

信息

毕强 张海涛 主编

检索



吉林大学出版社

信 息 检 索

主 编 毕 强 张 海 涛
编 写 黄 微

吉林大学出版社

图书在版编目 (CIP) 数据

信息检索/毕强, 张海涛主编。—长春: 吉林大学出版社,
2003. 5
ISBN 7—5601—2823—8

I. 信… II. ①毕… ②张… III. 情报检索
IV. G252.7

中国版本图书馆 CIP 数据核字 (2003) 第 032933 号

信 息 检 索

主编 毕强 张海涛

责任编辑、责任校对: 刘子贵	封面设计: 孙 群
吉林大学出版社出版 (长春市明德路 3 号)	吉林大学出版社发行 吉林农业大学印刷厂印刷
开本: 787×1092 毫米 1/16	2003 年 11 月第 1 版
印张: 24	2003 年 11 月第 1 次印刷
字数: 486 千字	印数: 1—1000 册
ISBN 7—5601—2823—8 / G · 399	定价: 37.50 元

前 言

21 世纪是信息世纪,在这新的世纪里,人们越来越认识到物质、能源、信息是构成现代社会发展的三大支柱。信息活动已成为人类的主要活动,信息已成为人们认识社会、创造社会财富和促进社会发展的主要动力。那么,人们在信息世纪将如何生存?如何提高生存质量?如何能跟上时代前进的步伐,推动社会进步?其中的关键问题之一是如何获取信息、利用信息,这就涉及到跨世纪阶段人们应掌握的一种基本技能——信息检索。

20 世纪以来,人类创造的信息量高速增长。据估计,70 年代以来全世界每年出版图书 50 万种以上、期刊 10 万种以上、专利约 50 万件、科技报告约 90 万件、会议文献 10 多万篇、产品样 50 多万种,每年发表的科技论文总数近 500 万篇,并呈指数式增长,真可谓信息浩如烟海。如何从这浩如烟海的信息中找出所需信息,就成为信息检索的重任。为此,原教育部于 1984 年发出过“教高一字 004 号”文件要求在高校开设《文献检索与利用》课。10 余年来,在文献检索教学尤其是手工文献检索方面已取得很大成绩。然而,自 80 年代以来,以缩微品、声像带、磁盘、光盘等形式记录的非纸信息急剧上升,伴随计算机进入多媒体时代、信息科技也步入多媒体发展时期,手工检索靠“手翻、眼看、大脑判断”的检索方式已难以全面适应当今信息的发展,计算机信息检索必然地提到了应用与发展阶段,以 Internet 为代表的全球性网络的广泛应用更进一步推动了这一发展,这既是对手工检索的扩展,也是时代的需要。本书是由吉林大学管理学院信息管理系与长春工程学院的教师共同编写完成的。全书由信息检索基础知识、传统的科技文献检索、网络信息检索和计算机联机检索四部分组成,分四篇共 14 章,其中第 1 章~第 10 章由张海涛撰写,第 11 章~第 13 章由黄微撰写,长春工程学院的孔祥琴、阎放和于德昌参加了第 5 章、第 6 章、第 7 章、第 9 章、第 11 章和第 14 章的编写,全书最后由吉林大学管理学院信息管理系博士生导师毕强教授统稿后完成。

本书的撰写和出版得到吉林大学出版社的热情支持,责任编辑给予了精心组织,热心帮助,提出了许多宝贵意见,在此深表谢意。

由于编写匆忙及编者水平所限,内容难免出现疏漏和错误之处,章节也不尽系统完善,故恳请同行专家学者和读者予以批评指正。

编 者

2002.11.8

目 录

第一篇 信息检索基础知识

第一章 信息检索概述	(1)
1.1 信息检索的发展历史	(1)
1.2 信息检索的概念与类型	(4)
1.3 信息检索原理	(7)
1.4 信息检索策略	(9)
1.4.1 检索策略的含义	(9)
1.4.2 检索策略的制定	(9)
1.5 信息检索效果评价	(11)
第二章 信息检索系统	(15)
2.1 信息检索系统概述	(15)
2.1.1 信息检索系统的类型	(15)
2.1.2 信息检索系统的结构	(16)
2.2 手工检索系统	(18)
2.2.1 手工检索工具概述	(18)
2.2.2 常用的中文检索工具	(20)
2.2.3 常用的外文检索工具	(21)
2.3 计算机信息检索系统	(25)
2.3.1 计算机检索系统的构成	(25)
2.3.2 计算机检索系统的特点	(26)
2.3.3 常用的计算机检索系统	(26)
第三章 信息检索语言	(28)
3.1 信息检索语言基础	(28)
3.2 分类检索语言	(31)
3.2.1 分类法概述	(31)
3.2.2 分类法体系结构	(34)
3.2.3 分类法介绍	(41)
3.2.4 分类标引的基本内容	(48)
3.3 主题检索语言	(49)
3.3.1 主题法概述	(49)
3.3.2 主题词表介绍	(53)
3.4 其他检索语言	(61)
3.4.1 自然语言	(61)
3.4.2 引文语言	(68)

