

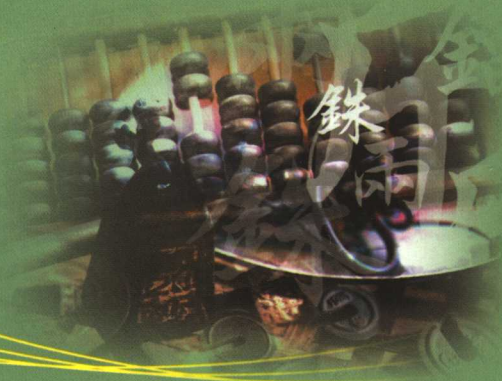


人文信息资源检索

RENWEN XINXI ZIYUAN JIANSUO

主 编 王立革

副主编 邹晓华 耿宇航 丁 丽



东北林业大学出版社

人文信息资源检索

主 编 王立革

副主编 邹晓华 耿宇航 丁 丽

东北林业大学出版社

图书在版编目 (CIP) 数据

人文信息资源检索/王立革主编. —哈尔滨: 东北林业大学出版社, 2006.8

ISBN 7-81076-925-1

I. 人… II. 王… III. 人文科学—情报检索 IV. G 252.7

中国版本图书馆 CIP 数据核字 (2006) 第 094839 号

责任编辑: 卢 伟

封面设计: 彭 宇



NEFUP

人文信息资源检索

Renwen Xinxin Ziyuan Jiansuo

主编 王立革

副主编 邹晓华 耿宇航 丁丽

东北林业大学出版社出版发行

(哈尔滨市和兴路 26 号)

东北林业大学印刷厂印装

开本 787 × 1092 1/16 印张 11.75 字数 267 千字

2006 年 8 月第 1 版 2006 年 8 月第 1 次印刷

印数 1-1 000 册

ISBN 7-81076-925-1

C·312 定价: 23.00 元

前言

1000多年以前,唐代诗人杜牧在吟诵他那“塞外音书无信息,道傍车马起尘埃”(《寄远》)的诗句时,也许根本不会想到,他“石破天惊”地吐出一个改变人类生存方式的词汇——“信息”。

“梦断美人沉信息,目穿长路倚楼台”,南唐李中的一首《暮春怀故人》,使我们感受到了古人对信息的渴望。李清照的那句“不乞隋珠与和璧,只乞乡关新信息”同样让我们体味到了信息在人类生活中的弥足珍贵。

时光飞逝,人类的脚步已迈进21世纪的第六个春天。正像人类在20世纪所畅想的那样,全球信息化浪潮一浪高过一浪。文献信息的数字化,数字信息的网络化,网络信息的海量激增和信息的有序组织、存储和检索,将是我们日益关注的前沿课题。对知识的收集和管理意义已经变得和对知识的探索创造一样重要,信息已经和物质、能源并列为人类社会发展的三大资源。

人文信息资源作为人类生存方式的内容和载体,从根本上影响和改变了人类的进程。不论是作为文献之邦的中国,还是科技发达的美英等国,都把人文信息资源的占有和利用看做立国富民的头等大事。人类越来越清楚地认识到:在竞争日趋激烈的社会环境中,信息意味着机遇和发展,谁最先获得了有价值的信息,谁就会在竞争中占据优势。

因此,提高当代大学生的人文素质,特别是信息素质已经成为国内外教育界的共识。在这方面,一些发达国家已经走在了我们的前头,从20世纪80年代末,我国开始在高等院校增设文献检索课,以使大学生适应信息时代的需要,尽快掌握信息检索的基本原理与方法,了解信息检索在人文研究过程中的作用和意义,培养人文专业学生的信息意识及计算机和互联网检索技术,使其能够在激烈的竞争中生存和发展。

信息是环境,而且是物质环境。每个大学都有其特定的学科结构、发展历程及建设规模和与之相对应的藏书体系,因此,尽管我们处于一个网络化的信息共享的社会,但各自的信息环境还是存在着一定的差异。为此,我们结合佳木斯大学文献信息资源,为文史、政法、经济、艺术、外语类专业的学生量身定制了这本《人文信息资源检索》。

本书在编写过程中注重教材的针对性、可行性和指导性。注重科学性、系统性、新颖性和实用性。力争做到检索理论与检索方法有机结合,手工检索与计算机检索有机结合,信息检索与信息利用有机结合,使教材更具有科学性和系统性。强调了基本技能、自学能力和研究能力的培养。

在编写本书的过程中,我们更加真切地感受到了信息、情报和文献对我们人类的重要性。对Google的创始人Larry Page所说的那句话“我们若能更妥善地搜寻资料,实在已经改变了世界”,感同身受。

