普通高等教育“十一五”国家级规划教材

高等学校图书馆学专业系列教材

第二版

信息检索

黄如花 主编

LIBRARY SCIENCE

WUHAN UNIVERSITY PRESS
武汉大学出版社

普通高等教育“十一五”国家级规划教材

高等学校图书馆学专业系列教材

第二版

信息检索

主　编　黄如花

副主编　林　佳　胡永生

编著者　黄如花　林　佳　胡永生　徐　轩　苏小波
韩　丽　陈振英　吴文辉　陈　朋

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

信息检索/黄如花主编. —2 版. —武汉:武汉大学出版社,2010.5
普通高等教育"十一五"国家级规划教材
高等学校图书馆学专业系列教材
ISBN 978-7-307-07671-6

Ⅰ.信… Ⅱ.黄… Ⅲ.情报检索—高等学校—教材 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2010)第 051624 号

责任编辑:严 红 周 亚 责任校对:刘 欣 版式设计:詹锦玲

出版发行:**武汉大学出版社** (430072 武昌 珞珈山)
(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.com.cn)
印刷:武汉中科兴业印务有限公司
开本:720×1000 1/32 印张:17 字数:300 千字 插页:1
版次:2002 年 4 月第 1 版 2010 年 5 月第 2 版
2010 年 5 月第 2 版第 1 次印刷
ISBN 978-7-307-07671-6/G·1555 定价:35.00 元(含一张 CD-ROM 光盘)

前　言

在信息社会里，信息素养对个人的学习、生活、工作与研究具有重要作用。2009 年 10 月 1 日，美国总统奥巴马签署了代号为 3195—W9—P 的“国家信息素养宣传月”(National Information Awareness Month) 的议案，指出：“我们还必须学习掌握在任何情况下获取、整理和评价信息的必要的技能，而非仅仅拥有数据”，“国家信息素养宣传月突出了所有美国人需要掌握擅长于在信息时代航行的必需技能。”该议案将信息素养的重要性提到了相当的高度。

信息素养是可以培养的，最直接、有效的途径是信息检索有关课程的学习。本书定位于信息检索教材，旨在从信息检索的基本概念、原理与方法出发，介绍国内外重要的综合性信息检索系统与专业性信息检索系统，阐述多媒体信息检索、专利等专题信息检索的技巧以及信息资源选择与评价的方法，分析了信息检索与利用中的费用与知识产权问题，介绍了个人文献管理软件、自动问答系统与自动翻译系统的使用方法，对信息检索的重要应用——科技查新的有关问题进行了论述。

本书系教育部“十一五”规划教材，是《网络信息检索与利用》(武汉大学出版社，2002 年) 的修订版。后者自 2002 年出版以来已重印 10 次，已被有的兄弟院系作为本科生教材或硕士生入学考试指定参考书，被 20 余所大学和省级图书馆作为文检课教学、培训、职称考试和业务竞赛的教材或指定用书。

本教材由黄如花负责全书的策划和提出编写大纲，林佳、胡永生参与策划并对大纲提出修改建议。各章节撰写的具体分工情况是：第一章、第二章的第一节与第三节、第五章、第八章、第九章的第一节与第四节、第十章和附录由黄如花编写，第六章第一、二、四节由林佳编写，第四章、第六章的六至九节由胡永生编写，第三章和第九章的第二、三节由徐轩编写，第七章由苏小波和陈朋编写，第六章第三、五节由陈振英编写，第二章第二节由韩

