

信息检索中文献漏检误检因素分析

俞 平

(遵义医学院图书馆, 贵州 遵义 563003)

摘 要: 文章分析了信息检索中文献漏检和误检的因素, 从数据库的选择、检索词的确定、检索策略的制定等方面提出了避免漏检误检的对策。

关键词: 信息检索; 漏检; 误检

中图分类号: G350 **文献标识码:** A **文章编号:** 1007-7634(2011)08-1234-04

Analysis on the Factor of Missing Retrieval and Erroneous Retrieval in the Information Retrieval

YU Ping

(Library of Zunyi Medical College, Zunyi 563003, China)

Abstract: This paper analyzes the factor of missing retrieval and erroneous retrieval in the information retrieval, proposes the countermeasures of missing retrieval and erroneous retrieval from the selection of database, the determination of index terms and the establishment of retrieval strategies and etc.

Key words: information retrieval; missing retrieval; erroneous retrieval

科研选题的新颖性判断是依赖于相关文献的查全, 只有掌握了全面的相关文献, 才能对所选课题的新颖性做出客观的评价^[1], 因此, 文献信息检索的查全率对科研人员的科研活动还是查新人员的查新质量都有着重要意义。随着信息技术的迅速发展, 检索手段由原来的手工检索发展到计算机检索, 为此, 在检索过程中需要检索人员熟悉这些检索工具, 系统掌握检索技术的基本原理及操作技巧, 否则会造成漏检和误检。笔者在查新实践中常常发现科研人员所提出的创新点已有文献报道而被否定, 分析其原因, 主要是不熟悉计算机检索的原理和操作技巧, 导致了关键文献的漏检, 从而影响科研课题创新点的成立。

1 数据库的选择

目前国内外数据库种类繁多, 既有综合性的, 也

有专业性的, 在文献收录范围、文献信息标引质量和提供的检索功能上都有很大的差异^[2], 任何一种数据库或检索刊物都不可能全面收录本专业及相关专业的文献, 仅选择一两种数据库或依靠一两种检索手段, 往往会造成文献的漏检和误检。如我们曾接到一课题为“利用多模态 MRI 新技术探讨小乳腺癌的早期诊断价值”的立项查新, 该课题的目的是研究磁共振动态增强、扩散加权 and 单体素质子波谱成像技术三者联合应用对小乳腺癌的早期诊断价值。查新后发现, 在检出的学位论文中有天津医科大学 2006 年的一篇题为“磁共振动态增强、扩散加权及单体素质子波谱成像对乳腺癌诊断价值研究”的学位论文, 经仔细阅读发现该学位论文采用的磁共振成像技术与课题申请者完全相同, 对比研究了磁共振动态增强、扩散加权及单体素质子波谱成像技术分别单独或三者联合检查对乳腺良恶性病变诊断的价值, 论文中研究的乳腺肿瘤的病灶最大直径为

收稿日期: 2011-03-10

作者简介: 俞 平(1967-), 女, 浙江富阳人, 副研究馆员, 主要从事科技查新及医学信息服务工作研究。

0.4cm~11.0cm,包括了小乳腺癌(小乳腺癌病灶直径为 ≤ 2 cm),因此,该课题的创新性是不能成立的。

分析这个案例,其失误产生于开题前的信息收集过程,申请人虽然在开题前检索了一些数据库文献,但忽略了对学位论文数据库的检索,以至于漏检了关键文献而使课题新颖性被否定。因此,在信息检索过程中,选择什么样的检索系统进行检索是保证查全率和查准率的关键,如果选择不当,直接影响检索效果。确定检索系统的一般原则是:

(1)合理选取数据库,以专业检索系统为主,综合性检索系统为辅。根据课题所属学科范围合理选择数据库进行检索,如检索医学文献必须首选CMCC、CBMdisc、MEDLINE等数据库;检索药学文献必须选药学文献数据库、美国《化学文摘》、《德温特药学文档数据库》、《药理毒理学数据库》、《国际药学文摘》等;检索中医药文献必须选中医药文献数据库、CMCC、CBMdisc等。同时再辅以综合性的数据库检索,如:《中国知网(CNKI)》、《维普信息资源系统》、《万方数据资源系统》等。

(2)重视学位论文、会议论文、成果、专利等的检索。科研项目一般有基础理论研究、应用研究、开发研究等性质,要根据课题的性质和要求选择相应的检索数据库。对于医学基础理论类、临床研究类项目可以选择以期刊文献类数据库为主的信息资源,因为医学基础理论研究、医学公益性技术研究成果大多以论文形式发表于期刊、会议论文集上,同时不能忽视学位论文、会议论文、成果等数据库,如:《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》、《中国学位论文全文数据库》、《中国重要会议论文全文数据库》、《中国学术会议论文全文数据库》、《国家科技成果网》等。医药、医疗器械产品的开发研究,则必须检索专利数据库,如中国专利信息网、国家知识产权局的《中国专利文献数据库》等。

