

《中国学术期刊(光盘版)》的检索效果与利用率研究

王永久

(东北大学图书发行中心 沈阳 110006)

摘要 通过对《中国学术期刊(光盘版)》的利用,根据“《中国学术期刊(光盘版)》全文检索管理系统”(CAJR)的功能和它对《中国学术期刊(光盘版)》(CAJ-CD)的要求,对《中国学术期刊(光盘版)》的检索效果和利用率进行了分析,针对CAJR和CAJ-CD在使用过程中存在的问题,提出了一些对策和建议。

关键词 CAJ-CD CAJ 功能 检索效果 利用率

《中国学术期刊(光盘版)》(以下简称CAJ-CD)是我国连续出版的大规模集成化、多功能全文数据库系统,它收录了国内近3000种学术期刊,现分9个专辑按月定期出版发行的综合性电子杂志,它必需通过支撑软件“《中国学术期刊(光盘版)》全文检索管理系统”(以下简称CAJR)及相应的硬件设备才能实现相应的各种功能。CAJ-CD和CAJR是目前我国收录范围最广,检索功能比较先进的全文数据库检索系统。笔者根据实际运用过程中的具体情况,对这一系统的检索效果和利用情况从以下几个方面进行分析。

1 CAJR的检索功能和对CAJ-CD的要求

CAJR根据CAJ-CD的特点和在实际应用的要求,具有比较科学和全面的检索功能,可从整刊、专项、全文任意词等多方面进行检索,其中的专项检索具有分类、关键词、作者、篇名等8个检索项目。整刊检索具有期刊查找、整刊浏览等功能。这种多检索渠道的系统功能无疑提高了检索查全率,而且检索速度较快。实践证明,对一张光盘采用一种途径从开始检索到屏幕显示检索结果最多也不到一分钟。此外,CAJR还具有期刊管理、光盘计到、定题服务、项目背景分析、科研能力评价、双向字典等其它功能。CAJR的功能之多速度之快和操作比较简单方便的特点,不仅使读者感到惊讶和赞叹,也给检索站工作效率的提高提供了可靠的系统保障。CAJR的这些强大功能要想得到有效的发挥,必然对CAJ-CD具有较严格的要求,因为数据的规范性和格式的统一性及关键词语的准确性对检索效果都具有较大的影响。

2 检索效果的评价指标及对CAJR检索效果分析

检索效果(retrieval effectiveness)包括技术效果和经济效果。技术效果主要是指系统的性能和服务质量;经济效果由完成这些功能的价值确定,主要指检索系统服务的成本和时间。(英)C.W.克萊弗頓(C.W. Cranfield)在分析用户基本要求的基社上,提出了收录范围、查全率、查准率、响应时间、用户负担和输出形式6项评价系统性能的指标。其中查全率和查准率是两个最主要也是最常用的指标。计算方法如下。

$$\text{查全率} = \frac{\text{检出的相关文献量}}{\text{系统中相关文献总量}} \times 100\%$$

$$\text{查准率} = \frac{\text{检出的相关文献量}}{\text{检出的文献总量}} \times 100\%$$

查全率和查准率因检索途径的不同而不尽相同,可根据用户对检索目标要求不同,选择相应的途径进行检索。这就对检索者在知识面及信息检索方面提出了更高的要求。在此,笔者以“绿色贸易壁垒对中国对外贸易的影响”为检索题目,以不同的检索词和常用的方式进行检索,其结果如下:

检索词:“贸易壁垒”CAJ-CD为政治经济法律 9901、9902、9903

检索方式	检出数目	相关数目	非相关数目	查准率 %	查全率 %	用时 (分)
分类	48,48,38	8,9,6	40,39,32	17.16	88.5	3
关键词	0,1,0	0,1,0	0,0,0	100	3.6	3
篇名	0,0,0	0,0,0	0,0,0	0.0,0		3
摘要	2,0,0	1,0,0	1,0,0	50	3.6	3
全文	33,37,30	8,11,9	25,26,21	28	100	5

检索词:“壁垒”CAJ-CD为政治经济法律 9901、9902、9903

检索方式	检出数目	相关数目	非相关数目	查准率 %	查全率 %	用时 (分)
分类	48,48,38	8,9,6	40,39,32	17.6	88.5	3
关键词	1,2,1	1,2,1	0,0,0	100	14.3	3
篇名	2,2,1	2,2,1	0,0,0	100	17.9	3
摘要	2,2,1	2,1,1	0,1,0	80	14.3	3
全文	93,129,119	8,11,9	85,118,110	8.2	100	5

