再论计算机检索技巧

——兼与万碧君等同志商榷

赵 勇

(中科院资源环境科学信息中心 兰州 730000)

万碧君等同志在《计算机检索技巧的探讨》(以下简称《探讨》)一文中,介绍和总结了常用的计算机检索的技巧,这对帮助那些刚开始利用计算机检索的人员提高检索效率是很有裨益的。但仔细阅读之后,笔者感到万碧君等同志的论文尚有遗漏和不妥之处,现结合该文谈谈一些实际工作心得,借以抛砖引玉,望同行多多指教。

计算机检索以其"检索速度快、费时少、检索面广"等特点而得到迅速发展和普及,但计算机检索是机器内部没有内涵的字符串匹配及逻辑运算的过程,缺乏人类的思辨能力,无法从学科内涵上选择文献,这就要求检索人员首先能充分理解课题的实质,选取准确表达课题的检索词,才能根据课题的性质、要求及实际检索情况,确定出符合要求的检索策略,保证较好的检索效果。

1 提高检全率的检索技巧

1.1 同义词检索 大多数数据库在前期制作过程中,标引人员对从文献中抽取的主题词按照特定的主题词表进行了规范,或通过同义词表(see/see also)将异型同义词指向规范词,保证使用规范词即可将主要相关文献检索出来。但对于一些新的概念,由于国内外尚未形成共识,主题词表也未能收录,这种情况就需要收集尽可能多的异型同义词,防止漏检的发生。

a.化学物质名称和化学物质表达式。对专业人员而言,化学物质表达式由于简洁、直观、准确而在文献中广泛使用,但有的化学物质表达式在计算机文本方式下无法准确表达(如有机化学中结构式、最简式的下标等),这就造成检索的无奈。在实际工作中,对比较成熟的化学物质来说,并且在相关文献中是主要论述的内容之一(即在标引中明确提出)的化学物质来说,用化学物质名称检索就可达到比较满意的效果。但对于一些特殊的情况(如文摘中出现了表达式,但题目、主题词均未出现名称的文献),或者有时为了扩大检索结果,需要使用化学物质表达式进行检索。并且有些数据库也提供了独特的检索途径,例如,DIALOG系统中的《化学文摘数据库》(CA)提供了用结构式检索的途径,只要用户终端有相应的输入设备即可检索(一般仿真终端则无法进行);也可以化学物质登记号进行检索(对成熟的物质,CA都赋予唯一的登记号),如用CR=2321-07-5就可将大部分有关Fluorescein(荧光素)的文献检索出来。又如INSPEC在自由词中也设置了相应的方式来解决上、下标问题,例如用/sup 3+/Al^{3+}表示Al^{3+}。DIALOG系统为仿真终端使用方便,在提供化学分子式检索途径的几个数据库中大部分都可用"元素符号+数字"的方式来进行分子式检索,如三氧化二铝用A1203即可进行检索。需要提醒的是,实际工作中必须按照相应数据库的规定格式输入(如词间空格等),否则检索结果会为零。

《探讨》一文谈到"当你输入主题词A1时,经机器匹配输出的命中文献中,除包含有主题词A1的文献外,还有含99%主题词al(含在etal中)的文献也同时输出",该文也提到"检索式包括单元词和由单元词组成词组的表达式"。其实对单元词检索的数据库(目前大部分数据库属此类型)而言,用A1检索,不会将含有al(如etal)的主题词的文献检出。单元词检索就是将输入的每个单词作为独立的单词(前后应有空格)进行检索,单元词检索的数据库是不会发生以上误检情况的。但是笔者也曾碰到这样的情况:国内某数据库系统有"同根词扩检"项目,选择该项,就会出现以上类似的误检现象。在实际工作中,应根据不同数据库的规定,确定相应的策略:是否选用元素符号进行检索,以防以上现象的出现。

b.缩写和全称。正如《探讨》一文所谈的,用缩写检索就会产生误检,用全称就有可能产生漏检。对相对成熟的概念,缩写和全称都有可能在文献中使用,如果只用其中一个检索都将造成漏检,因此检索时必须同时使用。同时,由于缩写词的多义性,在实际中应加以学科等其他限制,从而在尽量提高检全的同时减少误检的发生。如课题"计算机辅助设计",用"CAD+computer aided design"并加以其他限制(如分类、应用领域等),则既保证检出计算机辅助设计的相关文献,又基本可以排除"检验分析词典"、"弹药动力装置"、"循环辅助装置"及"冠状动脉疾病"(缩写均为CAD)等无关文献。

c.商品名、俗名等。有的物质在使用中,在国内外形成了广泛认可的商品名、俗名等。特别是在工程类、专利、商情之类的数据库中物质名、商品名、俗名等是共存的,如果漏选其中的一个,都将可能造成漏检。如聚四氟乙烯又叫泰氟龙,如果只用聚四氟乙烯检索,就可能将使用别名泰氟龙讨论聚四氟乙烯的文献遗漏。

