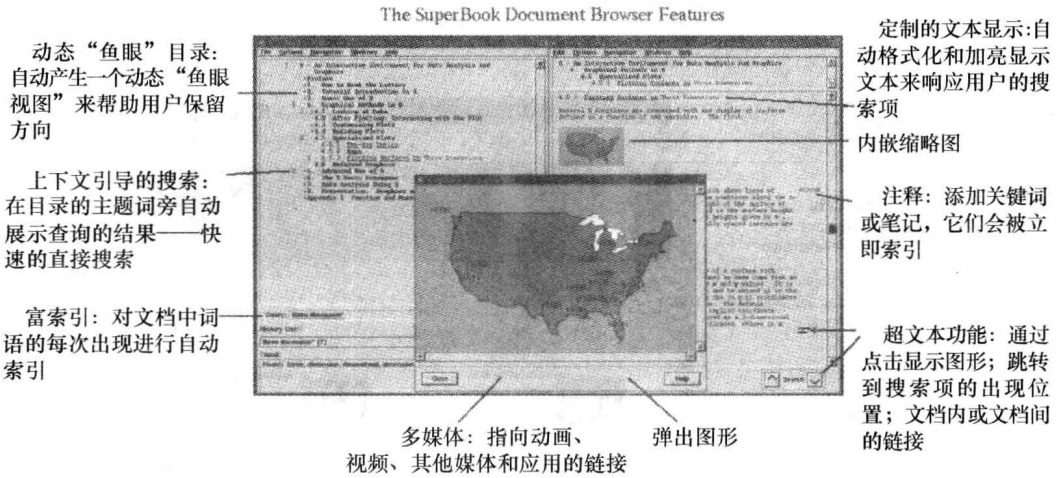


图 2-8　SuperBook 的界面，通过应用一个从大规模手册中制定的目录，在上下文中显示了检索结果［974］

越来越多的人认识到，严格的层次化分类组织对于信息结果的导航并不是理想的选择。层次结构强迫用户从一个特定类别开始，而大多数信息项可能会有许多不同类别的属性。层次化还经常假设信息项仅仅被放置在分类系统的一个地方，然而，计算机界面要比图书馆书架更为灵活。 35

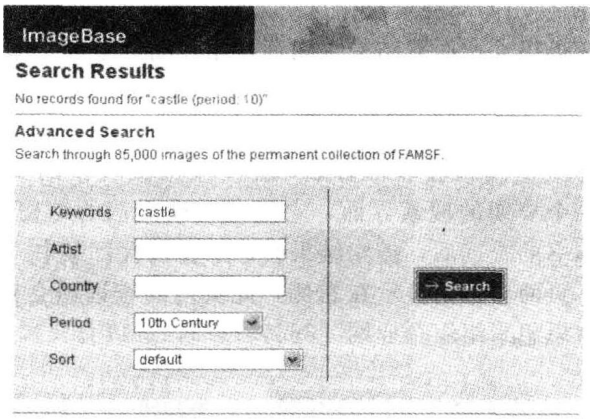
一种叫做分面元数据的表示方法，已经成为组织网站内容和搜索结果的主要方式。分面元数据是扁平分类和完全知识表示在复杂度上的折中，如果设计得合理，就很容易被用户所理解，相比于其他组织形式也更受青睐。不同于建立大规模的分类层次结构，分面元数据由分类集合所组成（扁平的或层次的），每一个都对应于和需要导航的文档集相关的不同分面（维度或特征类别）。在设计好分面之后，文档集中的每一项都被赋予分面中的若干个标签。

应用了层次化分面导航的界面会同时显示接下来要去的网页的预览和如何在浏览中返回前一个状态，同时将类别结构中的文本搜索无缝地结合进来。从而，用户所需的思考就减少了，因为提高了识别的召回率，同时又保证用户在每次操作时都给出逻辑合理但不常见的选择，最终还保证不会有空的结果集。这种方法提供了组织搜索结果和随后查询内容的一种方案，它可以作为探索和发现过程的重要结构。

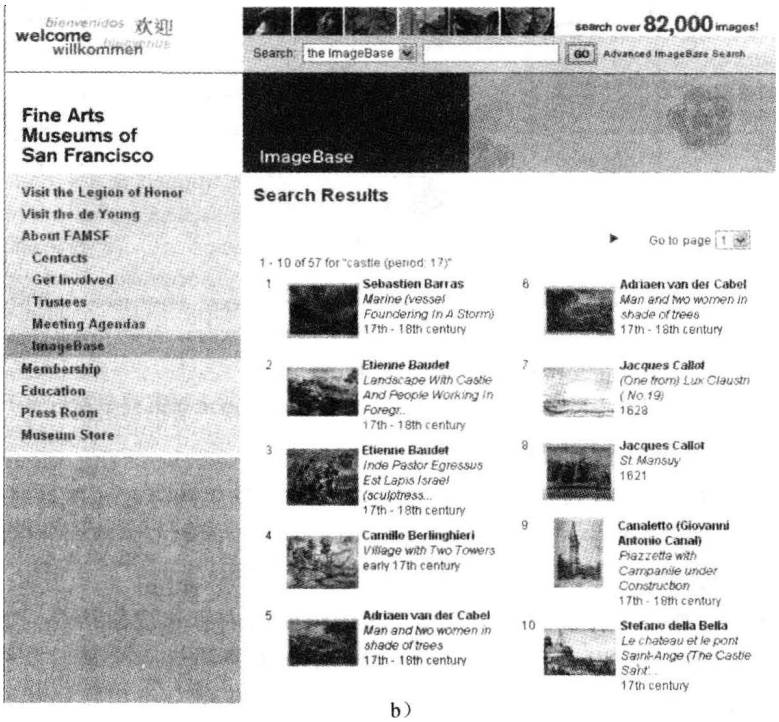
图 2-9a、b 显示了在一个假想的搜索会话中，一个典型图片搜索界面的搜索结果。用户

在“Advanced Search”（高级搜索）框中键入关键词“castle”（城堡），并试图选择“10th Century”（10世纪），但系统返回了错误信息，表示没有记录被找到。在一些试验和错误后，用户发现对于“17th Century”（17世纪）的搜索可以返回结果，并且该结果以固定的顺序显示，不允许组织和探索。

36



a)



b)

图2-9 在旧金山美术馆图片集上的典型图片搜索界面。a）用户在典型的“Advanced Search”（高级搜索）框中的两个字段输入查询后出现的错误信息。这种输入框的常见问题是产生空的答案集。b）通过“Advanced Search”（高级搜索）框搜索关键词“castle”和时间区间“17th Century”的标准搜索结果列表

图2-10显示通过应用分面导航的Flamenco系统，同种类的信息可以获得更好的可理解性[737，1746]。用户最初键入查询关键词“castle”（城堡），搜索结果显示了229个图片，左侧结果可以允许用户通过Media（媒体）类型、Location（地点）、Object（对象）（图片内可见）、Building（建筑物）类型（“castle”（城堡）是其中一种），或Author（作者）等信息来组织答

案的结构。由于用户能够选择超链接，并且查询预览［1281］可以显示在链接被点击后可以看到多少结果，因此空的答案列表不再是一个严重的问题。在这个例子中，用户首先选择“Media＞Prints”（媒体＞印刷品），然后再由“Location＞Europe”（地点＞欧洲）对结果进行组织，然后通过选择“Print”（印刷品）下面的分支重新进行组织。左侧的层次化分面元数据显示了剩下的 197 张图片属于哪个欧洲国家，以及出现的次数。选择一张图片会显示与其相关的元数据，并附有相关概念的链接，如“ruins”（废墟）和“hill”（小山）。图 2-11 显示了如何将类似的想法应用于数字图书馆目录，图 2-12 显示了它应用于黄页和网站预览的情形。

