

内容简介

本书对国际联机情报检索原理与方法作了系统阐述。内容主要涉及国际联机情报检索系统基本组成、数据库结构、DIALOG Version 2 检索功能、国际联机情报检索基本原理与检索策略,以及提高检索质量的方法讨论。

全书内容紧密结合具体检索实例,重点突出,原理清晰,可作高等院校情报专业教学用书和广大用户、专业检索人员和研究生的参考书。

国际联机情报检索原理与实践

邓汉成 著

*

兵器工业出版社 出版发行

(北京市海淀区车道沟 10 号)

各地新华书店经销

兵器工业出版社印刷厂印装

*

开本:787×1092 1/16 印张: 18 字数: 440 千字

1989 年 10 月第 1 版 1989 年 10 月第 1 次印装

印数:2000 定价:35.00 元

ISBN 7-80038-063-7/TP·4

目 录

第一章 绪 论	(1)
第一节 国际联机情报检索产生的背景	(1)
第二节 国际联机情报检索发展简况	(2)
第三节 国际联机情报检索系统的组成	(4)
第四节 国际联机情报检索系统的服务方式与检索范围	(6)
第五节 国际联机情报检索的主要优缺点	(7)
第二章 数据库结构	(9)
第一节 国际联机情报检索系统的数据库类型	(9)
第二节 文献记录的结构与特点	(10)
第三节 文档结构与特点	(16)
第三章 DIALOG 系统的基本检索功能	(21)
第一节 检索终端与国际联机检索系统的联机过程	(21)
第二节 DIALOG 系统的基本检索指令	(23)
第三节 DIALOG 系统的逻辑算符与截词符	(34)
第四节 检索结果的输出格式	(47)
第五节 检索结果的表格输出格式	(58)
第四章 DIALOG 系统的主要辅助检索功能	(65)
第一节 数据库词表的联机查询	(65)
第二节 联机检索中的字段与特征限定	(70)
第三节 联机检索中的范围检索	(92)
第四节 联机检索中的检索保留	(95)
第五节 多文档同时检索	(104)
第六节 定题检索与联机订购	(109)
第七节 其他功能	(116)
第五章 国际联机情报检索的基本原理与评价标准	(118)
第一节 检索的一般过程	(118)
第二节 计算机情报检索处理的逻辑过程	(120)
第三节 计算机情报检索的评价标准	(126)
第六章 联机情报检索策略	(130)
第一节 什么是检索策略	(130)
第二节 情报需求分析	(132)
第三节 数据库和检索途径的选择	(134)

第七章 概念组配——情报需求的概念表达	(146)
第一节 概念组配的含义与分类	(146)
第二节 概念组配与情报需求的关系	(150)
第三节 概念组面对概念组配的影响	(152)
第四节 概念组面及其逻辑关系的确定	(162)
第八章 标识组配——情报需求的标识表达	(170)
第一节 标识组配的含义	(170)
第二节 标识组配与概念组配的关系	(172)
第三节 检索单元的切题与匹配	(175)
第四节 标识组配的切题与匹配	(185)
第五节 标识组配的切题与匹配的调节	(195)
第九章 联机检索的信息反馈	(203)
第一节 联机检索前的信息反馈	(203)
第二节 联机检索中的信息反馈	(205)
第三节 联机检索后的信息反馈	(207)
第十章 国际联机情报检索中的质量要求	(209)
第一节 查全率与查准率	(209)
第二节 国际联机情报检索的质量要求	(215)
附 1 DIALOG 系统操作指令一览表	(220)
附 2 DIALOG 系统联机文档查询与多文档检索的文档主题范围与目录	(224)
附 3 DIALOG 联机情报检索系统数据库目录	(245)
附 4 ESA / IRS 联机情报检索系统数据库目录	(267)
附 5 ORBIT 联机情报检索系统数据库目录	(271)
附 6 BRS 联机情报检索系统数据库目录	(274)
附 7 STN 联机情报检索系统数据库目录	(278)

第一章 绪 论

当前,世界各国正面临着一次新的工业革命。其特征是计算机、生物工程、光导纤维、激光、超导材料、机器人……等高技术的广泛研究、开发与利用,社会生产力将发生一个新的飞跃。这对我国实现社会主义四个现代化也是一个极好的机会。

要利用好这个机会,信息沟通是一个很重要的前提。只有信息准确灵通、传递迅速,经济决策才有可靠基础,社会进步才能加快速度。英国的一位学者曾经预言:“信息技术是社会进步的关键,对中国来说,掌握信息技术现代化建设的进展可以加快十年。”国际联机情报检索的出现和应用,为沟通各种信息提供了极为方便的条件。由于它能极其迅速、准确地获取世界性的科技或其他情报,因此,它在我国一出现,就引起了人们的极大兴趣与关注,并在我国的经济建设中显示出越来越重要的作用。

第一节 国际联机情报检索产生的背景

国际联机情报检索是用户利用检索终端,通过国际通讯网络与国际联机情报检索系统的中央计算机相互进行数据传输(即所谓“人机对话”),从检索系统的数据库中查找用户所需情报的一种计算机情报检索。这种现代化情报检索技术的出现是人类知识急剧增长和现代科学技术发展的必然结果。