第二篇 科技文献及其检索

第四章 科技文献概述	(69)
4.1 科技文献的定义	(69)
4.2 科技文献的类型	(70)
4.3 科技文献的特点与功能	(75)
4.3.1 科技文献的特点	(75)
4.3.2 科技文献的功能	(76)
4.4 科技文献检索方法、途径及步骤	(76)
4.4.1 科技文献检索的方法	(76)
4.4.2 科技文献检索的途径	(78)
4.4.3 科技文献检索的步骤	(79)
4.5 科技文献检索的策略	(80)
第五章 科技图书及其检索	(83)
5.1 科技图书的概念及类型	(83)
5.2 工具书概述	(83)
5.2.1 百科全书	(84)
5.2.2 年鉴	(87)
5.2.3 手册	(89)
5.2.4 词典	(89)
5.2.5 名录	(91)
第六章 科技期刊及其检索	(94)
6.1 中文科技期刊文献的检索	(94)
6.1.1 《全国报刊索引》	(94)
6.1.2 《中国机械工程文摘》	(96)
6.1.3 其他中文科技检索刊物选录	(96)
6.2 外文科技期刊文献的检索	(98)
6.2.1 美国《工程索引》	(98)
6.2.2 英国《科学文摘》	(102)
6.2.3 美国《化学文摘》	(108)
6.2.4 美国《科学引文索引》	(115)
6.2.5 其他外文科技期刊检索工具	(119)
第七章 特种文献及其检索	(124)
7.1 专利文献的检索	(124)
7.1.1 知识产权与专利文献	(124)
7.1.2 中国专利文献的检索	(129)
7.1.3 美国专利文献的检索	(131)
7.1.4 《世界专利索引》	(133)
7.2 科技报告的检索	(137)

7.2.1	科技报告的类型和特点	(138)
7.2.2	美国四大科技报告的检索	(139)
7.3	会议文献的检索	(142)
7.3.1	会议文献概述	(142)
7.3.2	会议文献的检索工具	(143)
7.4	标准文献的检索	(145)
7.4.1	标准文献的概述	(145)
7.4.2	我国标准文献及其检索	(146)
7.4.3	国际标准文献及其检索	(147)
7.4.4	世界主要国家标准及其检索	(148)
7.5	学位论文的检索	(149)
7.5.1	学位论文的概述	(149)
7.5.2	我国学位论文的检索工具	(150)
7.5.3	国际学位论文的检索工具	(151)

第三篇 网络信息资源及其检索

第八章	网络信息检索概述	(155)
8.1	Internet 概述	(155)
8.1.1	Internet 的发展历史	(155)
8.1.2	Internet 在中国的发展	(156)
8.1.3	Internet 的基本内容	(156)
8.2	网络信息资源概述	(159)
8.2.1	网络信息资源的类型	(160)
8.2.2	网络信息资源的特点	(165)
8.3	网络信息检索的基本方法	(166)
8.3.1	网络信息检索的发展	(166)
8.3.2	网络信息检索的方法	(167)
8.3.3	网络信息检索的特点	(168)
第九章	网络信息检索工具	(169)
9.1	网络信息检索工具概述	(169)
9.1.1	网络信息检索工具的概念	(169)
9.1.2	网络信息检索工具的构成	(169)
9.1.3	网络信息检索工具的类型	(170)
9.2	网络信息检索工具的原理	(172)
9.2.1	搜索引擎的原理	(172)
9.2.2	目录型网络检索工具原理	(174)
9.2.3	网络目录与搜索引擎的比较	(176)
9.3	网络信息检索工具的检索功能	(177)
9.3.1	基本检索功能	(177)
9.3.2	高级检索功能	(178)

