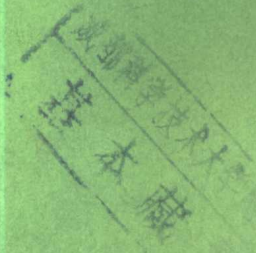


827866

79415

1042

国际联机 情报检索方法



中国科学技术情报研究所
国际联机情报检索服务部

一九八七年十二月

415
42

国际联机情报检索方法

编著者（按姓氏笔划为序）

王太和 尹国英 冯佩娟

杜宝荣 张曼华 徐 玢

葛树兰 谢湘音

中国科技情报所国际联机情报检索服务部

1988年1月

前 言

一九八三年十月中国科技情报所国际联机情报检索服务部正式开业。四年来,在对外开展ESA/IRS, DIALOG和STN系统的联机情报检索服务中,收到了一定的社会和经济效益。目前,我部的国际终端已向外延伸到十多个城市,用户遍布全国。为不断提高机检质量,用户培训已成为当务之急。但由于缺少培训教材,影响了该项工作的顺利开展。为此,我们编著了本教材。本书分两大部分,设十一章。第一部分:概论。概述了国际联机情报检索的概况和检索技术以及三大系统的检索方法。第二部分:常用数据库。介绍了世界专利索引(WPI),美国化学文摘(CA),世界产品市场信息(PTS),美国国家技术情报服务处数据库(NTIS),英国科学文摘(INSPEC),英联邦农业文摘(CAB),美国工程索引(COMPENDEX)等七个数据库的检索方法。

本书是一本通俗易懂,深入浅出的培训教材,而且也是一本可供机检人员和广大科技人员使用的工作手册。

本书第一章由王太和、尹国英编著;第二章和第四章由冯佩娟编著;第三章和第九章由徐玢编著;第五章和第七章由张曼华编著;第六章由谢湘音编著;第八章和第十一章由杜宝荣编著;第十章由葛树兰、张曼华编著。王太和进行了全书的整理编辑工作。

由于我们机检经验不足,业务水平有限,本书定有不妥和错误之处,敬请读者和有关专家批评指正。

中国科技情报所国际联机情报检索服务部

目 录

第一章 国际联机情报检索概况与检索技术.....	(1)
1.1 联机情报检索的国外概况	(1)
1.2 联机情报检索的国内概况	(2)
1.3 国际联机情报检索的一般过程	(2)
1.4 数据库结构	(4)
1.5 检索技术	(8)
第二章 ESA 系统的检索方法	(14)
2.1 ESA系统简况	(14)
2.2 基本检索方法	(14)
第三章 DIALOG 系统的检索方法	(23)
3.1 DIALOG系统简况	(23)
3.2 DIALOG系统各指令的功能及用法	(23)
第四章 STN 系统的检索方法	(36)
4.1 STN系统简介	(36)
4.2 基本检索方法	(36)
第五章 世界专利索引 (WPI)	(45)
5.1 世界专利索引简况	(45)
5.2 世界专利索引联机数据库	(45)
5.3 基本检索方法	(50)
附件 1 禁用词表	(66)
附件 2 各国专利说明书代号及文种	(67)
第六章 美国化学文摘 (CA)	(69)
6.1 美国化学文摘数据库简况	(69)
6.2 ESA系统化学数据库 (CHEMABS)	(69)
6.3 DIALOG系统化学数据库 (CA search)	(82)
附件 CA search的基本索引和辅助索引	(90)

第七章 世界产品市场信息 (PTS)	(92)
7.1 PTS数据库概况	(92)
7.2 PTS数据库的记录格式及内容	(95)
7.3 PTS数据库的主要检索字段和检索方法	(108)
第八章 美国国家技术情报服务处数据库 (NTIS)	(116)
8.1 NTIS数据库简况	(116)
8.2 NTIS数据库的报告号码	(118)
8.3 NTIS数据库的刊物来源	(120)
8.4 NTIS数据库的标引规则	(120)
8.5 检索方法	(121)
第九章 英联邦农业文摘 (CAB)	(127)
9.1 CAB数据库的概况	(127)
9.2 CAB数据库的检索途径	(134)
9.3 CAB数据库联机检索中应注意的几个问题	(137)
9.4 CAB数据库的检索实例	(137)
第十章 英国科学文摘 (INSPEC)	(139)
10.1 INSPEC数据库概况	(139)
10.2 INSPEC词表	(143)
第十一章 美国工程索引 (COMPENDEX)	(149)
11.1 简况	(149)
11.2 工程主题表	(149)
11.3 主题分类	(155)
11.4 检索方法	(156)

第一章 国际联机情报检索概况与检索技术

用户使用终端设备(包括显示终端或兼用微机,调制解调器和打印机),按规定的指令输入检索词或检索参量,通过通讯线路或通讯网络,查找近在眼前或远隔重洋的机读数据库系统中的文献资料,称为联机情报检索。而中国用户采取同样方法查找国外机读数据库系统中的文献资料,称为国际联机情报检索。

