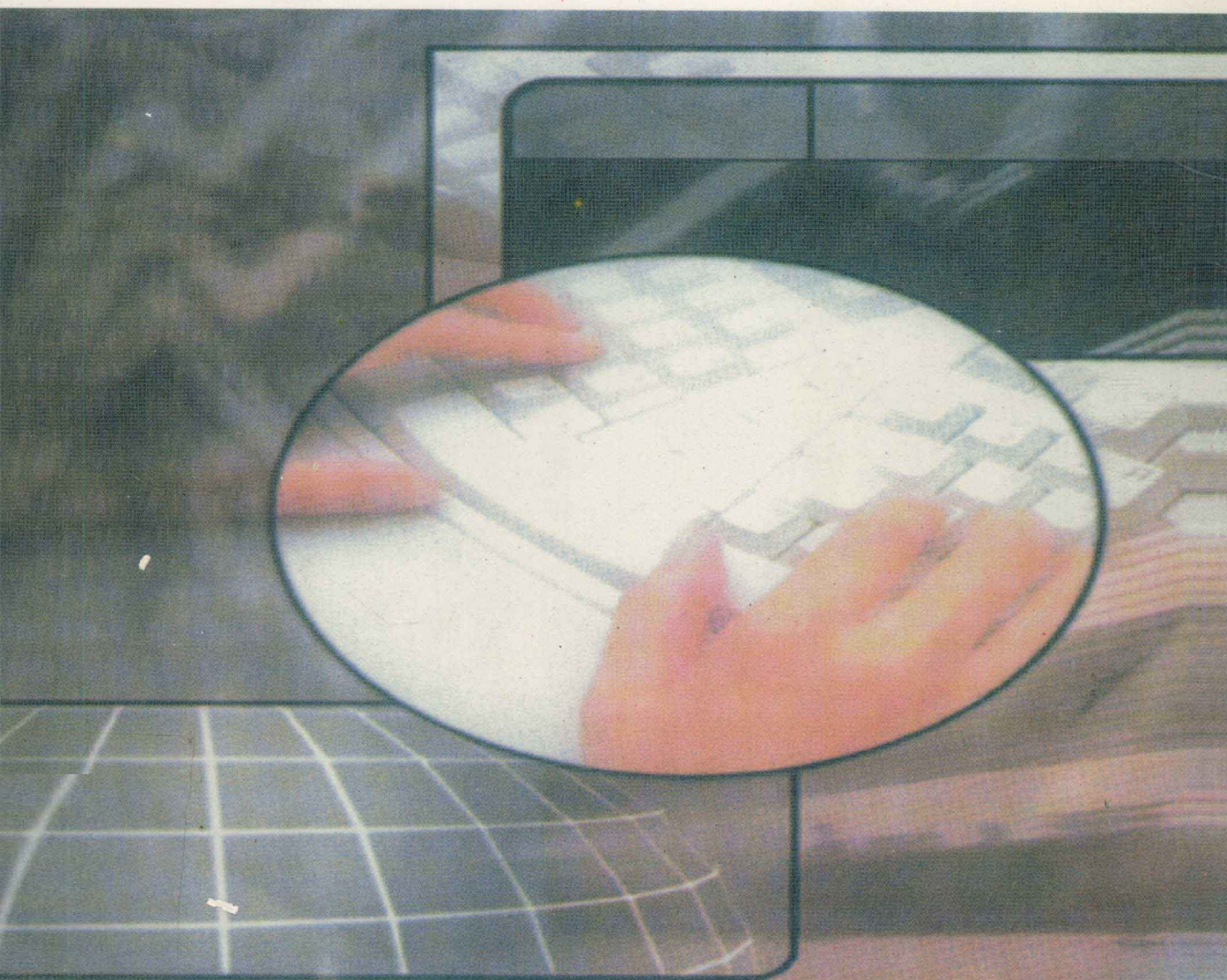


信息检索与利用

胡 伟 主 编

洪 滢 魏 浩 张凤坤 副主编



中国科学技术出版社

图书在版编目(CIP)数据

信息检索与利用/胡伟主编. —北京:中国科学技术出版社,2003.5

ISBN 7-5046-3523-5

I. 信… II. 胡 III. 情报检索 IV. G252.7

中国版本图书馆 CIP 数据核字(2003)第 036921 号

中国科学技术出版社出版

北京市海淀区中关村南大街 16 号 邮政编码:100081

电话:62179148 62173865

新华书店北京发行所发行 各地新华书店经销

北京印刷学院实习工厂印刷

*

开本:787 毫米×1092 毫米 1/16 印张:12.25 字数:314 千字

2003 年 5 月第 1 版 2003 年 5 月第 1 次印刷

印数:1—5000 册 定价 15.00 元

(凡购买本社的图书,如有缺页、倒页、
脱页者,本社发行部负责调换)

前 言

近几十年以来,信息检索已发展成为一门具有丰富理论的学科,并在实际中广泛应用,特别是高等院校已普遍开设信息检索课,随着信息技术的高速发展,以手工检索为主的教材已无法适应现代教学的要求,因此,有必要编写一部内容丰富、全面、新颖,注重理论联系实际,并且适宜高等院校信息检索课教学的教材,为信息检索课增添现代感和生命力。