在本书编写过程中,我们参阅了大量中外文资料,在此向引文作者表示敬意。

由于水平所限,书中纰漏或恐难免,不当之处敬请同行和读者斧正。

王立革
2006年春于佳木斯

目 录

第一章 信息与信息检索	(1)
第一节 信息及相关概念	(1)
第二节 信息检索	(4)
第三节 计算机信息检索	(8)
第四节 人文社会科学信息检索教学	(12)
第五节 知识与知识管理	(13)
第二章 人文工具书类型及其概述	(15)
第一节 人文文献工具书的性质和作用	(15)
第二节 常用人文文献工具书的类型	(16)
第三节 人文社会科学工具书的主要排检方法	(19)
第四节 数字化的人文社会科学工具书	(24)
第三章 字词检索	(26)
第一节 文字的产生与发展	(26)
第二节 中文字词的检索	(27)
第三节 外文字词的检索	(33)
第四章 书目与古今图书查找	(38)
第一节 古籍的查找	(38)
第二节 近代图书的查找	(41)
第三节 禁书、伪书的查找	(42)
第五章 哲学政治学信息检索	(45)
第一节 马列文献的检索	(45)
第二节 哲学文献的检索	(49)
第三节 政治学信息检索	(52)
第六章 法律法规类文献检索	(54)
第一节 中国法律法规信息检索工具	(54)
第二节 外国法学文献信息的检索	(57)
第三节 国际条约的检索工具	(58)
第四节 法律法规网络信息资源检索	(61)
第七章 经济与管理学信息检索	(65)
第一节 经济与管理学文献概况及分布特点	(65)
第二节 经济与管理学检索工具及利用	(66)
第三节 统计信息的查找	(70)
第八章 文学信息资源检索	(74)
第一节 文学专题书目检索	(74)

第二节 文学词典检索·····	(77)
第三节 文学资料汇编检索·····	(78)
第九章 历史地理信息检索·····	(81)
第一节 历史信息的检索·····	(81)
第二节 地理信息的查找·····	(88)
第十章 特种文献检索·····	(92)
第一节 会议文献检索·····	(92)
第二节 政府出版物检索·····	(96)
第三节 学位论文检索·····	(98)
第十一章 数据库检索·····	(102)
第一节 数据库及数据库检索·····	(102)
第二节 商业数据库检索·····	(105)
第三节 中文数据库检索·····	(118)
第十二章 Internet 信息检索 ·····	(128)
第一节 Internet 基础知识 ·····	(128)
第二节 Internet 主要服务及使用 ·····	(131)
第三节 搜索引擎·····	(133)
第四节 Internet 信息的收集选择开发和利用 ·····	(137)
第十三章 Internet 信息查询技巧 ·····	(140)
第一节 Internet 检索策略 ·····	(140)
第二节 Internet 上的电子报纸与新闻 ·····	(142)
第三节 Internet 上的工具书 ·····	(146)
第四节 Internet 专题信息的查找 ·····	(150)
第十四章 图书馆馆藏文献检索·····	(155)
第一节 认识图书馆·····	(155)
第二节 ILAS 与图书馆馆藏文献检索 ·····	(156)
第三节 TPI 与虚拟馆藏检索 ·····	(161)
第十五章 社会科学研究·····	(165)
第一节 社会科学研究·····	(165)
第二节 社会科学研究选题·····	(166)
第三节 社会科学论文的撰写·····	(171)
附录 I 汉语拼音和威妥玛式拼法详细对照·····	(173)
附录 II 日语假名英语音译对照·····	(175)
附录 III 俄语字母英语音译对照·····	(176)
附录 IV 各种纪年与公元纪年的换算·····	(177)
参考文献·····	(179)

第一章 信息与信息检索

第一节 信息及相关概念

我们所处的时代是一个前所未有的不断产生、传递和利用信息情报的时代。因此,有人称我们的时代为信息时代。那么,究竟什么是信息,什么是情报,它们与本课程大量讨论的文献有着什么关系,这是我们首先需要了解的问题。

一、信息

信息——物质存在的一种方式,一般指数据、消息中所包含的意义。可以使消息中所描述的事件的不定性减少。

不同的学科从不同的角度对信息这个概念有不同的解释。经济学、心理学、新闻学和哲学的定义不同于图书情报学对信息的定义。控制论专家 N·维纳从信息自身具有的内容属性给信息所下的定义被许多研究者所引用。国内在这方面也有许多争论。

信息的定义之所以呈现多样化,主要原因有三:第一,信息本身的复杂性,它是一个多元化、多层次、多功能的综合物;第二,信息科学是一门新兴学科,它的许多分支学科仍在随着社会、经济和科学技术的发展而发展,其内涵和外延不很确切;第三,人们出于不同的研究和使用目的,从不同的角度或层次出发,对信息概念就会做出不同的解释。

因此,许多学者建议将信息的概念分为不同的层次来解释。在诸多层次中,最重要的是两个层次:一是没有任何约束条件的本体论层次,即信息是一种客观存在的现象,是事物的运动状态及其变化方式的表征,不受主体意志的影响。不停运动着的事物不断产生本体论意义上的信息;二是受主体约束的认识论层次,即信息就是主体所感知或所表述的事物运动状态及其变化方式,是反映出来的客观事物的属性。