1.1 联机情报检索的国外概况

1963年美国国立医学图书馆等机构建立了世界上第一批文献编辑与检索系统,开创了计算机应用于情报检索的新时期。60年代末至80年代初,由于通讯、计算机和数据库技术的飞速发展,相继出现了美国的ORBIT、DIALOG和BRS,欧洲宇航局的ESA/IRS,欧洲共同体的DIANE,英国的INFOLINE,德美日联合开发的STN等著名的联机检索系统。据统计,目前已有100多个联机检索系统。印度,埃及等第三世界国家也已建立了自己的联机检索系统。

联机检索系统的兴起,标志着情报检索已从传统的手工检索向计算机检索过渡。这是情报工作现代化的重要一步。

联机检索的发展历程,虽然只有二十多个年头,但在世界性的信息革命中所起的作用十分巨大。目前尚在继续发展,其趋势是:

1. 文献型的数据库逐步向数值型和事实型数据库过渡。据了解,世界科技文献的年产量约为450万篇,其中期刊论文约200万篇,科技报告约20万篇,会议录约20万篇,其它约100万篇。发达国家的文献型数据库已接近百分之百覆盖科技文献,而且重复建库的现象相当严重。与此同时,对商业行情,产品信息,统计数据,厂商名录和理化性质等数值和事实型情报需求量越来越大。因此,此种类型的数据库发展很快。以1982年为例,在1845个商用数据库中,文献型的有762个,占41.3%,数值和事实型的有1083个,占58.7%。

2. 全文数据库进入实用阶段。以往的全文数据库都是直接把书本式文献资料的书目,文摘和正文输入数据库。而现在的全文数据库正在把图表,以及图表的说明都输入数据库,成为名副其实的全文数据库。例如,STN系统中的西德专利数据库,将采用光栏/向量扫描技术提供图文并茂的西德专利。

3. 情报媒体的小型化和高密度化。1972年荷兰飞利浦公司和美国音乐公司利用激光技术研制成世界上第一张光盘。光盘的结构与磁盘类似,信息收录在盘片表面,读/写头寻找其上的任一信息轨迹。光盘的存贮密度已达到 10^8 bit/mm^2 ,是磁盘存贮密度的50倍左右。换句话说,一张软盘大小的光盘可存贮几万篇、乃至几十万篇文献记录。为此,国际上的许多学者认为,它是70年代以来在信息技术领域里影响极为深远的一项新技术。目前,以光盘为存贮媒体的联机检索系统正在逐步进入实用阶段。例如,美国专利和商标局,已着手把美国

专利存贮在光盘上,并计划到1990年可通过国际通讯网络查找世界各国专利。

1.2 联机情报检索的国内概况

我国的联机检索起步较晚,与国外相比,大约落后10至15年。但在70年代后期和80年代初期出现了可喜的发展势头,在建立中文数据库,利用进口磁带建立西文数据库,以及建立国际终端等方面都取得了很大的成绩。

自1980年建工部建筑技术发展中心等部門在香港租用一台国际终端以来,至1987年底,国内已有50多台国际终端正在提供DIALOG, ESA, BRS和STN等系统的联机情报检索服务。据估计,已检课题达3万多个,为用户提供数百万篇文献记录,在申报专利,搞发明创造,引进国外技术,研制新产品等工作中,收到了良好的效果。联机检索作为一种获取信息的快速手段,也已越来越受到各级领导和广大科技人员的重视,其社会效益是不能低估的。据此,建工部建筑技术发展中心,中情所和水电部情报所的国际终端,曾获得了国家科委1985年度进步奖三等奖。

50多台国际终端的大致分布为:

1. 从地区上可划分为两部分。一是以北京为中心的北方地区服务网。北京的10多个单位的终端及其向外延伸的几十台终端,均通过ITACABLE国际通讯线路与ESA系统联机,并通过罗马分组交换站与美国的DIALOG, ORBIT和BRS, 德美日的STN, 欧洲共同体的DIANE, 英国的INFOLINE等系统联机。二是南方地区服务网。广州, 深圳, 福建, 厦门等情报所的终端, 经香港与美国的DIALOG, ORBIT, BRS等系统联机。

2. 从终端的分配上大致可划分为五个部分。一是中情所及其延伸的沈阳, 重庆, 成都, 武汉, 南京, 呼和浩特, 石家庄, 西安, 天津, 大连, 哈尔滨, 无锡, 安徽, 山西等省市的15台分终端。可称之中情所服务网。二是原兵器工业部情报所及其延伸的西安, 成都, 包头, 济南, 长沙等地的11台分终端。可称之国防口服务网。三是水电部情报所及其延伸的武汉, 郑州, 保定等地的几台分终端。可称之水电口服务网。四是广州, 深圳, 上海, 厦门, 福建等地的终端。虽然各自经营, 但可称之东南地区服务网。五是冶金部情报所、原机械部情报所、清华大学、核工业部情报所、科学院图书馆等独家经营的北京地区终端。

当然, 上面的划分法并不是科学的, 也许有不同意见, 但我们的目的只是让大家了解一下国际终端的分布现状。大家可以看到, 全国的国际终端已遍布近20个省市, 并已向网络化方向发展, 这不仅为国内用户提供查找国外文献资料的快速手段, 而更重要的是为将来的国内检索系统的联网创造条件。