一、现代社会与科学发展的客观需要

当代科学技术的发展异常迅速。据统计,近四十年来出现的科技成果远远超过人类历史2000年来的总和。科学知识的相应增长也快得惊人。例如,在19世纪,科学知识每50年增长一倍,到了20世纪中叶,就缩短到每10年增长一倍,到70年代每5年增长一倍,到80年代就缩短到每3年增长一倍。科学技术的发展不仅导致知识量的猛增,而且也引起学科内部结构的深刻变化:一方面,学科越分越细,形成了许多高度专门化的新型学科;另一方面,学科之间又相互交叉渗透,形成了许多高度综合的边缘学科。科学技术发展的这些特点也使情报的发展产生了巨大的变化,主要表现在:

1 文献种类和数量“爆炸性”的增长

以期刊为例,19世纪末,世界上还只有100多种期刊,到20世纪60年代就增加到一万余种,到了1980年猛增到5万余种。目前,全世界每年出版的各种图书可达五、六十万种之多(总数约100亿册)。文献量的增长更是惊人。现在,每年发表的国际科技会议论文可达10多万篇,申请专利100多万件,科技论文500多万篇……,平均七、八年就要翻一翻。难怪有人惊呼现在已处在“情报爆炸”、“信息爆炸”的时代。

2 文献分布异常分散

学科的交叉渗透引起文献的分布异常分散。例如,美国麻省理工学院曾经统计,就电工方面的文献而论,仅有50%刊登在1000多种电工杂志上,而其余的50%则分散在物理、机械、化工、生物等杂志上。

3 文献寿命明显缩短

科学技术的迅猛发展,新理论、新技术、新工艺、新材料和新产品不断出现,大大缩短了知识更新的周期。据苏联《发明问题》杂志的统计分析认为,各类文献的平均时效是:图书 10~20 年、科技报告 10 年、学位论文 5~7 年、期刊与连续出版物 3~5 年。而西方国家认为,有 80~90%文献的有效周期只有 5 年。

面对这种严峻的现实,如何提高人们准确、迅速识别、处理、吸收和有效利用情报的能力已成为一个十分突出的问题。国际联机情报检索技术就是根据人们的这种迫切需求产生的。

二、计算机、空间技术和现代通讯技术的发展,为国际联机情报检索技术提供了物质条件

计算机的出现与发展,为情报的自动、快速处理、存贮和管理提供了坚强的物质基础,通讯网络、数据传输等技术又为远距离检索提供了有利条件,而空间技术与通讯技术的结合——国际卫星通讯技术的出现,使完全超越地区限制的“国际联机情报检索”变成了现实。

第二节 国际联机情报检索发展简况

尽管国际联机情报检索的实际使用始于 70 年代,但计算机在情报检索中应用的研究可追溯到 50 年代初。从那时算起,计算机情报检索技术的发展大致可分为三个阶段:

一、脱机检索阶段 (1954—1964)

早在 1954 年,美国海军军械试验站就开始从事这方面的研究,并用 IBM-701 型计算机建立了世界上第一个科技文献检索系统。随后,1958 年美国通用电器公司、1963 年日本外务省、1964 年美国国立医学图书馆也相继进行了这方面的研究,并投入实际使用。但是,这一阶段主要是利用计算机来进行文献目录、索引和文摘的前处理工作,也进行一些脱机定题检索。

二、联机检索阶段 (1965—1972)

这一阶段,由于分时技术的发展、计算机网络的形成和检索软件包的建立,使用户能够借助检索终端直接与检索系统的中央计算机进行“人机对话”,实现了情报的联机检索。这一阶段,比较有代表性的检索软件有:1965 年美国洛克希德导弹和空间公司为美国国家航空和航天局 (NASA) 设计的 RECON 软件、美国系统发展公司 (SDC) 研制的 ORBIT 软件,以及 1967 年美国洛克希德导弹和空间公司在 RECON 软件基础上发展起来的 DIALOG 软件。其中 RECON 在 NASA 数据库、ORBIT 在美国系统发展公司的 ORBIT 联机情报检索系统、美国国家医学图书馆的 MEDLINE 检索系统、DIALOG 在美国洛克希德导弹和空间公司的 DIALOG 情报检索系统的联机检索中,均得到实际应用。但是,这一阶段的计算机网络主要是通过电话线路联结,因而,联机检索受到一定地区的限制。

三、国际联机检索的发展与普及阶段 (1973—现在)

由于国际卫星通讯网络的出现,使用户可以在终端前直接通过国际通讯网络与检索系统

进行联机检索。真正实现了不受地区限制的国际联机情报检索。

这种现代检索技术的出现,使更多地区的用户能够直接进行联机检索,大大方便与加速了世界情报资源的交流与利用,使计算机情报检索服务业得到迅猛的发展。这种发展从联机检索市场销量也可窥其一斑:据 1982 年报导,世界上已有十多万联机检索用户,1980 年,美国的联机检索市场的销售总额约为 10.63 亿美元,欧洲约为 1.23 亿美元。而至 1985 年,美国就增加到 28.8 亿美元,欧洲猛增到 20.53 亿美元。

情报计算机检索服务的发展,也促进了这一现代检索技术的迅速发展。据西德《1986—1988 年发展规范》称:目前可供联机服务的数据库已有 1851 个,其中科技数据库 816 个,经济数据库 1035 个。这些数据库中,美国 1145 个,法国 126 个,加拿大 128 个,西德 82 个,英国 121 个。联机情报检索服务机构也迅速增加,目前,已有联机检索服务系统 100 多个(其中美国 60 多个、日本 25 个、欧洲、加拿大 20 多个)。其中在国际上影响较大的有:美国洛克希德导弹和空间公司的 DIALOG 系统,美国系统发展公司的 ORBIT 系统(现已合并到英国的 PERGAMON 公司),美国 IBM 公司的 STAIR 系统,美国文献目录检索服务公司的 BRS 系统,美国国家医学图书馆的 MEDLINE 系统,法国信息交换中心的 MISTAL 系统,欧洲空间组织的 ESA-IRS 系统,英国图书馆自动化情报服务处的 BLAISE 系统,英国 PERGAMON 公司的 INFOLINE 系统和日本科技情报中心的 TOIS 系统等。