本教材的编制,本着“以人为本”的思想,从读者的角度来组织材料,在内容的编排上,划分清晰、准确,对知识点的介绍,既简单又易懂,做到读者见到书后就知道要怎样去利用它,尤其是在计算机检索方面,用简单易懂的语言对其进行了详尽的阐述,并且实例丰富,图文并茂,作者介绍各类数据库和网上资源的检索方法,十分注意引用实例予以深入浅出的说明,所举之例,多与学生的论文写作有关,令读者有亲切感。另外,它所介绍的数据库或者网站都是新出现的,具有推陈出新、信息及时的特点。此外,它是从图书、期刊、报纸和其他类型文献检索的角度来介绍各种实用数据库的,符合读者习惯检索的特点。因此,它不仅仅有更为详细的内容,更完备的体系,而且更容易被读者所理解,是一部具有实用性、新颖性,符合时代发展的新型教材。

本书执笔分工如下:第一章至第六章及第十五章由张凤坤编写;第七章至第十四章由洪艳编写;魏浩申修。全书由胡伟统稿、整理、定稿。在编写过程中,参考了有关论著,在书后参考文献中一一列出。对本书所引资料的原著者,我们表示深深的感谢。由于我们的编写水平有限,书中如有缺点、疏漏和欠妥之处,恳请广大读者和用户批评指正。

编 者

2003 年 3 月

目 录

第1章 信息资源检索基础知识	(1)
1.1 现代信息资源的特点	(1)
1.1.1 信息与信息资源	(1)
1.1.2 信息资源的变化及特点	(2)
1.1.3 文献信息检索的意义和作用	(2)
1.1.4 科技文献的类型	(3)
1.2 文献信息检索	(5)
1.2.1 信息检索的概念及原理	(5)
1.2.2 信息检索语言	(6)
1.2.3 信息检索的步骤	(8)
第2章 国内检索刊物及检索方法	(10)
2.1 概 况	(10)
2.2 《中图法》分类体系	(10)
2.3 《汉语主题词表》的体系结构	(12)
2.3.1. 《汉语主题词表》的主表和附表	(13)
2.3.2. 《汉语主题词表》的辅助索引	(14)
2.4 著录格式	(15)
2.4.1 题录或文摘的著录格式	(15)
2.4.2 索引著录格式	(16)
2.5 国内检索刊物检索实例	(17)
第3章 美国《工程索引》	(19)
3.1 发展概况	(19)
3.2 Ei 年刊(印刷版)	(19)
3.2.1. 文摘正文部分	(19)
3.2.2. 索引	(20)
3.3 Ei 词表	(21)
3.3.1 《工程用主题词表》	(22)
3.3.2 《工程信息叙词表》	(23)
3.4 Ei 的检索途径与实例	(24)
第4章 美国《化学文摘》	(27)
4.1 发展概况	(27)
4.2 CA 的编制格式与内容	(27)
4.2.1 文摘的编排及著录	(27)
4.3 CA 的索引体系	(29)

4.3.1 期索引(Issue Index)	(29)
4.3.2 卷、累积索引	(32)
4.3.3 辅助索引(Auxiliary Index)	(35)
4.4 检索实例	(37)
第5章 英国《科学文摘》	(40)
5.1 发展概况	(40)
5.2 编排的结构及使用方法	(41)
5.2.1 SA 期刊	(41)
5.2.2 SA 累积索引(Cumulative Index)	(45)
5.3 检索途径及实例	(45)
5.3.1 分类途径	(46)
5.3.2 主题途径	(46)
5.3.3 著者途径	(47)
第6章 专利文献及其检索	(48)
6.1 专利文献	(48)
6.1.1 专利基本知识	(48)
6.1.2 专利文献的类型及用途	(51)
6.2 国际专利分类法	(53)
6.2.1 IPC 的分类体系和结构	(53)
6.2.2 《IPC 关键词索引》	(54)
6.3 世界专利索引	(55)
6.3.1 概况	(55)
6.3.2 WPI 的主要分册	(55)
6.3.3 《优先案索引》	(57)
6.3.4 《文摘周报》	(58)
6.4 中国专利文献检索方法	(59)
6.4.1 中国专利及专利说明书的类型	(59)
6.4.2 专利文献检索方法及途径	(60)
第7章 电子信息资源概述	(63)
7.1 电子信息资源的产生与发展	(63)
7.1.1 电子信息资源的概念	(63)
7.1.2 电子信息资源的产生与发展	(63)
7.2 电子信息资源的种类与特点	(63)
7.2.1 电子信息资源的种类	(63)
7.2.2 电子信息资源的特点	(65)
7.3 电子信息资源的开发与利用	(66)
第8章 计算机信息检索概述	(68)
8.1 计算机信息检索的发展历史	(68)
8.2 计算机信息检索的原理与技术	(69)