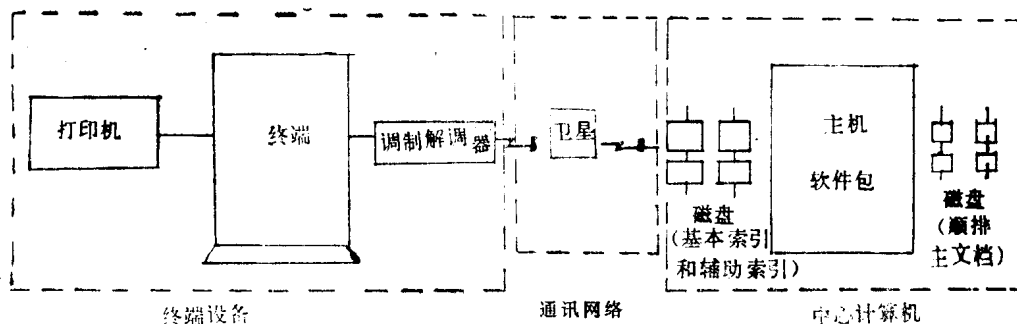
1.3 国际联机情报检索的一般过程

联机检索大体分为准备检索和执行检索两个阶段。准备检索主要是分析检索课题, 选择文档, 确定检索概念, 列出相应的检索词或检索参量, 然后使用布尔逻辑符编制检索提问式。执行检索主要是按照编好的检索提问式, 输入规定的指令, 检索所需的文献记录。本节只简单介绍联机检索系统的构成及联机的一般过程, 具体的检索方法将在下面各节论述。

1.3.1 国际联机情报检索系统的构成

目前, 国内用户所使用的DIALOG, ESA, STN等联机检索系统均有终端设备, 通讯

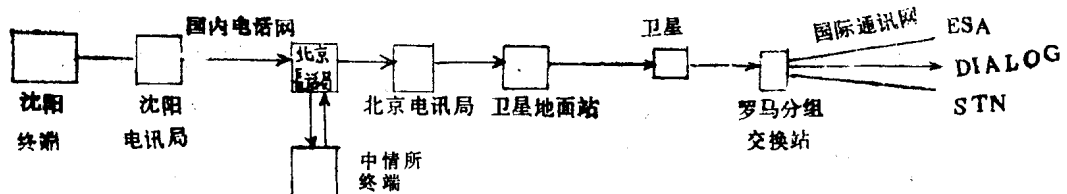
网络和中心计算机三个主要部分。其构成图如下：



简单地说，终端的功能是供机检人员在其键盘上输入检索策略，显示检索结果，并由联动打印机打印。调制解调器的功能是将终端和计算机可识别的数字信号变成可以在通讯线路或通讯网络中传输的模拟信号。计算机则在其软件包的支持下，供用户终端检索内存的各个数据库，并把检索结果反馈到用户终端。

1.3.2 国际联机情报检索信息的传输途径

为使大家易于了解，仅举沈阳终端为例，以图加以简单说明。



沈阳用户若要检索某个联机检索系统（如DIALOG）的某个数据库（如INSPEC）中的某个方面（如计算机）的文献记录，则在接通主机，进入该文档后，接着输入选词指令S和检索词COMPUTER，按动回车键。此时终端用户的提问信息通过上图所示的向右途径传输到主机，经计算机查找，在1至数秒的时间里将检索结果的信息经原途径向左反馈到终端的显示屏上，并由打印机打印。系统的每次响应，均按序给出相应的组号（Set号）和命中的文献记录篇数。终端用户可根据实际需要使用组配指令C或限制指令L等进行逻辑组配或年代、文种等的限制，亦可随机修改检索策略，直到检索结果令人比较满意时为止。这一过程，就是人们通常所说的人机对话过程。

1.3.3 国际联机情报检索的接机过程

首先打开终端设备（如拨号分终端，则包括拨通电话，如专线分终端，则包括上好线），接通北京电讯管理局中的通讯接口。其次输入通讯网络标识和网络密码，接通通讯网络；接着输入主机密码，接通主机；最后进档检索。

例如

```
WELCOME IN BEIJING
31104150002003Dbechi0003ul
COM
network: password:
× × × × × × × ×
enter system id
d
```

（以DIALOG系统为例）

打通讯网络标识

打通讯网络密码

打系统标识

Welcome to DIALOG

Dialog Version 2 , level 12 . 4 . 6 接通主机, 接机过程完结
 B32 进 FILE32, 即金属文摘
 ? S NICKEL 选 词
 1 42119 NICKEL 系 统 响 应
 ? S COBALT 选 词
 2 10913 COBALT 系 统 响 应
 ? S HYDROMETALLURGY 选 词
 3 450 HYDROMETALLURGY 系 统 响 应
 ? C (1+2) * 3 组 配
 4 50 (1+2) * 系 统 响 应
 ? t 4/3/1-10 脱 机 打 印
 ;
 ? PR 4/5/11-50 脱 机 打 印
 P303: PRINT 4/5/11-50 系统响应, 确认打印要求
 ? logoff 关 机

1.4 数据库结构

联机检索系统的软件包, 把每篇文献记录都准确地组织起来。首先, 给每篇文献记录一个唯一的存取号, 按年代或卷期的先后次序排列成目录文档。该文档是供联机和脱机打印输出的顺排文档。其次, 给每条文献记录的每个可检字段的每个可检词, 按它们的位置排序, 经抽词后排列成一个基本索引字典和若干个辅助索引字典。这些字典是供联机检索的倒排文档。