信息是广泛存在的,世间万物的运动,人间万象的更迭,都离不开信息的作用。据说上古时期,诺亚的方舟在洪水中飘荡许久,当放出的飞鸽衔回一束橄榄,意味着带回了洪水已退的信息。李太白的诗“日照香炉生紫烟,遥看瀑布挂前川,飞流直下三千尺,疑是银河落九天”给我们带来的显然就是庐山瀑布的信息;苏东坡的词“大江东去,浪淘尽,千古风流人物”给我们传递的是赤壁怀古的信息。

从自然界角度看,信息表征了物质的属性,地球昼夜的变化是一种信息,它反映出地球绕太阳自转的运动特性和状态;山的高度是一种信息,它反映出山的空间特性;树干的年轮是一种信息,它反映了树木成长的时间特性——树龄;闪电是一种信息,它反映了云层中所含能量的特性;花的香味也是一种信息,它反映了花分子结构的化学特性等。

由此我们可以看出信息并非是区别事物的本质属性，信息并非事物本身，而是事物发出的体现它本质属性的信号。

信息具有如下特征：客观性；动态性；相对性；依存性；可传递、亦可干扰性；可加工性；共享性。

二、知识

知识是人类的主观世界对客观世界的概括和反映，是大量有组织的信息，是关于事实和思想的有组织的陈述，是提供某种经过思考的判断和某种实验的结果。

知识是人类社会实践经验的总结，知识是人类通过信息对自然界、人类社会以及思维方式与运动规律的认识，是人的大脑通过思维重新组合的系统化的信息的集合。因此，人类不仅要通过信息感知世界、认识和改造世界，而且要根据所获得的信息组成知识。可见，知识是信息的一部分。

1. 知识的分类

根据经济合作与发展组织（OCED）出版的《以知识为基础的经济》报告将知识分为以下四类：

第一类“知事（Know - what）”，指关于事实方面的知识，也可理解为 Know - when、Know - where；即在什么样的时间（Know - when）、什么样的地点或条件下（Know - where）能解决什么样的问题。

第二类“知因（Know - why）”，指自然原理和规律方面的科学理论，知识的生产是在专门研究机构如实验室和大学完成的。

第三类“知道怎样做的知识（Know - how）”，指做某些事情的技艺和能力，被称为技术情报和商业秘密，其典型是企业开发和保存在其内部的技术诀窍或专有技术；

第四类“谁以及是怎样创造知识的（Know - who）”，侧重创造思想、方法、手段、过程以及特点等的了解。

2. 知识的来源

第一类是直接源于产生信息的客观事物；

第二类是通过信息载体或媒介（文献、电视、广播等）的传递、交流而间接获得。

三、文献

文献是记录有知识的一切载体。

目前有关文献的较权威的定义主要有两个，一是《文献情报术语国际标准（草案）》（ISO/DIS5127）的定义：“为了把人类知识传播开来和继承下去，人们用文字、图形、符号、声频、视频等手段将其记录下来，或写在纸上，或晒在蓝图上，或摄制在感光片上，或录到唱片上，或存贮在磁盘上。这种附着在各种载体上的记录统称为文献。”另一个是中国的国家标准，即信息是各种媒介和形式的信息集合，包括文字、声像、印刷品、电子信息、数据库等。

1. 文献的构成要素

（1）文献信息是文献的内容。

(2) 符号系统是信息的携带者。

文字已不再是表达思想的唯一手段,图形、声频、视频等同样成为表达思想、传递感情的重要手段。总的说来,文献中的符号系统指图画、文字、公式、图表、编码、声像和电磁信息等。

(3) 载体是符号赖以依附的“寄主”。

随着生产力的发展,存贮和表达人们思想的物质载体不同了,从金石、竹简、羊皮、丝帛、纸张发展到感光介质和磁性介质;从非人工材质的泥版、岩石、石板、兽骨、木板、竹片、兽皮、树叶、桦树皮等到各种人工材质,如无机材质(陶、砖瓦、瓷、玻璃等)、金属材质(青铜、铁、铝、金银等)、高分子材质(帛、纸、胶片、醋酸纤维等)、复合材质(磁带、光盘等)。纸质文献已经变得不利于快速传输信息、高效查阅和高密度存贮了,于是其他介质的文献应运而生,如光盘等。但纸质文献具有价格低廉、质地柔软、易于书写、携带和收藏等其他一些载体所无法比拟的性能,仍然是文献家族的主干。

(4) 记录方式是将文献的符号进入载体的方法和过程。

按记录方法分有手工记录、机械记录、光记录、电记录、声记录和磁记录。其中最常见的是印刷。随着科技的发展,文字记录可以转变为数据记录,并用电子方法存贮到磁介质上去。光学字符识别(Optical Character Recognition,简称OCR)技术就代表了这种崭新的技术,它通过光学方法对字符、标记表示的书面数据进行自动识别,转换成机器可以处理的信息,这样就实现了纸质文献信息向电子信息的转换,而且可以通过不同接口装置向不同设备输入数据,这些设备进而与通信网络相连,实现远距离数据传送。