我国的联机情报检索起步较晚,国际联机情报检索又更迟一步。1980 年初,由建工总局建筑研究院(现为中国建筑技术发展中心)情报所等单位在我国驻香港海外建筑工程公司设置了我国第一台国际联机情报检索终端,通过香港大东电报局与美国的 DIALOG 和 ORBIT 系统联机。但由于终端在香港,决策机构在北京,限制了决策人员与中央计算机的人机直接对话。

1981 年底,北方科技情报所开始与美国的 DIALOG 系统直接联机。由于检索终端设在北京,检索人员可以在终端旁与 DIALOG 系统直接进行“人机对话”。因而,有利于检索效果的提高。

1982 年 9 月,冶金部、石油部、化工部等部委情报所在北京实现了与 DIALOG 或 ORBIT 系统的直接联机。

同年年底,华东工学院在南京设置了我国在北京地区以外的第一台国际联机情报检索终端,与 DIALOG 系统进行了联机检索业务。

但是,当时的终端,除在香港的终端外,都是采用 50 波特传输速率的电传终端。因而,耗费通讯时间和机时较多,费用较贵。

1983 年 10 月,中国科技情报所开始使用 300 波特速率的数传终端与欧洲空间组织的 ESA-IRS 系统联机。

不久,北方科技情报所和水电部情报所也以 300 波特速率的数传终端与 ESA-IRS 系统实现了联机。

迄今,在我国,已有 20 几个省、市设置了 70 多个国际联机情报检索终端,与 DIALOG,ESA-IRS,ORBIT,BRS,INFOLINE,DMS/DRI,FIESTA 等国际联机情报检索系统进行了联机。为我国经济建设、科学研究、技术开发、管理决策等工作提供了大量极有价值的信息。

第三节 国际联机情报检索系统的组成

国际联机情报检索是借助检索终端通过国际通讯网络与国际联机情报检索中心的中央计算机联机来进行情报检索的。因此，一个国际联机情报检索系统应该包括（见图 1-1）：

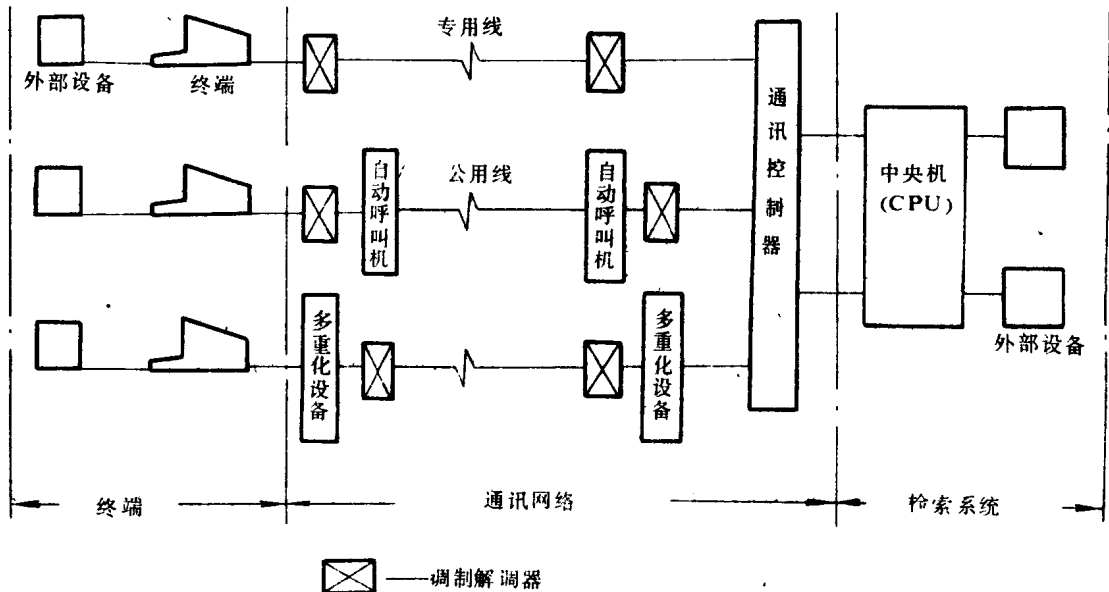


图 1-1 联机情报检索系统示意图

一、检索终端 (Terminal)

二、国际通讯系统 (International Communication System)

三、国际联机情报检索中心

(一) 检索终端

是用户与情报检索中心的中央计算机进行“人机对话”的设备。用户利用终端向中央计算机发送检索系统约定的指令或情报语言，并通过终端反馈中央计算机的响应。

目前，常用的检索终端设备有：

1. 屏幕显示终端。这种终端由一个键盘和一个显示屏幕组成。可以带一台打印机。数据通过键盘发送，同时显示在屏幕或打印到记录纸上。中央计算机的响应也显示在屏幕或打印到记录纸上。这类检索终端可以进行 300、1200 或 2400 波特速率的数据传输。