检索词:“壁垒+影响”CAJ-CD为政治经济法律 9901、9902、9903

检索方式	检出数目	相关数目	非相关数目	查准率 %	查全率 %	用时 (分)
分类	48,48,38	8,9,6	40,39,32	17.6	88.5	3
关键词	0,0,0	0,0,0	0,0,0	0	0	3
篇名	0,0,0	0,0,0	0,0,0	0	0	3
摘要	0,0,1	0,0,1	0,0,0	100	3.8	3
全文	75,110,96	8,11,9	67,99,89	10	100	5

表注:“用时”是指从开始检索到显示检索结果及更换光盘所用的时间,不包括在结果中人工检索用时

以上检索结果表明,查准率和查全率之间存在互逆关系,而且,检索词和检索途径的选择对检索结果的影响也很大。

入编CAJ-CD期刊的质量、范围和CAJR软件系统的科学性是实现其实用价值的基础,同时也是充分发挥其功能的保障。这方面的评价指标除了上面的查全率和查准率之外,还包括收录范围、响应时间和输出形式以及用户负担。现在我们就着重对系统的检索费用和输出形式两个方面进行分析。

3 系统的检索费用和输出形式

3.1 获取CAJ-CD信息的费用 获取CAJ-CD(下转第69页)

算符表示允许在相连两词间插入最多 n 个单词,但词序不能互换,如 design(1w)robot 相当于检索 design of robot 或 design robot 等。(N)算符表示该算符两侧的词必须紧密相连,但可颠倒次序。(nN)算符表示两词之间可插入 n 个单词,且两词次序可以颠倒。(L)为 Link 的缩写,表示两词间有一定的从属关系,且只限定在受控主题词索引字段查找,如 DE、CT 等字段(此算符很少用)。(S)为 Subfield 的缩写,表示两词必须同时出现在同一个子字段中,即同一句子或短语中,词序和中间插入词数不限。(F)是 Field 的缩写,表示两词必须同时出现在同一字段中,词序和词间插入词数不限,但需指明要查找的字段,如 design(F)DRAM /AB,表示凡在文摘中同时有 design 和 DRAM 的记录,均为命中文献。(P)算符限制较宽,所以其运算结果不如用(S)精确。(C)为 Citation 的缩写,表示符号前后二词同时出现在同一记录中,词序和出现的字段都不限,作用与布尔算符中 AND 相同。总的来说,在同一数据库中,对于相同的检索词 a 、 b ,如果使用不同的位置算符,检出记录有如下关系: $G_a(W)b \subseteq G_a(N)b \subseteq G_a(S)b \subseteq G_a(F)b \subseteq G_a(C)b$ 。式中 $G_a(W)$ 、 $G_a(N)$ 、 $G_a(S)$ 、 $G_a(F)$ 、 $G_a(C)$ 分别表示由检索词 a 和 b 分别能与不同位置组合在同一数据库中所命中的记录的集合。掌握了不同位置符之间的关系之后,适当选用位置算符非常重要,通常要根据以下几点来确定检索词的位置关系: a. 根据文献中常见的位置关系; b. 根据实际检索结果,重新选择位置算符; c. 根据不同检索目的,对于侧重点为查全的用户,在查准的基础上适当使用相对宽松的位置符,对于侧重点为查准的用户,应适当使用限制较严的位置符。

2.3 适当使用检索字段符提高查准率

位置算符只能限定词间的相对位置,但不能确定检索词在记录中的位置。尤其是在采用自由词进行全文查找时,需要用字段

(上接第 67 页)信息费用包括检索费用和检索结果输出费用。检索费用由于受多方面因素影响,集中体现在检索用时上,其中系统收录范围、光盘制作质量、软件功能和检索技巧等几个方面对检索用时有较大影响。检索结果输出费用因输出方式的不同而不同,网络传输和拷贝输出最为经济方便而且省时,打印输出效率低而且费用较高。通过实践表明,光盘质量和检索技巧对检索用时的影响最为突出。此外,CAJ-CD 的读取费用相对较高。虽然 CAJ-CD 具有信息容量大而投资相对较小的优势,但其也存在一些不可避免的弱点,首先它必须借助计算机和系统软件及相应的技术才能进行阅读,其阅读费用相对较高,而且具有一定的难度。尤其是 CAJ-CD 内容的属性决定了读者不可能采用走马观花式的阅读,占机时间的增加,必然使读者阅读此类资料的费用较高,这给其普及应用和利用率提高带来了一些负面的影响,如果费用由读者自己负担,一般来说就不利用 CAJ-CD 了。

