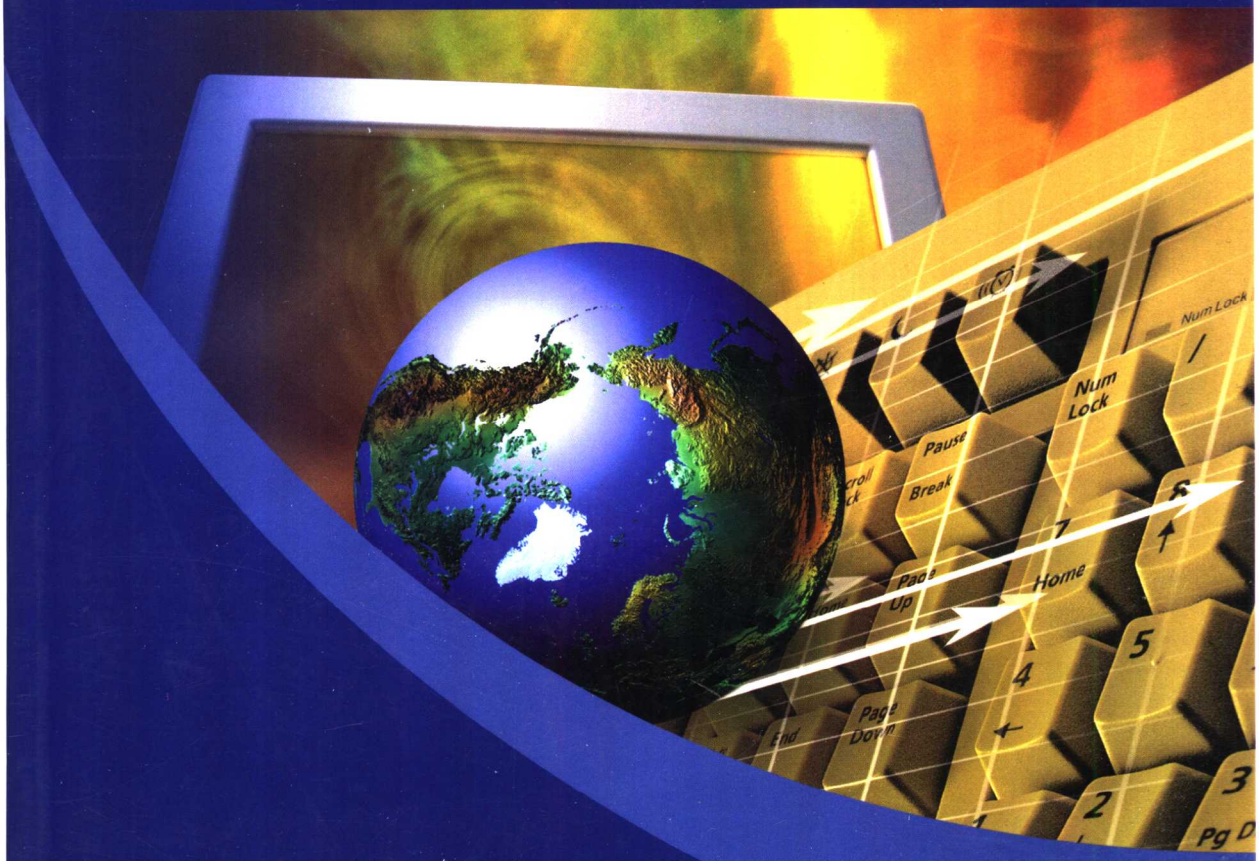
孙建军　成　颖　等 编著

信息检索技术



科学出版社
www.sciencep.com

孙建军　成　颖　丁　芹
李君君　宋玲丽　柯　青　编著

信息检索技术

科学出版社

北京

内 容 简 介

本书系统地介绍了信息检索的原理与技术。讨论的中心问题是如何能迅速地检索到相关信息。具体内容包括：信息检索的布尔模型、向量空间模型、概率模型，以及逻辑模型；文献自动处理技术：自动分类、自动聚类、自动文摘；查询的扩展与精化、相关性、Z39.50，以及搜索引擎等。

本书可作为高等院校信息管理与信息系统专业本科生和研究生教材，也可作为信息机构有关信息服务人员、咨询人员、管理人员的参考用书。

图书在版编目(CIP)数据

信息检索技术/孙建军，成颖等编著. —北京：科学出版社，2004

ISBN 7-03-014244-6

Ⅰ.信… Ⅱ.①孙…②成… Ⅲ.情报检索 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2004)第 087155 号

