普通高等教育“十一五”国家级规划教材

马文峰 著

XINXI JIANSUO JIAOCHENG

信息检索教程

国家图书馆出版社

普通高等教育“十一五”国家级规划教材
“十一五”规划高等学校核心课程教材

信息检索教程

马文峰　著

国家图书馆出版社

图书在版编目(CIP)数据

信息检索教程/马文峰著. —北京:国家图书馆出版社,2009.2
ISBN 978-7-5013-3969-3

Ⅰ.信… Ⅱ.马… Ⅲ.情报检索—教材 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2009)第 006869 号

书名 信息检索教程
著者 马文峰 著

出版 国家图书馆出版社 (100034 北京西城区文津街 7 号)
发行 010-66139745 66151313 66175620 66126153
66174391(传真) 66126156(门市部)
E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)
Website www.nlcpress.com
经销 新华书店
印刷 河北三河弘翰印务有限公司

开本 787×1092 毫米 1/16
印张 13.25
版次 2009 年 2 月第 1 版 2009 年 2 月第 1 次印刷
字数 300(千字)

书号 ISBN 978-7-5013-3969-3
定价 36.00 元

前　言

21世纪以来，信息检索的理论与实践有了长足的发展，主要体现在：随着信息环境的变化，获取信息的途径和手段发生根本变革；信息检索技术日渐成熟并不断应用于实践，促进信息检索水平和效益的提高；在实践的基础上，学术界积极开展理论研究，出版发表了相当数量的教材、专著和论文。与此同时，信息检索也显示出一些新的特征：各种类型的数据库逐渐替代印刷型工具书，成为主流的检索工具；计算机检索和手工检索、文科信息检索和科技信息检索的理论、方法与技术逐渐融合。

本书参考了国内外大量研究成果、相关资料和检索系统，将信息检索的全过程作为研究对象，从理论和实践两个层面阐述了信息检索的机理、方法、技术与应用，力图做到内容全面、系统、新颖、深入和实用。

本书分为文字和光盘两部分。

文字内容分为13章。

1至5章是基本理论部分。第1章、第2章论述了信息检索的基本概念、基本原理和基本方法。第3章、第4章、第5章分别概述了三大检索系统——工具书、数据库、搜索引擎的原理、结构、类型与功能，揭示它们产生、演变与发展的历史，使读者对信息检索基础知识和不同类型的检索工具有个整体性了解，以指导对各类检索工具的利用。

6章至12章是实践部分。主要介绍各类学术信息源——图书、期刊论文、学位论文、会议论文、报纸文章、术语信息、事实数值信息、专利与标准、网络学术资源的检索与利用。鉴于数据库是目前各类信息获取的主要途径，所以在实践部分，侧重介绍检索各类信息的重要数据库，内容包括收录范围、检索方式和检索规则等，同时介绍与之对应的重要、常用的印刷型检索工具，并列举出检索该类信息的其他数据库的基本情况；对于目前无法完全通过数据库获得的信息，则以介绍工具书为主。

13章主要阐述如何在信息检索理论与实践的基础上，设计并撰写出一篇有价值的、规范的学术论文，重点说明学术论文设计与撰写过程中需要注意的一些主要问题，以及应该遵循的学术规范。

光盘内容分为31部分。主要包括各类手工工具书介绍及相关图片，各类数据库的主要检索界面，以及文字部分相关内容的补充。

信息检索的理论与实践中尚有许多值得深入探讨的问题，因时间和学识所限，难免有缺漏错误，恳请读者赐教。

国家图书馆出版社的金丽萍女士和责任编辑王涛先生为本书的编辑出版付出辛勤的劳动，在此深致谢意。

作　者

2008年6月于中国人民大学

目　录

第1章　信息检索概述

1.1　信息与信息检索

1.1.1　信息与文献信息

1.1.1.1　信息

(1)信息的含义

信息一直是被多学科、多领域研究的一个基本问题。由于信息涉及的领域广、内容丰富，人们的研究点不同，对信息的界定也不尽相同。但一般可以概括为广义和狭义两种表述。

●广义的信息概念。广义的信息指的是事物运动的状态与方式。这一定义中“事物”泛指一切可能研究的对象，包括外部世界的物质客体，也包括主观世界的精神现象；“运动”泛指一切意义上的变化，包括机械运动、物理运动、化学运动、生物运动、思维运动等；“运动方式”是指事物运动在时间上所呈现的过程和规律；“运动状态”则是事物运动在空间上所展示的态势。广义的信息概念实际上是哲学意义上的信息的定义，其特征是不受任何条件的约束，不以人的意志为转移，具有最广泛的适应性和高度的抽象性及概括性。