8.2.1 计算机信息检索的基本原理	(69)
8.2.2 计算机信息检索的基本方法	(70)
8.3 文献数据库简介	(74)
8.3.1 数据库的定义	(74)
8.3.2 数据库的类型	(74)
第9章 期刊报纸类信息资源的检索	(75)
9.1 《中文期刊数据库》(维普)的检索	(75)
9.1.1 简介	(75)
9.1.2 检索方法	(75)
9.2 中国期刊全文数据库	(80)
9.2.1 简介	(80)
9.2.2 内容	(80)
9.2.3 检索步骤	(83)
9.3 《人大复印报刊资料数据库》	(90)
9.3.1 简介	(90)
9.3.2 检索方法	(90)
第10章 图书类信息资源的检索	(92)
10.1 书生之家数字图书馆	(92)
10.1.1 简介	(92)
10.1.2 书生之家的检索方法	(93)
10.2 超星数字图书馆	(97)
10.2.1 简介	(97)
10.2.2 超星数字图书馆主页的功能	(98)
10.2.3 “超星版”网络图书的特点	(99)
10.2.4 使用超星数字图书馆需注意的问题	(100)
第11章 资讯信息资源的检索	(101)
11.1 万方数据资源系统	(101)
11.1.1 简介	(101)
11.1.2 主要数据库简介	(106)
11.1.3 万方数据资源系统的使用方法	(109)
11.2 中国资讯网	(114)
11.2.1 概况	(114)
11.2.2 数据库简介	(115)
11.2.3 中国资讯网—中国教育网镜像站的检索步骤	(116)
第12章 DIALOG 检索系统	(121)
12.1 DIALOG 系统简介	(121)
12.2 DIALOG 基本操作	(121)
12.3 DIALOG 检索技术	(122)
12.4 DIALOG 常用的数据库	(124)

第 13 章 OCLC First Search 检索系统	(127)
13.1 OCLC 概况	(127)
13.2 OCLC FirstSearch 的检索	(127)
13.2.1 联接 OCLC	(127)
13.2.2 FirstSearch 数据库	(128)
13.2.3 FirstSearch 的检索方法	(130)
13.2.4 New FirstSearch 的检索方法	(132)
第 14 章 INTERNET 信息资源的检索	(135)
14.1 INTERNET 概述	(135)
14.1.1 Internet 的历史与发展	(135)
14.1.2 中国的互联网络	(135)
14.1.3 Internet 的基本概念	(139)
14.1.4 Internet 的接入方式	(142)
14.1.5 Internet 提供的主要服务	(142)
14.2 搜索引擎	(145)
14.2.1 什么是搜索引擎	(146)
14.2.2 搜索引擎的分类	(146)
14.2.3 搜索引擎的局限性	(148)
14.3 中文搜索引擎的简介与利用	(149)
14.3.1 搜狐	(149)
14.3.2 网易	(151)
14.3.3 新浪	(151)
14.4 英文搜索引擎的简介与利用	(152)
14.4.1 Yahoo!	(152)
14.4.2 Alta Vista	(153)
14.4.3 Infoseek	(154)
14.5 网络电子期刊的检索	(155)
14.5.1 电子期刊的产生与发展	(155)
14.5.2 电子期刊的特点与类型	(156)
14.5.3 网络电子期刊的检索	(158)
14.6 网络电子报纸的检索	(158)
14.6.1 电子报纸的类型	(159)
14.6.2 网络电子报纸的检索	(159)
14.6.3 网上免费电子报纸	(164)
14.7 网络电子图书的检索	(167)
14.7.1 电子图书的特点	(167)
14.7.2 网上电子图书的检索	(167)
14.8 网上其他信息资源的检索	(170)
14.8.1 专利信息的网上检索	(170)

14.8.2 科技报告的网上检索	(171)
14.8.3 标准信息在网上检索	(172)
14.9 Internet 检索技巧	(173)
第 15 章 特种文献信息检索	(177)
15.1 科技报告	(177)
15.1.1 科技报告的特点	(177)
15.1.2 科技报告的类型	(177)
15.1.3 美国政府四大科技报告	(178)
15.1.4 国外科技报告的检索工具及使用方法	(179)
15.2 学位论文	(180)
15.2.1 概况	(180)
15.2.2 检索学位论文常用的手检工具	(180)
15.2.3 《中国优秀博硕士学位论文全文数据库》(CDMD)	(180)
15.2.4 国外学位论文数据库	(181)
15.3 标准文献	(181)
15.3.1 标准和标准文献	(181)
15.3.2 标准的类型	(181)
15.3.3 中国技术标准及其检索	(182)
15.3.4 国际标准及基检索	(183)
15.4 会议文献	(184)
15.4.1 会议的类型	(184)
15.4.2 国内会议文献的检索	(184)
15.4.3 国际、国外会议的检索	(184)
15.5 参考工具书	(185)
15.5.1 百科全书	(185)
15.5.2 词典	(185)
15.5.3 年鉴	(186)
15.5.4 手册	(186)
15.5.5 名录	(186)
15.5.6 年表、历表	(187)
15.5.7 资料汇编	(187)
主要参考文献	(188)

第1章 信息资源检索基础知识

1.1 现代信息资源的特点

1.1.1 信息与信息资源

1. 信息

信息的英文名称是 Information。它来源于拉丁文, Information 是消息传递的意思。信息是物质存在的反映,普遍存在于自然界、生物界和人类社会中。人类真正认识和理解信息的概念和内涵是近百年的事情,但是作为一种客观存在,在远古的宇宙万物中,在史前的猎、牧部族之中,在农业社会和工业化社会中,在当今的文明社会中,信息一直在积极地发挥着人类意识或没有意识到的重大作用。

所谓“信息”是指应用文字、数据或信号等形式通过一定的传递和处理,来表现各种相互联系的客观事物在运动变化中所具有的特征性的内容的总称。

信息被看成是物质的一种属性。可以认为:信息是对客观世界中各种事物的变化和特征的反映,是客观事物之间相互作用和联系的表现,是客观事物经过感知或认识的再现。它具有价值性、可传递性、可存贮性、可加工性、时效性等特征。