9.3.3 其他检索功能	(180)
9.4 网络信息检索的策略与技巧	(181)
第十章 网络信息检索工具的使用	(185)
10.1 综合型网络信息检索工具	(185)
10.1.1 英文网络信息检索工具	(185)
10.1.2 中文网络信息检索工具	(193)
10.2 专门、专题型网络信息检索工具	(197)
10.2.1 地址、电话、电子邮件及相关信息的网络检索工具	(197)
10.2.2 地图及相关信息的网络检索工具	(199)
10.2.3 图像、影像及相关信息的网络检索工具	(202)
10.3 其他网络信息检索工具	(210)
10.3.1 FTP 的检索工具	(210)
10.3.2 Telnet 的检索工具	(215)
10.3.3 USENET 的检索工具	(217)
10.3.4 Mailing List 的检索工具	(219)
10.4 专业学术信息的网络检索工具	(222)
10.4.1 Ei Village	(222)
10.4.2 OCLC 检索系统	(230)
10.4.3 Web of Science 数据库	(236)
10.4.4 UnCover 数据库	(240)
10.4.5 其他的专业学术信息网络检索工具目录	(242)

第四篇 计算机信息检索篇

第十一章 计算机信息检索概述	(245)
11.1 计算机信息检索的基本概念	(245)
11.1.1 计算机信息检索的定义	(245)
11.1.2 计算机信息检索的类型	(246)
11.1.3 计算机信息检索的研究范围与对象	(247)
11.1.4 计算机信息检索的相关学科及领域	(248)
11.1.5 机检与手检的区别	(249)
11.2 计算机信息检索的发展	(250)
11.3 计算机信息检索系统的构成	(252)
11.4 数据库结构及类型	(256)
11.4.1 几个与信息数据库相关的概念	(256)
11.4.2 数据库的组成及类型	(257)
11.4.3 数据库的记录格式	(258)
11.4.4 数据库的文档结构	(260)
第十二章 计算机信息检索的基本原理与技术	(262)
12.1 计算机信息检索的策略与步骤	(262)
12.1.1 计算机信息检索的要点	(262)

12.1.2	分析检索课题	(263)
12.1.3	选择检索系统和数据库	(264)
12.1.4	选择检索词,构造检索提问式,确定检索策略	(264)
12.1.5	计算机信息检索的策略模式	(268)
12.1.6	计算机信息检索的随机调整措施	(271)
12.2	计算机信息检索中的检索技术	(272)
12.2.1	布尔检索	(272)
12.2.2	截词检索	(275)
12.2.3	限制检索	(277)
12.2.4	原文检索	(278)
12.2.5	加权检索和聚类检索	(282)
12.3	计算机信息检索中的文档检索技术	(285)
12.3.1	顺排文档检索技术	(285)
12.3.2	倒排文档检索技术	(292)
第十三章	联机信息检索	(305)
13.1	联机信息检索概述	(305)
13.1.1	联机信息检索系统的构成与特点	(305)
13.1.2	国际上著名的联机检索系统简介	(306)
13.2	联机信息检索的基本原理	(308)
13.2.1	联机信息检索系统的检索服务方式	(308)
13.2.2	联机数据库	(310)
13.2.3	联机信息检索的运行步骤	(315)
13.3	DIALOG 联机信息检索系统检索方法	(318)
13.3.1	DIALOG 系统的常用检索指令和算符	(318)
13.3.2	数据库的选择—Dialindex 及 Dialog Finder 的使用	(326)
13.3.3	多重数据库一次检索—Dialog OneSearch 的使用	(333)
13.3.4	基于 DialUnit 的 DIALOG 检索策略	(341)
第十四章	光盘信息检索	(347)
14.1	光盘检索系统概述	(347)
14.1.1	光盘的特点	(347)
14.1.2	光盘检索的特点	(348)
14.2	光盘检索系统的组成及基本检索方法	(349)
14.2.1	光盘检索系统的组成	(349)
14.2.2	光盘检索的基本方法	(352)
14.3	《Chemical Abstract》光盘数据库检索方法	(355)
14.3.1	《Chemical Abstracts》光盘数据库(CA on CD)	(355)
14.3.2	“《Chemical Abstracts》第十二次累积索引及文摘”光盘数据库	(363)
参考文献		(371)

第一篇 信息检索基础知识

第一章 信息检索概述

本章为概述, 主要从信息检索的发展历史入手, 讲述信息检索的概念与类型、信息检索的原理、策略及检索效果评价等有关知识。

21 世纪科学技术迅速发展, 同时产生了大量的信息, 有人说这个世纪将是信息世纪。那么, 人们在信息时代将如何生存? 如何提高生存质量? 如何能跟上时代前进的步伐, 推动社会进步? 其中的关键问题之一是如何获取信息、利用信息, 这就涉及到跨世纪阶段人们应掌握的一种基本技能——信息检索。