●狭义的信息概念。狭义的信息概念主要指进入认识领域和传播领域，可以被理解或被接受的消息、情报、知识、事实、数据等。这一定义包括以下要点：其一，信息是指已被人类认识、理解、开发利用的信息。自然界、人类社会时时会生成信息，但未进入人的认识领域、未被使用的就不能被看做信息。其二，信息是认识过程中不确定性的消除或减少。信息的作用是消除信息接受者认识过程中不确定性，即消除了他对某种情况的不了解状态，这就是信息论的奠基人香农认为的“信息就是不定性的排除”。

综上，可以将信息概括为：信息是物质的一种属性，是客观事物的存在方式或运动状态的表征与反映；就其存在领域而言，信息包括自然信息和社会信息两大类，自然信息是在自然界中传递的信息，社会信息是在社会领域内流通的信息；信息必须通过主体的认识才能被反映和揭示，信息必须通过一定的方法加以表示才能得以传播。本书所讲的信息主要指在社会领域中流通的、经过主体认知并以一定方式表示及传播的信息。

(2)信息的类型

信息可按多种形式划分。按人类对信息的认识逻辑层次，信息可划分为语法信息、语义信息和语用信息。三者是密切相关、互为作用的，反映了人们从语法、语义到语用的这一逐步深化的认识顺序和认识过程。

●语法信息。语法信息是指主体对事物运动状态和方式的直观描述，表现为一连串的符号或语言，并不涉及信息的内容解释和实际效用。语法信息是最基本和最简单的信息层次。

●语义信息。语义信息指主体对事物运动状态和方式含义的逻辑表述，也即是说要揭示信息内容真实而准确的含义，研究这些含义的表达方法。语义信息以语法信息为基础，它是从内容角度反映信息特征。

• 语用信息。语用信息即主体事物运动状态和方式含义的逻辑表述不仅要反映事物的运动状态和方式，而且要揭示其对人类的价值和效用。语用信息反映信息的功能与效用，通常说某种信息“有价值”、“有用”等，即是对信息的语用性的判断。语用信息以语法信息、语义信息为基础，是最复杂的信息层次。

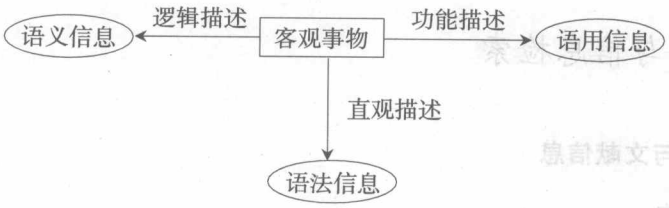

图 1-1　信息的逻辑层次

(3)信息的特征

信息具有多种特征，主要有以下几点：

1)寄载性。各种信息必须借助于各类物质形式的载体(文字、图像、胶片、磁带、磁盘、声波、电波、光波等)才能够表现，才能够被接受和共享。从某种意义上说，没有载体就没有信息。多媒体信息是各种形式信息的综合表现，它集声音、文字、图像于一身，使多种类型的信息得以集中表现。

2)可传递性。可传递性指信息可通过一定的渠道进行传输。信息传输即信息由信源(信息发送者)经信道传递至信宿(信息接受者)的交流过程。传输包括空间和时间上的传输。空间传输即信息的利用不受地域的限制，能由此及彼；时间传输即信息的传递不受时间限制，可以由古及今，信息的积累就是一种时间传输。信息在时空中交流与传输才能为人们所共享。

3)可处理性。可处理性是指信息可通过一定的方式进行加工与处理。如可以根据需求对信息进行分类、标引、组织等有序化处理；可以对信息进行筛选、浓缩、概括，以形成内容具有完整性、准确性、针对性、精炼性的信息；经过处理的信息可以被检索与利用。信息处理可以提高信息的可利用性。

4)可转换性。可转换性是指信息可利用一定的技术进行形式和内容的重新变更。如可以进行格式化转换，可复制、编辑、修改、更新，可移动存储位置等，使信息从一种形态转变为另一种形态。信息转换扩展了信息的获取与利用范围。

5)价值性。存储在某种物质载体上的信息经过加工处理或转换，就是一种资源，具有效用性。信息价值性的核心是信息可以被提炼为知识。知识是抽象化、系统化、理论化的信息，知识的主要功能是能够产生新的信息，信息既是知识的原料又是知识的载体。没有信息，就不会形成知识。信息不形成知识，就缺乏长期使用价值。

6)共享性。信息作为一种资源，可以由不同个体或群体在同一时间或不同时间共同享用。这是因为，信息传递与实物传递有本质区别，实物传递，一方有所得，必使另一方有所失。一个苹果两人分享，一人可得半个；四人分享，每人所得只有四分之一。而信息传递，不会因一方拥有而另一方失去知和用的可能，也不会因使用次数的累增而耗损信息内容。相反，信息可共享的特点，使信息资源能够发挥最大的价值效应，同时信息的共享性还使信息能够再生，并且使其增值。