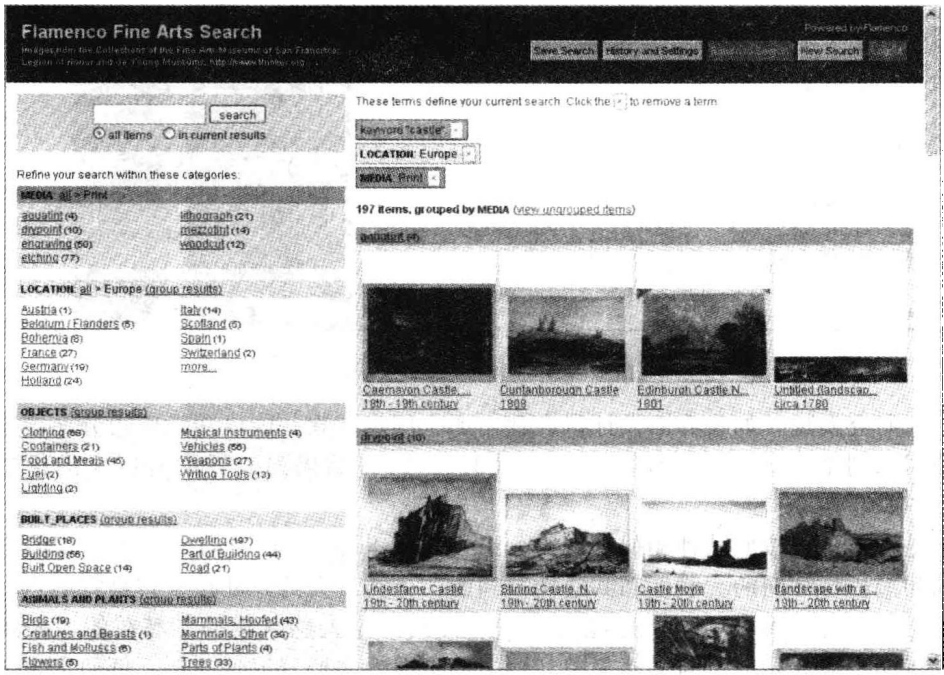


图 2-10　Flamenco 界面的分面导航，应用在旧金山美术馆图片集的一个子集上

37

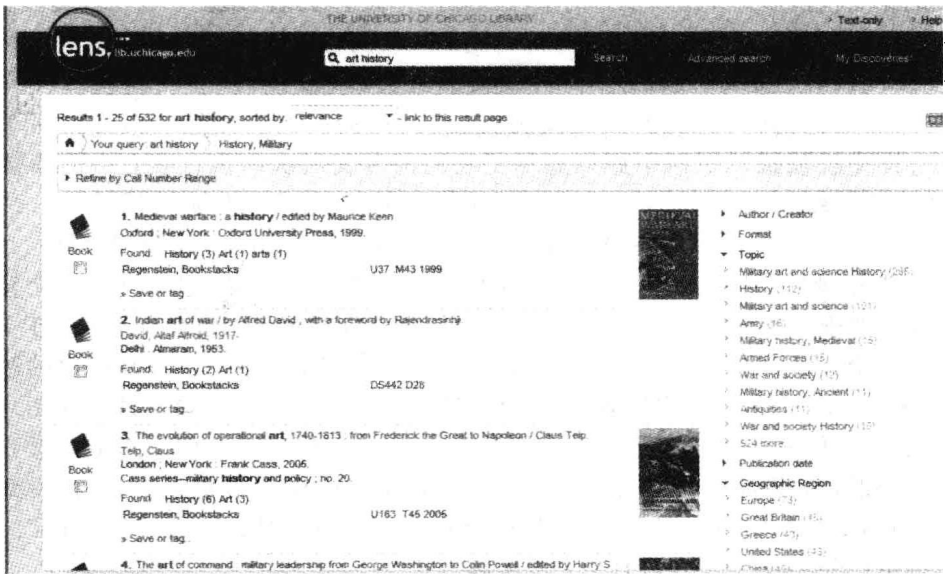


图 2-11　芝加哥大学数字图书馆的分面导航，来自 AquaBrower

图 2-12 yelp. com 的分面导航

可用性研究发现，如果界面设计得合理，那么用户喜欢并能成功地应用分面导航。在用做集合搜索和浏览时，分面界面相对于标准的关键词-结果列表界面有巨大的优势［737，
38 1746］。

聚类是指将条目按照某种相似度进行分组（见第 8 章）。在文档聚类中，相似度一般通过计算词和短语特征间的关联性和共通性得到的。聚类最大的好处在于它是完全自动的，很容易应用于任何的文本集。聚类还能反映一组文档中令人感兴趣的、潜在的、未知的新趋势，并能将那些彼此相似但与文档集中其他文档不同的文档分组到一起，例如出现在主要语种为英语的文档集中所有用日语写的文档。

聚类的缺点包括形式和结果质量的不可预测性、标记分组的难度，以及聚类层次化的反直觉性。有些算法［862，1764］在占主导作用的短语间建立簇（cluster），来构造可理解的标签（见图 2-13），但是每个簇的内容不一定是彼此连贯的。

图 2-14 显示了 Vivisimo 的 Clusty 系统在查询“senate”（参议院）时，搜索引擎结果的聚类输出。图中扩展显示了两个簇以表示了它们的分支层次结构。最上层的簇被标记为“Biography，Constituent Services”（传记，选举服务），其子簇分别被标记为：“Photos”（图片）、“Issues/news”（出版物/新闻）、“Visiting Washington”（访问华盛顿）、“Voting record”（投票记录）、“Virginia”（弗吉尼亚州）和“Maine”（缅因州）等。每个簇代表什么并不是非常明确；如果它代表的是美国参议院，那么在其他簇中也会有很多关于美国参议院的页面。无论如何，最上层的标签并不能代表具体的信息。下一个最高层次的簇标签为“Senate Committee”（参议院委员会），选择它后则会显示相应的组成文件（在图片右侧），
39 从美国参议院主页（关注的不是其下属的多个委员会）到某些具体的美国参议院委员会的网页，再到堪萨斯州和柬埔寨的页面。第三个主要的簇“Votes”（投票），也扩展到一些如“Constituent Services”（选举服务）、“Obama Budget”（奥巴马预算）、“Expand”（扩张），以及“Senate Calendar”（参议院日历）的子簇。