简言之, 信息检索 (Information Retrieval), 就是根据信息需求, 从大量的信息中查找出所需的特定信息的过程。信息需求是由需要信息的用户产生的, 而信息的来源, 除了传统意义的文献, 主要的还有联机数据库和网络。所以, 信息检索可以被看做是联系用户与各类信息的中间环节, 是用户与各类信息的中介或桥梁。

1.1 信息检索的发展历史

信息检索起源于 19 世纪前期。随着近代科学团体的涌现, 集体研究效率的提高, 文献量逐渐增多, 导致了一种社会分工的出现, 即对所有发表的文献, 及时地进行收集、加工和整理, 并提供一定的手段, 方便人们查找文献——信息检索工作便由此产生了。

19 世纪末、20 世纪初, 出现了覆盖各种专业领域的多种检索工具书, 其中有一些经后来的长期发展调整, 逐步成为世界闻名、享有很高声誉的检索工具, 如美国的《工程索引》和《化学文摘》、英国的《科学文摘》等。

直到 20 世纪 40 年代以前, 信息检索还只有手工检索一种方式, 且发展速度已较为缓慢。到了四五十年代, 出现了一些半机械化、机械化的检索操作方式, 例如各种穿孔卡片检索工具。这些检索工具的诞生, 一方面打破了完全依赖手工操作的检索方式, 另一方面也产生了组配的检索思想, 为计算机信息检索的发展, 提供了逻辑基础。

1951 年, 世界上出现了最早的计算机检索系统。自 60 年代起, 随着计算机性能的提高, 储存量的扩大, 价格的同步下降, 为计算机信息检索的发展, 提供了坚实的物质基础。计算机信息检索逐步从脱机信息检索, 发展到联机信息检索, 再到如今的 Internet 网络信息检索, 渐渐超越手工信息检索而成为现代的主要检索方式。

70 年代, 随着卫星技术和通信技术的发展, 联机信息检索突破了地域的限制, 走向全球化。一些著名的联机信息检索系统, 如 DIALOG、ORBIT、BRS、ESA-IRS 等, 开始向全世界提供商业化的联机信息检索服务。与此同时, 在这段时期内, 许多手工信息检索工具也针

对新的社会需求和信息技术的背景,进行了相当程度的调整,进一步趋于成熟。

80年代,联机信息检索从主要为科学研究提供学术信息,逐渐走向主要为工商经贸领域的企业用户提供商业信息服务。各大联机信息检索系统都强化了面向企业用户的商业化措施,推出了一些新的商业信息检索服务。同时,在联机检索领域内,也出现了多起收购和兼并检索服务机构的事件,使得联机信息检索的经营更为激烈。

从70年代初直至80年代中期,联机信息检索的市场不断扩大,市场销售额的年增长率达到20%~30%以上。1987年,DIALOG系统率先推出光盘数据库检索服务。与昂贵的联机信息检索费用相比,光盘成本低廉,能够大规模生产,便于随身携带和传播,加之个人计算机的普及,使得光盘信息检索很快受到了用户欢迎。联机信息检索机构因而开始了对用户及其服务的调整,其发展速度也趋于平稳。

90年代,网络检索、多媒体检索以惊人的速度迅速崛起,对“传统”(相对于因特网而言)信息检索提出了严峻的挑战。1992年,因特网向全世界的社会公众开放,并步入商业应用时期。

1994年春,网上出现了第一个Web搜索引擎。到1998年,仅仅4年之后,网上的搜索引擎数量已达400个之多。网络检索以极低的费用、海量的信息、迅速的存取以及对多媒体功能的支持,对联机信息检索、光盘信息检索造成了强大的冲击,很快改变了计算机信息检索的发展格局。

综上所述,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发展过程。

(1) 手工检索

手工检索是指通过手工方式利用检索工具来处理 and 查找文献的过程,如利用文摘、索引、目录、参考工具书等。它是一种传统而又基础的检索方式。

手工检索的主要优点有:

- 1) 几乎不需要特殊设备,检索方法简单、灵活;
- 2) 可以边查边思考,随时修改检索策略;
- 3) 无检索经费或费用较低。

主要缺点是:

- 1) 效率低,检索速度慢,所需时间较长,特别是进行专题检索和回溯检索时需要查阅大量工具书,费时费力;
- 2) 在进行复杂问题的多途径检索时,需要反复查找若干检索工具;
- 3) 查全率一般较低。