2. 文献的类型

按文献载体分有印刷型文献、缩微型文献、机读型文献、声像型文献。

按文献加工程度分有一次文献、二次文献、三次文献。

按文献的出版形式分有图书、期刊论文、会议文献、科技报告、专利文献、学位论文、标准文献、政府出版物、产品样本、科技档案等。

四、信息、知识与文献的关系

1. 区别

信息:有形的,独立于行动和决策,经过处理改变形态,物质产品,与环境无关,可以复制;

知识:无形的,与行动和决策相关,经过处理改变思维,精神产品,环境改变含义,经过学习才能转让,无法复制。

2. 联系

知识是有组织的大量的信息,获得知识有赖于获得信息。

信息是知识得以形成和传播的中介,而不是知识本身。知识是经过精心研究、领会后的有用信息,是人类对信息加工处理后的产物。

并非所有信息都可成为知识,在信息时代,源于众多复杂客体的大量信息,只有借助于现代化的信息手段,并通过掌握现代信息科学技术的认知主体,才能真正转化为知识。

第二节 信息检索

一、信息检索的含义

信息检索的概念有狭义和广义之分。狭义的检索 (Retrieval) 是指依据一定的方法, 从已经组织好的大量有关信息集合中, 查找并获取特定的信息的过程。这里的信息集合, 不是通常所指的信息本身, 而是关于信息的信息或线索。如果真正要获取信息所记录的内容, 那么还要依据检索所取得的线索索取原文。

广义的检索包括信息的存储和检索两个过程 (Storage and Retrieval)。信息存储是指工作人员将大量无序的信息集中起来, 根据信息源的外表特征和内容特征, 经过整理、分类、浓缩、标引等处理, 使其系统化、有序化, 并按一定的技术要求建成一个具有检索功能的工具或检索系统, 供人们检索和利用。而检索是指运用编制好的检索工具或检索系统, 查找出满足用户要求的特定信息。

二、信息检索的基本原理

信息检索的基本原理是: 通过对大量的、分散无序的信息进行搜集、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志达到一致, 以便有效地获得和利用信息源。其中存储是为了检索, 而检索又必须先进行存储。

文献信息的存储和检索的全过程可用图 1-1 表示。

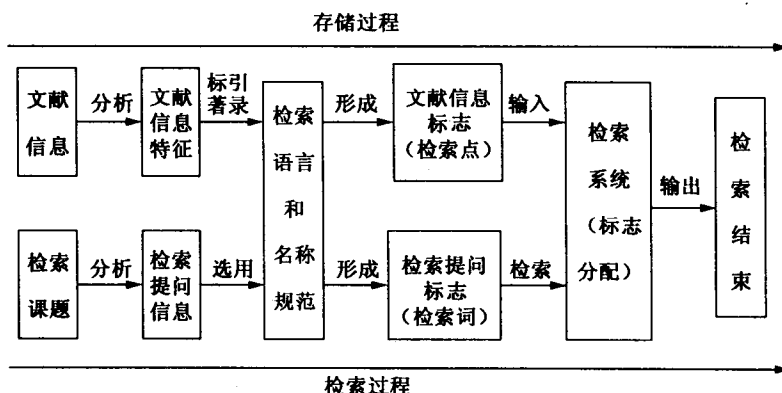


图 1-1 文献检索基本原理示意

存储过程主要是对信息源进行标引, 将其外表和内容特征 (如文献的标题、作者、来源和主题等) 用特定的检索语言转化为一定的标志 (如主题词、分类号和类目名称等), 再将这些标志按一定的顺序编排后输入检索系统, 从而为检索提供有规可循的途径。

为了保证文献信息能存得进、取得出, 就必须使文献存储所依据的规则与文献信息

检索所依据的规则尽量做到一致。也就是说,为了检索过程的顺利进行和达到较高的检索效率,除了在存储和检索过程的各个环节必须依据一定的方法和规则外,还必须有一致的检索语言和名称规范作为存储人员和检索人员的共同依据。

三、信息检索系统

信息检索是通过对信息系统的操作来实现的,信息系统包含信息集合的载体、技术设备和语义规则。按信息的存贮媒体和技术手段来划分,把信息系统分为两大部分:手工检索系统和计算信息检索系统。

1. 手工检索系统 (manual retrieval system)

手工检索是计算机检索的基础和先声。现在计算机检索的基本理论和检索方法都是从手工检索发展而来的。而且,目前手工检索工具在所有的检索工具中仍占有相当大的比重,某些学科和科研课题的检索利用手工检索也较为方便;同时,手工检索与计算机检索相比查准率更高,也不需要什么费用,所以,了解重要的手工检索工具及其检索方法仍有其必要性。