2. 信息资源

在了解什么是信息资源之前,我们应该先了解一下什么是信息源。信息源是信息产生的始点。人类对于自然界和社会的任何知识,无不来源于实践。科学研究人员、工程技术人员和一切参加物质资料生产的人们所从事的科技活动、生产活动及其成果,都是信息发生的源泉。也就是说只有自然界和人类社会才是信息的真正源泉。信息源按照其物质载体的不同可分为以下几种类型:

(1) 人脑信息源 指人们通过思维、交谈、讨论、报告等形式不断沟通情况,发出并传递信息,也称口头信息源。其优点是:信息的传递直接、及时,针对性强,并且与信息接收者可以进行双向沟通,实现信息的及时反馈,对接收者来说信息的使用价值较大。但是,直接获取口头信息的机会是有限的,也不便为以后的研究进行信息积累。所以,在获取、传播、利用口头信息源的过程中,人们通常又把他转换成文献信息资源。

(2) 实物信息源 实物信息源是由产品展览、商品的陈列和展销等方式发出的信息来源,人们通过实地参观、考察和举办展览来加以交流传播。实物信息往往是直接为生产服务的技术情报,具有真实、直观、易检验、易仿制的特点,同引进技术和设计相比,花钱少且见效快,是值得重视的信息源。但是实物信息只有经过复杂的分析和研究才能解析出来。

(3) 文献信息源 文献信息源是一种经过加工的信息源,是将人类知识用文字、图形、音频、视频信号等手段记录在一定的物质载体上进行交流传播。文献信息源便于广泛传播、系统积累、长期保存和直接利用,是获取信息的最基本、最重要的来源。

而信息资源的概念则是随着现代信息技术(特别是计算机技术)和信息资源管理理论的发展和普及而为人们所接受的。信息源不等于信息资源,信息源是蕴含信息的一切事物,信息资源是可利用的信息的集合,是高质量、高纯度的信息源。美国有学者对“信息资源”的理解概述

为:信息资源=文献信息;信息资源=数据;信息资源=多种媒介和形式的信息(包括文字、图像、声音、印刷品、电子信息、数据库);信息资源=信息活动中各种要素的总称(包括信息、设备、技术和人等)。由此可以看出信息资源还是一个发展中的概念,是一个具有丰富内涵的术语。

1.1.2 信息资源的变化及特点

1. 载体形式与传递方式发生变化

现代信息技术对信息资源载体形式及传递方式的影响是最直接、最根本的。随着现代信息技术的不断发展与应用,以及信息载体的磁性化、机读化,特别是电子计算机和通讯卫星用于图书信息,给信息资源的快速传递与交流提供了非常方便的条件。十五世纪末,哥伦布发现新大陆,时隔半年,这个消息才传递到西班牙女王那里。一百多年前,美国总统林肯被刺的消息,在12个星期后才传到英国的王宫,而1981年里根总统被刺的消息和现场,通过卫星传播,5分钟全世界都知道了。目前,以印刷型信息资源、磁带型信息资源、光盘型信息资源、网络型信息资源为主且这四种信息资源并存的格局已经形成,而且印刷型信息资源的比例逐渐减少,而光盘型信息资源和网络型信息资源的数量日益扩大。现代信息技术的发展、加速了信息转换,丰富了信息资源载体形式和传递方式。

2. 数量和内容发生变化

现代科学技术和社会经济的发展为信息资源内容的拓展和数量的增大提供了动力。目前,信息资源的类型除了传统的印刷型以外,各种视听型、缩微型、机读型大有与印刷文献相抗衡的趋势。他们同时并存,各自发展,共同竞争,互相补充。

3. 分布发生变化

在以印刷型文献为主的时代,信息资源的社会分布主要还是以图书馆、情报所、档案馆为主要分布点,信息服务部门的信息主要来自于出版社、报刊编辑部、新华书店和图书进出口公司,如今,随着科学技术的不断发展,人类知识的迅速增长,信息资源的社会分布异常分散,数量众多的信息资源广泛地分布在各类社会机构,社会组织及家庭中。信息服务部门的信息除上述来源外,还可以来自计算机硬件和软件公司、数据库开发公司。此外,个人也利用信息网络传播自己的研究成果和其他信息,也是信息来源之一。

以上几个方面就是信息资源的变化及特点。我们研究探讨信息资源的变化及特点,其目的在于充分认识和掌握信息资源发展变化的一般规律,来不断强化完善信息交流工作,从中来提高信息交流与传递的效率。

1.1.3 文献信息检索的意义和作用

1. 文献信息检索的意义

二十一世纪是知识经济占主导地位的时代。知识经济时代的核心是知识的生产、传播和利用。在知识经济社会,要求人们应该普遍具备获取信息和利用信息的能力。这种能力将成为劳动者基本素质的重要组成部分,而且将成为人们在今后高新技术和激烈竞争为特征的社会中把握机遇迎接挑战的必备能力之一。因此,人们不论是进行科学研究,还是发展市场经济,不论是教书育人,还是探索新的事物领域,首先都要掌握足够的信息,成功者的智慧正是在于能最敏捷、最有效地掌握信息,谁先掌握了有价值的信息,谁就能率先把握住事物的主动权。因为创造和成功在很大程度上取决于对信息的综合与利用。文献信息检索对于任何一个现代人来说,都是至关重要的事情。因为只有掌握了获取信息的科学方法,才能在日新月异的信息海洋里捕捉到有价值的信息。