丽编写，第六章第十节由吴文辉编写。全书由黄如花审稿和统稿。武汉大学信息管理学院研究生宋琳琳、熊惠霖、刘鎏、张路漫参与部分章节的资料收集工作。

针对信息检索技术的新变化和用户获取信息的变化，也得益于广大读者提出的宝贵意见，本版有以下几个主要变化：重构了教材体系，针对信息检索工具的“集成化”和用户获取信息的一站式需求，去掉了初版中电子图书、电子期刊、电子报纸、书目信息与全文信息检索等章，将其内容整合到国内外重要的综合性信息检索系统和科技查新中；增加了更多信息检索原理、模型等理论知识和信息检索的应用，增加了新的信息检索系统或检索工具以及原有检索工具的新功能，每章后补充了思考题；删除了一些生活性强的内容，如旅游、求职与招聘、英语学习等信息的检索；补充了多个与信息检索的教学、研究和应用相关的附录，使教材的系统性更强、内容更新颖、更有针对性和参考价值。为使本书更好地与信息检索实践和文献检索课程的需要相结合，特邀清华大学、武汉大学、复旦大学、浙江大学、北京师范大学、中南财经政法大学和中南民族大学等高校图书馆从事参考咨询和文献检索课程教学的同行参编。同时，为了便于广大文献检索课教学人员的备课，本教材附上配套的课件光盘而非教材的电子版，并拟通过本书主编主讲的“信息检索”课程网站（http：//jpkc. whu. edu. cn/jpkc2010/xxjs）提供本领域最新资料与授课录像，使之成为真正的立体化教材。

本书在写作过程中广泛吸取了国内外大量相关研究成果，在此，谨向这些文献的作者致以诚挚的谢意！本书的出版得到了武汉大学出版社的帮助，责任编辑严红编审为本书的出版付出了辛勤劳动，提出了许多宝贵的修改意见，在此谨表谢忱。

信息检索技术和检索工具日新月异，加之编著者学识、水平有限，书中疏漏和不妥乃至错误之处在所难免，敬祈专家学者和读者批评指正，以便今后的修订和补充。

目　录

图表目次

第一章 信息检索概述

第一节 信息检索的基本概念

一、信息检索的含义

信息检索作为一种实践活动由来已久。但作为一个比较规范、正式的学术术语，信息检索（Information Retrieval，简称 IR）这个术语 1950 年由美国信息科学的先锋 Calvin Northrup Mooers（1919—1994）① 首先提出，这也是他 1978 年获得美国信息科学协会荣誉奖（American Society for Information Science's Award of Merit）的原因之一②。

广义的信息检索是指将信息按一定的方式组织和存储起来，并根据信息用户的需要找出有关信息的过程。所以，它的全称又叫信息存储与检索（information storage and retrieval），即包括信息的“存”与“取”两个环节。广义信息检索的其他表述有：信息检索是对信息项（information items）进行表示（representation）、存储（storage）、组织（organization）和存取（access）③。

狭义的信息检索则仅指该过程的后半部分，即从信息集合中找出所需信息的过程，相当于“信息查询”或“信息查找”（information search）。“信息

① *Calvin Northrup Mooers* [EB/OL]. [2009-11-28]. http://web.utk.edu/~alawren5/mooers.html

② *Award of Merit* [EB/OL]. [2009-11-28]. http://www.asist.org/awards/award_of_merit.html

③ Ricardo Barza Yates, Berthier Ribeiro Neto, et al.. *Modern Information Retrieval* [M]. Boston: Addison Wesley Longman Publishing Co. Inc., 1999: 1

检索的含义很广，但作为一个学术研究领域，可界定为：信息检索是从文档集合（通常存储在计算机中）查找满足某种信息需求的具有非结构化性质（通常指文本）的资料（通常是文献）。"① 可见，这也是从狭义的角度界定的。在通常情况下，人们讲"信息检索"是从狭义的角度而言的。

信息检索与文献检索的主要区别在于：文献检索是以获取文献信息为目的的检索，信息检索则收集、组织、存储一定范畴的信息，并可供用户按需要查询文献中的信息或知识单元，比文献检索更深入。

二、信息检索的种类

根据检索手段的不同，信息检索可分为手工检索、光盘检索、联机检索和网络检索。网络检索是信息检索的发展方向，因而本书以网络检索为主。

根据检索对象形式的不同，信息检索又可分为：

文献型信息检索（document retrieval）：是以文献（包括题录、文摘和全文）为检索对象的检索。凡是查找某一主题、时代、地区、著者、文种的有关文献，以及这些文献的出处和收藏处所等，都属于文献型信息检索的范畴。完成文献型信息检索主要借助于各种书目型数据库。

数值型信息检索（data retrieval）：是以数值或数据为对象的一种检索，包括文献中的某一数据、公式、图表，以及某一物质的化学分子式等，数据检索分为数值型与非数值型。完成数据型信息检索主要借助于各种数值数据库和统计数据库。

事实型信息检索（fact retrieval）：是以某一客观事实为检索对象，查找某一事物发生的时间、地点及过程的检索，其检索结果主要是客观事实或为说明事实而提供的相关资料。完成事实型信息检索主要借助于各种指南数据库和全文数据库。