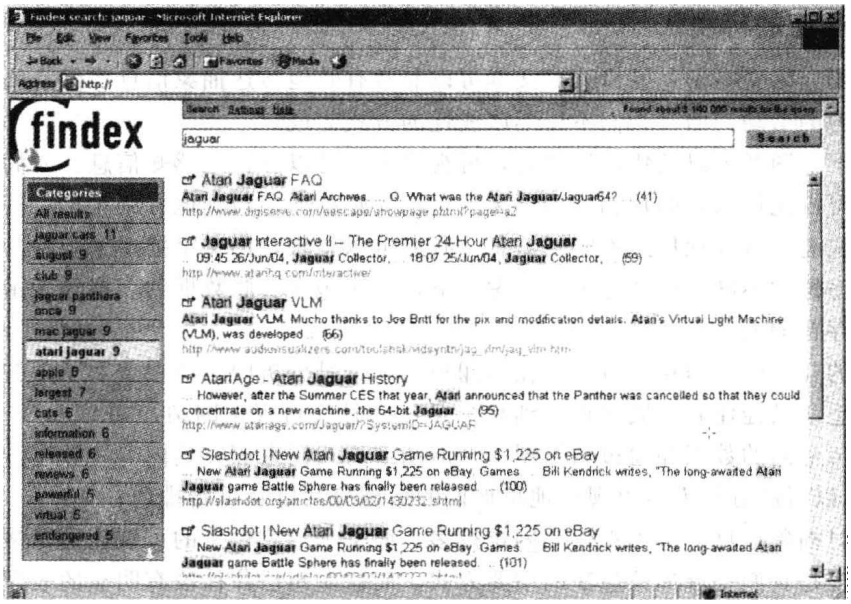


图 2-13　应用 Findex 聚类［862］产生的输出，来自 FindEx. com Inc. © 2010 和许可方

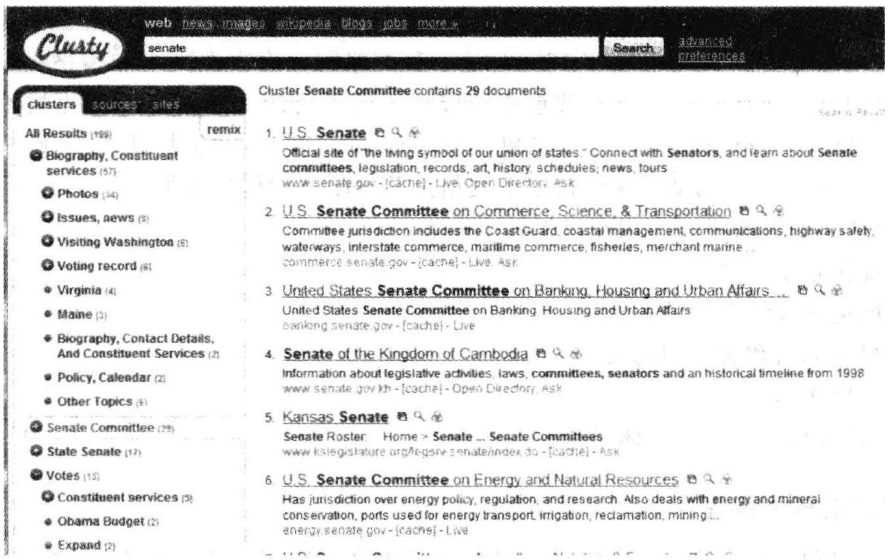


图 2-14　查询“senate”的聚类输出，来自 Clusty. com

话题的混合性和分组之间的重叠，对文档聚类是很典型的。可用性结果表明，用户不喜欢聚类产生的无规律的分组，而是更喜欢可理解的、并通过统一的层次颗粒度来表示的层次化结构［1301，1376］。

分面界面相对于聚类的一个缺陷是感兴趣的类别必须是预先知道的，所以数据中一些重要的新趋势可能不会注意到。尽管建立分面层次化结构的尝试正在不断推进之中［1536］，但到目前为止，最大的缺陷是在大多数情况下，类别的层次化结构是手动建立的，而自动建立类别只取得了部分的成功。

2.4 搜索界面的可视化

本书主要介绍文本信息的检索。文本可以非常有效地传达抽象信息，但是阅读甚至浏览文本都是费力的活动，而且人们不得不以线性的方式完成。

40 ~ 41

相比之下，图像可以被快速地浏览，可视化系统可以并行地感知信息。人们可以很好地理解图像和可视化信息，图片和图形更迷人也更有感染力。与其他方法相比，信息的可视化表示可以更快速和有效地传达不同的信息。我们可以想象，用文字描述一张脸与显示一张脸的图片会有怎样的不同，或者还可以考虑一下，一张包含关联数据的表格与表示相同信息的散点图之间存在的差异。

在过去的几年中，信息的可视化在新闻报道和金融分析中已经很普及了，关于信息可视化的创新性想法已经蓬勃发展并扩展到整个 Web。社交可视化网站，如 ManyEyes [1637]，允许用户上传他们的数据并通过条形图、气泡图或折线图来探究其内容，数据分析工具，如 Tableau，可以帮助分析人员可视化地将他们的数据分片以及重新排列。

然而，对抽象信息进行可视化要困难得多，而文本形式信息的可视化更是格外有挑战性的任务。语言是我们交流抽象想法的主要方式，而这些想法往往没有明显的表现形式。词语和概念没有内在的顺序，这使得词语很难通过坐标来画出有意义的图。

尽管有这些困难，但研究人员还是试图通过信息可视化技术来表示信息获取过程的各个方面。除了应用图标和颜色的加亮显示之外，他们还常常使用线条、圆圈和画布形式的空间布局来作为信息视图。Sparklines [1606] 是一种缩略图，内嵌在文本和表格里显示。对于可视化抽象信息，交互性似乎是非常重要的属性；最主要的交互信息可视化技术包括平移（panning）和缩放（zooming）、变形视图（distortion-based）（包括焦点加上下文（focus plus context)），以及使用动画来保持上下文信息并使闭塞的信息可见。

搜索可视化的实验主要应用在以下各个方面：

- 可视化布尔语法
- 可视化查询结果中的查询项
- 可视化词语和文档间的关系
- 可视化文本挖掘

接下来将对每一项进行具体讨论。

2.4.1 可视化布尔语法

正如上文所提到的那样，布尔查询的语法对于大多数用户来说是有困难的，也很少应用于 Web 搜索中。很多年来，研究人员已经实验了如何通过可视化布尔查询来使查询更容易地为人所理解。一个常用的方法是可视化显示韦恩图（Venn diagram)；Hertzum 和 Frokjaer [755] 发现简单的韦恩图表示可以获得比布尔语法更为准确的结果。这种想法的一个更为灵活的版本可以在 VQuery 系统上看到 [851]（见图 2-15)。每一个查询项用一个圆圈或椭圆表示，圆圈间的交集代表查询项之间的 AND 运算（逻辑合取)。VQuery 通过画布活动区域内的圆圈集合表示逻辑析取，通过活动区域内对圆圈的取消选定来表示逻辑非。

42

布尔查询的一个问题是，它们很容易最终产生空结果或者太多的结果。为了纠正这个问题，过滤流可视化允许用户为查询设计不同的分量，然后通过图形流的方式显示在应用每个

运算符后有多少搜索结果 [1755]。布尔查询的其他可视化表示包括垂直和水平的排列区块 [60]，以及将查询项的分量表示为重叠的“魔术”镜头 [566]。

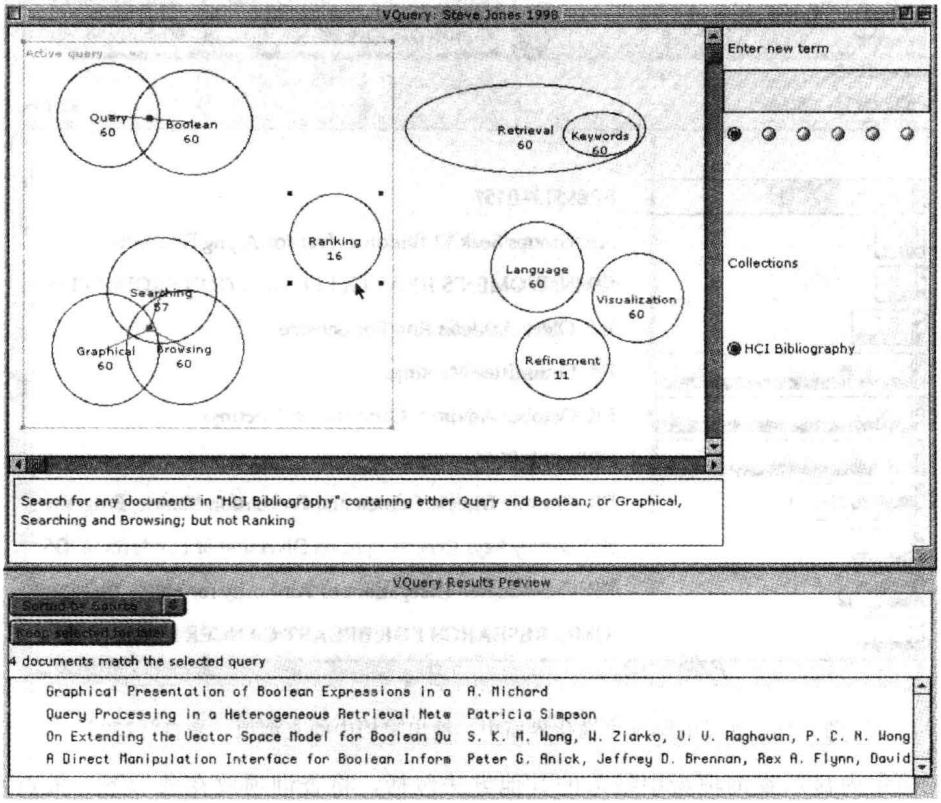


图 2-15　VQuery [851] 为布尔查询制定的韦恩图界面

2.4.2　可视化查询结果中的查询项

如上文所讨论的那样，理解查询项在检出文档中所扮演的角色有助于对相关性的评估。在标准的搜索结果列表中，通常会选择那些包含查询项的句子作为摘要，而这些查询项出现在标题、摘要和网址时会被加亮或黑体显示。从可用性方面看，这种加亮显示的方法已经证明是有效的。

