

提高信息检索效率的途径——提高查全率与查准率

沈阳航空航天大学图书馆 刘 晓

[摘 要]查全率和查准率是评价文献检索效果的两项重要指标,本文通过对影响查全率和查准率的因素分析,从检索途径、数据库选择等方面探讨了提高查全率和查准率的方法。

[关键词]信息检索 检索效率 查全率 查准率

信息检索效率是指全、准、快、便、省(检全率、检准率、检索方便性、检索成本与效益),最主要的是全和准。^[1]1955年,美国的佩里(J.W.Perry)和肯特(A.Kent)最先提出了查全率和查准率的概念。上世纪50-60年代,英国学者克里维顿(C.M.Cleverdon)在他著名的Cranfield试验中首次将查全率(Recall)和查准率(Precision)作为信息检索系统效率的评价指标。^[2]半个世纪过去了,在高性能计算机、数据库、数据仓库和互联网等信息技术广泛应用于信息检索领域的今天,查全率与查准率仍是衡量文献信息检索效果的重要数据指标。

1. 信息检索中的查全率与查准率

在某个数据库或信息检索系统中,对某个单主题、多主题或者复合主题进行检索时,检出一定数量的文献,其中有的是检索主题的相关文献(t),剩下的是非相关文献(f)。而在整个数据库或者信息系统中对于检索主题存在一个相关文献信息的总量,除去检索出的相关文献,还存在没有检索出的即漏检的相关文献(d)。则查全率(R)= $t/(t+d) \times 100\%$,查准率(P)= $t/(t+f) \times 100\%$ 。

2. 信息检索中查全率与查准率之间的关系——互顺和互逆关系

2.1 从两者的计算公式来看,我们也可以尝试推论一下两者的关系。设某一数据库的某一相对稳定的时刻,包含信息总量为Z,此时对于检索主题的相关信息量t+d固定不变。假如我们设定一种极限的状况,检索出的信息总量t+f也为Z,当查全率最大的时候,R=1,d=0,查准率P=t/(t+f)=Z/Z,此时,P的大小只随t而变化,P可能很大,也可能很小,此时P与R并不存在可逆关系,反之亦然。或者简单的作下代换,得到P=(t+d)/(t+f) × R=(t+d)/Z × R,由于(t+d)/Z ≥ 0,P与R则呈现出一种互顺关系。也就是说在检索中可使查全率和查准率同时提高。

2.2 然而,相对于互顺关系大多数学者更倾向于互逆关系,许多检索实例也证明了两者的互逆关系。如英国学者克里维顿(C.M.Cleverdon)著名的克兰菲尔德Cranfield试验。对于实际检索而言,理想的状态当然是查全率与查准率都为100%,然而,实际上这是根本不可能的。对于一个已经确定数据库的某一时刻来说,对于要检索的主题,它应该存有这一专业领域的文献信息集合。此时要想检索到更多的信息,势必要对检索范围放宽,检索途径要宽,选择检索词时泛指性要强,逻辑运算符应选择具有扩检意义的“逻辑或”关系(or),这时避免了大量漏检相关信息,然而这样做必然使同时检索出不相关的信息增多,也就是“噪声”增加,势必降低检索的准确性。相反为了检索结果更加准确势必要减小检索范围,严格控制检索途径,精确选用检索词,使用“逻辑与”关系(and)作为运算符,此时虽然得到的信息与所检索的主题匹配度较高,但是漏检的相关信息量就会大大提高,全面性必然受到影响。

随着计算机和网络技术普及,电子信息资源成几何级数增长,用户对电子信息资源的需求也日益增长。为了迎合这一趋势,数据库建设也蓬勃发展起来,数据库的规模越来越大,内容越来越多,数据库运营商都想以大而全吸引用户。相对的在一个日益庞大的数据库中某一主题的相关信息所占的比例就越来越小,这对查全率和查准率都产生了影响,实际应用中多体现为查全率上升、查准率下降。

3. 实际工作中提高信息检索的效率

查全率和查准率对检索系统的评价是通过该系统的检索命令和检索式在该系统后台进行匹配而得到的结果中反映出来,是检索系统质量和检索策略质量的综合体现。^[3]信息检索技术首先涉及的是计算机对用户的需求识别与转换技术,包括提问式的构造技术、自然语言技术、信息过滤技术、跨语言检索/机器翻译技术等。一直以来人们都在探索利用新的信息技术,来解决查准率的问题。语言学和数据挖掘技术以及智能代理技术在相继应用于网络信息检索。^[4]这些技术改善了传统检索中简单的“字”、“词”匹配,实现概念匹配、用户行为分析,从而提高了文献标引与用户需求之间的匹配度。目前信息系统基于文本分析、自动分词技术的自动标引,作为检索人员都无法控制。那么从检索者的角度,如何提高文献检索的效果,可以注意以下问题。

3.1 检索前充分准备。工作在正式检索之前应对所检索的主题有较深入的理解,努力做好知识储备和准备工作。对主题内容进行分析,与委托人进行交流,分析出检索词,检索词的从属词、同位词甚至英文翻译等。如“工艺参数对钛合金镀铬质量的影响”这一课题。通过与委托人交流,需要查询的是航空航天工业中,钛合金镀铬质量的问题,而钛合金属于轻有色金属及其合金,镀铬是在金属制品表面镀上一层致密的氧化铬薄膜,可以使金属制品更加坚固耐用,在电镀工业中一直占重要的地位。仅仅通过简单的分析,我们在检索时为了提高查全率就应该同时考虑以“轻有色金属”和“电镀”为检索词。