手工检索工具的组成主要由三部分组成:

文摘部分:这部分是对收录文献的详细介绍,包括文摘号、文献标题、作者、作者工作单位、文献来源(如期刊、会议录、专利等)、文摘。

索引部分:这部分将所有文献按一定的文献特征(如主题词、分类号、作者等)以一定的次序(如字母顺序、拼音顺序等)排列,每一个文献特征对应一个或几个文摘号或其他号码。从索引中获取的文摘号等可以过渡到文摘部分从而了解该文献。重要的索引包括主题索引、分类索引和作者索引等。

附录部分:包括主题词表(供选择主题词用)、缩略语表、期刊来源索引等。

有些检索工具将前两部分放在一起(如《工程索引》),有些索引部分与文摘部分分开(如《化学文摘》的卷索引),有些索引对文献不做详细介绍(如《最新目次》中收录的文献就没有文摘),也有些索引没有附录。

2. 计算机检索系统 (computer - based retrieval)

与手工检索相比,计算机检索的特点是:

(1) 速度快、效率高,仅几分钟就可以从成千上万条记录中找出所需信息。

(2) 检索范围广,可以迅速而方便地浏览相关学科或主题的所有数据库中的记录,在网络中,几乎每一台个人计算机都可以成为信息源。

(3) 检索不受时空的限制,只要拥有相应的软件和硬件设备,就可以在任何地方借助光盘和通信网络查询所需信息。

(4) 由于数据更新快,可以及时获得最新信息。

(5) 检索辅助功能完善,使用方便,检索软件可采用菜单驱动,几乎所有检索系统都有查询服务或提供操作演示盘,界面友好的程度在不断提高。检索结果的输出方式丰富多样,并可以按要求做排序、统计、绘图等加工。

在后面的章节里将对计算机检索有深入论述,这里不再多谈。

四、信息检索的方法和策略

1. 检索方法

检索方法是为实现检索计划或方案所提出的检索目的而采取的具体操作方法或手段的总称。检索获取知识信息的方法主要有直接检索和间接检索两种。

直接检索是通过浏览各种出版物上发表的学术论文、专利说明书等,以了解有关学科或专题发展动态的一种最简单的检索方法。其优点是能立即明确判断所包含的知识信息是否具有针对性和实用价值;不足是存在较大的盲目性和偶然性,查全率较低。

间接检索是借助于各类检索系统,从数量庞大的信息集合中,迅速、准确地查找特定课题有关知识信息的常用检索方法,其优点是所获得知识信息的全面性和准确性都较高,这也是人们普遍采用的方法。

间接检索方法的全过程,一般可分为信息需求分析、制定检索策略、实施检索策略和评价检索效果。在此主要对检索策略做一个详细的解释。

2. 检索策略

检索策略是为达到检索目标而制定的具体检索方案或对策。制定检索策略一般包括:检索系统的选择、确定检索途径或检索单元、拟定检索程序。按检索系统使用的检索设备和检索手段有手工检索策略和上机检索策略之分。

(1) 手工检索策略。

第一步,选择检索系统。要考虑选择与信息需求结合紧密、学科专业对口、覆盖信息面广、量大、报道及时、揭示信息内容准确、有一定深度、索引体系完善的手工检索系统为最佳,如文摘类和题录类检索系统。

第二步,确定检索途径。检索途径由所选择的检索系统提供,主要有:分类途径、主题途径、著者途径、号码途径等。其中分类和主题途径是最常用的检索途径。分类检索途径以学科体系为入口进行检索,具有族性检索的特点,查全率较高。但一般只能满足单维概念的检索,对多维概念的检索,查准率较低。若信息需求范围较宽,泛指性较强时,宜选用分类途径。主题检索途径以叙词或关键词为入口进行检索,具有特性检索的特点,查准率较高,能满足多维概念的检索,并能及时地反映新兴学科、交叉学科和边缘学科的发展。若信息需求范围窄、专指性要求强时,宜选用主题途径。无论哪种检索途径,都有各自的特点,究竟选择哪一种,既要根据检索的目的与要求,还要根据所选检索系统可提供的检索途径来确定。当选择的检索系统提供的检索途径较多时,各种途径可交叉运用产生互补效应,从而使检索效果更接近需求。

最后,拟定检索程序。在手工检索过程中,是由人的手查、眼阅,检索提问与存储标志之间的比较是靠大脑随时思维作判断而完成的,检索需求往往只存于人的大脑,因此,检索程序不必写成书面的表达语句,可以边查边思维,根据需求随时修改检索策略。

(2) 计算机检索策略。

第一步,选择检索系统。除了要考虑所选择计算机检索系统是否包含有与信息需求结合紧密、学科专业对口、覆盖信息面广、量大、报道及时、揭示信息内容准确、有一定深度的数据库外,还要考虑系统的检索功能是否完善。如著名的 Dialog 系统,不仅数