(2) 脱机批处理检索

1946年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(RSL)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。

与以前的手工检索相比,脱机批处理有许多优点:

- 1) 批处理可同时进行多项检索;
- 2) 可处理检索关系相当复杂的检索词汇;

3) 一次输入作业, 生产多种输出的多种服务能力。它可用于生产普通印刷索引、专题书目、回溯检索和定题检索。

但批处理也有以下缺点:

- 1) 用户不能在检索过程中与主机进行“对话”和浏览文献;
- 2) 不能在检索的同时修改检索策略, 用户必须事先把可能的途径都考虑周全;
- 3) 不能及时获得检索结果;
- 4) 批处理是委托式检索, 信息需求和查询结果之间有一定误差。

(3) 联机检索

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。

1962 年, 美国麻省理工学院的技术情报加工系统 (TIP) 开始创建, 1964 年建成。1965 年, 系统发展公司 (SDC) 在高级研究计划局的支持下进行了第一次全国规模的联机检索网的表演; 为 13 个单位提供 20 万份国外文献的检索, 这也就是今天 SDC 联机情报检索系统的前身。1967 年以后, 许多联机检索系统相继出现。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。1970 年洛克希德火箭公司的 DIALOG 系统和系统发展公司的 ORBIT 系统相继建成。美国的 MEDLARS 也于 1970 年发展了联机系统, 成为 MEDLINE。欧洲航天文献中心经过较长时间的试验和发展, 于 1969 年初步建成联机检索系统, 是欧洲空间组织情报检索中心 ESA/IRS 的前身。随着计算机硬件的发展, 各系统的规模不断扩大, 技术装备不断改进, 文献的存储、编辑、排版、检索过程都实现了自动化。

联机检索有以下几个特点:

- 1) 用户通过检索终端与通信网络直接与远程中央计算机相连, 检索远程数据库内文献信息, 这几乎是同时的、直接的, 无需委托;
- 2) 检索过程是“人一机对话”式, 可及时修改检索策略, 及时显示、浏览文献信息;
- 3) 可根据用户的不同需求进行各种输出, 及时取得检索结果;
- 4) 检索指令复杂, 多为专业检索人员才能熟练使用。90 年代后联机检索随着计算机、数据库、通信网络技术的发展而迅猛发展。

(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 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, 倍受人们的青睐。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。

光盘检索系统除可提供追溯检索、定题服务外, 还可用于“自建库”和作联机检索预处理。光盘检索的费用大大低于联机检索。

概括起来, 光盘检索主要有以下特点:

- 1) 在通信不发达地区、联网较困难的地区, 使用网络信息服务比较困难。而使用光盘

检索,可以免去联机检索必须支付的联机系统使用费,只需支付数据库生产者收取的数据库费即可。

2) 有些出版商出于保密、版权及控制等方面的考虑,还不可能将收费信息产品很快上网。在这种情况下,CD-ROM 就成为获取此类信息的最佳途径。

3) CD-ROM 可以替代或补充图书情报机构的印刷本收藏,其多媒体的书籍、游戏和参考资料是用户最受欢迎的类型之一。

(5) 网络信息检索

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 年美国提出 NII (国家基础设施建设) 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,每个节点联接多个检索终端,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作息资料,使人类实现信息资源共享成为可能。除了传统的文献作息,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 信息检索的概念与类型

1. 信息检索的含义

1950 年,莫尔斯 (Calvin N. Mooers) 首次提出信息检索 (Information Retrieval) 一词。其后,随着信息检索理论和实践的更新发展,人们对信息检索的认识也在不断深入。对于信息检索,主要存在时间性通讯、信息处理和文献查找三种角度的认识。

(1) 时间性通讯角度的认识

莫尔斯在 1950 年发表的《把信息检索看做是时间性的通讯》一文,不仅首次提出了信息检索这个概念,并认为“信息检索是一种时间性的通讯形式”。换言之,通过信息检索得到了一些文献,从而使得著者与读者(信息检索的用户)之间建立起了一种通讯。

按照这种通讯角度的认识,莫尔斯强调了在通讯双方中,信息发送者必须尽可能发送一切信息,是时间性通讯的被动一方,而信息接受者是主动活跃的一方,正是接受者才决定什么时候接受以及接受什么信息。因此,信息检索的问题就在于,如何把一个可能的用户指引向所储存的信息。