2. 文献信息检索的作用

(1) 增强自学能力和研究能力 随着科学技术的飞速发展,知识老化、更新的速度也在加快。个人获取的知识(包括在校学到的知识)不可能是一劳永逸的。需要不断地进行知识的更新,不断完善知识结构。通过文献信息检索,能及时把握住科学技术发展的脉搏和新理论、新观点。从而为今天的学习和今后的工作打下坚实的基础。

(2) 避免重复劳动 通过文献信息检索,能摸清该学科或该课题的现有水平和发展方向,避免因重复劳动造成人力、物力、财力上的浪费(无价劳动)。

(3) 为科学研究服务 文献信息资源汇集了人类生产和科学实验的事实数据、理论方法;也反映着科学技术的发展水平和动向;还是许多成功经验和失败教训的文字记录。是一巨大的知识宝库,通过文献信息检索,能及时总结前人的经验,在前人成就的基础上有所发展,有所创新。

1.1.4 科技文献的类型

文献是记录有知识的一切载体。凡是用文字、图形、符号、声频、视频记录下来,具有存贮和传播知识功能的一切载体都称为文献。

科技文献是指记录有科技知识的一切载体。科技文献按不同的划分标准有不同的类型。

1. 按载体形式划分

(1) 印刷型 以纸张为媒体,以手写、石印、油印、胶印、铅印、影印等为手段来记录知识、信息,这是传统文献的基本形式,也是目前的主要形式。其优点是读取方便,流传广泛,不受时空的局限;其不足是存贮信息密度低,占据空间大,笨重,尤其在当今信息爆炸时代,其缺点显得尤为突出。

(2) 缩微型 以感光材料为媒体,以缩微照像为记录手段的文献,也称缩微复制品,包括缩微胶卷、缩微胶片(或缩微平片)、缩微卡片等。其优点是信息存贮密度高,文献体积小;可节省95%以上的存贮空间,便于收藏、保存和传递,能安全贮存珍贵资料,便于管理,比印刷型文献经济实惠,某种程度上弥补了印刷型文献的不足。但此种文献的阅读必须借助缩微阅读机或其他辅助设备才能就室阅读,不便携带,保存条件要求严格,难于普及。

(3) 视听型 又称声像型。以磁性材料或感光材料为存贮介质,借助特殊的机械设备,直接记录声音和图像,并通过视听设备存贮与播放信息知识的文献形式。如唱片、录音带等。依感官接受功能划分,视听型文献又可分为视听资料、听觉资料和视听合一的多媒体资料三种类型。其优点是,声情并茂、形象逼真、直观性强、动静交替、易于接受,在其传递信息知识方面,有着其他文献不可替代的优势。其缺点是必须借助一定的设备才能使用。

(4) 机读型 这是一种通过编码和程序设计,把文字资料转换成数字语言和机器语言,以磁性材料为存贮介质,以打字、穿孔或光学字符识别装置为记录手段,输入计算机,存贮在磁盘、磁带、光盘上,阅读时再由计算机将其内容按要求输出的文献。机读型文献近年来有了新的发展,其优点是存贮量大,查找文献快速方便;其不足是相应设备投入较大,短期内难以广泛应用和普及。

2. 按出版类型划分。

(1) 图书(Books) 图书大多是对已发表的科技成果、生产技术知识和经验通过选择、比较、核对、组织而成的。该类型文献内容成熟、定型,论述系统、全面、可靠。但图书出版周期较长,知识的新颖性不够。图书一般包括下面几种类型:专著(Monograph)、丛书(Series of Monograph)、教科书(Textbook)、词典(Dictionary)、手册(Handbook)、百科全书(Encyclopedia)

等。

(2) 期刊(Periodicals 或 Journal, Magazine) 期刊一般指具有固定题名, 定期或不定期出版的连续出版物。期刊上刊载的论文大多数是原始文献, 包括许多新成果、新水平、新动向, 其特点是出版周期短, 报道文献速度快, 内容新颖, 发行及影响面广。

(3) 科技报告(Technical reports) 科技报告是科技人员围绕某一专题从事研究取得成果以后撰写的正式报告, 或者是在研究过程中每个阶段的进展情况的实际记录。其特点是内容详尽专深, 有具体的篇名、机构名称和统一的连续编号(即报告号), 一般单独成册。科技报告的种类有: 技术报告(Technical reports)、札记(Notes)、论文(Papers)、备忘录(Memorandum book)、通报(Bulletin)等。科技报告是二次大战期间及战后迅速发展起来的, 目前全世界每年都有大量的科技报告产生, 估计约有 10 万件, 其中以美国政府研究报告(PB、AD、NASA、DOE)为主。

(4) 政府出版物 是各国政府部门及其所属的专门机构发表、出版的文件。其内容可分为行政性文件(如法令、法规等)和科技文献(科技报告、科普资料等)两大类, 其中科技文献约占 30%~40%。

(5) 会议文献(Conference papers) 会议文献指在国内外重要学术会议上发表的论文和报告, 此类文献代表某学科领域的最新成就, 反映该学科领域的发展趋势。会议文献分为会前文献(论文预印本和论文摘要)和会议后文献(会议录)

(6) 专利文献(Patents) 专利文献指由专利局公布出版或归档的所有与专利申请案有关的文件和资料。专利文献的种类有发明专利、实用新型专利和外观设计专利文献。

(7) 学位论文 学位论文是高等学校、科研机构的毕业生、研究生为获得学位所撰写的论文。学位论文探讨的问题往往比较专深, 一般具有一定的创造性。根据学位的不同一般分为学士学位、硕士学位、博士学位论文。