2. 微机终端。配上一定的通讯软件，就可以将微机作检索终端。检索时，数据通过键盘发送，也可预先存入内存或外存设备，由微机直接发送。发送的数据直接显示在屏幕或打印到记录纸上。中央计算机的响应也显示在屏幕或打印到记录纸上，或直接存入内存或外存设备。使用不同的调制解调器和通讯规程，微机终端可以进行 300、1200、2400 或其他波特速率的传输。目前，在我国多半使用 IBM-PC 系列或兼容机作终端。

3. 电传终端。普通的电传机也可用作检索终端。其操作与一般电传发送相同。发送前先将数据在纸带上进行穿孔，用穿孔纸带在电传机上直接发送，也可以通过电传机的键盘进行发送。发送的数据直接打印在记录纸上，如果电传机配有显示屏幕，也可直接显示在屏幕上。中央计算机的响应也可打印在记录纸或显示在屏幕上。电传机终端一般只能进行 50 波特速率的传输。

(二) 通讯系统

检索终端通过通讯网络与中央计算机联机。联机检索的通讯系统，一般包括通讯线路、自动呼叫应答机、调制解调机、通讯控制器和多重化装置等设备。

1. 通讯线路。目前，在我国国际联机有三种通讯方式：

(1) 各分终端通过专用线或公用长话线与北京各检索中心的终端相接，再由这些中心通过北京邮电大楼接北京通讯卫星地面发射站，通过 ITACABL 国际通讯网络与欧洲的 ESA / IRS 系统联机，或者转而通过罗马分组交换站与美国的 DIALOG、ORBIT 或英国的 INFOLINE 等国际联机情报检索系统联机。

(2) 通过公用长话线或专用线与香港大东电报局相接，由该局进入 TYMNET 或 TELNET 等国际通讯网络与美国的 DIALOG 或 ORBIT 等系统联机。

(3) 用电传机作终端，直接通过电传线路与各检索系统联机。

专用线路质量好，使用方便，但租用费用昂贵。公用线路使用费用便宜，但使用质量不稳定，特别目前我国许多地方的公用线路通讯质量还不高，用这样的线路进行数据传输往往出现较多的误码，影响检索效果。电传线路的传输速率限定在 50 波特，因此，检索时占用较多的机时与通讯时间，检索费用较高。

2. 调制解调器。在通讯线路中，信号以电脉冲形式进行传输。因此，从终端或中央计算机发送的数据信号进入通讯线路之前必须将其转换成脉冲信号；而这些脉冲信号进入终端或中央计算机之前又必须将其转换成数据信号。调制解调器就是将数据信号与通讯传输信号之间进行变换与反变换的设备。

3. 自动呼叫应答器。仅用于公用线路。联接在调制解调器与通讯线路之间。其作用与自动电话拨号机相类似。许多单位就用自动电话拨号机。

4. 多重化装置。当多个终端共用一条通讯线路时，可以在通讯双方各配置一个多重化装置，解决在同一条线路上传输多个信号的问题。

5. 通讯控制器。一个由程序控制的通讯控制设备。用来控制线路、数据传输和中央计算机的接口等。

(三)、国际联机情报检索中心

主要由中央计算机、数据库和一些其他外部设备组成：

1. 中央计算机 (Computer)。是检索系统的核心部分。通过一定的检索软件使其进行情报的存贮、处理、检索和整个系统的运行和管理。如 DIALOG 系统在 1980 年用一台 IBM-3033 联结 TYMNET 卫星通讯网络，一台 NAS-AS9000 与 TELENET 相联 (现已有四台主机)，每秒可处理 1400 万条指令。

2. 数据库 (Databases 或 Databanks)。检索系统存贮文献或数据的记录和索引的场所。目前，主要用磁盘或磁带来存贮这些信息。如 DIALOG 系统使用的 IBM3330I 磁盘机和 IBM2321 磁卡机等。迄今，该系统已用这些设备存贮了 38 万多兆字符数相当于 1 亿 2 千多万条记录的信息。

3. 其他外部设备。例如操作台、高速打印机等。DIALOG 系统用三台 XEROX-9600 / 9700 高速打印机，处理每天联机检索的脱机打印记录。

不难看出，只需一台检索终端、调制解调器和一台自动电话拨号机 (如果使用公用通讯线路)，就可以建立一个分终端检索系统。

第四节 国际联机情报检索系统

的服务方式与检索范围

一、服务方式

主要有下面几种方式:

1. 追溯检索 (Retrospective Searching)

这种检索是用户利用检索终端与检索系统的中央计算机直接进行“人机对话”，从数据库中查找一定时间范围内的情报。这是联机检索中最常用的检索方式。

2. 定题检索 (Selective Dissemination of Information)

这种检索是将用户预先确定好的检索提问式长期保留在检索系统的中央计算机中，中央计算机在相应数据库更新时，自动按保留的检索提问的要求将存贮到数据库中有关的最新情报记录检索出来提供给用户。这对长期从事某一专题研究或开发，而又需要长期跟踪其最新动态的用户特别有用。

3. 联机订购原文

目前，许多国际联机情报检索系统也拥有一些全文数据库，但大部分数据库还只存贮二次文献记录。如果，在国内找不到原始文献而又迫切需要时，可以通过检索终端向检索系统订购原文的复制件。

4. 光盘服务

随着光盘技术的发展，一些国际联机情报检索系统还开展了光盘服务。例如，从 1987 年开始，DIALOG 系统已经可以向用户提供部分数据库的光盘盘片，称之为“DIALOG On Disc”系列产品。并可提供与之配用的软件和所需设备，盘片可以每季度更新一次。这种局部系统可以独立进行检索。如果不能完全满足情报需求时，用户可以随即通过国际通讯网络与 DIALOG 系统直接联机进行补充检索。转接时，完全由通讯软件自动控制，无需再由用户重新编写与键入检索提问式。