(3)从内容方面来说,为保证查全率,应该选择收录量多,收录年限长,更新频率快的检索系统;要保证查准率,应从文献的学科范围、收录时间等考虑。

2 检索词的确定

大部分信息检索系统是用自然语言组织的,可采用关键词进行检索,这极大地方便了用户^[1]。但是,关键词未经词形和词义的控制,不显示词的等级关系和相关关系,使得检索语言中存在大量的反映

同一概念的不同形式的词,不利于文献的查全^[4]。因此,在选取检索词的时候,必须从多种角度考虑,对关键词加以扩展才能尽量避免漏查。包括:同义词、近义词、全称、简称和缩写、上下位词,同时还要注意外来词的音译变化、检索词的易错形式和自造词等。(以下涉及到的中文检索为CBMdisc检索,检索时间为2011年1月初)。

2.1 同义词的检索

信息检索中的同义词所指的范围相对宽泛,主要包括:同义词、近义词、相关词、全称、简称、缩写、别名等,检索时要将表达同一概念的同义词、近义词、相关词用布尔逻辑算符“OR”联结起来,以提高相关文献的查全率。

以课题“局部晚期鼻咽癌新辅助化疗前后Ki67、VEGF表达变化及意义”为例,“癌”和“肿瘤”是同义词,“新辅助化疗”又叫“诱导化疗”、“术前化疗”,蛋白因子“Ki67”表达形式还有“Ki-67”,“VEGF”是“血管内皮生长因子”的缩写,检索时如果只采用其中一种表达形式,检索结果是不同的,只有将全部的同义词、全称、缩写等不同表达形式用“OR”连接起来检索,命中文献数才最大。如下:

- (1)新辅助化疗 1815 (1)鼻咽肿瘤 12020
- (2)新辅助化学治疗 12 (2)鼻咽癌 11808
- (3)诱导化疗 713 (3)鼻咽瘤 6
- (4)诱导化学治疗 6 (4)(1)OR(2)OR(3) 13007
- (5)术前化疗 892
- (6)术前化学治疗 7
- (7)(1)OR(2)OR(3)OR(4)OR(5)OR(6) 3203
- (1)VEGF 12408 (1)"Ki-67" 2553
- (2)血管内皮生长因子 12862 (2)Ki67 703
- (3)(1)OR(2) 14415 (3)(1)OR(2) 2776

2.2 反义词的检索

同义词和反义词是意义相反的一组词组,在文献检索中我们发现,有些反义词是从相反的角度反映同一个问题,造成相同主题文献的分散,所以,检索时对反义词同时进行检索,可以提高信息检索的查全率。如:“瑞特恩对低促性腺激素型少精不育症的治疗”,“少精”一词除了检索同义词“弱精”、“无精”外,还要检索反义词“生精”。

2.3 上位词和下位词的检索

词义间的关系除相同和相关关系外,还有属分

关系,即有上位词和下位词的存在,上位词是泛指概念,内容含义更广泛,下位词是专指概念,内容含义狭窄。在信息检索时,要弄清楚当前检索词的上位概念和下位概念以及这些概念之间的关系,通过检索上位词和下位词,扩大检索范围,提高相关文献的查全率。

如课题“心肌缺血大鼠心肌微血管内皮细胞损伤机制研究”中“心肌缺血”一词,下位词有心绞痛、冠心病、心肌梗死、心肌顿抑、急性冠状动脉综合征等,其中心肌梗死又简称心梗,与心肌梗塞是同义词,分别检索这些词,命中文献数量为:心肌缺血(20159篇)、心绞痛(29628篇)、冠心病(60026篇)、心肌梗死(55569篇)、心肌梗塞(20044篇)、心梗(3813篇)、心肌顿抑(392篇)、急性冠状动脉综合征(2874篇);将这些词全部用“OR”连接检索,则检出135343篇文献。显然,如果只是简单地将“心肌缺血”作为检索词来检索,将会漏检大量文献。又如,在检索“中药银杏达莫对冠状动脉无复流现象的作用”这一课题时课题中,检索词用银杏达莫(或银杏黄酮甙、银杏苦内酯、白果内酯、双嘧达莫,均为银杏达莫的有效成分)与无复流(或慢复流、无再流,为无复流的同义词)组配,则检索不到文献,用银杏达莫的上位概念中药或中草药来扩大检索,则检出13篇相关文献。

由于上位概念词所指过于宽泛,没有明确性,虽然能保证查全,但也容易造成误检,要提高信息检索的查准率,则应避免使用外延广泛的上位词,尽量选择专指度强的主题词、下位词、以及规范的专业术语,这样能减少误检。总之采用上位词和下位词检索,应根据检索课题的具体情况灵活运用,以达到查全和查准的最佳检索效果。

2.4 使用单元词检索,慎用短语和词组

单元词是表达事物概念的最小的不可拆分的词语,检索时要尽量使用确切表达概念的单元词。如检索“高能聚焦超声”,检出206篇、用“高能 AND 聚焦超声”,检出227篇,多出21篇。