3.2 CAJ-CD 的输出方式 由于 CAJ-CD 不提供全文信息下载(存盘)功能,在网络日益普及的今天,用户获得的信息不能通过 E-mail 发送,也不能进行全文存盘处理,在一定程度上限制了期刊的充分利用。

4 CAJ-CD 在使用过程中出现的问题

入编期刊的编排质量主要取决于编排格式的统一和论文中

符限定检索的字段范围。DIALOG 系统中常用的字段标识符有: AB 文摘,AN 存取号,AU 作者,CS 分类号,CO 期刊代号,CS 出版机构,SO 文献出处,TI 篇名,TN 期刊名称等。使用时,描述文献外部特征的字段符常作为检索提问表达式的前缀(如 AU=)出现,而描述文献主题内容的字段符限定检索范围,可节省机时,提高效率及查准率。例如,在实际工作中,经常会遇到一些用户需要查找自己的文献被 Ei 或 Sci 收录情况,在查找这类信息时,经常使用“作者”与“出版机构”即“AU”与“CS”的方法进行检索,既快又准。另外 Ei 和 Sci 中对作者名的收录格式不同,例如作者名为: Li Dongcai, Ei 中的收录格式为: Li, Dong - cai,而 Sci 中的收录格式为: LiDC。因此,在检索时要首先了解其收录格式,否则会产生误检。另外为了提高查准率,在检索时通常对关键的检索词进行加权处理,如:在检索液体洗涤剂助剂文献时,如果采用 s detergent()builder Ai and liquid 的检索策略,就意味着重点在于 detergent builder 而 liquid 是次要的,这样检出的文献应与 detergent builder 密切相关;如果采用 liquid Ai and detergent()builder 的检索策略,则意味着 liquid 是重点,detergent builer 是次要的,这样检出的文献应与 liquid 密切相关。由此看来,词和重点的选择是查找成功的关键,而其选择又带有高度的主观性并需要相当大的技巧。

总之,检索文献时不能单一地运用一种检索途径,而要将位置算符、检索字段符、布尔逻辑算符、前缀和后缀等结合起来灵活使用,避免漏检和误检,以达到查准、查全的目的。

参考文献

- 1 DIALOG ONDISC. Knight - Ridder Information, Inc. 1994
- 2 丁自汉等. 科技情报检索. 西安:西安交通大学出版社,1993

(责编:王京韵)

关键词语的规范化,升级后的 2.5 版本在打印输出格式的统一方面有了很大的提高。但由于入编期刊较多,作者的水平也参差不齐,著录项目上要全面、准确,关键词语要规范。在这些方面上的严格要求,能够使检索效率有更快的提高。

5 对策与建议

要加强宣传,对读者进行培训,并使他们掌握这一系统的检索技术。增加网络检索终端,降低 CAJ-CD 的价格,提高质量,改进系统性能;提高检索技巧和检索效率,缩短检索用时,这是提高 CAJ-CD 利用率的有效途径之一。

对一个目标的检索来说,应根据用户的需求而采取相应的方式,以提高检索效率。资料表明,目前国外一些信息检索系统能达到的查全率在 70% 左右,查准率在 50% 左右,如果检索词恰当,检索方式正确,CAJR 的检索效果超过这个水平是有可能的。

参考文献

- 1 王 岩,吉智文.《中国学术期刊(光盘版)》的使用情况分析. 情报学报, 1998;(4)
- 2 韩改样. 学术期刊出版规范化与其光盘的检索效率. 情报学报,1998;(4)
- 3 孙 平,任其荣. 科技信息检索. 北京:清华大学出版社
- 4 上海科学技术情报研究所编. 科技情报检索手册. 上海科学技术文献出版社

(责编:王京韵)