1.1.1.2　文献信息

(1)文献的含义

文献一词,在我国首见于《论语·八佾》:“子曰:夏礼,吾能言之,杞不足征也;殷礼,吾能言之,宋不足征也。文献不足故也。足,则吾能征之矣。”大意是,孔子能讲解夏、殷的典章制度,但杞、宋两国的典章制度因缺乏足够的文献而无法证实。孔子所说的文献,南宋著名学者朱熹在《四书章句集注》中解释为:“文,典籍也;献,贤也。”即是说,文献一词包含典籍和贤人两种含义。典籍指历朝有关典章制度的文字资料,贤人指熟悉典籍的学者的言论。随着时代的发展,书籍文章的增多,文献中“贤”的意义逐渐消失,其含义仅侧重于“文”,专指典籍资料,内容包括一切有价值图书资料。

到了现代,随着科学技术的发展,新的文献载体材料不断涌现,信息知识的记录方式也不断增多,文献概念的外延也在不断扩大。国际标准化组织《文献信息术语国际标准》对文献的定义是:“在存储、检索、利用或传递记录信息的过程中,可作为一个单元处理的,在载体内、载体上或依附载体而存储有信息或数据的载体。”我国颁布的国家标准《文献著录总则》给出的定义是:“文献是记录有知识的一切载体。”

上述两个定义,将文献含义规定得非常广泛。可以说,现代文献的外延囊括了记录信息与知识的所有载体,不仅包括纸质载体,也包括古代的甲骨、金石、简策以及现代的胶片、磁带、光盘等。只要记载有知识、信息或数据,无论其形态如何,都可以称之为“文献”。

据此,我们可以将文献理解为:通过一定的技术手段,以文字、符号、图形、声音等方式,将知识或信息记录在某种物质载体上,以便长期保存和广泛传播的固态精神产品。图书、报刊、录音带、录像带以及因特网上的资源都可以视为文献。

(2)文献形态的演变

文献载体的发展演变,大体经历三个阶段。

第一阶段:是文字与天然实物载体结合的手写文献。

这是最早的文献载体形式。古代中国两河流域和埃及是最早产生文字的地方,通过文字将信息记录在某种实物上,就形成了文献。如我国刻写于龟甲兽骨上的甲骨文献,刻铸在青铜器或石头上的金石文献,书写在竹片或木片上的简策文献,记录于丝织品上的缣帛文献,以及外国古代的泥版书、纸莎草书、蜡版书、羊皮书等。这些文献的载体基本上是自然物的原始状态,都属非纸质文献类型;记录方式是靠手工刻划和抄写来完成。

第二阶段:由手工刻写文献转为纸质印刷文献。

即以纸为载体,以油印、石印、胶印等印刷技术记录信息和知识而形成的文献形式。纸发明于我国东汉,自晋代以后,纸逐渐成为我国和世界各国最主要的书写材料。各种天然文献载体逐步退出文献生产的历史舞台。纸质载体以其重量轻、载量大、易携带、价格低等特有的优势,独居文献载体主体地位近两千年,至今仍是传播知识信息的主要载体。这是人类信息存储和传播技术的一次质的飞跃。

第三阶段:以感光材料、磁性材料、光学材料为载体的数字型文献。

纸质文献虽然便于阅读,但存储密度低,受时空局限,严重阻碍了文献的快速传递和资源共享。所以,满足文献信息高效率需求的新型文献载体——数字文献在20世纪应运而生。20世纪初,以光学缩摄技术为记录方式,以感光材料为载体的缩微制品(缩微胶卷、缩微平片、缩微胶套)是重要的文献形式。缩微文献体积小、存储容量大,保存寿命长,是当时存储珍贵文献的首选文献形式。20世纪中期,以数字形式将图、文、声、像等信息记录在光、磁等存储介质

上，由计算机设备输入和输出的机读形数字文献开始出现。这是文献形态演化的过程中一次深刻的革命。它极大地提高了文献传播的速度和效率。

与传统纸型载体截然不同的是，数字载体是以虚拟形式在覆盖全球的网络上传递知识信息，可以远程利用，基本不受时空因素的影响；在信息存储容量、传递速度及检索的便捷性等方面，具有传统纸型文献载体所不具备的特有性能；同时，数字载体可以将知识信息通过多媒体（文字、图像、声音）形式加以传播。因此数字载体将成为传递知识信息的主流媒体，但这是一个比较漫长的演变过程。

（3）文献构成要素

文献主要由信息内容、信息符号、物质载体、记录方式和载体形态五个要素构成。前四者是基本要素，载体形态是辅助要素，是前四者的外在表现形态。

1）信息内容。即文献所记录的各种内容，例如数据、情报、知识等都是信息内容。没有信息内容就不成其为文献，信息内容是文献最主要的构成因素。

2）信息符号。也称记录符号，主要指记录和表达信息内容的标识符号，如文字、图形、代码、声频、视频等。信息符号是表达信息内容的手段。信息内容只有用具有特定含义的符号表示出来，才能为人们所识别。