人们已经设计出了许多实验性的可视化界面来明确这种关系。最有名的是 TileBars 界面 [732]，其中文档用水平布局图（horizontal glyph）显示出来，命中的查询项在布局图中的相应位置标出（见图 2-16）。它鼓励用户将查询项拆分成不同的分面，每一行有一个概念，每一个文档表示内的水平行说明了每一个主题下查询项出现的频度。较长的文档被分为子话题的片段，其方法或者通过段落或章节分割标记，或者通过一项叫做 TextTiling [731] 的自动段落划分技术。颜色的灰度表示了查询项出现的频度。可视化显示表明了不同的查询主题在文档中重叠的情况。 [43]

人们还设计出了一些其他的 TileBars 显示形式，例如为每个查询项显示一个正方形的简化版，在这个版本中，使用了颜色分层来表示查询项的频度 [770]。两个更为精简的版本在文档图形的按钮中用灰色显示了查询的命中结果 [741]，或者在饼图中用彩色显示命中结果 [51]，但这些视图并不能显示查询项的位置重叠情况。

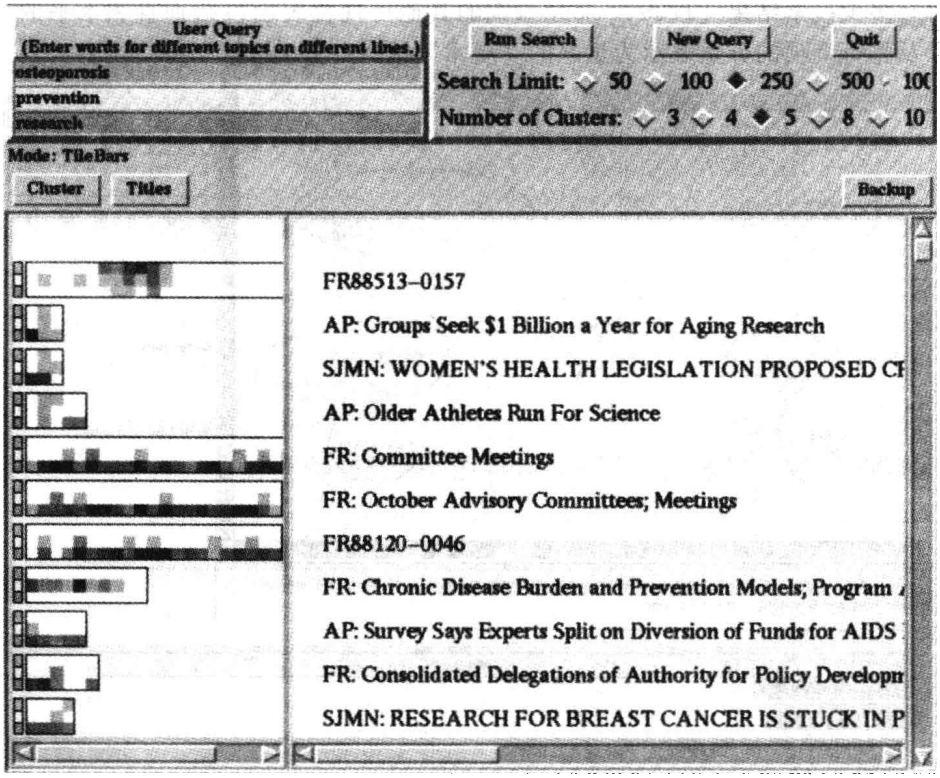


图 2-16 在 TileBars 可视化界面中，检出文档中的查询项，来自 [732]

在文档集内显示查询项命中结果的其他方法包括，将查询项放在条形图、散点图和表格中。Reiterer 等人 [1342] 的可用性研究比较了五种视图：标准的 Web 搜索引擎形式的结果列表视图；包含了标题和文档元数据，用图形显示在文档内部的查询项命中位置，并通过高度来表示频率的列表视图（见图 2-17）；彩色的 TileBars 类型的视图，其中的文档标题显示在图像旁边；像 Veerasamy 和 Belkin [1630] 那样的彩色柱状图视图；表示相关性得分及其发布日期的散点图视图。

44
~
45

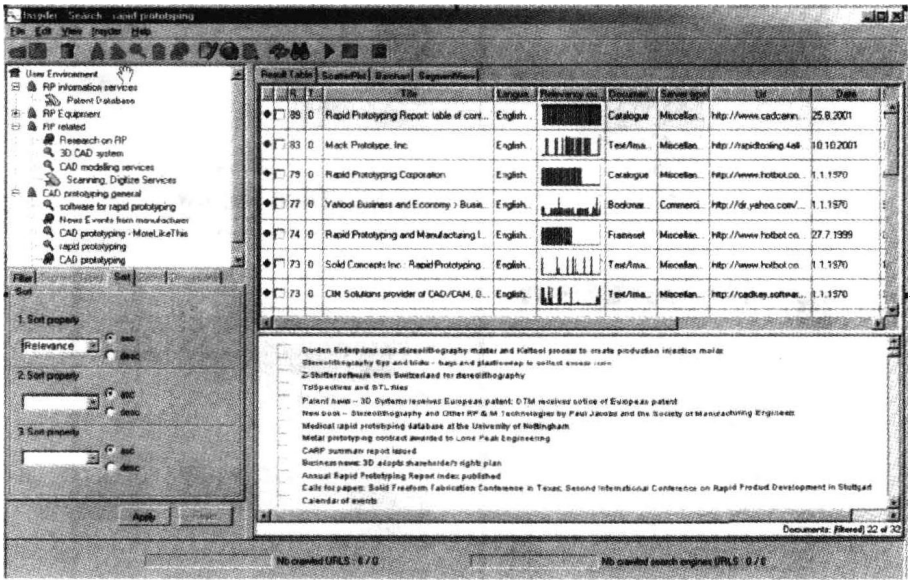
在被问及主观感受时，40 个参与者大体上都会首先选择字段可排序的视图，然后是 TileBars，最后是网页样式的列表。柱状图和散点图则有很多负面反应。对于任务有效性而言，其他方法和网页样式的列表没有明显的不同（除了柱状图之外，其效果会差很多）。所有其他方法都比网页样式列表花费更多的任务时间。最后一点说明了可视化搜索中一个常见的结果——即使在可视化被认为有助于任务的情况下，它通常也比只有文本的界面花费更长的搜索时间。这可能是因为从解释图像转换到阅读文本需要花费一些时间，因为它们属于不同的认知功能。

另一种在文档内部显示命中查询项的想法是显示**缩略图**——文档视觉外观的缩略版（见图 2-18）。一项应用缩略图的实验发现它们在改进搜索结果方面不会比空白区域更好 [468]，另一项实验发现参与者更容易错误地认为显示了缩略图和标题的文档是相关的（相比于那些只显示缩略图的情况）[523]。这两项研究都显示了缩略图对用户有主观上的作用。