责任编辑：李 敏 / 责任校对：钟 洋

责任印制：钱玉芬 / 封面设计：东方上林

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2004 年 10 月第 一 版 开本：B5 (720×1000)

2004 年 10 月第一次印刷 印张：30 1/2 插页 1

印数：1—3 000 字数：615 000

定价：49.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

前　言

目前，信息检索研究可以分为实务与原理两部分，图书馆、情报以及档案学界主要侧重于信息检索实务方面的研究与应用，成果颇丰，据不完全统计，国内已经有好几百种信息检索实务类的教材与教程。信息检索原理方面的研究则主要由计算机、数学及情报学界完成，20 世纪 90 年代以后，该领域再次吸引了众多学者的目光，许多原先从事人工智能、自然语言处理等领域的学者也加入到信息检索的研究队伍中，近些年成果也非常丰富，但是对该领域系统总结的专著或者教材尚不多见，本书旨在完成该任务。

本书共 12 章，分别从信息检索模型、技术、标准以及应用 4 个方面系统地介绍了信息检索原理方面的最新进展。

引言（第 1 章）部分从宏观的角度鸟瞰了信息检索原理领域的全貌，构建了全书的框架，预测了原理研究的方向。

模型部分（第 2、3、4、5 章）系统总结了布尔模型、向量空间模型、概率模型以及逻辑模型，考虑部分兄弟院校可能会以本书作为教材使用，因此，在介绍这些模型时，我们兼顾了历史与进展。介绍这部分内容的目的，是让读者对信息检索原理有一个初步的、同时也是较为全面的理解与掌握。

技术部分（第 6、7、8、9、10 章）主要介绍信息检索领域的自动分类、自动聚类、自动文摘、查询扩展和精化以及相关性等研究，其研究比较深入，已经处于即将从实验室进入实用阶段或者已经部分实用的技术。通过这些章节的学习后，希望读者能对信息检索领域的主流技术有一个概要的认识，并能通过这些技术进行信息检索的相关研究以及应用工作。

标准部分（第 11 章）介绍了目前已经实际应用的标准——Z39.50，通过标准的学习，希望读者能够认识到标准对于信息检索研究与应用的意义。

应用部分（第 12 章）通过搜索引擎系统地阐述了前面章节中各种原理、技术以及标准的综合运用，旨在使读者深化前面内容的学习，从而做

到原理、技术、标准与应用的有机结合。

本书由南京大学孙建军、成颖、丁芹、李君君、宋玲丽和柯青合作编写。编写工作的具体分工如下：孙建军组稿，提出大纲，第1章由孙建军、柯青撰写，第2、3、4章由丁芹撰写，第5、7、12章由李君君撰写，第6、8章由成颖撰写，第9、11章由宋玲丽撰写，第10章由孙建军、成颖撰写，最后由孙建军审订统稿。柯青协助完成了大量的校对工作。

本书得到了南京大学“985”工程的资助，特此致谢。

在本书写作过程中得到了科学出版社李敏编辑的大力支持与帮助，在此表示衷心的感谢。

国内外众多的信息检索研究为本书提供了良好的基础，本书的顺利完成也得益于参阅了大量的相关作者的成果，在此我们向这些文献的作者表示诚挚的谢意。

由于我们的学识、水平有限，书中难免存在一些缺陷以及不足，恳请专家与读者批评指正。

孙建军

2004年8月

目　录

1 引　言

1.1 信息检索的起源与发展

信息检索是指信息用户为处理解决各种问题而查找、识别、获取相关的事实、数据、文献的活动及过程。作为人类社会活动不可分割的一部分，信息检索有着悠久的历史。信息检索研究则是伴随着科学技术的发展和信息数量的剧增而兴起的研究领域。英国科学家詹姆斯·马丁认为：人类的科学知识在19世纪是每50年增加1倍，20世纪中叶是每10年增加1倍，在20世纪70年代就已经缩短到每5年增加1倍；同时，信息分散，交叉引用频繁，人类信息的生产能力超过了人类对信息的处理、组织和吸收能力，从而产生了信息爆炸的危机。人们越来越关注如何从浩如烟海的信息源中迅速而准确地查找到学习和研究所需要的资料，因而，信息检索的战略地位也就显得日益重要。其主要研究范围包括：信息检索理论、信息检索语言、信息检索系统的建构及评价、信息检索技术与方法等。