这种认识突出了用户需求的重要性,对于如何做好信息检索服务工作,至今依然具有很

强的理论和实践意义。

(2) 信息处理角度的认识

从信息处理的角度来看,信息检索的基本问题,是如何处理信息和信息的结构。这种认识偏重于信息管理,认为信息不仅限于文献的范围,图像、声音、数据等也都能反映信息,并把信息检索视为计算机科学技术的一个分支。

把信息检索看做是一种信息处理的认识,强调了如何构造以及利用什么形式来构造信息结构的问题。在当今因特网迅速发展,网络信息浩如烟海的情况下,这种认识对于信息检索工具的设计和组建,具有指导意义。

(3) 文献查找角度的认识

简言之,从这种角度来看,信息检索就是查找出含有用户所需信息的文献的过程。在信息检索领域,这是一种传统的主流观点,支持者众多。例如:英国著名学者维克利(B.C.Vickery)认为,“信息检索是从汇集的文献中选出特定用户在特定时间所需信息的操作过程”;美国著名信息专家兰卡斯特(F.W.Lancaster)认为,“信息检索是查找某一文献库的过程,以便找出那些某一主题的文献”。

关于这种认识,兰卡斯的经典表述是“信息检索系统并不检索信息”。因为信息是无形的,必须依附于文献而存在。虽然信息检索的最终结果是满足用户的信息需求,但检索的直接对象还是文献,当用户阅读文献并理解其内容时,用户的信息需求才被满足。

之所以产生上述三种不同的认识,主要有两方面的原因。一方面,认识任何事物都可以有多种角度,由此产生各种不同看法也很正常;另一方面,信息检索本身的技术发展,也使人们对于信息检索的认识不断发生变化。

实际上,信息检索具有广义和狭义两重含义。广义地讲,信息检索包含信息储存和信息查找两个过程。信息储存是对文献进行收集、标引及著录,并加以有序化编排,编制信息检索工具的过程;信息查找是从大量的信息中查找出用户所需的特定信息的过程。实施检索的主要方法就是利用各种检索工具。狭义地讲,信息检索仅仅指信息查找的过程。

由于文献是信息的主要来源,无论用户所需的是数据、事实,还是关于某一主题的信息,它们都必须从各种类型的文献中获取,所以,就这种意义而言,信息检索与文献检索基本上是等同的。

2. 信息检索的类型

由于用户的信息需求多种多样,信息检索技术也在不断发展变化,进而产生了多种类型的信息检索。

(1) 按检索对象区分

按照检索的查找对象,信息检索分为数据、事实及文献检索。

数据检索:以数据作为检索对象,查找用户所需要的数值型数据。其检索对象包括各种调查数据、统计数据、特性数据等。例如:查找某一企业的年销售额、某一国家的人口数、物质的属性数据等。

事实检索:以事实作为检索对象,查找用户所需要的描述性事实。其检索对象包括机构、企业或人物的基本情况等。例如:查找某一企业的名称、地址、业务经营范围,查找某一人物的生平。

文献检索:以文献作为检索对象,去查找含有用户所需信息内容的文献。其检索对象是

包含特定信息的各类文献。例如：查找有关“信息组织”的文献、有关“现代企业制度的建立”的文献等。

其实，对于数据检索或事实检索而言，由于用户所需的数据或事实不能脱离文献而单独存在，这些数据或事实也是反映信息的一种形式，或者说是反映信息的一种特殊形式。因此，数据或事实检索也是以文献为依托的，换言之，它们也可以被视为文献检索的特例。它们与文献检索的差异主要在于：对于数据或事实检索，检索人员只需把有关的数据或事实提供给用户，就能直接满足用户的信息需求；而对于文献检索，检索人员通常是把查找出来的相关文献提供给用户，由用户阅读、理解、吸收和利用文献的信息内容，从而满足其信息需求。

（2）按检索方式区分

按照检索的操作方式，信息检索分为手工信息检索和机器信息检索。

手工信息检索：以手工操作的方式，利用检索工具书进行信息检索。手工信息检索是信息检索的传统方式，已经历了一个多世纪的发展历程。其优点是便于控制检索的准确性，缺点是检索速度慢、工作量较大。

机器信息检索：以机械、机电或电子化的方式，利用检索系统进行信息检索。机器信息检索是从本世纪 40 年代以后逐渐发展起来，电子计算机诞生之后，以强大的储存能力、不断提高的处理性能以及同步降低的价格，很快便成为机器信息检索的主流和代表。因此，机器信息检索主要就是指计算机信息检索。其优点是检索速度快、能够多元检索、检索的全面性较高；相对手工信息检索而言，其缺点主要是需要借助相应的设备进行检索。