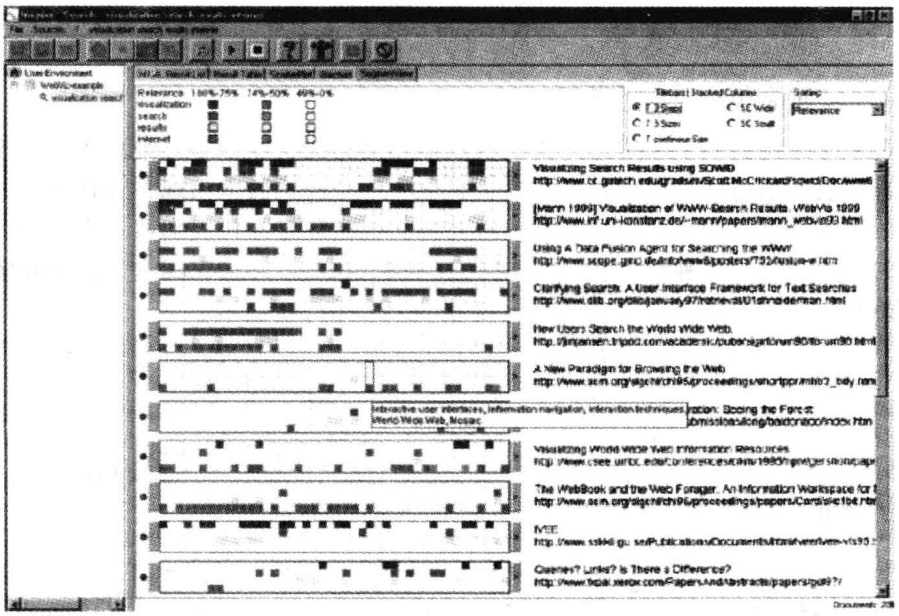
46

负面的研究结果可能源于缩略图大小的问题，更新的结果表明增加缩略图的大小可以提高搜索结果的可用性 [858]。一项相关的研究表明通过加亮显示缩略图内部的查询项，使其

更大并且更可见，对于某些种类的任务来说可以提高搜索结果的可用性［1720］（见图 2-18）。在 SearchMe 搜索引擎中，已经开始在搜索结果中使用更大的缩略图和相应增大的文本，而这是通过封面流（Cover-flow[⊖]）动画来表示的。



a）



b）

图 2-17　a）字段可排序的搜索结果视图，包括一个柱状图风格的图形界面，表示了查询项命中的位置，这是一个简化版本的 TileBars，来自［M. A. Hearst，*Search User Interfaces*，Cambridge University Press，2009，figure 10.17a］［735］；b）彩色版的 TileBars 视图，来自［1342］

⊖　即由苹果公司首创的将多首歌曲的封面以 3D 界面形式显示出来。——译者注

图 2-18 带文本增强功能的缩略图，来自 [1720]

2.4.3 可视化词语和文档间的关系

很多可视化系统的开发人员已经提出将词语和文档放置在一个二维画布上的想法，其中符号的邻近代表词条与文档语义相关。这个想法的一个较早版本出现在 VIBE 的界面中，其中查询项放置在一个平面上，包含查询项组合的文档被放置在代表那些查询项的图标的中间（见图 2-19）[1228]。这种想法的一个现代版本在 Aduna Autofocus 产品中出现，另外在 VIBE 的基础上，Lyberworld 项目 [744] 制作出了一个 3D 版本。

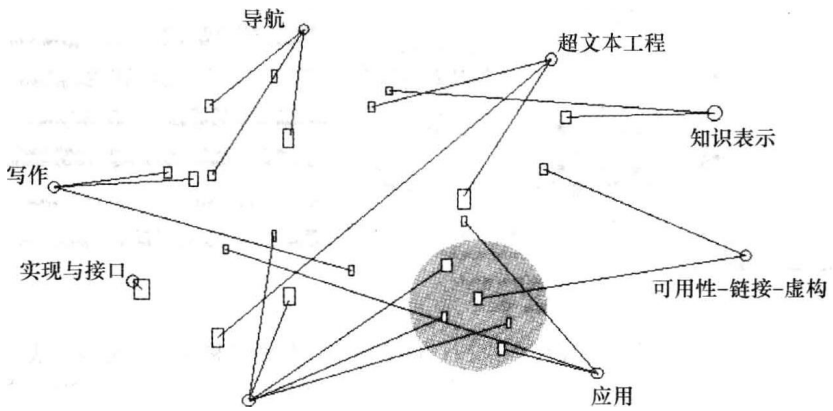
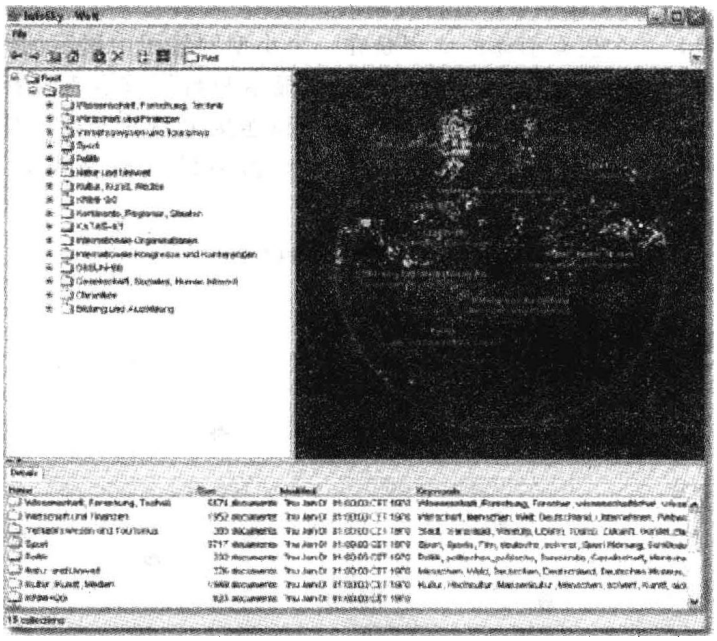


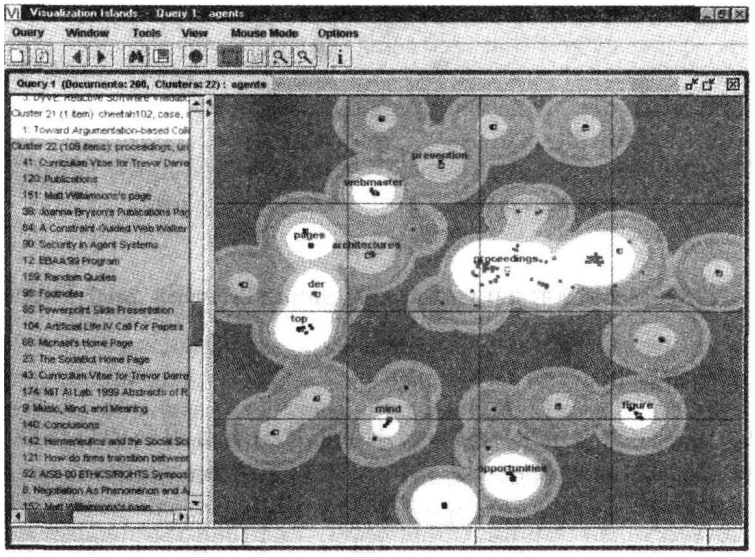
图 2-19 VIBE 的显示，其中查询项在二维空间内展示，而文档根据其文本来排列，源自 [1228]

这种想法的另一个形式是将文档或词语从一个高维的词项空间映射到二维的平面，然后文档或词语通过 2D 或 3D 显示在这个平面上 [53，662，758，1688，1700]。这种形式的聚

类能够用在根据查询检出的文档，或者在一个预处理过的文档集内加亮显示与查询相匹配的文档。图 2-20 是两个基于这种星空（starfield）理念的显示结果。



a）



b）

图 2-20　用一个 2D 或 3D 映射的标志符号来表示文档的想法已经提出了多次。这里显示两个例子：a）InfoSky，源自论文 Evaluating a system for interactive exploration of large, hierarchically structured document repositories，*Proceedings of the IEEE Symposium on Information Visualization*，pp. 127-133（Granitzer，M.，Kienreich，W.，Sabol，V.，Andrews，K. and Klieber，W. 2004），© 2004 IEEE [662]；b）xFind 中的 VisIslands，源自论文 Search result visualisation with xFIND，*Proceedings of User Interfaces to Data Intensive Systems*，pp. 50-58（Andrews，K.，Gutl，C.，Moser，J.，Sabol，V. and Lackner，W. 2001），© 2001 IEEE [53]