3.2 选择合适的数据库。当代电子资源数据库繁多,然而,规模、质量各不相同,侧重也有所不同。对不同主题进行检索时,选择也应该灵活、准确。首先应选择专业数据库,如检索航空航天领域的主题就可以检索航空工业文摘数据库、美国航空航天学会(AIAA)全文电子期刊及

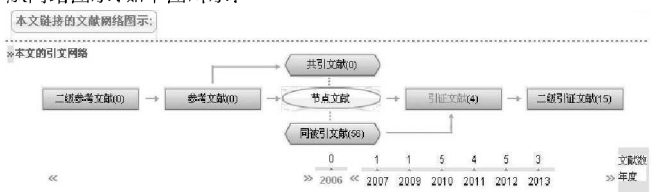
会议论文数据库、AV-DATA全球航空法规数据库,同时辅助检索综合性的大型数据库,如CNKI、SpringerLink外文库等。同时,为了满足科研需要,应尽量选择能够提供全文的数据库。此外,各大数据库多提供电子期刊文献,而在实际检索时还应注意学位论文和会议论文等的检索,以免造成大范围的漏检。

3.3 使用模糊检索。“模糊检索”是与“精准搜索”相对应的一个概念,由模糊的语言及关系运算符、逻辑运算符等组成的约束条件下,从数据库中检索出在某一程度上满足约束条件的记录,这种检索称为模糊检索。^[5]在模糊检索条件下系统可以智能识别词组中的单词并进行拆分搜索,模糊检索时拆分所得的结果约等于两词用and连接。模糊检索并不等同于布尔逻辑运算。实际上模糊检索是在模糊数学基础上建立起来的检索技术理论是布尔逻辑的扩展。用布尔逻辑只能根据逻辑值“非真即假”来考察系统中是否还有符合检索提问的文献。可是,检索词在归属或不归属文献之外还存在“部分归属”文献的情况,如一个词的多种形态,在精确匹配系统看来他们是不同的,而从用户的角度看他们是相同的。

3.4 多途径构建检索式。数据库专业检索中通常有责任者、题名、主题词、文摘、全文、分类、刊名、机构、单位等多种检索途径。用同一检索词在主题词与全文中检索,结果可能相差百倍。为了尽量减少漏检和错检,可适当的控制检索途径。而对于不同的检索课题或同一课题的不同检索阶段,通过责任者、机构等不同的检索途径,限定文献类型和时间可起到不同的检索效果。在日常检索中,检索人员常常忽视责任者、机构、单位等与文献主题无直接关系的途径,实际上通过这些途径检索往往有出人意料的效果,对于某一专业主题,通过对具有权威性的专家作者和单位机构的检索,能够减少因为标引质量不高引起的漏检。

3.5 多用“逻辑或”,减少“逻辑与”。无需多言,只看一个实际的例子。在万方数据知识平台用“钛合金”和“镀铬”在主题中以“逻辑或”检索得到期刊论文、学位论文等结果合计15001条,而以“逻辑与”检索,结果只有5篇文献。

3.6 充分利用数据库或信息系统提供的相关内容。对于现在的数据库或信息系统来说,简单提供文献的匹配查询已经不能适应市场竞争的需要,很多数据库提供网络式增值服务,即对选中的任何一篇相关文献提供形成网络的链接服务。以中国知网为例,以“钛合金镀铬”为检索词检索,得到一篇文章,此时在文章题录下方就有该文章链接的文献网络图示,如下图所示:



此外,还列出了系统自动匹配的【引证文献】、【相似文献】、【同行关注文献】、【相关作者文献】、【相关机构文献】以及【文献分类导航】。通过这些链接,就可以建立检索到的相关文献的立体关系网络,从而不断扩展检索范围。

4. 结语

随着互联网的发展和电子信息资源的不断增加,检索工具也更为广泛,但是各种检索工具都有自己的信息收集方式、检索算法和结果排序方式,信息组织没有统一的规格,索引方式也各不相同,标引质量更是参差不齐。但是,从整体上看,检索技术正朝着智能化检索、全文检索、基于知识的检索、多媒体检索、多语言检索等方向发展。^[6]只要采取合理的措施和科学的策略,就能够提高用户的信息需求的满意度,使查全率和查准率都达到理想状态,从而实现检索效率最佳化。

参考文献

- [1] 燕燕,赵蓉英.国内外网络全文数据库比较研究[J].情报科学,2004,(2):228-231.
- [2] 范晴,徐建华,宋震.查全率与查准率关系初探[J].情报杂志,2002,(9):41-42.
- [3] 余丹.关于查全率和查准率的新认识[J].西南民族大学学报(人文社科版),2009,(2):283-285.
- [4] 郭家义.网络信息检索效率研究[J].图书与情报,2003,(2):60-62.
- [5] 李茂表.一类广义模糊检索功能的研究[J].厦门大学学报(自然科学版),1997,36(5):674-677.
- [6] 王颖春.影响网络信息检索效率的因素与对策[J].现代情报,2007,(9):5-7.