1.4.1 文献记录结构

每篇文献记录一般由篇名TI(TITLE), 文摘AB(ABSTRACTS), 规范词CT(ESA; CONTROLLED TERMS)或DE(DIALOG; DESCRIPTORS), 自由标引词UT(ESA; UNCONTROLLED TERMS)或ID(DIALOG; IDENTIFIERS)等四个主要字段, 以及若干个辅助字段(例如作者AU(AUTHOR), 文献出处SO(SOURCE)等等)所组成。每篇文献记录均有一个存取号AN。

(AN) 77537463
 (TI) DETECTION and CONTROL of AIR POLLUTION and WATER
 TI1 TI2 TI3 TI4 TI5 TI6 TI7 TI8
 POLLUTION
 TI9
 (AU=) HOLLANDER, W. E.
 (AB) ... DETECTION of AIR POLLUTION POLLUTION
 AB3 AB4 AB5 AB6 AB18
 ... AIR POLLUTION WATER POLLUTION
 AB25 AB26 AB33 AB34
 (DE) DESCRIPTORS: AIR POLLUTION DETECTION AND CONTROL / POROUS
 DE1 DE2 DE3 DE4 DE5 DE6
 METERIALS
 DE7
 (ID) IDENTIFIERS: DETECTION / AIR POLLUTANTS / WATER POLLUTANTS
 ID1 ID2 ID3 ID4 ID5

计算机自动地给TI, AB, DE, ID这四个主要字段中的每一个单元词(包括小写的禁用词)排序, 给出位置标识, 以此确定词间的位置关系, 从而才有检索时常用的位置算符(W)(nW)(S)(F)(C)等。现举例说明如下:

(W)WORDS邻词位置符。AIR(W)POLLUTION表示两词相邻, 词序不能颠倒。TI字段的TI5与TI6; AB字段的AB25与AB26; DE字段的DE1与DE2符合条件。

(S)SENTENCE句子位置符。TI字段, 或者AB字段中用句号隔开的一段完整的话, 或者DE与ID字段中用斜线隔开的每个标引词组, 均称为一个句子。如AIR(S)POLLUTION表示两词在同一个句子中。TI字段的TI5与TI6, TI5与TI9; AB字段的AB25与AB26; DE字段的DE1与DE2均符合条件。

(F)FIELD字段位置符。字段即为一个完整的著录项。如AIR(F)POLLUTION表示两词在同一个字段中。TI字段的TI5与TI6, TI5与TI9; AB字段的AB5与AB8, AB5与AB26, AB26与AB25, AB34与AB5, AB5与AB34, AB25与AB34等均符合条件。

(C) CITATION引文位置符。一篇文献记录即为一篇引文。如AIR(C)POLLUTION表示两词在同一篇引文的四个主要字段的任一字段中。TI5与TI6, TI5与TI9, TI5与AB8, AB25与DE2等等均符合条件。

从上面可以看出, 从(W)→(S)→(F)→(C)词间的位置关系越来越松, 因此在检索时选择恰当的位置符来搭配检索词, 对提高查准率或查全率十分重要。

1.4.2 基本索引

计算机自动地从TI, AB, DE, ID这四个主要字段中抽词, 按字顺排列成一个庞大的机内字典, 每个词均有位置标识, 并给出相应的存取号。

基 本 索 引

:		
AIR	TI5 AB5 AB25 DE1 ID2	77537463
:		
AIR POLLUTANTS	ID2 ID3	77537463
:		
AIR POLLUTION DETECTION AND CONTROL	DE1 DE2 DE3 DE4 DE5	77537463
:		
DETECTION	TI1 AB3 DE3 ID1	77537463
:		
POLLUTANTS	AB6 ID3 ID5	77537463
:		
POLLUTION	TI6 TI8 AB18 AB26 AB34 DE2	77537463
:		
WATER	TI8 AB33 ID4	77537463
:		
WATER POLLUTANTS	ID4 ID5	77537463
:		
ZONE		
:		
2000		
:		

应当指出:

①这是从一篇文献记录的TI, AB, DE, ID这四个主要字段中抽取的若干个可检单元词或词组, 在基本索引字典中的排列情况。如果一个文档有300万篇文献记录, 则要从每篇文献记录的上述四个主要字段中抽取所有的可检单元词或词组, 构成可供检索的基本索引字典。该字典的规模之大是可想而知的。

②因为基本索引字典中的单元词或词组, 均摘自TI, AB, DE, ID这四个主要字段, 故称这四个字段为基本索引字段。在检索时使用对应的后缀码/ TI, / AB, / DE, / ID对检索词进行后缀限制, 以提高查准率。如S AIR(W)POLLUTION/ TI, DE, 选择的词组被限制在TI, DE字段中。