3）物质载体。记录信息内容的物质材料，也即信息内容存储的依附体，或是信息内容传播的媒介体。如甲骨、金石、竹木、缣帛、纸张、胶卷、胶片、录音带、录像带、磁带、磁盘、光盘等。信息内容只有负载于一定的载体材料，才能进行传播。

4）记录方式。即信息符号所表示的信息内容被存储到载体材料上的方式。文献的记录方式主要有手刻、书写、印刷、拍摄、录制、计算机输入等。

5）载体形态。文献的载体形态即文献的信息内容、信息符号、物质载体、记录方式综合为一体的外在表现形式或呈现方式，如印刷型、声像型、缩微型、数字型等。文献只有通过一定的形态呈现出来，才能有效利用。

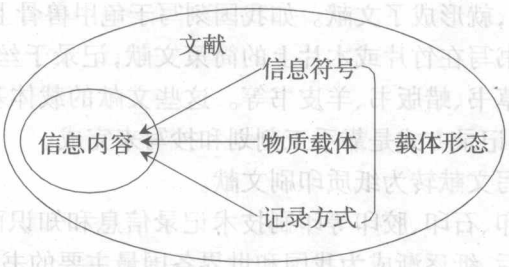

图 1-2　文献五要素

1.1.1.3　信息与文献信息的关系

（1）文献信息的含义

“文献信息”这一术语是在 20 世纪 80 年代初开始广泛使用。文献信息主要指以文献为物质载体的信息，也就是说文献信息是从文献实体结构中抽象出来的内容，它是借助于文献这种载体显示出的信息，通过文献进行存储和传播。无论来自自然界的信息，还是来自社会的信息，只要借助于文献而传递的内容，都属于文献信息。

文献信息与文献既有同一性，又有区别性。

• 两者的一致性：文献必须包含有信息，信息必须依附于一定的载体，文献信息就是文献中所记录的信息，两者在本质上没有什么区别。

• 两者的区别性：文献是一个信息实体，是文献信息的储存者；而文献信息是指文献中的信息内容进行传播交流，从而产生社会效应和思维效应的一种动态信息。人们利用文献，实质是利用文献中的信息和知识，文献信息是作为文献的价值内涵而存在。因此，文献概念侧重于物质属性，而文献信息则侧重于信息属性、价值属性。两者的关系见图 1－3。

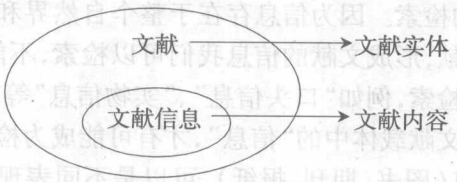

图 1-3　文献与文献信息的关系

(2)信息与文献信息(知识、情报)的异同

文献信息是信息的物化形态，是以文献为物质载体或传播媒介的信息，是社会信息的重要组成部分。文献信息是信息的下位概念，信息包含文献信息。

与信息及文献信息密切相关的概念还有知识和情报。

知识可以从不同角度来理解，按知识的可呈现程度，可分为显性知识和隐性知识。显性知识是指存储在各类物质载体中的客观知识，即系统化、理论化的信息，存储于文献载体中，就是文献信息。但并不是所有的文献信息都是知识。隐性知识也称主观知识，一般依附在人的头脑中，如个人的经验诀窍、判断联想、解决问题的思维方法等。主观知识加以表达就是社会信息，如果对其进行编码和记录，就转化为文献信息。知识和信息、文献信息是交叉关系。

情报通常指有着明确接受对象的那部分信息或者知识，具有很强的针对性。信息和知识在特定场合都可能成为情报，但并不是所有的信息和知识都是情报。信息或知识被记录在物质载体上被物化后，就是文献信息，而文献信息为满足特定需求被激活后就具有情报价值；如果某些情报(如口头表达、传递的情报信息)以文献形态存在时，就转化为文献信息。情报与信息、知识、文献信息也是交叉关系。