除了上面几种基本服务方式外，许多国际联机检索系统还可提供建立私人数据库、电子邮寄、用户培训、旅行指南和该系统的出版物等服务。

二、检索范围

有些国际联机情报检索系统，如美国国家医学图书馆的 MEDLINE 系统，完全存贮有关医学方面的文献，因而比较专一。而有许多检索系统，例如，DIALOG, ORBIT, BRS 和 ESA-IRS 等系统，存贮的情报内容范围极其广泛。例如，DIALOG 系统，目前拥有 260 多个数据库，其中涉及各种科学技术的数据库约占 34.4%，涉及金融、市场和商业方面的数据库约占 34.7%，涉及社会科学方面数据库占 17.2%，综合性数据库占 1.0%，训练用数据库占 ~9%，其他占 3.7%。因而，可检索的主题范围非常广泛。一些检索系统声称，可以提供以下所列方面的情报：

1. 及时了解与你专业有关的发展情况；
2. 编制所有专业的科技文献目录；
3. 跟踪了解与你研究课题有关的信息；

- 4.收集书写报告、演讲、备忘录及建议所需资料;
- 5.核对已出版的文献,避免研究课题重复;
- 6.审定财政预算;
- 7.汇集产品产销量数据;
- 8.预测经济发展趋势;
- 9.查找专利文献;
- 10.收集消费者对产品的评述意见;
- 11.了解产品竞争者的资料;
- 12.撰写各种论文;
- 13.制定新教学大纲;
- 14.了解世界各大图书馆馆藏情况;
- 15.了解世界金融、市场最新动态;
- 16.查阅技术标准与规格。

第五节 国际联机情报检索的主要优缺点

国际联机情报检索克服了地理位置的限制,又兼有一般计算机检索的特长,使其具有许多优点。主要表现在:

一、主题广泛

如前所述,许多国际联机情报检索系统拥有各种学科领域的数据库,而且,一个检索终端可以与许多国际联机情报检索系统建立业务联系。因而,在同一个检索终端可以检索各种主题范围的课题,满足各种专业和行业的用户需求。

二、使用方便

用户无需长途跋涉、走访各大图书、情报机构,在检索终端旁与有联系的国际联机情报检索系统联机,就可从数据库中查找自己所需的情报。使用十分方便。

三、查找迅速

计算机检索的速度大大超过传统手工检索。正常情况下,在检索终端旁,从拨号到与检索系统主机接通,只需一分钟左右。在一个存贮 20 多年有 100 多万篇文献记录的数据库中,检索一个语词,几秒钟内就可得到主机的响应。在这样的数据库中检索一个中等的课题,从联机到脱机,一般也只需要几分到十几分钟。这是一般手工检索所无法比拟的。

四、检索全面

如前所述,当今各学科的文獻资料分布异常分散。一般只有 50%左右的文献资料刊登在相应学科的杂志上,而其余部分则分散在其他学科的杂志上。而对边缘学科、新型学科更是如此。如果用传统的手工检索方法把这些极其分散的资料收集起来,简直是大海捞针,异常困难。而计算机检索可以很容易做到这点。它可以在检索系统的不同数据库中进行搜索,只需几分钟到十几分钟就可以得到具体结果。

五、内容新颖

国际联机情报检索系统都定期更新数据库中的信息。对不同数据库,更新周期有所不同。例如,《工程索引》(COMPENDEX)数据库每月更新一次,美国《化学文摘》(CA Search)数据库每二周更新一次,《世界专利索引》(WPI)每周更新一次。一些时事新闻、商业数据库更新的周期更短,例如,DIALOG系统的《商业与工业新闻》数据库每天更新一次,而《商业电讯》(BUSINESSWIRE)数据库每15分钟就更新一次。因此,国际联机情报检索系统可以向用户及时提供最新的世界性的各种信息。

六、反馈及时

计算机检索系统可以联机查询数据库中的检索词、或联机显示、或打印检索的记录。因此,可以立即判断检索的质量,及时调整检索提问式,以便获得较高的查准效果。但是,这需要花费一定的通讯时间和计算机机时,这意味着需要花费更多的检索费用。因此,在我国,目前一般不使用系统的这种功能。

国际联机情报检索也有其不足之处。主要有:

1. 检索费用较高

国际联机情报检索费用不仅包括文献记录的打印费和计算机检索时用去的机时费,而且还包括国际和国内通讯费。因而,需要较多的检索费用。一般情况,检索一个课题平均需要人民币100元左右。但是,如果用传统的检索方法需要外出检索,所花去的旅差等费用并不一定会低于联机检索所花去的费用。再说,国际联机情报检索具有能够迅速查找最新的世界性信息的能力,更是传统检索方法无法相比。而这些,对引进外资、外贸出口和及时掌握世界水平和动态等方面的需求尤为重要。

2. 检索质量较难控制

计算机不具备人脑的思维能力,完全按照输入的检索语词或其他标识进行机械“匹配”来命中文献。对一定检索系统,检索的查准与查全质量完全取决于所用检索语词及其组配关系。在以后将会看到,这是一个比较复杂的问题,这就需要检索人员应该具备较高的计算机情报检索知识与经验,和较广的专业知识。

第二章 数据库结构

数据库是计算机情报检索系统的重要组成部分。它是在计算机存储设备上（例如，磁盘、磁带或光盘等）合理存放的相互关联的数据的集合。通常，由存贮情报记录及其索引的若干文档组成。不同数据库，存贮不同主题内容、情报数量、时间范围和情报类型的信息，以及不同的标引方式，因而提供不同的检索范围和检索途径。所以了解国际联机情报检索系统的数据库特点，对提高检索效果有着重要意义。