数据库所包含的文档多,而且学科的涵盖面也大。

第二步,确定检索单元。检索单元是表达信息需求的基本单元,也是与系统中有关数据库进行匹配运算的基本单元。检索单元选择得当与否,会直接影响其检索效果。检索单元可分为两类,一类是表示主题概念的检索词,如叙词(经规范化处理的自然语言词汇)和自由词(未经规范化处理的自然语言词汇),是常用的检索单元;另一类是某些特殊的符号,如分类号、代码、作者姓名等。检索词的选择与确定,主要遵循下列两个原则:

一是根据检索课题所涉及的学科专业和技术内容选词。少数检索课题,可直接选用课题名称中的主要概念作检索词。例如,检索“夏商周断代史”方面的信息,选用“夏代”、“商代”、“周代”、“断代史”这两组4个检索词,就基本能表达该课题所需信息的要求。但大多数检索课题,须从专业、技术的角度对研究内容进行仔细分析,才能找出全面确切表达主题概念的检索词。

二是对检索词进行处理。在机检系统中,计算机只能从词形上辨别所扫描的记录是否符合检索要求,不可能像人的大脑那样去考虑检索词的含义。因此,要考虑所选择的检索词是否与数据库相容,须将检索词做如下处理:用有关数据库的词表或相应的印刷型词表进行比较对照,使其从概念上、词形上与词表上的词表达一致。此时,检索词作叙词使用。如选择的检索词无词表可查,或未在词表中反映出来,此时,检索词作自由词使用。自由词作检索词时,要注意尽量从专业技术角度出发,尽可能使用国际上通用的术语。

第三步,拟定检索程序。在计算机检索过程中,检索提问与存储标志之间的对比匹配是机器进行的,拟定检索程序的核心是构造一个既能表达检索课题需求,又能为计算机识别的检索提问式。检索提问式是检索策略的具体体现,是计算机检索的依据。构造检索提问式主要使用布尔逻辑算符、位置逻辑算符、截词符、限制符等,将检索词进行组配,确定检索词之间的概念关系或位置关系,准确地表达课题需求的内容,以保证和提高信息的查全率和查准率。

布尔逻辑算符常用的有三种,其表达形式一般为 AND 或 *、OR 或 +、NOT 或 -。布尔逻辑算符的作用是把检索词连接起来,构成一个检索式的基本框架。对某些较为复杂的检索需求,仅用布尔算符还达不到要求,检索结果会出现较高的漏检和误检。

位置逻辑算符常用的有六种,其表达形式在不同的检索系统中用不同的符号表示。位置逻辑算符的作用是对复合检索词进行加工修饰,限定词与词之间位置关系,可弥补布尔逻辑算符只是定性规定检索词的范围,使检索结果的查准率提高。

截词是指在检索词的合适位置进行截断,按截断的位置分为后截词、前截词、前后同时截词和中截词。其中较常用的是后截词和中截词,其表达形式在不同的检索系统中用不同的符号表示。截词符的作用是对检索词进行截词处理,解决一个检索词的单、复数问题,词干相同而词尾不同的问题以及英美词汇拼写差异的问题等,既可节省输入的字符,又可达到较高的查全率。

限制符的作用是限制检索词或检索式在数据库记录中出现的字段位置。数据库可供检索的字段通常分为主题字段和非主题字段。主题字段主要有标题(TI)、文摘(AB)、

叙词 (DE) 和标志词 (自由词) (ID) 四个字段。非主题字段主要有作者 (AU)、语种 (LA)、年代 (PY)、刊名 (JN) 等字段。

在构造检索提问式中所使用的各类算符的确切含义和具体用法, 将在本书后续章节中做专门叙述。

第三节 计算机信息检索

计算机信息检索就是利用计算机对信息进行存贮与检索。相对于手工检索, 计算机检索无论在检索途径、检索速度, 还是在检索范围、检索时差方面, 均有很大的优势, 因而得到了广泛的应用。目前, 运用计算机检索系统进行文献信息检索已越来越普遍, 而且, 随着科学技术及计算机检索技术的进一步发展, 计算机信息检索将逐步取代手工检索, 而成为文献信息检索的主要手段。

一、计算机信息检索的发展阶段

计算机信息检索的发展过程是与计算机技术及其他现代科学技术的发展过程紧密相关的。计算机信息检索始于 20 世纪 50 年代初, 大体经历了三个发展阶段。

1. 脱机检索阶段

20 世纪 50 年代中期到 60 年代中期。自 1946 年 2 月世界上第一台电子计算机问世以来, 人们一直设想利用计算机查找文献。进入 20 世纪 50 年代后, 在计算机应用领域“穿孔卡片”和“穿孔纸带”数据录入技术及设备相继出现, 以它们作为存贮文摘、检索词和查询提问式的媒介, 使得计算机开始在文献检索领域中得到了应用。