③从TI, AB字段中抽取的为可检单元词。OF, WITH, ON等禁用词不在抽取之列, 但留出相应的位置, 给出位置标识, 例如上述文献记录结构中的TI2, TI4, TI7, AB4等只表示禁用词AND与OF的位置, 而不抽入基本索引字典。但是, 从DE, ID这两个标引词字段中抽取的不仅是除禁用词外的所有可检单元词, 而且还有每个标引词词组, 在这些词组内可包含禁用词。检索时可参考词表的用词, 在不加任何位置符的情况下进行直接选词。例如S AIR POLLUTION DETECTION AND CONTROL, 表示在DE或ID字段中选词。当禁用词为布尔算符AND等时, 系统规定打上引号后方可输入。再如S AIR(W)POLLUTION, 在基本索引的各个字段中选词; S AIR POLLUTION, 则只在基本索引的DE, ID字段中选词。很明显, S AIR(W)POLLUTION/ DE, ID类似S AIR POLLUTION。

1.4.3 辅助索引

为了增加文档的检索点, 文献记录不仅有基本索引字段, 还有辅助索引字段, 如作者(AU), 文献出处(SO), 分类代码(CC), 国际专利分类号(IC), 化学物质名称登记号(RN)等等。文档不同, 辅助索引的多少及用法也不同, 使用时应参考文档的简介。

计算机从文档的每篇文献记录的各个辅助索引字段中, 分别抽取可检词或可检参量, 构成一个一个倒排的辅助索引, 供辅助检索之用。现以作者(AU)辅助索引为例作一说明。

辅 助 索 引 (AU)

:	:
ARON, E. F.	87456543
:	85441721
:	:
HOLLANDER, W. E.	82113715
:	77537463
:	76556726
:	74445216
:	:
SMITH, E. N.	86432561
SMITH, F. R.	84314561
:	:
WERNER, E. G.	85443270
:	:
ZANER, G. H.	80435712
:	:
ZIMMERMANN, E. G.	87524321
:	:

作者姓名按字顺排列，一般姓在前，缩写的首字母名在后，每位作者的后边，均有对应的存取号。例如，若一位作者被收录的文献记录有10篇，则有对应的10个存取号。检索某作者的文献记录时，应使用前缀码AU= 。例如，S AU=HOLLANDER, W. E. 查找该作者的文献记录。

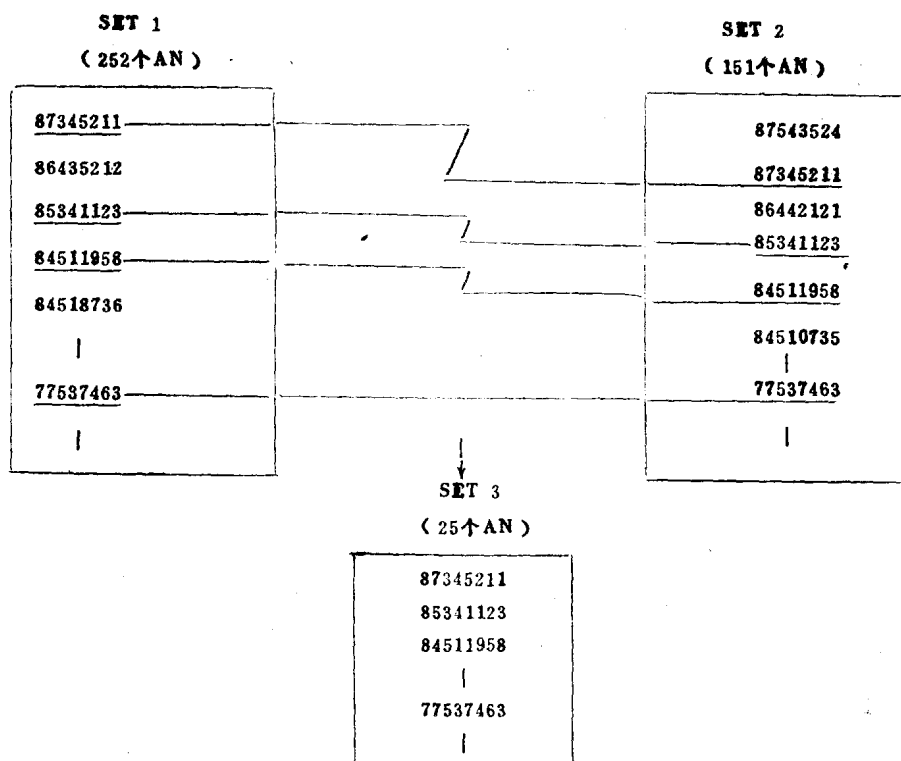
与此相似，应使用前缀码查找其它辅助索引，例如，S RN=9001—83—3 查找化学物质名称登记号为9001—83—3的那种物质的文献记录。又如，S IC=C06B—021/00 查找国际专利分类号为C06B—21/00属下的文献记录。

总之，在检索过程中，若直接选择检索词或检索参量，则扫描的是文档的基本索引字典。若使用前缀代码选择检索词或检索参量，则扫描的是该前缀码代表的辅助索引字典。无论是基本索引字典，还是辅助索引字典均录在磁盘上，并能进行修改和定期更新。