第一节 国际联机情报检索系统的数据库类型

目前，大多数国际联机情报检索系统都存有四种类型的数据库：文献型数据库、词典型数据库、数值型数据库和全文型数据库。

一、书目型数据库 (Bibliographic Database)

记录文档中的记录主要存入的是原始文献的书目 (bibliography)。例如，原始文献的篇名、作者、文献出处、文摘、关键词、……等等。DIALOG 系统的《科学文摘》数据库就是一种书目型数据库。其记录存贮的文献书目见图 2-1。

二、词典型数据库 (Dictionary Database)

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，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的情报。图 2-2 是一篇 DIALOG 系统《索马斯地区厂商录》(THOMAS Regional Industrial Suppliers™) 数据库的记录实例。介绍了一家的名称、地址、电话和产品目录。这是一种词典型数据库。

三、数值型数据库 (Numeric Databank)

数据库的记录存入各种调查数据或统计数据，例如 DIALOG 系统的《PTS 美国时间系列》(PTS U.S TIME SERIES) 数据库提供了有关美国人口、国民经济总产值、人均收入、就业、主要原材料和产品的生产与消耗、能源、车辆、建筑、工业生产和价格系列等大约 500 项自 1957 年来的历史数据和到 1990 年的预测时间系列。其文献记录参见图 2-3。

四、全文型数据库 (Complete Text Database)

这类数据库的文献记录存入了原始情报的正文 (见图 2-4)。通过这类数据库可以直接检索出原始情报的全文，大大方便了用户，也提高了情报的有效利用。但是，目前大多数的全文型数据库的记录中还没有能包含原始文献中的图表。同时，全文型数据库需要占用更多的存贮空间，因而，目前的国际联机情报检索系统中，多半还是一些涉及金融、商业、市场或时事等方面情报的全文型数据库，例如 DIALOG 系统的《工商时报》(Commerce Business Daily) 就属此类数据库，而涉及科学技术方面的这类数据库还不多见。不过，近年来

这方面已有了可喜进展。例如, 联邦德国第四情报中心 (FIZ Karlsruhe) 已采用光栅 / 向量扫描技术建立专利全文数据库, 并准备欧、美、日三方合作建立从 1920 年开始的全文专利数据库。

除上述四种基本类型的数据库外, 有些国际联机情报检索系统的一些数据库同时兼有几种类型。例如, DIALOG 系统的《统计数据》(CENDATATM) 数据库是一种兼有全文、数值两种类型的数据库, 《国际 DUN'S 市场概况》(International DUN'S Market Identifiers) 为一种词典-数值型数据库, 《药物情报全文》(Drug Information Fulltext) 为一全文-词典型数据库, 而《PTS 世界工业产品市场与技术概况》(PTS PROMT) 则是一种文献-数值型数据库, 等等。

随着科技、经济与政治发展, 全文型、词典型与数值型数据库的需求量日益增多。但是, 目前的许多国际联机情报检索系统, 仍以书目型数据库为主。以美国 DIALOG 系统为例, 在 1987 年的 260 多个数据库中, 文献型数据库约占 58%、词典型数据库约占 24%、全文型数据库约占 10%、数值型数据库占 8%。不过, 词典型、数值型和数值型数据库的总数已占全部数据库的 42%。而且, 这一比重还在不断增大。也可以根据文献内容的范畴和形式(如学科、专题或文献形式等)来区分数据库。在 DIALOG 系统中, 按学科范畴存贮文献记录的数据库很多, 例如《化学文摘》数据库 (CA Search), 它是化学或化工为研究目的的重要文献的集合。按专题存贮文献记录的数据库如《世界铝文摘》(World Aluminum Abstract), 它专门收集铝的性质、制造、分析和应用等方面的文献, 按文献类型存贮文献记录的数据库如《美国政府研究报告文摘通报》(NTIS) 它主要收集了美国各种政府研究报告的二次文献。

这些数据库中, 有的是一种印刷出版物的机读版本, 例如, CA Search 数据库, 是美国《化学文摘》(CA) 的机读版本, COMPENDEX 数据库是《工程索引》(EI) 的机读版本, 等等; 有的是几种印刷出版物构成的机读版本, 例如 INSPEC 数据库包括了印刷出版物《物理学文摘》(Physics Abstracts)、《电工与电子学文摘》(Electrics and Electronics Abstracts)、《电脑与控制文摘》(Computer and control Abstracts) 和《情报技术》(IT Focus) 收集的文献; 有的则没有对应的印刷出版物。

有许多数据库, 按不同年分又分成若干子库(也叫“文档”)。例如, DIALOG 系统的《科学文摘》(INSPEC) 数据库分成 File 12 (1969-1976) 和 File 13 (1977-现在) 两个子库。《化学文摘》(CA Search) 数据库, 除有 File 399 (1967-现在) 总库外, 还分成 File 308 (1967-1971), File 309 (1972-1976), File 310 (1977-1981), File 311 (1982-1986) 和 File 312 (1987-) 等五个子库。

第二节 文献记录的结构与特点

文献记录 (Records) 是构成数据库顺排文档的基本单元, 也是数据库存贮情报款目和标引的载体。每篇记录都表述了一原始情报的内容和外表特征。

尽管对不同范畴与内容的数据库, 其文献记录标引的内容和形式有很大差别。但是, 对每一种记录, 总可以将其分成三种基本字段: 检索系统的存取号、基本索引和辅助索引字段。