(8) 技术标准(Technical standards) 技术标准是一种规范性的技术文件。按使用范围, 技术标准可分为国际标准、区域性标准、国家标准、行业标准和企业标准。

(9) 科技档案 科技档案是指单位在技术活动中所形成的技术文件、图纸、图片、原始技术记录等资料, 包括任务书、协议书、技术指标、审批文件、研究计划、方案、大纲、技术措施、调研材料等, 它是生产建设和科研活动中用以积累经验、吸取教训和提高质量的重要文献。科技档案具有保密和内部使用的特点, 一般不公开。

(10) 产品样本 产品样本是国内外生产厂商或经销商为推销商品而印发的企业出版物, 是用来介绍产品的品种、特点、性能、结构、原理、用途和维修方法、价格等。全世界每年出版的产品样本, 据不完全统计约有 7080 万种。

3. 按加工的层次分

(1) 零次文献(灰色文献) 这是近几年最新提出的说法。所谓零次文献是指那些非正式出版物或非正式渠道交流的文献。

(2) 一次文献(原始文献) 是著者在科学研究、生产实践中根据科研成果、发明创造撰写的文献, 称为一次文献, 一次文献是文献的主体, 是最基本的信息源, 是信息检索的主要对象。一次文献包括专著、报刊论文、研究报告、会议文献、学位论文、专利说明书、科技档案、科技档案、技术标准、科技报告等。

(3) 二次文献 它是将分散、无序的一次文献按照一定的原则进行加工、整理、提炼、组织、使之成为便于存贮、检索的系统。如目录、题录、文摘、索引等。它从主题、分类、著者、题名等

不同角度提供查找一次文献的线索。

(4) 三次文献 是在二次文献的基础上,选用一次文献的内容,经过综合研究和归纳分析形成的述评性和综述性文献。如综述、专题述评、手册、辞典和年鉴等。

1.2 文献信息检索

1.2.1 信息检索的概念及原理

1. 信息检索的概念、类型

信息检索(Information Retrieval)是指将信息按一定的方式组织和存贮起来。并针对用户的需要找出有关信息的过程。因此,从广义上讲,信息检索包含了信息的存贮和检索两个不可分的部分,它的全称一般又称“信息存贮与检索”(Information Storage and Retrieval)。而我们通常所讲的信息检索是指狭义概念的信息检索,即从检索系统中查找出所需信息的过程。检索系统是指包含文献集合的载体的技术设备。按文献的存贮媒体和技术手段来分,文献检索系统有两种手工检索和计算机检索。

手工检索也叫手检,主要是利用一些书本、卡片式的检索工具来进行检索。如:书目、索引、文摘和各类工具书等。检索的过程主要是由人脑和手工操作的配合来完成。即靠人脑的思考、比较和选择来完成。

计算机检索也叫机检,主要是指利用计算机检索系统来进行检索。如计算机设备、终端、通讯设施、数据库和检索、应用软件等。检索的过程是在人机的协同作用下完成的,检索的本质不变,发生变化的只是文献的媒体形式、检索手段、存贮方式和匹配方式。

信息检索根据检索对象(或具体内容)的不同,可分为三种类型:文献检索、数据检索和事实检索

(1) 文献检索(Document Retrieval) 凡以文献为检索对象的叫文献检索。即利用相应的方式和手段。在存贮文献的检索工具或文献数据库中,查寻出所需的文献的过程。例如:查找某一研究课题一定年限内的有关文献,或对一项发明创造进行文献查新,或从事新产品开发时需要查找有关最新研究动态等等,均属文献检索。文献检索是一种相关性检索,检索结果是文献线索(文摘、题录)一般要阅读文摘后才能决定取舍。例如:检索有关大型发电机组方面有哪些文献?就是文献检索。

(2) 数据检索(Data Retrieval) 凡是以数据为检索对象的叫数据检索。检索的结果是数值性数据。例如查找一条数据、图表、价格、设备的型号与参数以及某种化学物质的分子式等。均属于数据检索。例如:检索大型发电机组的功率、转数各是多少?就是数据检索。

(3) 事实检索(Fact Retrieval) 凡以事实为检索对象的叫事实检索。检索的结果是某一事实。如某一个人或团体机构的基本情况,某一事件发生的时间、地点、过程等。例如:检索世界上最大水力发电机组是哪几个?就是事实检索。从信息检索的三种类型看,文献检索与数据检索和事实检索之间在原理方法和实践方面,并无本质差异,数据和事实检索要通过文献检索才能获得。因此,文献检索是最基本形式,是获取信息的主要手段。

2. 信息检索的原理

信息检索是按照一定的目标进行的信息活动。我们在检索之前必须对检索工具(或文档)编排原理有一个基本的了解。信息检索工具是把大量文献进行分析后按一定的特征排列组织

的文献集合体,信息检索就是根据一些既定的标志从文献集合中选出相关的文献信息。因此,信息检索应包括信息存贮和检索两个过程。如果没有信息存贮,信息检索就无法进行,没有信息检索信息存贮就毫无意义。这就是信息存贮和信息检索相互依存的关系。另外,信息检索只有了解信息存贮是怎样把文献存入检索工具(或文档)中去的,才知道应该如何取出所需文献信息。信息存贮也只有充分了解信息用户的需求,才能有效进行存贮。