1.4.4 检索和输出

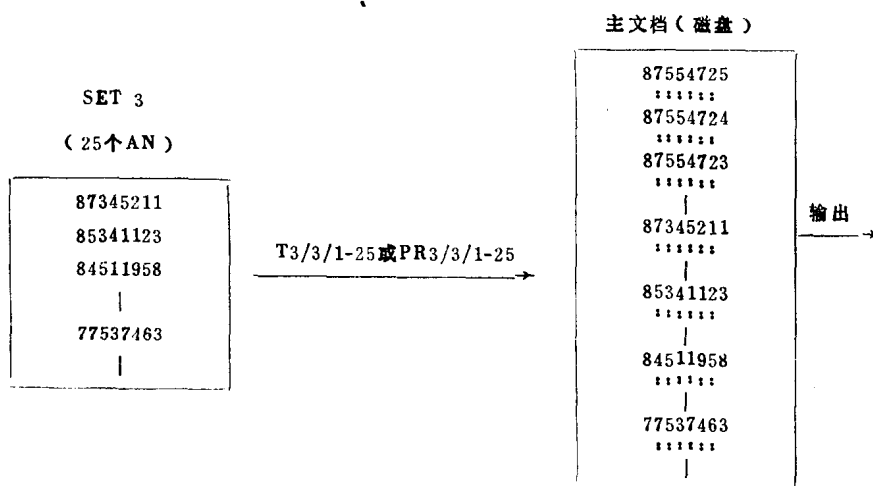
1. 检索。就是使用选词指令S，组配指令C，限制指令L等检索指令进行选词、组配、后缀限制等。

例如 S AIR(W)POLLUTION
1 252 AIR(W)POLLUTION
S SENSITIVE(S)DETECTION
2 151 SENSITIVE(W)DETECTION
C 1 • 2
3 25 1 • 2



每次输入检索词或检索参量后, 系统立即响应, 给出组号(SET号)和命中的文献记录篇数, 而存贮在工作区内的是与之对应的存取号。组配, 实际上是对比有关SET号中的存取号, 按 AND, OR, NOT 的布尔逻辑关系, 把对比后挑出的存取号存贮在另一个工作区内, 构成另一个SET号。例如, 上一检索实例有 3 个SET号, SET 1 有 252 个AN, SET2 有 151 个AN, 经逻辑与(AND)组配后, 把SET1与SET2中相同的AN挑出来, 构成SET3, 有25 个 AN。当然在检索过程中, 选词、组配、限制等要复杂得多, 但基本原理相同。

2. 输出。简单地说, 就是使用联机打印指令T或脱机打印指令PR, 把命中的文献记录按规定的打印格式打印。例如: T3/3/1-25和或PR3/5/1-25, 即分别采取第 3 种格式(目录)和第 5 种格式(全记录), 联机打印和脱机打印SET3中的第 1 至第25篇文献记录。不管联机打印, 还是脱机打印, 都是从收录全部文献记录的主文档(顺排文档)中, 按命中文献记录的存取号(AN)进行打印。



总之, 检索和输出是两码事, 检索的对象是文档的基本索引字典或辅助索引字典, 是倒排档。而输出, 即联机或脱机打印的对象是主文档, 是收录文献记录的顺排档。这是十分重要的两个不同的概念。

1.5 检索技术

检索技术, 是检索情报时使用的策略、方法、程序、经验及其有关技术的总称。

1.5.1 用户情报需求分析和情报提问

用户情报需求分析和情报提问与检索的目的直接相关。不同的用户有不同的情报需求,不同的检索课题所涉及的学科和专业也各不相同。即使是同一用户,在不同时期也有不同的情报需求。各种类型的检索课题对检出文献的需求也各不相同。它们有广度上的需求,也有深度上的需求;有现实的需求,也有潜在的需求。从所需文献的类型来看,对会议论文、专利、研究报告及期刊文献等的需要也各有所侧重。根据各类用户的各种情报需求以及在情报检索过程中的一些特点,我们可将用户的情报需求分为下列几种类型。

1. 普查型。用户需要系统地收集某一主题范围的文献资料,以编写教材、出版专题论文选编、从事基础理论或应用理论研究。他们要求检索的查全率高。这样的情报需求带有横向普查和纵向追溯的特点。

2. 探索型。用户需要了解和掌握国外的最新动态或研究成果,以便选择新的研究课题和应用新的技术。他们对查全率和查准率要求不一定很高,但要求提供的情报是新的及时的。

3. 攻关型。用户需要检索结果能解决科研或生产中的关键问题,或是满足申请专利、发明等的需要。他们要求有较高的查准率,而不一定需要检出大量的文献。

要想检索成功,就必须把情报需求与情报提问有机地结合起来。因此,检索员不仅必须正确理解用户的需求,而且要用系统能识别的检索语言把用户的需求准确无误地表达出来。换句话说,也就是将用户的情报需求转化成准确、完整的情报提问。这是取得检索成功的关键一环。如果提问比需求更专指,就必然在查全率上有所漏失;若是提问比需求更泛指,那末在查准率上就有所失误;假如提问偏离了需求,就必然造成查全率和查准率的失误都很大;假若提问与需求完全无关,必然导致检索失败。只有需求与提问一致或接近时,才能取得较好的检索结果。

由于联机检索的绝大多数用户对联机检索的功能和检索方法并不熟悉,因此不能正确地表达检索意图的情况是经常出现的。特别是一些课题的实质性内容往往很难从课题的名称上看出来,还有一些隐含的概念和相关的问题也不是一下就能明确的,这就必须进行主题分析。只有这样才能保证提问接近用户的需求。