信息与文献信息以及知识、情报的关系如图 1－4 所示。

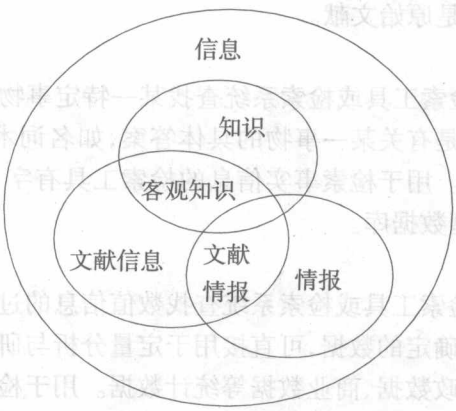

图 1-4　信息与文献信息(知识、情报)的关系

1.1.2 信息检索的概念、类型与意义

1.1.2.1 信息检索的概念

所谓信息检索，是根据特定的需求，借助于某种检索工具或检索系统，运用一定的方法，从信息集合中查找出所需信息的过程。

信息检索可以称为文献信息检索，“信息检索”中的“信息”，应该理解为“文献信息”，信息检索的实质是文献信息的检索。因为信息存在于整个自然界和社会之中，有的信息能形成文献，有的信息不能形成文献，形成文献的信息我们可以检索，不能形成文献的信息无法对其进行分析标引，就不能提供检索，例如“口头信息”、“实物信息”等就属于非文献信息。

换句话说，只有存储于文献载体中的“信息”，才有可能成为检索的对象。这里的“信息”，可以是任何出版形式的信息（图书、期刊、报纸），可以是不同表现形态的信息（文本、图形、图像、动画和声音），可以是具有不同含义的信息（数值、事实），也可以是不同程度的信息（知识、情报）。

根据检索手段的不同，信息检索包括手工检索和计算机检索两种形式。手工检索是以手工方式、利用印刷型工具书查找文献信息的过程，计算机检索是通过计算机及网络设备，利用数据库检索系统或搜索引擎检索系统检索文献信息的过程。

1.1.2.2 信息检索类型

根据检索内容、检索目的及检索工具的不同，信息检索可以分为文献检索、事实信息检索和数据信息检索三大类型。

（1）文献检索

文献检索是指利用检索工具或检索系统查找文献的过程，包括文献线索检索和文献原文检索。

文献线索检索是指利用检索工具或检索系统查找文献的出处，检索结果是文献线索，包括书名或论文题目、著者、出版者、出版地、出版时间等文献外部特征。检索工具如书目、索引、文摘印刷型工具书，以及计算机检索系统中的书目型数据库、索引/题录/文摘型数据库等。

文献原文检索是指利用检索工具或检索系统获取文献原文的过程。这是计算机全文数据库检索系统所提供的一种检索类型，在全文数据库系统中，不仅可以检索到文献线索，而且可以直接获取全文，检索结果是原始文献。

（2）事实信息检索

事实信息检索指利用检索工具或检索系统查找某一特定事物的过程。事实信息检索是一种确定性检索，检索的结果是有关某一事物的具体答案，如名词术语、概念、定义、事件、事实，或某一机构、人物的状况等。用于检索事实信息的检索工具有字词典、百科全书、年鉴、手册，以及术语型数据库和指南型数据库。

（3）数据信息检索

数据信息检索指利用检索工具或检索系统查找数值信息的过程。数据信息检索也是一种确定性检索，检索的结果是确定的数据，可直接用于定量分析与研究，如各种科学数据、人口数据、管理数据、金融数据、财政数据、商业数据等统计数据。用于检索数据信息的检索工具有统计年鉴、统计资料汇编，以及各类数值型数据库。

1.1.2.3 信息检索的意义

(1)是获取信息知识的捷径

据测算,人类知识总量在19世纪每50年增加1倍,20世纪初每10年增加1倍,20世纪70年代每5年增加1倍,20世纪80年代以来几乎是每3年增加1倍。据联合国教科文组织统计,人类近30年来所积累的科学知识占有史以来积累的科学知识总量的90%。

绝大多数有用的知识信息都储存在各种文献中。加利福尼亚大学伯克利分校的研究表明,2002年中,全球由纸张、胶片以及磁、光存储介质所记录的信息生产总量达到5万亿兆字节,约等于1999年全球信息产量的两倍。如果以馆藏1900万册书籍和其他印刷出版物的美国国会图书馆为标准,5万亿兆字节信息量足以填满50万座美国国会图书馆。据统计,当今全世界每年出版大约100万种新书,期刊10万种,报纸6万种,每年发表的科技论文约600万篇。2005年,中国出版图书22.2万种,报纸2100多种,杂志9000多种,音像制品3.5万种,电子出版物6152种。

在汪洋般的文献中,如何找到所需文献信息并加以合理有效地利用,如何以最少的精力充分占有文献资料,是一个非常实际的问题。据美国和日本20世纪60年代统计,一个科学工作者,在其整个科研工作中,用于翻检文献的时间约占50%。当今时代,科学家即使夜以继日地阅读有关文献,也只能浏览5%。解决这一问题的关键在于掌握信息检索方法与技能,它可以帮助人们快、准、全地获取所需知识,最大限度地节省查找时间。