在实际工作中,同义词的选取是比较困难的,由于检索

人员专业知识所限,不可能将所有的同义词都选取,同时用户也可能由于种种原因(如用户认为有些名称的别名对检索人员来说是常识,不需提示;或不了解商品名、俗名等别名)而造成遗漏。所以在机检前要多与检索人员交流,同时应查阅相关工具书,使尽可能多的异型同义词都能被利用,保证较高的检全率。

1.2 利用上位概念或下位概念的检索技巧 对命中文献太少的检索策略,可考虑用上位概念或几个下位概念逻辑或组配来扩大检索结果。例如"液体推进剂"可以用上位概念"推进剂"检索,也可以由下位概念"二甲推进剂"与"推进剂"等用逻辑或连接检索。前者可以将标引为"推进剂"的相关文献命中,但也会将"固体推进剂"、"混合推进剂"等无用的文献检索出来;后者可以将标引为各个具体推进剂的相关文献检索出来,但同时检出了大量只讨论具体推进剂的文献。用这种方法检索时误检率很高,在实际工作中要根据具体情况慎重使用。

《探讨》一文中"相关词"的讨论,笔者认为似有不妥。从文章内容及结构看,作者讨论的"相关词"应该是指异型同义词,但所举例证却不是为了证实这个观点。我们知道,"专家系统"是对数据库技术的完善和发展,"软件"是"数据库"的上位概念。"数据库"、"专家系统"、"软件"并非是"同义的相关词",而是上位概念和发展应用基础等概念,并不属于作者所讨论的"相关词"。

1.3 截断技巧 前截断由于技术难度大,检索时间长,实际应用的并不多见(上文提到的同根词检索就是一个特例)。由于英式英语和美式英语拼写形式有差异,或有些相关外文词区别只是一两个字母,所以中截断在实际检索中的作用是比较重要的。如用"wom? n"就可以替代"women"和"woman",用"fib?? board"替代"fiberboard"和"fibreboard",从而使检索式表达简洁,可以提高检索效率(因为有的联机数据库是按检索词数目的多少来收费的,例如美国的 OCLC 系统)。后截在《探讨》一文已有较详尽的说明,这里不再赘述。顺便说一下,有限截断中只截取一个字符用"? ?",即两个? 之间有个空格(《探讨》中未详细提示,易产生误解)。

需要指出的是,中文数据库由于主题词还不规范,如果不用截断技术,往往漏检率较大,所以在实践中对命中文献数量不是很大的情况,应使用后截断来保证查全。用户可进行二次选择,就可以有效地排除无关文献。

2 提高检准率的检索技巧

2.1 逻辑与和逻辑非 笔者很同意万碧君等同志对逻辑与和逻辑非的概念解释,但文中所举例证似乎让人有点费解。我们知道,Al 和 Aluminium 是同一物质(铝)的不同表达方式,如果执行检索式"Al * Aluminium",则命中的文献将是在标引中既要出现 Al 拼写形式又要出现 Aluminium 拼写形式的文献,有关只用其中之一表达方式的文献将会被排除,这样漏检率将会很高。笔者认为在实践中应使用"Al + Aluminium"进行初步检索,然后加其它限制(应用领域、方法等)。对单元词检索的数据库来说,查准率和查全率都较令人满意。但对于上文中提到的"同根词扩检"的数据库来说,利用以上检索式就有可能造成大量误检,故应将 Al 从检索式中删除,以降低检全率来提高检准率。

《探讨》一文认为"但实际上却不是这样的,减号'-'在此并没有起逻辑非算符的作用,而却起了逻辑或的作用",不知作者特指的是哪种数据库,但据笔者实践经验和所接触的数据库来看,减号"-"一般是起逻辑非算符的作用。对于单词间的破折号,检索时需用引号加以标注,以便与减号相区别(这是对于词组检索的数据库而言,对单元词检索的数据库则需要用位置符(W)连接各单元词来检索);至于逻辑或的作用,笔者到目前尚未发现此类例证。

2.2 词位置检索和固定词组检索 需强调的是词位置检索只能用于单元词检索的数据库,而固定词组检索也只能用于词组检索的数据库,二者不能混用。在 DIALOG 系统中,较常用的词位置检索还有限定在同一子字段的(S)位置符,对用逻辑与连接的单元词(或词组)有时命中文献仍较多,则可考虑使用(S)位置符,使词间逻辑联系相对紧密。

2.3 其它检索途径 在命中文献数目较多的情况下,可限定在题目、主题词等较窄检索途径中检索,使检中文献更加符合要求。当然,漏检的文献也相应地多了。

检全和检准之间存在着一种类反比例的关系,其中任何一项的提高,就有可能导致另一项的降低,因此在实际检索中要根据课题性质,决定首先要保证检全还是检准,然后根据以上技巧,用准确表达课题的检索词(或词组)制定合适的检索策略,为用户提供较满意的检索结果。

由于笔者接触的只是部分国际、国内联机和光盘数据库的计算机检索工作,必定有许多疏漏之处,请万碧君等同志和各位读者能不吝赐教。

参 考 资 料

1. 万碧君等. 计算机检索技巧的探讨. 情报杂志,1997;(5)

2. 辛瑞杰等. 联机检索原理与方法;DIALOG 系统用户指南. 黑龙江省科学技术情报研究所,1992

(责编:亦愚)