例如,《枸杞同源四倍体新物种新颖性调查》这样的检索课题,如果完全按照课题上的词,一个不漏地选作检索词:枸杞、同源四倍体、新物种、新颖性进行检索,或按用户所提的检索词:宁夏枸杞、同源四倍体、未受精子房培养、培养基、低温予处理进行检索,就可能造成一些不必要的失误。而换用:枸杞、宁夏枸杞、中国枸杞、同源四倍体和四倍体这样几个检索词进行检索,检索结果甚佳。经分析证明该项科研成果属世界首创。具有国际水平,报刊上登载了这一消息。

1.5.2 数据库的选择与比较

选择数据库是检索前必不可少的一个重要环节。数据库选择不当,犹如买东西走错了商店,纵使你对课题分析得完全正确,关键词选得很恰当,检索策略编制得特别好,也难于检出用户所需要的文献。

选择数据库的主要依据是检索员将他们所掌握的有关数据库的知识同他们所理解的用户情报需求和情报提问进行有机的匹配。要搞好匹配,主要应考虑以下几个问题。

1. 数据库的可检索性,即数据库的收录范围及特点。这是判别数据库检索价值的重要因素。要搞好匹配首先要了解数据库收录文献的专业范围、来源、类型、数量、时间范围

以及与其他数据库相比较的特点和交叉情况。

2. 数据库的标引与词表因素。包括数据库所用词表的控制程度、词表中词的专指度、标引的网罗度、确切性与一般性,以及所提供的检索项目和辅助的检索手段等。

3. 数据库的大小与增长率及其更新周期。在各种检索系统中,各个数据库的规模相差十分悬殊,因此要熟悉各数据库的专业特长、入藏总量、起止年代、更新周期和更新量。以化学文摘为例,目前收录的文献记录已达600万篇以上。在ESA系统中,它是一个文档,即2号文档。而在DIALOG系统中,既有以年代划分的308、309、320、310、311五个文档。又有一个统一的399文档。它收录有150多个国家的14000种期刊,26个国家公布的专利说明书以及世界各国出版的新书、会议记录、学位论文和技术报告,其收录的时间范围是从1967(66卷)起至现在,每两周更新一次,每次更新16,000篇。

4. 数据库的检索费用。在各系统中,每个数据库的检索费和每条记录的价格均不一样,检索时要注意比较。以便选用最合适、最便宜的数据库,这样既可达到检索的目的,又可节省检索费用。

检索一个课题,往往需要在一个以上、有时甚至需要在5—6个数据库中查寻。遇到这种情况,应根据课题与数据库的相关性,选择最有希望得到较好检索效果的数据库。现以DIALOG系统为例,检索《兰皮书的定义、形成和作用》的课题,宜在8个有关数据库中查找,它们是:

- ①美国政府研究报告文摘通报;
- ②社会科学引文索引;
- ③社会学文摘;
- ④图书馆与情报科学文摘;
- ⑤美国政府出版物索引;
- ⑥美国政府科学文摘;
- ⑦美国国会出版物索引;
- ⑧书评索引。

进行多文档检索时,为节省机时,宜利用DIALOG系统的总索引文档(411文档),就可一次性显示各文档中有没有或有多少所需的文献记录。

1.5.3 检索词的选择与调节

在国际联机检索中,常用的检索词分为两类:一类是主题词,另一类是自由词。主题词来源于各数据库使用的词表,如果没有词表,可参照相应书本式检索工具的用词规律;自由词多来源于用户根据自己课题需要提出的词。但一般应优先选择主题词作为最基本的检索词。由于词表本身的动态变化及其结构的不完善等因素影响着选词的准确性,为了检索的专指性经常需要选用自由词配合检索。此外,还应考虑选用与主题词和自由词相应的同义词、广义词、狭义词、相关词,以及有关的化学物质登记号和作者、机构等,甚至还要考虑到这些词的英、美不同拼法等等。这样才能既保证查全率,又满足查准率的要求。

对课题的概念进行全面地、准确地分析是选择检索词的前提。根据检索语言表达情报提问所涉及的概念的能力,即词表的网罗度与专指度。在把概念转化成检索词时,要求从词表中选择最适合的词。选词时既要考虑概念的内涵深浅,又要考虑概念外延的宽窄,不仅要从句面上择词,更要从词的含义上进行择义。其一般的选词原则是:

1. 要从词表规定的专业范围出发, 选用各学科内具有检索意义的基本名词术语;
2. 避免选用频率低的词;
3. 多选用基本词汇进行组配;
4. 一般不选用动词和形容词。

选词失误的原因, 一般有以下几种:

1. 所选检索词与词表不符, 这是最常见的失误。出现这种情况的原因主要是用户没有词表或不会利用词表, 或是检查员不熟悉词表。

2. 所选检索词专指性太高, 检不出有关文献。要避免这种现象, 就要将这类检索词扩展到上位词或增加相关词。

3. 所选检索词太泛指, 不能准确表达提问的实质, 因而需要缩小检索范围, 增加专指性的词。对于一些词频很高的词, 如: “分析”、“研究”、“应用”、“方法”、“发展”、“设计”等词, 要尽量少用或不用。必须用时, 也应与其它能表达主要检索特征的词一起组配, 或增加一些限制条件再使用。

4. 选词不全面、不准确, 这是造成失误的一种普遍现象。例如: 《固定床反应器》课题, 靠查普通字典编出来的检索词然后列出检索式, 进入ESA系统的file2中检索。

S fid(w)bcd(s)reactor? ?