(2)是科学研究的向导

科学研究是一种创造性的劳动。科学研究对某一课题或某一领域的认识及判断应是前所未有的。如果重新去发现他人早已知晓的真理,在已有的研究成果中周旋,这种研究就是毫无价值的,白白浪费了时间和精力。据统计,我国科研项目重复率达40%,而另外60%中部分重复率又在20%以上;与国外重复的也约占30%左右,其中大部分是国外已公开的技术,因而造成了人力、物力、财力的严重浪费。

要进行有价值的科学研究,必须以前人已提供的知识为起点,必须全面获取相关文献信息,了解各学科领域出现的新问题、新观点,这只能依赖信息检索才能实现。通过信息检索,可以了解研究课题的历史和现状,及时掌握国内外有关科学技术的发展水平、研究方向,从而确定自己的研究起点和研究目标,避免重复研究。

(3)是终身教育的基础

终身教育被认为是21世纪的生存概念。终身教育这一术语是1965年由联合国教科文组织成人教育局局长法国的保罗·朗格朗(Parl Lengrand)正式提出。他认为,将人的一生分为教育期和工作期,前半生的时间用来积累知识,后半生一劳永逸地使用知识,这是毫无科学根据的。他提出教育应当贯穿于人的一生,成为一生不可缺少的活动。其后,终身教育思想在世界各国广泛传播,对国际教育改革产生了重要的影响。

终身教育的思想有着深刻的社会根源。如前所述,人类知识的总量呈指数增长,但另一方面知识的陈旧速度也明显加快。据测算,18世纪知识陈旧的速度为80—90年,19—20世纪为30年,近50年缩短为15年。知识的陈旧速度比知识汲取的速度快得多,一个人从大学只能获得10%的知识,而人们原有知识以每年5%的速度不断"报废",如果不随时进行知识的更新和补充,10年后就有50%知识变得陈旧和老化。学校不再是学习知识的唯一的最后场所,信息化社会把所有的人转变为终身受教育者。

然而,无论接受何种形式的终身教育,首先应具备接受终身教育和继续学习的能力,这种

能力在很大程度上就是获取利用新知识的能力，就是对新知识的敏感力和接受力。而绝大多数知识都储存在各种文献中，所以要培养获取利用新知识的能力，就必须掌握信息检索的方法。掌握了这种方法，有了终身不断接受教育的能力，就可不断地丰富自己、完善自己，以适应社会发展的需要。

1.2 信息检索原理

1.2.1 信息检索原理概述

信息检索包括信息存储和信息检索两个过程。

信息存储即是对信息的组织，是按照既定的标准，从信息源中选择合适的信息，根据一定的规则和标准，对信息的外部特征和内容特征进行标引与著录，并以科学的方法加以有规律的排列，使分散的文献变成有序的集合，存储在各种检索工具或检索系统中。

信息获取即检索查找信息，是根据特定的需求，依据一定的信息检索语言和相应的检索技术，提出检索需求（检索提问），通过存储信息的检索工具或检索系统，将所需信息查找出来。

信息检索的实质是将反映特定信息需求的提问概念与信息存储系统中检索标识概念进行比较匹配，从中找出与提问概念特征一致的信息。信息存储和信息获取密不可分，存储的目的是为了检索，要检索必须先对信息进行存储。信息存储是信息获取的基础，信息获取是信息存储的逆过程。其原理如图 1-5、图 1-6 所示。

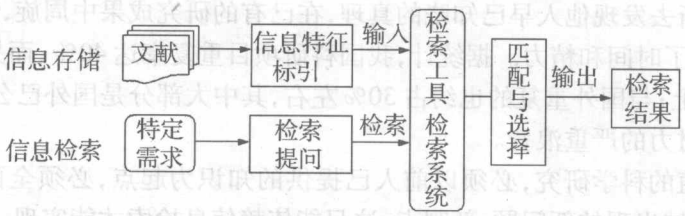

图 1-5 信息检索原理

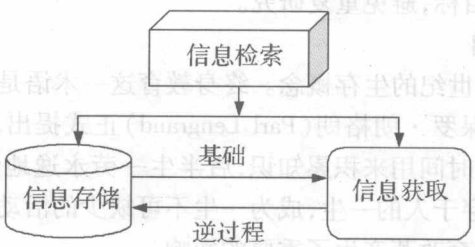

图 1-6 信息检索原理

1.2.2 信息存储过程

信息存储一般包括信息选择、信息著录、信息标引、信息整序等环节。

1.2.2.1 信息选择

信息选择是根据用户需要，从各类信息源中将符合既定标准的一部分文献挑选出来的活

动。由于信息量庞大,信息内容复杂,所以信息选择需要遵循一定的标准和方法。

信息选择的标准主要有相关性标准、可靠性标准、先进性标准和适用性标准。相关性是指在信息源中选择出与用户需求有关的信息;可靠性即指要鉴别信息的真实性;先进性一般指信息内容的新颖性和信息内容的领先水平;适用性指信息内容要符合用户的需要及使用的程度。