1954 年, 美国海军兵器中心首先采用 IBM-701 型计算机建立了世界上第一个科技文献检索系统, 实现了单元词组配检索, 检索逻辑只采用“逻辑与”, 检索结果只是文献号。1958 年, 美国通用电器公司将其加以改进, 输出结果增加了题名、作者和文献摘要等项目。1964 年, 美国化学文摘服务社建立了文献处理自动化系统, 使编制文摘的大部分工作实现了计算机化, 以后又实现了计算机检索。同年, 美国国立医学图书馆建立了计算机数据库, 即医学文献分析与检索系统, 不仅可以进行逻辑“或”、“与”、“非”等种运算, 而且还可以从多种途径检索文献。

这一阶段主要以脱机检索的方式开展检索服务, 其特点是不对一个检索提问立即做出回答, 而是集中大批提问后进行处理, 且进行处理的时间较长, 人机不能对话, 因此, 检索效率往往不够理想。但是, 脱机检索中的定题服务对于科技人员却非常有用, 定题服务能根据用户的要求, 先把用户的提问登记入档, 存入计算机中形成一个提问档, 每当新的数据进入数据库时, 就对这批数据进行处理, 将符合用户提问的最新文献提交给用户, 可使用户随时了解课题的进展情况。

2. 联机检索阶段

从 20 世纪 60 年代中期到 70 年代初。由于计算机分时技术的发展, 通信技术的改进, 以及计算机网络的初步形成和检索软件包的建立, 用户可以通过检索终端设备与检索系统中心计算机进行人机对话, 从而实现对远距离之外的数据库进行检索的目的, 即

实现了联机信息检索。

这个时期,由于计算机处理功能的加强,数据存贮容量的扩大和磁盘机的应用,为建立大型的文献数据库创造了条件。例如美国的 DIALOG 系统(DIALOG 对话系统)、ORBIT 系统(书目情报分析联机检索系统)、BRS 系统(存贮和信息检索系统)、欧洲的 ESA-IRS 系统(欧洲航天局信息检索系统)等都是在此时期开始研制并逐步发展起来的,并且均在国内外或组织范围内得到实际应用。

可以说,联机检索是科技信息工作、计算机、通讯技术三结合的产物,它标志着 20 世纪 70 年代计算机检索的水平。

3. 网络化联机检索阶段

20 世纪 70 年代初到现在。由于电话网、电传网、公共数据通信网都可为情报检索传输数据,特别是卫星通信技术的应用,使通信网络更加现代化,也使信息检索系统更加国际化,信息用户可借助国际通讯网络直接与检索系统联机,从而实现不受地域限制的国际联机信息检索。尤其是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数据库的数据。这种联机信息系统网络的实现,使人们可以在很短的时间内查遍世界各国的信息资料,使信息资源共享成为可能。

可以说,联机网络和检索终端几乎遍及世界所有国家和地区,使得国际联机信息检索的发展达到了相当高的水平,开展商业性国际联机检索服务的大机构已达 200 余家,像美国的 DIALOG 信息公司已成为全世界最为著名的网络联机检索服务机构。

计算机信息检索的实现,大大方便和加速了信息资源的交流和利用,并对社会经济的发展 and 人们的科研方式产生了深刻的影响,从而也极大地促进了科技的进步。

二、计算机信息检索系统的构成

计算机信息检索系统主要由三个部分构成,即硬件部分、软件部分和信息数据库。

1. 硬件部分

硬件是指以计算机主机为中心的一系列机器设备,包括主机、外围设备以及与数据处理或数据传送有关的其他设备。

主机可以是大型、中型计算机,也可以是微型机,它负责整个系统的运行和管理,一般应具有较高的运算速度和较快的响应速度,同时还要有较强的逻辑运算能力及对大量文献或数据的处理能力。

外围设备包括外部存贮器、输入输出设备等。存贮器用以存贮原始数据、程序和计算结果。外部存贮器主要包括磁带、磁盘、光盘等。输入输出设备包括各种必要的数据录入装置和输出装置,输入设备有数据录入设备、键盘、鼠标器、光笔、缩微扫描器等;输出设备有显示终端、打印机等。

其他与数据处理或数据传送有关的硬件设备还有运算器、控制器等。运算器主要用于对数据进行运算、比较、判断、组合等加工处理,从而完成计算机的计算工作,其运算结果传送到存贮器中。控制器主要控制输入输出设备、内外存贮器和运算器,使之按

照程序的规定,产生所需要的命令,并控制命令的执行和信息处理。

广义地讲,硬件设备还应包括电源设备和通信设备及网络设备。

2. 软件部分

计算机软件又称计算机程序,是指控制计算机进行各种作业的一系列指令和进行“人机对话”及各种数据的存贮和传输的“翻译”规则。软件都存放在磁盘或磁带上,也可固定存放在主存储器中,一旦需要就可取出并送到控制器中,控制器则按程序中的指令操纵各个部件进行工作。软件包括操作系统软件、语言编译软件、应用软件和用户软件等。没有这些软件,计算机就不能运行。