结果等于零。而换用:

S fided(s)bed(s)reactor? ?

在同一系统中, 同一文档中就取得了有关文献145篇。

总之, 应由用户和检索人员共同选定检索词, 为了更有把握, 除确定一些基本的检索词之外, 还需要选择一些备用的词。一旦用户对检索结果不满意, 可随机进行分析比较, 并立即进行调节, 决定增词、减词或换词, 以便取得最佳的检索效果。

1.5.4 检索策略的编制

所谓检索策略, 就是在分析情报提问的基础上, 即根据检索课题的要求, 确定检索途径和检索用词并明确各词之间的逻辑关系。在选定检索词之后, 就可以用布尔逻辑算符和检索指令将情报提问中各有关概念之间的关系用检索式表达出来, 并可用括号来体现各检索词之间的完整性和运算的优先次序。这种使用布尔算符组配的检索式, 即称为检索策略。

无论手工检索还是计算机检索都存在如何编写和编好检索策略的问题, 但它们又各有不同。手工检索的检索策略, 只在检索者的脑子里, 不必写成书面的语句, 而且是边查、边看、边思考, 可以灵活地改变这种策略。而用计算机检索, 提问与文献标识之间的对比匹配工作是用机器进行的。因此必须事先拟订周密的检索策略。

从大量检索实践中所获得的经验教训表明, 检索策略编制的好坏, 直接影响着检索效果。编制检索策略应着重探讨以下几个问题。

1. 在编制检索策略过程中, 首先必须弄清用户的提问要求, 包括检索课题的内容、查找文献的类型、文种和时间的范围等。

2. 根据上述要求, 确定选用哪个检索系统。是用ESA系统, 还是用DIALOG系统, 或是用STN系统。然后确定选用哪个(或哪几个)文档。

3. 对用户的提问进行概念分析, 并把各组面概念转换成词表中的词或自由词。应参照有关词表选出能代表各个概念的检索词或利用用户的自由词。有时用户的检索课题要求很专

指,需要用非规范化的词或用若干自由词组合的多元词组进行查找。

4. 运用布尔逻辑运算符和位置逻辑算符等计算机可识别的语言和符号,将用户的提问要求正确地表达出来。从表面上看,任何情报需求,不论复杂到什么程度,经过主题分析,变成几个明确的概念之后,便可用布尔逻辑组配的原理和方法将其转化成布尔表达式,也即检索策略。在编制检索策略时,为了提高检索的准度和广度,往往需要用位置逻辑符(W)或(nW)或(N)或(S)或(F)等适当地搭配检索词。实践证明,位置逻辑符用得是否合适,对检索质量的影响极大。

此外,在编制检索策略时,应注意某些技巧问题。例如:用布尔“逻辑与”(AND)组配检索词时,应将词频低的词放在AND的左边,词频高的词放在AND的右边;同样,用布尔“逻辑或”(OR)搭配检索词时,应将词频低的词放在OR的左边,词频高的词放在OR的右边,等等。

1.5.5 评价检索效果的方法与提高检索效率的途径

检索效率,不仅是检索人员经常考虑的问题,而且是广大用户十分关心的问题。评那末检索服务效果的准则是什么呢?影响情报检索效果的因素很多,概括地说,主要就是查全率和查准率。

所谓查全率是衡量符合提问要求的文献中实际检出文献的数量,而查准率是衡量检出的文献中有多少符合情报提问的文献数量,即:

$$\text{查全率 } R = \frac{\text{检出的相关文献量}}{\text{库中存贮的相关文献量}}$$

$$\text{查准率 } P = \frac{\text{检出的相关文献量}}{\text{从库中检出的文献量}}$$

查全率R与查准率P结合起来,可以表示检索系统的筛选能力。

一个理想的检索过程就是能最大限度地满足用户的情报需求。一个好的检索策略应该能为用户提供全面而准确的相关文献。然而,由于查全率与查准率之间倾向于互逆关系,所以要提高查准率而限制检索范围时,查全率就会相应降低,反之,要提高查全率而放宽检索范围时,查准率也会随之降低。

总之,当检索泛指时,可达到较高的查全率,而当检索专指时,可达到较高的查准率。

世界上任何一个检索系统都不可能有一百之百的查全率和一百之百的查准率。一般的检索系统的平均查全率为60~70%,查准率为40~50%。

在实际检索中,查全率是用户无法估计的,因为他们不知道计算机中究竟存贮有多少相关文献,因此我们通常所说的查全率和查准率也只能是个估算。尽管在机检实践中评价查全率与查准率存在着一定的困难,但仍可以根据用户的反馈信息或通过用户对比手检结果或其他途径进行估计。

提高查全率。简单地说,要从扩检入手,常用的方法有如下几种:

1. 降低检索词的专指度。从有关词表或根据检出文献中的关键词选定专指度低的上位词或相关词、同义词和近义词,以提高所检文献的网罗度。例如,查找“飞机场的软土地基的处理”。“飞机场”一词用了二个同义词Airpoft, Airfield;“软土”一词用了十个同义词,如Soil, Clay, Stugde等;“地基”一词用了三个同义词Foundation, Base, Su