20世纪中叶以前，信息存储和传播主要以纸质介质为载体，信息检索活动也围绕着文献的获取和控制展开。因此，信息检索研究关注的是如何检索、利用文献中记载的信息，从而导致文献检索成为信息检索的同义词。早期的文献中不使用“信息检索”这一概念。50年代以后，社会信息传播与存储载体呈现多元化，人们不再拘泥于纸质载体研究信息检索，于是开始广泛使用情报检索一词。由于汉语中“信息”一词较“情报”一词的含义更为宽泛，加上英文information可以理解为“信息”和“情报”，近年来人们越来越倾向于将情报检索研究和文献检索研究归为信息检索研究这一更具兼容性的概念，以便将各种不同的检索综合起来，使该研究领域取得更多、更为实用的研究成果，对信息检索实践起到更全面的指导作用。

随着科学技术的发展，尤其是计算机的应用，信息检索经历了从手工检索到机械检索再到计算机化检索的过程。

1.1.1 手工检索

手工检索是指仅用手工的方式来处理和查找文献工具，如文摘、索引、目录、参考工具书等。它是一种传统而又基础的检索手段。

手工检索因其不需要特殊设备，查找简单、灵活，而且用户可以随时修改检索策略，检索费用较低等优点而在某些部门领域仍然使用，但是，利用手工检索往往

费时较多、效率低下、查全率也较低。

1.1.2 脱机批处理检索

20世纪40年代中期世界上第一台电子计算机问世后，50年代初就有人开始研究其在信息检索领域的应用。50年代中期至60年代中后期是信息检索的脱机批处理阶段。当时，计算机硬件发展很快，但还没有连接通信网，也没有远程终端装置，不能提供问答服务(Q-A)的检索方式，只能进行现刊文献的定题检索(SDI)和过期文献的追溯检索(RS)，同时利用计算机编辑出版检索性刊物。所谓脱机批处理方式，是指定期由专职检索人员把许多用户课题汇总，批量处理提问要求并把结果提供给用户。在美国，这个时期出现了3个重要系统：1959～1963年美国武装部队技术情报局的ASTIA(即后来的国防文献中心DDC)系统；1962年美国国家航空和航天局的NASA系统；1964年美国国家医学图书馆创建的医学文献分析与检索系统MEDLARS。MEDLARS不仅可以进行逻辑“或”、“与”、“非”多种运算，而且可以从多种途径检索文献。

脱机批处理能同时进行多项检索，对复杂的检索词也具有处理能力，因此，在生产普通印刷索引、专题书目、回溯检索和定题检索服务等方面得到广泛使用。但是，它也有许多不足之处，如缺乏与用户的交互过程，检索结果获得不及时以及信息需求和检索结果之间存在一定误差等，这些缺点限制了脱机批处理的发展。

1.1.3 联机检索

20世纪70年代计算机分时系统的出现，通信技术的改进，使得多终端、远距离两地检索信息的技术得以推广，计算机检索技术从脱机阶段进入联机信息检索时期。所谓联机检索，就是用户使用终端设备，通过通信线路与中央计算机连接，直接与计算机对话进行检索，结果由终端输出。第一个大规模的联机检索系统是美国NASA的RECON系统，1969年全面投入运行。随后许多著名的联机检索系统相继出现，如1970年洛克希德火箭公司的DIALOG系统，美国的MEDLARS于1970年发展的MEDLINE联机系统等。

联机检索无需委托，直接面向最终用户，在检索过程中是“人机对话”方式，具有很强的交互功能，而且能及时取得检索结果，但是检索指令复杂，需要依赖专业检索人员。

1.1.4 光盘检索

利用国际联机检索系统检索到的电子文献信息具有较高的使用价值，但国际联机检索费用昂贵，一般用户难以承受。人们开始努力寻求一种低廉的存储、检索电子信息的方式，光盘存储技术则适应了这一要求。CD-ROM光盘是20世纪80年

代在计算机技术、激光技术等现代新科技成果的基础上发展起来的新型电子出版物。它具有信息存储密度高、容量大、读取速度快、存储的信息类型多等优点，备受人们的青睐。光盘技术与光盘产品的发展相当迅速，品种和数量激增，而且更新换代快，功能日益完善。光盘塔和光盘网络的出现和广泛应用提高了单张光盘的利用率，使光盘的多用户检索和共享成为现实。