一、存取号 (Accession Number)

这是计算机检索系统为各数据库的每篇记录规定的能被计算机识别的特定号码。在同一数据库中, 每篇文献记录只有一个存取号。存取号一般由 6~9 位数字组成, 出现在每篇联机记录的左上角位置。在 DIALOG 系统中称这种号码为 DIALOG 存取号。

二、基本索引字段 (Basic Index Fields)

这是记录中主要用来表达文献内容特征的字段。一般情况, 记录中表达文献内容特征的字段有:

1. 篇名字段 (Title Field, / TI)。这一字段记录了原始文献的篇名。一般直接引用原文篇名, 也有使用编辑修改后的篇名。

2. 文摘字段 (Abstract Field, / AB)。这一字段记录了描述原文主题内容的简明提要。通常由标引人员提供。对一些英文文献, 也有采用原文中原有的摘要。

3. 叙词字段 (Descriptor Field, / DE)。在 ESA / IRS 系统中称为受控词字段 (Controlled term Field)。这一字段中标引了有关原始文献主题的叙词或其他规范词。这些规范词都是由文献标引人员根据原始文献主题内容标引的能够表达文献主题内容的规范语词 (规范化的单元词或多元词)。这些语词都刊印在相应的规范词表中。例如, 《科学文摘》数据库 (INSPEC) 的 INSPEC 叙词表、《工程索引》数据库 (COMPENDEX) 的 SHE 标题词表, 等等。

有些数据库, 还把叙词分为主叙词 (major descriptors) 和次叙词 (minor descriptors)。用主叙词来表达原始文献的核心主题内容特征; 用次叙词表达原始文献中比较次要的主题特征。在联机记录中, 主叙词都标有星号作为其特征标志。

4. 自由标引词字段 (Identifier Field, / ID)。按原义, Identifier 意为限定词, 按习惯仍称为自由标引词。在 ESA / IRS 系统中就叫自由标引词 (Free Index terms) 或非控词 (Uncontrolled terms)。这类词也是由标引人员根据原始文献的主题内容标引能够表达文献主题内容的语词 (单元词或多元词)。但他们不是规范词, 因而不一定出现在规范词表中。这类词相当于一般所说的关键词。

除了上面几种常见的基本索引字段外, 还可以有其他类型的字段。例如, 全文数据库记录的正文字段 (Text Field) (见图 2-4)。

基本索引字段中的篇名、文摘 (或正文)、叙词和自由标引词等字段, 从不同角度表达了原始情报的主题内容。叙词字段中的语词是标引人员按照一定的情报语言规则标引的, 能够表达原始情报主题内容的一些规范语词。因而, 叙词字段中的每一语词 (多元词或单元词), 通常都有确切的概念含义和表达一确切的主题内容。自由标引词字段与叙词字段有类似的作用。而篇名、文摘或正文等字段都是用自然语言对原始情报主题内容的描述。这些字段在整体上可以表达原始情报的主题内容, 而其中的每个语词则不一定能够表达一个确切的概念, 因而, 不一定能够确切代表原始情报的某一主题内容。

基本索引字段中的叙词和自由标引词有单元词和多元词之分, 在联机记录中这些词都用分号隔开。篇名和文摘全以单元词作基本元素。其他一些基本索引字段, 有的只包括多元词, 有的也存有单元词。为了叙述方便, 把数据库记录的基本索引字段中的叙词字段和自由标引词字段里标引的单元词称为标引单元词; 把以词组形式作基本索引的多元词称为标引多

元词 (主要是叙词或自由标引词字段中的多元词); 而把全部基本索引字段中的全部单元词 (不包括系统规定的非用词 Stop words) 称为全单元词。不难看出, 全单元词包括组成标引多元词的单元词和叙词或自由标引词字段中的标引单元词。除标引单元词外的其他全单元词称为自由单元词。

三、辅助索引字段 (Additional Index Fields)

辅助索引字段主要是一些表达文献外表特征的字段。例如, 作者字段 (Author Field, AU=)、期刊名称字段 (Journal Field, JN=)、出版年份字段 (publication Year Field, PY=)、语种字段 (Language Field, LA=), 等等。也有许多辅助索引字段可以表达文献的内容特征。例如, CAS 登记号字段 (CAS Registry Number Field, RN=)、产品代码字段 (Product Code Field, PC=)、事件代码字段 (Event Code Field, EC=), 等等。

在计算机情报检索系统中, 数据库记录的辅助索引字段是一个种类繁多的字段。不同数据库, 记录中包含的辅助索引字段很不相同。即使在同一数据库中, 不同种类文献记录中包含的这类字段也不完全相同。对不同数据库, 记录中包含这类字段的数目也很不相同, 少的只有一二个; 多的可达八九十个之多。

图 2-1, 2-2, 2-3 和 2-4 分别为 DIALOG 系统的书目型、词典型、数值型和全文型数据库的四篇记录实例:

SAMPLE RECORD

√DIALOG Accession Number

0249373

SF=

→C79022216

←

AN=

WHICH MICROPROCESSOR SYSTEM FOR THE BEGINNER? CRITICAL FACTORS

FOR THE CHOICE

←

/ TI

AU=

→RUNGE, H.

←

CO=

JN=

→FUNKSCHAU (GERMANY) VOL.50, NO.26 1338-9 15 DEC.1978 Coden: FUSHA?