这些视图都很容易计算并且能够直观地显示。然而，现有的评价对它们的可用性提出了
47
~
48
负面的意见 [662，773，910，1400]。主要的问题是，在这些视图中文档的内容是不可见的，而且 2D 表示不像语义那样可以进行复杂的交流。

利用这种思想，一个更有前途的应用是在一个小型网络图中展示词典中的索引项，例如 Visual Wordnet（见图 2-21）。这种方法通过只显示与目标节点直接相连的节点从而使较大的 WordNet 数据库得到简化。对于这种节点和连接关系的视图，其应用情况还没有在已发表的研究中进行过评价，但它在用于组织搜索结果时并没有获得很好的结果 [1548]。

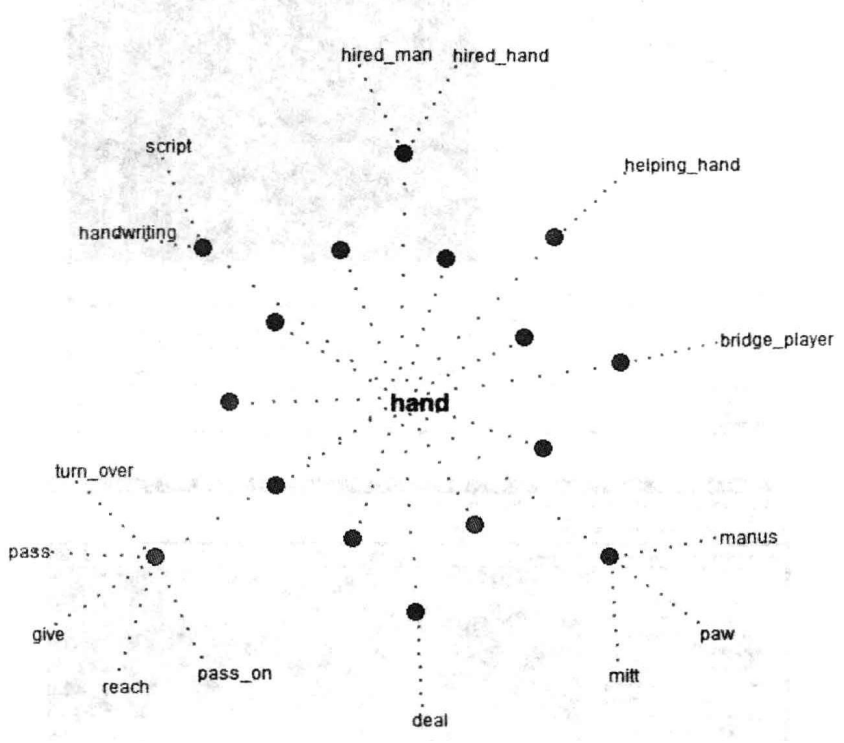


图 2-21 WordNet 同义词典的可视化表示，来自 http://kylescholz.com/projects/wordnet/

2.4.4 文本挖掘的可视化

2.4.3 节说明了对于搜索结果来说，搜索可视化并没有很强的可用性。事实上，可视化似乎对文本数据的分析和探索更为有用。大多数搜索系统的用户对文档中的词语如何分布以及文档集中最常见的词语并不感兴趣，但这些是计算语言学家、分析学家以及奇特词语爱好者感兴趣的活动。像 Word Tree 那样的可视化系统 [1669] 会显示一部分文本词汇索引，使得用户能够看到哪些词语和短语会常常出现在给定词语的前后（见图 2-22），另外还有 NameVoyager 系统 [1670]，它显示在不同的年代中，美国婴儿名字的出
49
现频率（见图 2-23）。

对搜索界面进行可视化，有时是为了方便信息分析师。图 2-24 显示的是 TRIST 信息“分类”系统 [854，1303]，它的作用是帮助信息分析师完成工作。系统将搜索结果用文档图标表示；数以千计的文档可以显示在一起，系统支持多维链接，从而使我们能够发现文档间的特征与相关性。图标的颜色用做显示哪些文档是用户之前已经看过的，图标的大小和形状分别表示文档的长度和类别。这看起来是个有效的系统，它的设计者连续两年在

IEEE 视觉分析科技竞赛（IEEE Visual Analytics Science and Technology，VAST）中获胜[679]。

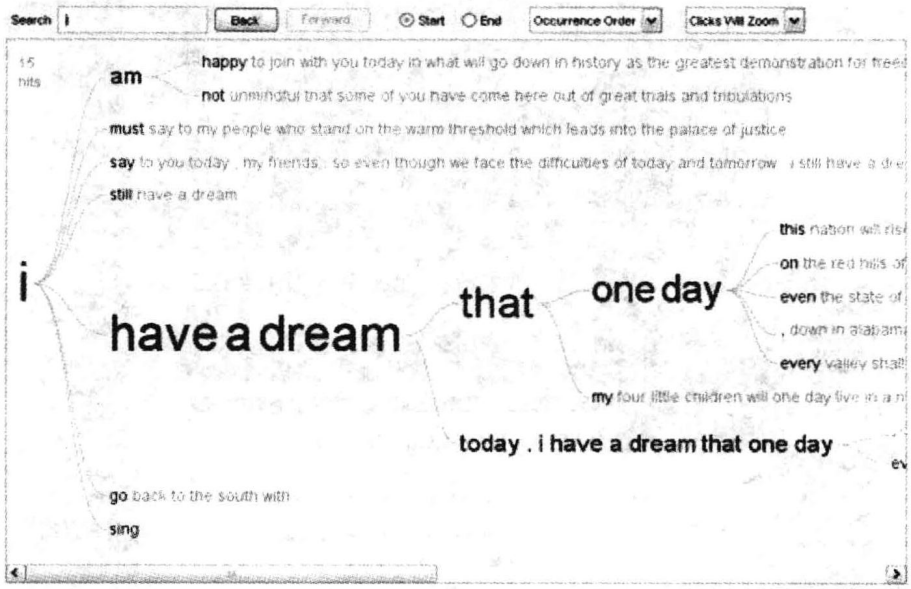


图 2-22　Word Tree 可视化系统在 Martin Luther King 的演讲“I have a dream”（我有一个梦想）上的演示，来自 The word tree, an interactive visual concordance, *IEEE Transactions on Visualization and Computer Graphics*, 14（6）, pp. 1221-1228（Wattenberg, M. 和 Fernanda, B., 2008）, © 2008 IEEE [1669]