信息选择的方法包括比较分析法、核查法、引用摘录法和专家评估法。比较分析法即对不同的信息进行比较分析,鉴别信息的优劣;核查法是通过对有关信息所涉及的问题进行审核查对来优化信息的质量;引用摘录法即是根据信息相互引证的次数来判断信息质量的高低;专家评估法是通过有关专家来评价信息的水平价值,判断其可靠性、先进性和适用性。

1.2.2.2 信息描述

信息描述是对信息实体的外部特征和内容性质进行分析、选择和记录的过程,也称为信息著录。信息的外部特征如题名、责任者、出版或发表机构、出版时间、信息编号等,任何载体形态的信息实体都有其直接反映的形式特征;信息的内容性质主要指信息所属的学科和主题等。对上述特征项逐一进行的客观描述,就是著录。

信息描述需要遵循一定的规则。不同的信息类型和不同的检索系统其著录项目不尽相同。

对传统文献信息描述的标准有很多,国际上影响最大、使用最广泛的标准有《国际标准书目著录》(ISBD)和《英美编目条例》(AACR)。我国信息著录的国家标准是《文献著录总则》(GB 3792.2-83),该总则规定的基本著录项目分为9个大项目,依次为:题名和责任者项;版本项;文献特殊细节项;出版发行项;载体形态项;丛编项;附注项;文献标准编号及有关记载项;提要项。

对网络信息特征的描述标准主要是"都柏林核心集"(Dublin Core),其描述项目包括15个元素项:

《都柏林核心集》描述项目

元素名称	基本定义
题名(Title)	由作者或出版者给出的被描述信息的名称
主题(Subject)	揭示信息内容的主题词(关键词、分类号)
描述(Description)	对信息特征的说明,包括文摘、目次、文本
语种(Language)	描述信息内容的语种
来源(Source)	信息出处信息
关联(Relation)	该信息与其相关资源的联系
覆盖范围(Coverage)	信息内容涉及的时间范围和空间范围
创建者(Creator)	创作信息内容的主要责任者
出版者(Publisher)	提供该信息利用的责任者
其他责任者(Contributor)	对信息内容作出贡献的其他责任者
权限(Rights)	对版权、权限管理与使用有关的信息
日期(Date)	信息创建日期,包括出版、发行、修订日期等
类型(Type)	信息内容的特征和类型
形式(Format)	信息的物理或数字化格式
标识符(Identifier)	标识信息的唯一性符号,如URL、ISDN、DOI等

1.2.2.3 信息标引

信息标引是指根据一定的规则和方法,对信息内容特征进行揭示的过程。信息的内容特

征一般指某一学科或主题的内容。标引即是对信息内容进行分析，提取出信息内容所反映的学科分类概念和一定数量的主题概念，然后用标引语言（从检索的角度称为检索语言）进行标识，作为存储和检索的依据。充分而有效的揭示信息内容是检索信息的前提条件。

信息标引通常分为分类标引和主题标引两种类型（详见第2章）。

（1）分类标引

分类标引是按学科属性来揭示信息内容特征的方法。通过分类标引可将具有共同学科属性的信息类聚在一起，并依据各类信息之间的学科关系，把属于该学科的所有信息组织成一个有层次、有条理的整体。

分类标引的工具是分类法（分类语言或分类表）。国内外著名的分类法如《国际十进分类法》（UDC）、《美国国会图书馆分类法》（LCC）、《中国图书馆分类法》（CLC）。分类标引的过程，就是根据选择的分类法，对标引对象进行分析，确定所属分类类目，并将标引对象的学科特征及相关信息抽取出来，用分类法规定的符号代码予以标识。分类标引实质上就是对信息进行分类。经过分类标引，原来分散无序的信息就组织成了一个有序的学科体系。

（2）主题标引

主题标引是按主题名称（或关键词）来揭示信息内容特征的方法。主题是信息内容所涉及的事物，表达主题的语词成为主题词（主题标识）。通过主题标引，可以按字顺把同一主题的信息集中在一起。

主题标引的工具是主题法（主题语言或主题词表/叙词表）。著名的主题法如《美国国会图书馆标题表》（LCSH）、《汉语主题词表》等。主题标引的过程，即是对信息内容进行主题分析，确定主题概念，然后按照一定的词汇控制方式，用主题法中选择相应的主题词（标题词、叙词等）进行标识；或者采用自由标引方式，直接从已有的描述标引对象的语句中选择合适的关键词作为标识。与分类标引相比，主题标引可以把分散在不同学科中的信息集中在一个相关主题下。

信息标引通常与信息描述同步进行。将分类标引和主题标引的结果同该信息的其他描述项目汇总，其结果就形成款目。一个款目就是一种信息的缩影。在数据库中，信息的外部特征项与内部特征项通常称为字段，一条记录就是对信息实体不同特征的描述。