←

PY=

TC=

→Treatment: PRACTICAL APPLIC

←

PY=

DT=

→Document Type: JOURNAL PAPER

←

PY=

LA=

→Languages: GERMAN

←

PY=

DESCRIBES SOME OF THE MAIN CHARACTERISTICS OF HOME

COMPUTER PACKAGES AT PRESENT AVAILABLE.THE READER IS WARNED

AGAINST ONE PARTICULAR OFFERING, WHICH REQUIRES EXPENSIVE

EXTRA EQUIPMENT AND GIVES POOR VALUE IN OTHER RESPECTS

←

/ AB

Descriptors: PERSONAL COMPUTING; MICROCOMPUTERS; COMPUTER

SELECTION AND EVALUATION

←

/ DE

Identifiers: CHOICE; CHARACTERISTICS; HOME COMPUTER PACKAGES;

VALUE, MICROPROCESSOR SYSTEM

←

/ ID

CC=

→Class Codes: C7830

←

PY=

图 2-1 书目型数据库记录实例

DIALOG 系统《科学文摘》(INSPEC) 数据库文献记录实例

DIALOG Accession Number

↓

→ 0045477

CO=, / CO=

→Superior Scale & Instrument Corp

ST = 621 E. Elizabeth Ave.
 CY = Linden, NJ 07036
 ZP =
 TE = TELEPHONE: 201 862 1188
 ORGANIZATION TYPE: Dist, Serv

PRODUCTS:

PN → (only underlined information) DYNAMOMETERS (Dist.: Dillon, Martin Decker)
 INSTRUMENT REPAIR & CALIBRATING SERVICES (Serv.)
 LOAD CELLS (Dist.: Dillon, Martin Decker)
 MACHINERY—PACKAGING (Dist.: NCI, Flexway)
 SCALE RENTALS (Serv.)
 SCALE REPAIR SERVICE (Serv.)
 SCALES—INDUSTRIAL (Dist.: NCI Martin Decker, Dillon, Flexway) ← / DE
 SCALES—POSTAL (Dist.: NCI, Pelouze)
 SCALES—PRECISION (Dist.: NCI, Ohaus)
 SCALES—TRUCK (Dist.: NCI, Martin Decker)
 SHIPPING ROOM EQUIPMENT & SUPPLIES (Dist.: NCI, Pelouze)
 TESTING EQUIPMENT (Dist.: Dillon)
 WEIGHING SYSTEMS (Dist.: NCI, Martin Decker)
 PA =

图 2-2 词典型数据库记录实例

DIALOG 系统《索马斯地区厂商录》数据库记录实例

DIALOG Accession Number

JO =
 → 261931 Qk Froz Fd 85 / 10 / 00 P40 United States
 / DE → Frozen dinners. sales.

YEAR	MIL	\$
1957		55.
1958		88.
1959		100.
1960		128.
1961		151.
1962		173.
	.	
	.	
1982		688.
1983		926.
1984		1230.

YR = →
 GR = → GROWTH RATE = 12.0%
 CC = → CC = 1USA PC = 2038220 EC = 652
 PC =
 EC =

图 2-3 数值型数据库记录实例

DIALOG 系统《PTS 美国时间系列》(PTS U.S Time Series) 数据库记录实例

DIALOG Accession Number

1994770 DATABASE: MI File 47 *Use Format 9 for FULL TEXT*

/TI → Products of the year.

AU = → Gottlieb, Carrie

JN = → Fortune v112 p106(7) Dec 9 1985 → PY =

CO = → CODEN: FORTA : → PD =

SF = → illustration; photograph

AVAILABILITY: FULL TEXT Online

LI = → LINE COUNT 00077

/NA, NA = → NAMED PEOPLE: Goeken, John → innovations

/CO, CO = → COMPANY NAME(S): Coca-Cola Co. → innovations; USA for Africa → marketing; Etak Inc. → technological innovations; Abbott Laboratories → innovations; Electro-Nucl
economics Inc. → innovations; Salomon Brothers Inc. → securities; First Boston Corp → se
curities; Yugo America Inc. → manufactures; Minolta Corp. → technological innovations,
American Cyanamid Co. → technological innovations; International Business Machines
Corp. → innovations; Nike Inc. → technological innovations

/DE → DESCRIPTORS: new products → surveys

COPYRIGHT Time Inc. 1985

/TX → TURNING OUT successful new products is almost as tough as picking a winning combination in the lottery, and a lot more expensive. Mark S. Albion, a professor of marketing at the Harvard Business School, figures that during the past decade the average cost of bringing out a new product has jumped fourfold to \$ 80 million. By next year he expects the figure to reach \$ 100 million. For all the effort, 80% of new products launched in the U.S. fail.

to do it. The company decided to retire its flagship Coke formula on the basis of taste tests that did not discriminate between Coke junkies—the six-pack-a-day types—and occasional sippers. Coca-Cola management not only ignored its best customers, but also failed to anticipate how the public would react to the end of a 99-year symbol of American spirit. For getting people riled up, for sparking general interest in product loyalty, and for providing a rare look at how a major marketer can make a mistake in the quest for a new product and try to correct it, New Coke, and its successor—predecessor Coca-Cola Classic, lead FORTUNE's list of Products of the Year.

图 2-4 全文型数据库记录实例

DIALOG 系统《杂志索引》(MAGAZINE ASAP™) 数据库记录实例

从上面各记录可以看出: 0249373, 0045477, 261931 和 1994770 分别为它们的 DIALOG 存取号。

《科学文摘》数据库记录的基本索引字段有文摘 (/AB)、叙词 (/DE)、自由标引词