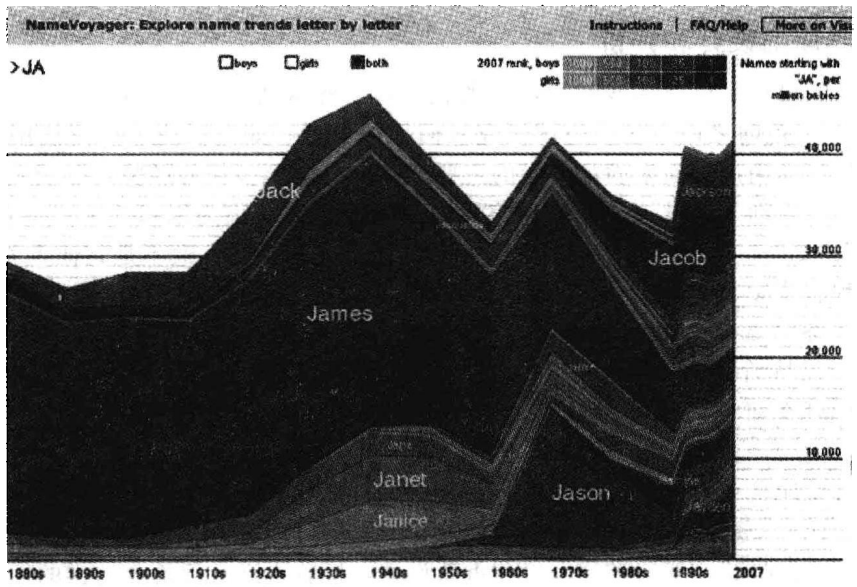


图 2-23　一个可视化演示的例子，针对一段时期内以 JA 开头的婴儿名字的相对普及度，来自 babynamewizard. com

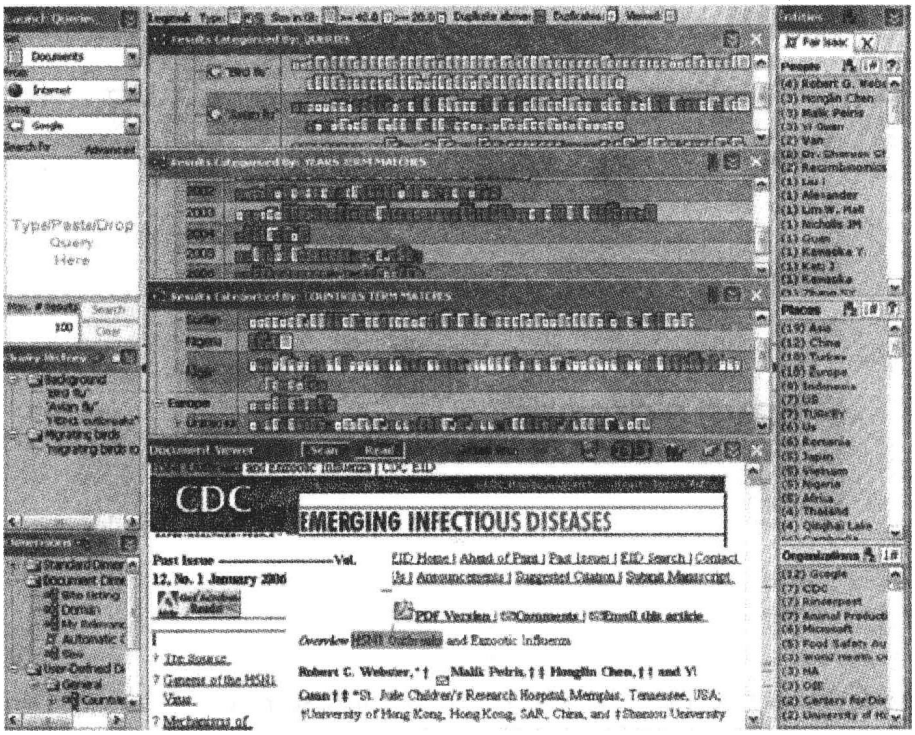


图 2-24 输入 Avian Flu 相关查询之后，TRIST 界面的显示，来自 Avian Flu case study with nSPace and GeoTime，*Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*（*VAST* '06），pp. 27-34（Proulx，P. et al. 2006），© 2006 IEEE [1303]

2.5 搜索界面的设计和评价

用户界面的设计是一项实践活动，它的技术包含在人机交互（Human-Computer Interaction，HCI）的范畴之内。这个领域研究人们如何思考、如何反应、如何使用技术，以及如何设计出最好的用户界面来满足人们的需要和倾向。基于多年的经验，已经制定出一些实践和指导方针，帮助人们设计成功的用户界面。这些实践统称为以用户为中心的设计（user-centered design），它围绕着用户的行为和思考过程完成设计，而不是其他无关的因素。

设计过程首先要预期用户的目标（goal）是什么，然后设计一个界面来帮助用户通过
50 完成一系列任务（task）来完成目标。在信息获取领域，目标的范围可能会非常广泛，从寻找管道工到保持对商业竞争对手的关注，从写作可发表的学术论文到调查一宗欺诈指控。信息获取任务可以用来完成这些目标。这些任务覆盖了从询问具体问题到彻底研究某个主题。

用户界面设计是一个不断改进的过程，其中的目标和任务通过对用户的研究来说明，然后构建初始设计——这通常会基于现存的设计，但也有可能包含一些新的想法。这些初始设计由预期用户进行测试，然后进行评价并重新设计，接着再进行评价，这样的循环需要重复很多次。

评价用户界面的过程通常与评价排序算法或爬虫技术不同。爬虫可以通过一些硬性的量化指标，如覆盖度和新鲜度来评价。排序算法可以通过精度、召回率和速度来评价。但是，用户界面的质量是由用户对它的反应来决定的。与量化指标相比，主观反应即使不能说是更

重要的，至少也是一样重要的，因为如果用户需要在两个系统中选择一个，那么他们会使用他们偏爱的那个。偏爱的原因可能是由多个因素决定的，包括速度、熟悉度、审美观、偏好特性，或者主观感知的排序准确性。通常更受青睐的选择会是用户熟悉的那个。尤其是在搜索界面中，一个新的界面必须要在主观感知上比旧的质量更好，用户才会转而使用新的界面。这一现象可以解释为什么自从搜索引擎第一次出现之后搜索结果的显示方法就没有什么明显的变化。

如何能够最好地评价用户界面，取决于当前处于开发周期中的哪个阶段。当开始一个新设计或新想法的时候，通常会使用简易（discount）可用性的方法。简易方法会向一些潜在的用户显示若干种不同的设计，然后让他们指出哪些部分是有前景的，而哪些是没有前景的。通常在这个任务中会使用草图上的原型设计，因为它们可以快速地开发，也很容易舍 51
弃，因为有证据显示，比起完整的成品，用户更愿意去批评那些明显未完成的东西［163，1344］。通常这个设计-测试-再设计的过程，在找到一个可接受的交互原型起始点之前会循环若干次。

接下来，应当进行非功能性的交互设计开发，并由少量的参与者进行测试，获取他们的主观反应，并确定哪些元素在设计中有良好的效果，以及哪些因素会产生混淆或者效果不佳。如果一个参与者看着屏幕而不知道该做什么，那么就意味着需要重新设计了。

另一种常见的折扣评价方法是*启发式评价*（heuristic evaluation），其中可用性专家“走查”（walk through）设计，然后评价其是否符合设计准则。这种评价方法在寻找可用性问题时有很好的效果，因为经验丰富的专家可以准确地预见潜在的问题，但是这一方法应当与目标用户的响应相结合。

在多次设计迭代后，通常就能设计出一个交互系统，此时可以在一个较正式的实验中进行测试，由研究的参与者来执行，并基于一系列的指标，将新的设计和一个有竞争力的基准进行对比，或者比较两种候选的设计方案。在评价搜索界面时，最重要的是让参与者有足够 52
的动力去完成那些任务［288，1524］。如果我们要求那些不关心照相机镜头或者对于照相机镜头非常了解的人去对照相机镜头做广泛的搜索，那么很可能不会产生有实际意义的结果。为了确保参与者的积极性，研究参与者应当对查询和信息集合的主题充满兴趣，另外应该选择那些会最终使用这个系统的人，或者接近的替代人员（例如，护理专业的学生通常要比执业护士更愿意测试一项设计）。

正式的实验通常旨在产生研究结果，用于发表以及在更为广泛的群体中应用，也有一些机构会在其内部进行正式的研究。正式的研究需要遵循科学领域中的实践方法，如招募研究参与者，对控制条件和实验条件进行比较。正式的实验应当谨慎地进行设计，需要考虑到那些潜在的干扰因素，比如要平衡那些参与竞争的设计的显示顺序，以及实验人员要避免对某个设计表现出偏见。只有达到这样要求的研究才能回答诸如“某个设计元素的优劣”或者“新的特性是否比现有的系统更好”之类的问题［735］。