关于从检索对象的角度对信息检索类型的划分，出现了一种新的三分方法，即文本检索、数值检索、音频与视频检索。相对于早期对信息检索概念的细分方法，新的三分法比较全面地反映了信息检索概念的基本内涵和最新发展②。

---

① Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *An Introduction to Information Retrieval* [M]. Online edition. Cambridge : Cambridge University Press, 2009: 1

② 赵丹群编著．现代信息检索：原理、方法与技术［M］．北京：北京大学出版社，2008：2

三、信息检索的原理

信息检索的基本原理是：通过对大量的、分散无序的文献信息进行搜集、加工、组织、存储，建立各种各样的检索系统，并通过一定的方法和手段使存储与检索这两个过程所采用的特征标识达到一致，以便有效地获得和利用信息源。其中，存储是检索的基础，检索是存储的目的。文献信息的存储和检索的全过程可用图 1-1 表示。要完成这种匹配与选择，要做好以下三个方面的工作。

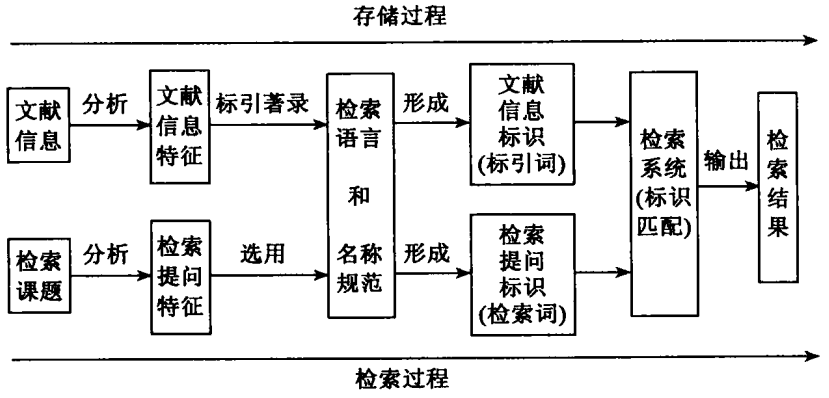

图 1-1　信息检索的基本原理示意图

1. 文献替代

即将表示文献资源特征的元数据替代它指代的资源，文献替代过程实际上是对原始文献的外表特征（包括题名、著者、出处等）和内容特征（包括分类号、主题词、摘要等）进行描述的过程，这项工作通常称为著录，著录的结果是将原始文献制成它的替代文献——二次文献。

2. 文献整序

文献整序就是对替代文献进行标引，给出文献标识（如分类号、主题词等），将所有替代文献按其标识进行有规律的组织排列，形成可检索的信息资源集合。

3. 文献特征标识与检索提问标识的匹配

检索者在查找所需文献时，只要以该系统所用的标识作为提问标识，与系统中的文献特征标识进行比较，并将文献特征标识与提问标识一致的文献线索从检索系统中检出，检出的部分就是检索的结果。

四、信息检索语言

为了使检索的过程，即文献标识和信息提问者的对比进行得顺利，两者都需要用一定的语言（即检索语言）来表达。检索语言的主要目的就是把信息的存储与检索联系起来，把标引人员与用户联系起来，以便取得共同理解、实现交流。信息检索语言是人们在加工、存储和检索信息时用来描述信息内容和信息需求的词汇或符号及其使用规则构成的供标引和检索的工具。

第二节　信息检索的历史

最早的信息检索主要是依靠信息分类。早在2000多年前，我国的汉代就有简单的图书分类法，如《七略》。随着社会的发展，信息量越来越大，简单的分类已不能完全解决快速查找有用信息的问题，特别是随着科技期刊的大量出现，对于大多数人来说，已没有时间将所有期刊上的所有文献都阅读或浏览一遍，而且就一个读者来说，一本期刊中也不可能每篇文献都有阅读价值。因而出现了文献索引，读者可以根据自己的需要查找相关文献。书目、索引、文摘、目录等检索工具也不断出现。这些印刷版的工具主要根据文献的内、外部特征，从题名、著者、主题词等途径提供手工检索。