从信息检索要求来看不外乎有两种:一是要查出已知文献外表特征;二是要查出具有所需内容特征的文献。所以,这就必须要求标引人员在进行信息存贮时要按照文献的外表特征和内容特征分别进行标引。以形成满足信息检索的两种需要的各种不同的检索途径。

文献的外表特征一般指文献的篇名(题目),作者的姓名、专利号、报告号、出版者或研究机构等。文献的内容特征是指文献所论述的主题,所属学科的分类等。标引时可按其特征进行有序排列。如按外表特征排列。有篇名、作者及号码的检索途径。按内容特征排列有主题、分类等检索途径,其存贮和检索的原理如图 1-1 所示。

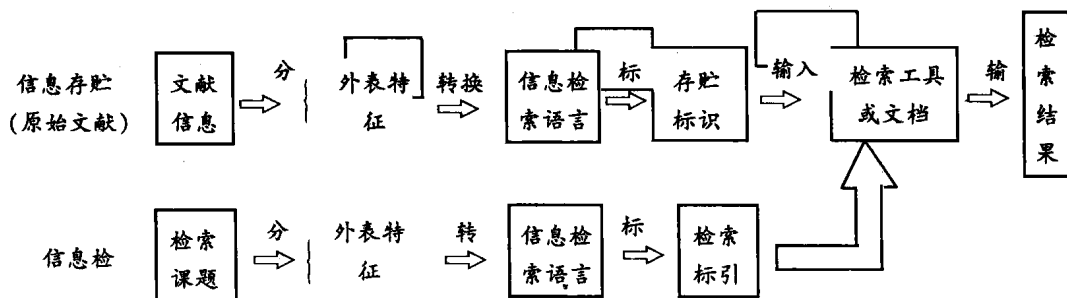


图 1-1

由图 1-1 可以看出,信息存贮时,标引人员要对各种文献进行分析,以形成其外表特征及内容特征,并用信息检索语言的语词(标识)标示出来,存入到检索工具或文档中。人们在检索时,一般对其检索课题的内容特征或外表特征是不清楚的。因此,也要对检索课题进行分析,以形成检索课题的内容特征或外表特征,并用信息检索语言的语词(标识)标示出来,从检索工具书中或输入到文档中,查找用该语词标引的文献线索,从而检索出所需信息的文献。

可见,信息的存贮与检索是意义不同的两个过程。存贮是积累和组织过程;检索是查寻和选取的过程。

1.2.2 信息检索语言

检索语言是信息存贮与检索过程中所使用的语言,又称标引语言、索引语言等。它是一种在自然语言的基础上经过处理而形成的能够表达和组织文献特征的人工语言。

检索语言不同于自然语言,它必须具备唯一性。也就是说它表达的概念必须只有一种解释,不能一词多义或多词一义,也不能模棱两可。检索语言只有具备了单一性,才能保证表达的概念意义的唯一性,也才能保证标引和检索的一致性。

检索语言是沟通信息存贮与信息检索两个过程的桥梁,在信息检索中起着媒介作用,没有检索语言就不能顺利完成信息检索,也就不能检索到所需的文献。检索语言按其结构、特征可划分为三类:

1. 表达文献特征划分

按表达文献特征划分为描述文献外表特征的检索语言和描述文献内容特征的检索语言两种。

2. 按组配方式划分

(1) 先组式检索语言 指在检索之前表达文献主题概念的标识已事先拟定好的检索语言。如标题词语言、体系分类语言等。

(2) 后组式检索语言 指表达文献主题概念的标识,在编制检索语言词表和标引文献时不预先固定组配,而是检索时,根据实际需要,按照组配规则临时进行组配的检索语言。如叙词语言等。

3. 按结构原理划分

按结构原理可分为分类语言、主题语言、代码语言、引文语言等四种。

(1) 分类语言 分类语言又可分为等级体系分类语言、组配分类语言和混合分类语言。

① 等级体系分类语言:它是采用一定的等级体系结构,按学科、专业集中文献,并从知识分类角度,用分类号作为表达文献内容标识的一种检索语言。等级体系分类语言采用对文献论述的事物,按照学科知识的逻辑次序进行层层划分、层层隶属的方法来形成一个严格有序的等级结构体系,将分类标识组织成具有隶属并列关系的概念。如《中国图书馆分类法》就是一个典型的体系分类语言,分类表则是这种语言的具体体现。

② 组配分类语言:基于概念的可分析性和可综合性,用科技术语进行组配的方法描述文献内容的语言。

③ 混合分类语言:即组配分类语言和体系分类语言的结合,两者侧重点不同,形成体系一组配分类语言和组配一体系分类语言,如《国际十进分类法》。

(2) 主题语言 是指应用经过或没有经过规范处理的自然语言的语词作为描述文献内容概念的标识,并按自然语言的字顺组织标识的一种检索语言。主题语言包括关键词语言、标题词语言、单元词语言、叙词语言四种。

① 关键词语言:是以关键词作为文献内容标识和检索入口的一种主题语言。关键词是指从文献的标题、正文或摘要中直接提取未经规范化处理的自由词汇作为标识语言。不受词表控制,可自由组配,以表达文献的内容特征。其最大的优点是能用计算机进行自动抽词标引,适合于计算机自动编制各种类型的关键词索引。

② 标题词语言:是传统的主题语言。它是用规范化的自然语言,即经过标准化处理的名词术语作为标识,来直接描述文献的内容特征并将全部标识依字顺排列的标识语言。如美国《工程索引》的“Subject Headings for Engineering”(工程用文题词表)即为典型的标题词语言。