1.2.2.4　信息整序

信息整序也即信息的组织排序，就是将信息描述和信息标引的结果系列化，也即组织成相应的检索工具和检索系统。对信息的特征描述和内容揭示，形成表示某信息的记录标识，但这只是一个款目，还需要将所有信息的记录标识（一系列款目）按照一定的方法组织排列成有序的信息集合，才能为用户获取信息提供方便。

不同检索工具和检索系统有着不同的组织排序方法。常用的有分类组织法、主题组织法、字顺组织法、号码组织法、时空组织法。分类组织法是依照学科类别和事物类别特征组织信息的方法；主题组织法是按信息的主题特征组织排列的方法；字顺组织法是根据表示信息语词符号的音序或形序来组织排列信息的方法；号码组织法是根据信息所赋予的号码次序或大小顺序组织排列的方法；时空组织法是按照时间顺序或地理位置来组织排列信息的方法。信息整序的手段可以分为人工组织和计算机自动组织。单纯的人工组织效率低下，现在人工组织的自动化、智能化的程度越来越高。信息整序的形式（或类型）主要有工具书、数据库和网络搜索引擎，这三种类型的检索系统是信息整序的主要形式，也是人们检索获取信息的重要途径

（详细见第3章、第4章、第5章）。图1-7表示了信息存储的主要过程。

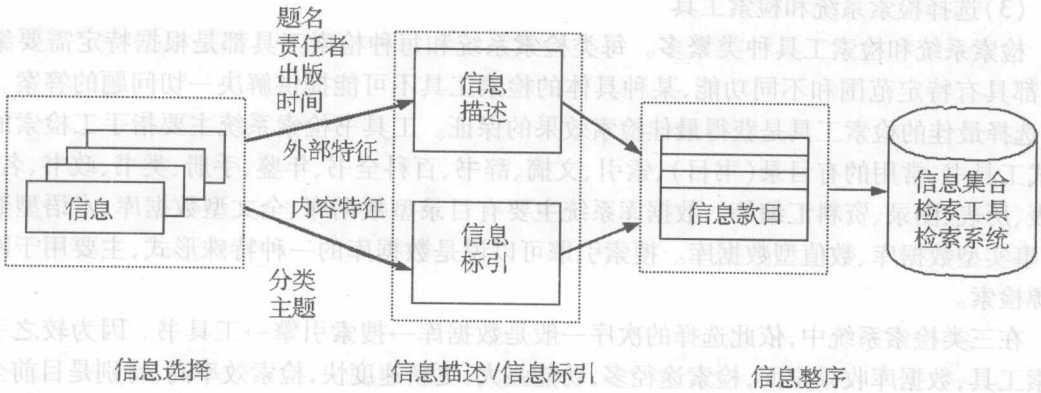

图1-7　信息存储过程

1.2.3　信息检索过程

信息检索是信息存储的逆过程，是从经过信息整序所形成的信息集合中查找出符合需求的原始信息。一般包括分析检索需求、确定检索标识、选择检索系统和检索工具、选择检索途径、检索匹配、检索结果输出等过程（图1-8）。

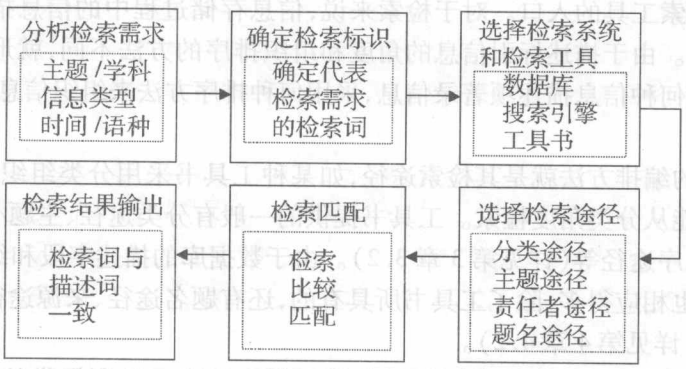

图1-8　信息检索过程

（1）分析检索需求

检索需求是根据信息查找的需要所拟定的问题。分析检索需求，就是要辨明检索问题的内容和要求，即：检索问题所属的主题及学科范围；检索信息的类型是文献类、事实类还是数据类；检索信息所属的时间范围（即查找的年代）和语种等。分析检索需求的目的是为了确定检索词和选择相应的检索工具。

（2）确定检索标识

明确检索要求后，就要将检索问题转换成检索标识。所谓检索标识，是检索问题包含的、具有检索意义的语言，也即能够代表检索需求的检索词，向检索工具发出“提问”。它包括：所属学科、所属类型、所属主题词和关键词、作者、时间范围等。每一检索问题可能都包含一个或多个甚至一系列的标识，应该提取出主要的、有检索意义的标识，分析各个术语概念之间的逻