这种研究能够发掘重要的主观性结果，如新的设计是否要明显好于基准系统，但是搜索界面本身的特点使得我们难以通过一小部分参与者来精确地找出可量化的差异。这个困难是由很多因素决定的，比如任务和查询对于系统行为的巨大影响——在许多搜索系统的研究中，无论是交互式的还是批处理式的分析，任务上的差异都具体表现在系统和参与者上的差异。另一个问题是人们提交的搜索可以有很多不同的形式，以至于很难直接量化地比较它们的输出。最后，时间变量在某些时候对于评价交互式的搜索会话并不是一个合适的指标，因为让搜索用户在搜索过程中理解他们的主题能够带来很大的好处，但是这可能会比其他设计

方案花费更多的时间。

面对这些问题，最近几年有两种评价搜索界面的方法开始逐渐流行起来。一种是进行纵向研究，意思是参与者长时间地使用一个新界面，并监控和记录他们的使用情况 [518, 1471]。评价是基于日志中记录的客观指标以及对参与者的问卷调查和当面访谈。在某些情况下，一种新方法的优势只有在用户使用了一段较长的时间后才会有明显的感受，所以只有长时间的研究才能发掘这种优势。例如，文献 [862] 中的研究发现，随着时间的推移，用户会根据界面相应地改变他们的搜索模式，该研究还揭示了在怎样的情况下新的特性是有用的，而什么时候它们是不需要的。另外，开始很吸引人的界面（比如说令人印象深刻的图形）随着时间的推移有可能会变得让人厌倦，导致我们希望重新使用之前熟悉的界面。

在过去几年里流行的另一项主要评价技术是在使用率较高的网站上进行的大规模实验。这种方法经常称为**水桶测试**（bucket testing）、**A/B 测试**，或者**平行航班测试**（parallel flights）[921]。一个每天接收到成千上万甚至是几百万个查询的搜索引擎可以进行以下的研究：随机选择一个用户子集，向他们展示新的设计，将他们的反应和同样随机选择出来的继续使用现有界面的控制组用户进行对比 [61, 922, 919]。这与正式的可用性研究不同，因为其中的参与者并不晓得自己参与了这项研究，网站会在被挑选出来的访问者没有了解和同意的情况下向他们显示新的设计（大多数网站的用户协议都允许这类服务）。

53

这种形式的研究通常能在 24 小时内完成，不过通常建议整个流程为 1～2 个星期。有些性能评价，如哪个链接被点击等，在这样的测试中是尤其有信息性的。例如，显示查询建议的界面可以和不显示建议的界面进行对比，并记录下查询建议被点击的频率。另一个例子是，在文本结果列表中插入多媒体结果的影响，可以通过周围链接点击情况的变化以及新信息被点击的频率来进行评价。将控制条件下和实验条件下的用户行为进行比较，有些时候，我们可以比较，对于相同的查询，分别处于这两种条件下的用户行为，因为用户的数量是如此之大。这种研究的一个潜在缺点是有些熟悉网站的控制组的用户很可能会在一开始对他们不熟悉的界面给出负面的反应，所以有些研究会在评价过程中考虑初始反应状态的因素。这项技术通常也没有考虑到主观信息的因素，很多实验会通过后续调查来研究主观的反应。

2.6 趋势和研究问题

本章介绍了很多关于提高用户搜寻信息时的人机交互体验的方法。这仍然是一个快速发展的领域，界面的进步很有可能会带来更好的搜索结果以及更有效的信息创建者和使用者。未来最重要的前进方向是社交搜索、移动搜索界面、多媒体搜索，以及面向自然语言的查询，Hearst [735] 曾对此有过详细的论述。

2.7 文献讨论

Hearst 的《Search User Interfaces》[735] 对本章介绍的主题进行了深入的讨论。另一本可供参考的关于信息搜寻行为的书是 Marchionini 的《Information Seeking in Electronic Environments》[1082]。Lesk 的《Understanding Digital Libraries》[1005] 也有对搜索界面的讨论。

也有很多关于 HCI 和界面设计的书，包括 Shneiderman 等人的《Designing the User Interface》[1472]，Kuniavsky 的《Observing the User Experience：A Practitioner's Guide to User Research》[948]。一本古老但依旧出色的关于用户界面评估的书是 Nielsen 的《Usability

Engineering》[1204]。

很多书都介绍了如何设计网站，而与本章内容最相关的书是 Morville 和 Rosenfeld 的《Information Architecture for the World Wide Web》[1157] 第 3 版，其中有两章内容与搜索有关。Kalback 的《Designing Web Navigation，Optimizing the User Experience》[863] 讨论的是网站导航的设计。另外，有许多网站都致力于可用性和用户界面设计的研究与讨论。 54

现在有很多非常好的书介绍如何利用信息的可视化进行设计，包括 Few 的《Now You See It：Simple Visualization Techniques for Quantitative Analysis》和《Information Dashboard Design：The Effective Visual Communication of Data》[562，563]，以及 Tufte 的《The Visual Display of Quantitative Information》[1605]，不过这些书并不是只关注文本或搜索的可视化。

更多关于 Web 搜索引擎界面以及网页内容可视化的参考文献将在 11.7 节中进行介绍。 55

第3章

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

信息检索建模

3.1 信息检索模型

信息检索中的建模是一个以产生排序函数为目标的复杂过程。该排序函数能根据给定的查询给文档打分。这个过程包含两项主要的任务：1）构想出表示文档和查询的逻辑框架；2）定义一个针对给定查询计算文档排名的排序函数。这个逻辑框架通常基于集合、向量，或者概率分布。它直接影响了文档排名的计算，这些排名然后被用来对给定查询所返回的文档进行排序。

考虑到排序可能是检索系统中最重要的过程，我们将广泛、深入地讨论信息检索建模。我们首先确立建模和排序之间的关系。然后，我们形式化地描述检索模型，并列出检索模型的分类体系。

3.1.1 建模和排序

如第2章所述，信息检索系统通常采用索引项来索引和检索文档。严格地说，一个索引项是一个关键词（或者一组相关联的词），可以独立地表达某种意思，通常扮演名词性的角色。从更广义的形式看，索引项可以是文档集内文档正文中的任何一个词。

基于索引项检索的主要优势是可以高效地实现，并且可以简单地用查询进行查阅。简单性是重要的，因为这减少了用户制定查询的精力。然而，仅用几个词来表达查询意图限制了

所能表达的语义。这样也无怪乎检索出的文档经常与用户的查询不相关。如果考虑到大部分用户没有受过如何合理制定查询的训练，那么这个问题会由于这些隐患而变得更糟。其直接的后果是搜索引擎的用户对于许多查询的返回结果感到不满。

为了检索出查询的答案，任何检索系统都要处理一个核心问题——预测哪些文档会被用户看做是相关的，哪些会被他们看做是不相关的。这个问题本来就困难。而且，由于不同的用户可能对于何为相关、何为不相关有各自的看法，因此这个问题自然地包含了一定程度的不确定性和模糊性。对于这种情况，系统要实现一个预测算法，希望该算法能和大部分用户对大部分查询与答案相关与否的看法相似。这个预测算法基本上就是排序函数，用来对检出文档建立一个简单的排序，排在前面的文档更有可能是相关的。所以排序函数是检索系统的核心。

排序算法是根据文档相关性这一概念的基本前提来执行的。关于文档相关性的不同前提条件会产生不同的检索模型。正如我们在这里讨论的，所采用的信息检索模型决定了什么是相关的、什么是不相关的（例如，系统所实现的相关性概念）。我们的讨论包括了过去数年中提出的各种关键的信息检索模型，为本书中其余大部分章节提供了概念性的基础。

3.1.2 信息检索模型描述

信息检索模型是由形成排序算法的基础前提决定的。我们的描述如下。

定义 一个信息检索模型是一个四元组 $[\boldsymbol{D}, \boldsymbol{Q}, \mathcal{F}, R(q_i, d_j)]$，其中

关于此电子书的说明

本人由于一些便利条件，可以为您提供各种中文图书的PDF电子版，保证质量清晰。只要图书不是太新，文学、法律、计算机、经济、医学、工业、学术等方面的图书，都可以帮您制作，如果您有这方面的需求，可以通过QQ联系我，我的QQ号是 3330972307 。