信息检索技术经过先组式索引检索、穿孔卡片检索、缩微胶卷检索、脱机批处理检索发展到今天的联机检索、光盘检索、网络检索，其发展经历了由低级到高级的过程，检索技术也从传统的线性检索向超文本支持的非线性检索发展。现在是手工检索、联机检索、光盘检索、网络检索并存，但以网络检索为主，网络检索也最有发展前景。

一、手工检索（1876—1945）

信息检索起源于参考咨询工作。读者被要求独立使用图书馆提供的书目和索引工具，查询所需的文献和情报，这时“信息检索”作为一项行为已经出现，只是具有分散性和非专业性，而且缺乏必要的重视和研究，未能形成专业化的情报检索系统。正规的参考咨询工作由美国的公共图书馆和大专院校图书馆于19世纪下半叶首先发展起来。20世纪初，多数图书馆成立了参考咨询部门，主要利用图书馆的书目工具来帮助读者查找图书、期刊或现成答案。索引成为独立的检索工具，书目、文摘开始编制并用于专题文献检索。“信息检索”从此成为一项独立的用户服务工作，并逐渐从单纯的经验工作向科学化方向发展。

手工检索操作简单、费用低廉、查准率高，但效率很低，查全率不能保证。随着科学技术的发展，文献信息在不断增加。传统的利用印刷型文献进行手工检索的方式已不能适应信息的急剧增长，更跟不上时代发展的步伐。

二、机械信息检索（1945—1954）

机械信息检索系统是20世纪50年代开始的用各种机械装置进行情报检索的机械系统，是手工检索向计算机信息检索的过渡阶段。1954年，现代情报学创始人美国的万尼瓦尔·布什（V. Bush）博士在“As we may think”一文中首次提出利用机械、电子技术实现情报检索的设想。他描述了一种叫做“Memex”的机器，用于非线性检索。他与美国农业部图书馆馆员拉尔夫·肖共同制造了一台快速检索机——布什·肖检索机。它利用光电原理，对复制在胶卷上的文档进行检索。胶卷的边缘上有黑白点作编码，当遇到检索内容时就停下来。

机械信息检索系统利用当时先进的机械装置改进了信息的存储和检索方式，通过控制机械动作，借助机械信息处理机的数据识别功能代替部分人脑，促进了信息检索的自动化。但它并没有发展信息检索语言，只是采用单一的方法对固定的存储形式进行检索的工具，而且过分依赖于设备，检索复杂，成本较高，检索效率和质量都不理想。机械信息检索系统很快被迅速发展的计算机情报检索系统取代。

三、脱机批处理检索（1954—1965）

自1946年第一台计算机问世不久，信息工作者就将这一新的技术与信息工作相结合，逐步建立了一种崭新的以计算机为核心的现代化信息系统。将计算机用于书目信息检索最早是在20世纪50年代提出来的①，1954年美国海军军械实验中心利用IBM701机将有关海军军械的4000篇技术报告进行了计算机存储与检索的试验，建立了世界上第一个计算机文献信息检索系统。

脱机信息检索系统是计算机检索初期使用的一种检索系统。它是利用单台计算机的输入输出装置进行检索，用磁带作存储介质的系统。使用该系统查找文献时，计算机只能顺序检索磁带上记录的信息，每检索一次都必须从头到尾读一遍磁带，很费时间。因此，不得不以批处理方式来实施检索。亦

① Charles P. Bourne. On-line Systems: History, Technology and Economics [J]. Journal of the American Society for Information Sience, 1980, Vol. 31, No. 3: 155-160

即由系统工作人员集中一批用户的信息要求，预先制定好检索策略，以机读形式存储在检索系统的计算机存储器中，定期地检索数据库新增加的内容，然后把命中的文献信息分发给用户。由于在检索过程中用户不直接与计算机接触，因此称之为脱机检索或定题检索，所用的系统称为脱机检索系统。

脱机批处理信息检索存在3点不足：一是地理上的障碍，用户与检索人员距离较远时，不便于检索要求的表达和检索结果的获取；二是时间上的迟滞，检索人员定期检索，用户不能及时获取所需信息；三是封闭式的检索，检索策略一经检索人员输入系统就不能更改，更不能依据机检应答来修改检索式。