计算机信息检索是在手工信息检索的基础上发展而来的，随着计算机技术和通信技术的不断创新，将来必定会逐步取代甚至是完全取代手工信息检索。但是，就目前而言，把两者之间的关系，简单地理解为取代与被取代的关系，未免有失偏颇。

信息检索作为信息服务业的重要领域，在相当长的时期内，为了满足用户的多种需求，给予用户更多的选择，向用户提供多样化的服务形式，使得手工信息检索还将与计算机信息检索共同存在，相互补充、相互促进。

（3）按检索要求区分

按照用户对检索的要求，信息检索分为强相关检索和弱相关检索。

强相关检索：强调检索的准确性，向用户提供高度对口信息的检索，也称为特性检索。

这种检索注重查准，只要检索得到的文献信息能够满足用户的需求即可，通常对于检索结果的数量多少不作要求。

弱相关检索：强调检索的全面性，向用户提供系统完整信息的检索，也称为族性检索。

这种检索注重查全，要求检索出一段时间期限内有关特定主题的所有信息，为了尽时能避免漏检相关信息，一般对于检索的准确性要求较低。

要注意的是，这是两种检索要求相当极端的检索类型。实际上，较多的时候，用户对于检索的要求介于两者之间，既要求查找对口的信息，又希望得到所有的信息。这是用户很自然的检索要求，可分为两种情况：一是虽然用户对查准和查全都有所要求，但要求都较低，要完成这样的简单检索（brief search）比较容易；二是用户不仅对查准和查全都有要求，而且要求都较高，要完成这样的复杂检索（complex search）相当困难。因为在信息检索中，查准和查全常常不能兼顾。

(4) 按检索性质区分

按照检索的运行性质, 信息检索分为定题检索和回溯检索。

定题检索: 查找有关特定主题最新信息的检索, 又称为 SDI 检索。其特点是只检索最新的信息, 时间跨度小。定题检索在文献信息库更新时运行, 即每当文献信息库加入新的文献信息时, 就运行一次定题检索, 从而查找出特定主题的最新信息。这种检索非常适合于信息跟踪, 以便及时了解有关主题领域的最新发展动态。因此用户一旦向检索服务机构订购定题检索, 一般就会在较长时间内多次运行, 由检索服务机构持续地向用户提供最新信息。

回溯检索: 查找一段时期内有关特定主题信息的检索, 也称为追溯检索。其特点是既可以查找过去某一段时间的特定主题信息, 也可以查找最近的特定主题信息。与每个定题检索需要多次运行有所不同, 每个回溯检索一般只运行一次, 从已有的文献信息库中查找出特定主题的信息, 并提供给用户。

在检索实践中, 用户利用最多的是回溯检索, 大多数的检索课题都属于回溯检索。同时, 定题检索也发展得很快, 受到了各种领域用户的欢迎, 尤其是在工商经贸领域, 定题检索更是受到企业用户的广泛运用。

(5) 按检索的信息形式区分

按照检索的信息形式, 信息检索分为文本检索和多媒体检索。

文本检索: 查找含有特定信息的文本文献的检索, 其结果是以文本形式反映特定信息的文献。这是一种传统的信息检索类型, 在信息检索中依然占据着主要地位。

多媒体检索: 查找含有特定信息的多媒体文献的检索, 其结果是以多媒体形式反映特定信息的文献, 如图像、声音、动画、影片等, 是在网络环境下发展起来的全新检索类型。

在因特网迅速发展的今天, 网上存在着大量的多媒体文献, 用户常常需要查找特定的图像、声音、动画等。多媒体文献的信息处理, 如标引、著录和有序化编排等, 与传统的文本检索截然不同, 向信息检索方式及其理论提出了新的挑战。比如: 用户可能需要查找包含某种颜色或色彩组合的特定图像, 或者是含有特定图案的动画等。

1.3 信息检索原理

广义地讲, 信息检索包含信息储存和信息检索两个过程。信息储存是对文献进行收集、标引及著录, 并加以有序化编排, 编制信息检索的工具的过程; 信息检索是从大量的信息中查找出用户所需的特定信息的过程。而实施检索的主要方法就是利用各种检索工具 (见图 1.1)。

(1) 信息储存

信息储存的工作内容, 主要是由标引人员通过对原始文献的阅读分析, 对文献中的信息进行鉴别、提炼和浓缩, 并采用特定的方式予以整理、保存起来。它大致有如下几个步骤:

1) 选择文献。根据信息检索系统的主题、性质及任务等, 结合原始文献本身的研究水平、角度及其信息质量, 对原始文献进行适当的评价, 从中筛选出符合要求的文献。

2) 文献的概念分析。对所选文献进行仔细的主题分析, 提炼出文献所论述的内容主题, 归纳为代表文献内容的若干主题概念, 并确定这些主题概念之间的关系。

3) 词汇转换。把文献的主题概念转换为适当的文献标识 (或称为标引词), 并以这些标识来表达文献的主题内容。这种转换需要严谨地建立在两个依据之上: 一是必须以对文献的

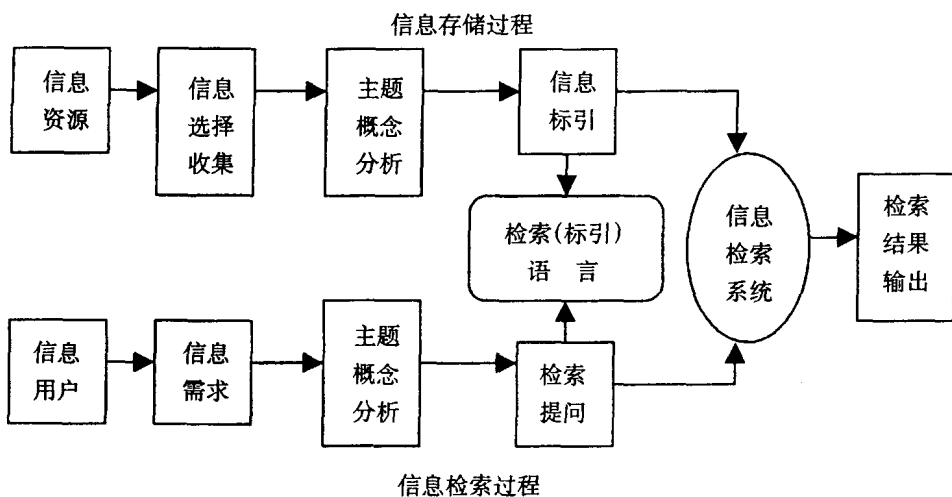


图 1.1 信息检索的原理

主题概念分析为依据，二是必须以信息检索语言为依据。前者主要决定转换什么的问题，即需要对文献中的哪些信息主题做出转换；后者主要决定怎样转换的问题，即把主题概念转换为哪些标识。

4) 信息检索工具的编制。概括地讲，检索工具是信息检索系统的核心和概括，它主要包括两个有序化的序列，即文献序列和文献标识序列。

文献序列是由文献描述体或文献本身按照一定的方式组织形成的有序化序列，构成文献库。文献描述体是对原始文献内容的浓缩，常见的有文摘、题录等，这是信息检索所采用的传统和主要的方式。其主要作用是，使用户能够对文献内容有较为全面和准确的了解，进而做出是否需要获取原始文献的选择。随着计算机技术和通信技术的发展，现在已经有越来越多的信息检索系统采用全文本的方式，直接把原始文献本身组织为有序化的序列，尤其是因特网的迅猛发展，为全文本检索拓展了更大的发展空间。

文献标识的序列，是由文献标识按照特定的顺序形成的有序化序列，构成文献库的索引。最常见的排列方式为字顺，即按照字母顺序或汉语拼音，排列为文献标识的序列。其作用主要是依靠字顺组织，提供对文献标识的快速查找，并与提问标识加以比较，据此做出文献是否与提问相符的判断。这个标识比较的过程，也称为检索的匹配。

(2) 信息检索

信息检索的工作内容，主要是由检索人员接受用户的检索提问，对提问进行细致的主题分析，提炼出用户检索的主题概念，并编制出相应的检索策略。

其工作步骤如下：

- 1) 用户提问。在特定的条件下，用户会把头脑中信息需求转变为具体的检索行为。
- 2) 提问的概念分析。分析检索提问，识别检索的真正主题内容，把检索主题分解为若干概念，并明确这些概念之间的关系。
- 3) 词汇转换。把检索提问的主题概念转换为相应的提问标识（或称为检索词），并以这些标识来表达检索提问的主题内容。其依据同样有两个方面：一是对提问的主题概念分析，二是信息检索语言。
- 4) 检索的实施。根据所得到的提问标识，在文献标识序列中，按照其排序的规则，迅速地进行查找，并对文献标识与提问标识进行匹配比较。如果文献标识与提问标识相同，那