3. 数据库

数据库是计算机信息检索的重要组成部分。它是以二进制代码形式在计算机存储设备上(如磁带、磁盘和光盘等)合理存放的相关数据的信息集合,通常由存储信息记录及其索引的若干文档组成。

目前,大多数计算机信息检索系统都存有四种类型的数据库:文献型数据库、数值型数据库、词典型数据库和全文型数据库。

(1) 文献型数据库,又叫书目型数据库。记录文档中主要存入的是原始文献的书目,例如原始文献的篇名、作者、文献出处、文摘、关键词等,主要用于查询各种文献资料的书目性线索、文摘等。

(2) 数值型数据库。数据库的记录存入各种调查数据或统计数据。它是一种以自然数值形式表示、计算机可读的数据集合,这些数据是从文献中分析、概括、提取出来,或在调研、观测及统计工作中直接获得的数据。数值型数据库主要用于查询各种有关的数字、参数、公式等。

(3) 词典型数值库。这类数据库的文献记录主要介绍一些有关公司、团体或名人的情况,供用户查询某一事物发生的时间、地点、过程或简要情况,或诸如化学物质名称、结构、俗称和化学物质登记号之类指南性信息,又称之为指南型数据库。

(4) 全文型数据库。这类数据库的文献记录存入了原始信息的正文。通过它可以直接检索出原始信息的全文,而不必进行二次检索,从而大大方便了用户,也提高了信息的有效利用。

三、基本检索方法

1. 布尔检索

利用布尔逻辑算符进行检索词或代码的逻辑组配,是现代信息检索系统中最常用的一种方法。常用的布尔逻辑算符有三种,分别是逻辑或“OR”、逻辑与“AND”、逻辑非“NOT”。用这些逻辑算符将检索词组配构成检索提问式,计算机将根据提问式与系统中的记录进行匹配,当两者相符时则命中,并自动输出该文献记录。

下面以“计算机”和“文献检索”两个词来解释三种逻辑算符的含义。

(1) “计算机”AND“文献检索”,表示查找文献内容中既含有“计算机”又含有“文献检索”词的文献。

(2) “计算机”OR“文献检索”,表示查找文献内容中含有“计算机”或含有“文

献检索”以及两词都包含的文献。

(3) “计算机” NOT “文献检索”，表示查找文献内容中含有“计算机”而不含有“文献检索”的那部分文献。

检索中逻辑算符的使用是最频繁的，对逻辑算符使用的技巧决定检索结果的满意程度。用布尔逻辑表达检索要求，除要掌握检索课题的相关因素外，还应在布尔算符对检索结果的影响方面引起注意。另外，对同一个布尔逻辑提问式来说，不同的运算次序会有不同的检索结果。布尔逻辑算符使用正确但不能达到应有检索效果的事情是很多的。

2. 截词检索

截词检索就是用截断的词的一个局部进行的检索，并认为凡满足这个词局部中的所有字符（串）的文献，都为命中的文献。按截断的位置来分，截词可有后截断、前截断、中截断三种类型。

不同的系统所用的截词符也不同，常用的有？、\$、*等。分为有限截词（即一个截词符只代表一个字符）和无限截词（一个截词符可代表多个字符）。下面以无限截词举例说明：

后截断，前方一致。如：comput? 表示 computer, computers, computing 等。

前截断，后方一致。如：? computer 表示 minicomputer, microcomputers 等。

中截断，中间一致。如：? comput? 表示 minicomputer, microcomputers 等。

截词检索也是一种常用的检索技术，是防止漏检的有效工具，尤其在西文检索中，更是广泛应用。截断技术可以作为扩大检索范围的手段，具有方便用户、增强检索效果的特点，但一定要合理使用，否则会造成误检。

3. 原文检索

“原文”是指数据库中的原始记录，原文检索即以原始记录中的检索词与检索词间特定位置关系为对象的运算。原文检索可以说是一种不依赖叙词表而直接使用自由词的检索方法。

原文检索的运算方式，不同的检索系统有不同的规定，其差别是：规定的运算符不同；运算符的职能和使用范围不同。原文检索的运算符可以通称为位置运算符。从 RECON、ORBIT 和 STAIRS 三大软件对原文检索的规定，可以看出其运算符主要是以下 4 个级别：

记录级检索，要求检索词出现在同一记录中；

字段级检索，要求检索词出现在同一字段中；

子字段或自然句级检索，要求检索词出现在同一子字段或同一自然句中；

词位置检索，要求检索词之间的相互位置满足某些条件。

原文检索可以弥补布尔逻辑检索、截词方法检索的一些不足。运用原文检索方法，可以增强选词的灵活性，部分地解决布尔检索不能解决的问题，从而提高文献检索的水平和筛选能力。但是，原文检索的能力是有限的。从逻辑形式上看，它仅是更高级的布尔系统，因此存在着布尔逻辑本身的缺陷。

4. 加权检索和聚类检索

(1) 加权检索。