四、联机检索（1965—1991）

联机检索系统是脱机检索系统的进一步发展，进入20世纪60年代后，随着计算机磁盘存储介质的出现，以及通信、操作系统等技术的发展，人们可以在磁盘上建立起可以直接存取的随机文件，建立了一台主机通过通信线路带动多个终端的联机检索系统，检索者借助检索程序通过终端以“即问即答”的方式与计算机对话，查找所需的信息。在检索过程中，用户还可以浏览有关信息，随时修改提问，直至得到满意结果。这种系统具有分时的操作能力，能够使许多相互独立的终端同时进行检索。联机检索系统主要由用户终端、通信网络、计算机及数据库组成。

20世纪60年代中后期，对联机信息检索进行研究开发试验。1965年，美国系统发展公司（SDC）研制成功联机信息检索软件——书目信息分时联机检索（Online Retrieval of Bibliographic Information Time Shared，ORB IT），标志着联机信息检索系统阶段的开始。1966年，美国洛克希德导弹与宇航公司（Lockheed Missile & Space Company Inc.）研制了第一个“人-机”对话的信息检索系统，即著名的Dialog系统，正式开展文献检索。

1969年欧洲空间研究组织（European Space Research Organization，ESRO）建立了ESA-IRS系统。

进入20世纪70年代后，联机检索基本上结束了自己的内部实验性应用，开始投入商业化运营，面向社会公众提供服务。

20世纪80年代以后，随着空间技术的发展，信息检索进入了一个信息—卫星通信—计算机三位一体的新阶段，即国际联机检索阶段，也有人称之为网络时期。1983年，美国、联邦德国和日本共同开发创建了国际科技信息网络（The Scientific and Technical Information Network，STN）。国际联机信息检索使信息检索超出了一个地区、一个国家的范围而进入了国际信息

领域，促进了全球信息资源的共享。Dialog 系统在 1980 年就已经向 40 多个国家的 1000 多个用户终端提供 100 多个数据库的国际联机信息检索服务。

国际联机信息检索是指商业性的计算机数据库检索服务机构（亦称联机卖主）通过国际（卫星）通信网络，为世界各地的用户终端提供人机对话式的检索的服务方式；用户则利用终端设备，通过国际（卫星）通信网络，与世界上任何国家的大型计算机检索系统的主机联结，从而检索世界各国存储在计算机数据库中的信息资料。

国际联机检索的优点有：检索速度快，效率高；检索范围广泛、全面；检索途径多，方便、灵活；检索内容新，实时性强；检索辅助功能完善（人机对话、检索结果输出方式灵活，输出格式多样等）。其缺点有：检索费用高；对检索系统及其文档（数据库）的收录、标引、特点等问题较难了解、熟悉；检索技术和技巧不易掌握。

其间，由于国际联机信息检索费用高，出现了光盘检索。光盘即高密度光盘（Compact Disc），是不同于磁性载体的光学存储介质，用聚焦的氢离子激光束处理记录介质的方法存储和再生信息，又称激光光盘。光盘按功能可分为 3 类：只读式光盘（Compact Disc-Read Only Memory，CD-ROM）、写读光盘（Write Once Read Memory，WORM）和可擦写光盘（Optical Random Access Memory，ORAM）。

光盘数据库的检索软件及数据装在盘片上，任何一台安装有光驱的 PC 机，只要装上光盘数据库，即可成为光盘数据库检索系统。光盘检索系统简单易学，投资低，并且不需支付通信费，因此，虽然它比联机检索系统的出现晚十余年，但目前在我国的普及率要大大高于联机检索。但光盘数据库更新速度慢，因此，还不能完全代替国际联机检索。

光盘检索的优点有：光盘存储容量大而体积微小；使用方便，不需通信联系，不受时间限制；使用方便、易于操作；价格低（一次购买，无限次检索，也不需要昂贵的联机检索通信费用）；使用寿命长，用户易接受；机房无特别要求，投资少，要求设备简单，可随地安装。光盘检索的缺点表现在：信息获得比国际联机慢（回溯检索须多次换盘）；信息更新不及时。

我国于 1966 年引进法国布尔（Bull）计算机，着手情报检索的试验和研发。1974 年 8 月在国务院直接领导下，国家 748 工程（汉字信息处理系统工程）全面启动，“汉字信息处理情报检索系统”则由中情所、国防科委情报所、四机部情报所、北京图书馆等联合组织攻关。《汉语主题词表》编制和“机器翻译”研究工作由此开始；1975 年 10 月我国正式加入世界科学