③ 单元词语言:是以单元词作为文献内容的标识和检索的主题语言。单元词即指一个个最小、最基本且概念不可再分的词汇单位,能用来描述文献所论及或涉及的事物——主题的那些单词,未经规范化处理的或只作少量规范化处理的自然语言。

④ 叙词语言:是以叙词(即主题词)作为文献内容的标识和检索依据的主题语言。所谓叙词,是从文献题目、正文或摘要中抽取出来的,用以表达文献内容的一种采用表示单元概念的规范化语词的组配来对文献内容进行描述的后组式词汇型标识系统的标识语言。叙词受词表控制,词表中词与词之间无从属关系,都是相对独立的概念单元。如我国编制的《汉语主题词表》就是典型的叙词语言。

(3) 代码语言 为了提高检索语言的语义能力而提出的一种号码或代码系统。如专利号、报告号、登记号等。

(4) 名称语言 一种以文献信息中所包含各类名称作为标识和检索的语言。如刊号、书号、篇名、作者姓名等。

1.2.3 信息检索的步骤

信息检索主要是根据既定的课题要求,利用检索工具、按照一定的方法和步骤,把符合需要的文献挑选出来的过程。诚然,合理的检索方法、步骤可以提高检索效率和检索质量。信息检索的步骤与其检索的具体要求密切相关,一般分以下几个步骤:

分析检索课 → 选择检索系 → 确定检索途 → 制定检索策 → 实施检索

1. 分析课题

在检索课题确定之后,首先要对检索课题进行认真的分析研究,明确检索的目的、范围,确定所需信息的内容、特征、年代范围以及文献信息量和类型等。这是制定检索策略的根本出发点,也是检索效率高和成败的关键。这样经过课题分析选择出能代表主题概念及相关事物中最本质的检索途径与标识、为正确地进行检索作好准备。总而言之,分析检索课题,是整个信息检索过程的重要准备工作,检索课题分析的越深刻、越细致,检索工作就越顺利,检索效率就越高,检索的效果就越好。

2. 选择检索系统

一个计算机检索系统可提供多个可检索数据库。选择检索系统时,首先必须对各种检索系统所覆盖的学科范围有清楚的了解,然后从文献的类型、文种、出版的时间等方面来考虑选用哪种检索系统。一般以综合性为主,专业性为辅,以保证查全率。检索系统和数据库的选择还可通过三次文献的检索,例如《书目指南》、《工具指南》、《数据库回录》、网上的 yahoo、sohu 等工具来指引二次文献检索系统。

3. 确定检索途径

检索途径也就是检索点,指的是通过文献的何种特征检索文献。而文献的特征一般可分为外表特征和内容特征。外表特征是指文献的书名、著者、序号等;内容特征是指涉及学科主题、分类号以及分子式等。因此,检索途径可分为主题途径、分类途径、著者途径、号码途径(如专利号、报告号)等。

4. 制定检索策略

检索策略是指检索前制定的概念组配和执行顺序的方案。如计算机检索中制定的检索式,手工检索中的检索词。两者的不同之处在于机检的检索式是将各个检索概念(一般为表达主题内容的检索词)之间的逻辑关系、位置关系等用检索系统规定的各个组配符也称算符(operator)连接起来,成为机器可识别和执行的命令形式。手工检索的检索词是一个个孤立的词汇。词与词之间的逻辑、位置关系的调整都贮存在检索者的脑海中,是无形的。

5. 实施检索

检索策略制定完成以后,就可以利用各种手工检索刊物或计算机检索文档实施检索。一般检索的方法与检索课题的要求(如检索课题的性质、目的及信息源的类型等)密切相关,归纳起来,常用的检索方法主要有以下三种:

(1) 常用法 是指利用各种检索工具进行检索的方法。常用法根据查找方式的不同,又分为顺查法、倒查法和抽查法。

① 顺查法:这是一种以检索课题起始年代为起点,按时间顺序由远及近逐年查找的方法。这种方法漏检和误检少,检准率高,适用于科研课题复杂、研究范围宽、研究时间久的课题,但费时、费力,工作量大。

② 逆查法:这是一种与顺查法正好相反,逆时间顺序,由近及远的查找方法。此法多用于新课题,或老技术新发展的课题,对近期状况比较重视,从近期的资料开始查起,一直到满足课题需要为止。

③ 抽查法:是根据学科发展的特点,抓住某学科发展迅速,文献发表较多的年代抽出几年或几十年逐年查找的方法。这种方法省时间检索效率高。

(2) 追溯法 是利用已有文献后面所附的参考文献,逐一追查原文,再从这些原文后面所列的参考文献逐一追查,不断扩大检索的线索,从而获得一批相关文献的查找方法。这种方法不必利用大量的检索工具,只利用原始文献所引用参考文献跟踪追溯。但漏检和误检比例较大,检索效果不高。

(3) 综合法 是上述两种检索方法的结合。也叫循环法。指先利用检索工具查出一批有用文献,然后利用这些文献所附参考文献追溯查找的方法。如此循环交替,直到满足需要为止。

以上几种信息检索方法,究竟采用哪一种方法最为合适要根据检索课题内容、检索时间要求和各个学科检索工具的特点而定。而且,在实际检索的过程中,有时检索策略要经过多次反复的调整、直至得到满意的结果为止。

这是检索文献的最后一步,也就是当根据课题查出所需要的文献线索后,就可以根据文献出处索取原始文献了。