利用光盘检索系统费用大大低于联机检索，利用 CD-ROM 存储信息方便、易于携带，除可提供追溯检索、定题服务外，还可用于“自建库”和做联机检索前预处理。

1.1.5 网络化联机检索

国际联机检索和光盘检索为我们提供了大量的信息资源，但各自又都有着或多或少的缺点，例如联机检索费用昂贵，指令复杂，而光盘检索得到的信息又不十分及时等。因此，极有必要产生一种新型的信息检索方式。1993 年，美国政府提出国家基础设施建设(NII)计划，兴建以 Internet 为雏形的信息高速公路，网络资源如潮水般涌来。在信息爆炸的当今社会，单个计算机所能完成的工作和所存储的信息都极为有限，而把单机连起来的计算机网络则能在局部或更大范围内实现通信和信息共享。由于电话网、电传网、公共数据通信网都可能为信息检索传输数据，世界各大检索系统纷纷进入各种通信网络，每个系统的计算机成为网络上的节点，每个节点连接多个检索终端，各节点之间以通信线路彼此相连。网络上的任何一个终端都可联机检索所有的数据库的数据。这就是网络化联机信息检索。网络联机信息检索是联机信息检索的高级阶段，它的实现使人们可以在很短的时间里查遍全球的信息资料，使人类实现信息资源共享成为可能。除了传统的文献信息，网络信息源还包括电子论坛、各种软件资料、图像文件、声音文件等。值得指出的是，网络信息环境的出现，使得信息检索研究的对象和范围不断扩大，研究队伍也突破了原有的以图书情报领域的专家学者为主的框架，众多的信息公司加入到研究开发信息检索系统的行列。可以这么说，网络使计算机信息检索技术进入一个崭新发展阶段，而网络信息检索又使得网上信息源利用率提高，信息组织更为有序和高效。

1.2 信息检索技术的研究内容

随着网络信息资源的日益丰富和复杂化，为满足不同用户能够检索到所需信息，检索系统朝着自然语言检索、用户界面友好的方向发展，这给信息检索技术提出了更高的要求。因此，当前信息检索技术的研究主要包括以下方面：

1.2.1 检索模型研究

信息组织是实现信息检索的基础,原始的文档中包括文本、图像、视频、音频等数据,不能直接进行检索,需要从这些原始数据中抽取逻辑视图,支持信息检索。用户用查询来表示他的信息需求。检索系统根据查询的表示,搜索文档集,获取与用户查询相关的文档。信息检索的匹配是相似性匹配,查询的结果按序返回。以上过程实际上涉及3个重要的处理:文档集的逻辑表示、查询的表示、相似匹配及其排序。对这些检索的因素和过程建模,就产生了各种不同的信息检索模型。

我们把信息检索模型定义如下:

一个信息检索模型是将文档表示、查询以及它们之间关系进行建模的框架,它由三元体

$$F[D,Q,R(q_i,d_j)]$$

表示。其中,D 是文档集中的一组文档逻辑视图(或称为文档的表示);Q 是一组用户信息需求的逻辑视图(表示),这种视图(表示)被称为查询;$R(q_i,d_j)$是一个排序函数,该函数输出一个与查询 $q_i \in Q$ 和文档表示 $d_j \in D$ 有关的实数。这样就在文档之间根据查询 q_i 定义了一个顺序。

信息检索中4个传统模型是:布尔模型、向量空间模型、概率模型和逻辑模型。近些年来,研究人员对于每种传统的模型都提出了各种不同的改进模式,如在基于集合论的检索模型中,提出了模糊布尔模型和扩展布尔模型;在代数型模型中,衍生出广义矢量模型、隐含语义索引模型和神经网络模型等3种;在概率检索模型中,发展出推理网络模型和信念网络模型等。除了涉及文本的内容之外,模型还应该涉及文本的结构。在这种情况下,就应该还有表示文本结构的结构模型。对于文本的结构模型,主要有两种类型:非重叠链表模型和邻近节点模型。

对这些检索模型可以用图1-1来表示它们的层次关系[①]:

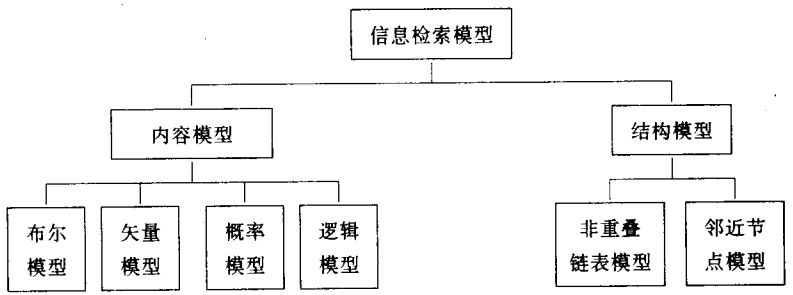


图1-1 信息检索模型的分类

① 李国辉,汤大权,武德峰. 信息组织与检索. 北京:科学出版社,2003. 91~105