2.5 忌用自造词检索

检索词应该涵盖行业内公知公用的技术术语,自创词汇或自己命名的产品名称、产品型号,都不宜用作检索词^[5]。如在查新中我们遇到的“穿支蒂网状供血皮瓣”、“瑞特恩”就是项目委托人自创词汇或自己命名的药品名,如果以此为检索词,是检不到文献资料的,即使有文献,也是误检或者委托人发表的

相关文献。

2.6 注意错别字的检索

网络信息资源中,错别字现象层出不穷,即便是CBM这种生物医学领域最重要的文献数据库检索系统也不例外,如:

细胞迁移 1377篇 (错)细胞迁徙 4篇
针灸 61234篇 (错)针炙 232篇

2.7 外来词的检索

检索外来词时要注意外来词的译写变化,尽量找全不同的音译形式才能做到查全。如:

格林-巴利综合征 1800篇
格林巴利综合征 369篇
吉兰-巴雷综合征 364篇
吉兰巴雷综合征 9篇
古兰-巴雷综合征 6篇

以上这些词都是同一种病“Guillain Barre syndrome”的不同音译形式。

在实际检索中,检索词选择过多过少都不利,放大查全但不准,缩小限定则不全,这就需要检索人员在对课题学科理解透彻的基础上进行归纳整理,根据词的同类、隶属、相关等关系列出一系列检索词,应用有关主题词表、分类词表查核相关检索词,试检筛查并随机调整、修正、扩充,选择适当的检索词进行检索,防止漏检误检。

2.8 截词检索

截词检索就是利用检索词的词干或不完整的词形进行检索,查找含有该词干的全部检索词的记录,它可以起到扩大检索范围、提高查全率、减少检索词的输入量、节省检索时间、降低检索费用等作用。由于目前的中文数据库基本上都可以进行字词的任意一致检索,只要直接键入检索词,即可自动实现截词检索。因此截词检索常用于西文文献检索中。在西方语言中,词干相同的词,其基本含义通常是一致的,如某一概念的名词、动词、形容词、分词、单数、复数、英美拼法等形式,进行文献检索时,在词干前后可能变化的字符处加上截词符作为检索标志,代替多个同义词、近义词进行检索,以提高查全率;同时也可以减少检索词输入量,简化检索步骤。

截词方式有多种,按截断的位置来分,截词有前截断、中间截断、后截断等;按截断的字符数量来分,可分为有限截断和无限截断两种。不同的检索系统

所采用的截词符有不同的规定。如:Medline 光盘数据库用“?”和“*”作为截词符,?代表0个或1个字符,可在一个单词中或末尾使用,如检索细胞用“cell?”可同时检索到细胞的单复数形式:cell 和 cells,“wom?n”可检出“woman”、“women”;*代表0个或多个字符数,查找糖尿病,用“diabet*”,可检出diabetic、diabetes、diabetogenic、diabetology 等不同词性的同根词,提高了文献的查全率。

3 检索策略的制定

检索策略是指处理文献信息检索时制定的检索计划和方案,检索策略的具体体现是检索式,检索式是指计算机检索中用来表达信息检索需求的检索提问式,它控制着检索过程,关系到文献的查全率和查准率^[6]。因此,设计合理的检索式成为控制和提高检索质量的关键。针对不同的数据库,不同的信息需求,有不同的检索策略,其检索式的构造也各有不同。

3.1 理解课题主要内容及技术路线,弄清主题概念之间的逻辑关系

对检索课题进行主题分析,弄清楚主题概念之间的逻辑关系,选择逻辑运算符进行概念组配,用以表达信息检索的需求。逻辑运算符也称为布尔运算符,有“逻辑与”、“逻辑或”、“逻辑非”三种,分别用“and”、“or”、“not”表示,也有的数据库分别用“*”、“+”、“-”表示,“and”和“not”缩小检索范围,“or”扩大检索范围。几乎所有的数据库都支持布尔逻辑运算符,在实际运用中我们可根据不同的检索需求来确立不同的逻辑组配关系,并选用正确的逻辑运算符来表达,如果逻辑组配关系确立不当,将产生漏检及误检,直接影响检索质量。

如:在检索“超声速度向量技术评价TIA患者颈动脉管壁运动及斑块稳定性研究”课题时,速度向量技术又叫速度向量成像技术、速度矢量成像,缩写为VVI,是逻辑或的关系,动脉、管壁、斑块之间也是逻辑或的关系,因此检索时可初步制定为“(速度向量 OR 速度矢量 OR VVI) AND (动脉 OR 管壁 OR 斑块)”,CBM中检到文献72篇。通过阅读文献我们发现,VVI也是起搏器的一种类型,还是心脏单腔起搏模式的简称,如果只用上面的检索式检索,会出现很多误检的文献,因此,检索时调整为:“(速度向量 OR 速度矢量 OR VVI) AND (动脉 OR 管壁 OR 斑

块) NOT 起搏”,检到文献39篇,避免了误检。

3.2 合理使用位置算符

位置算符用来限定检索结果中词与词之间的位置关系,可以提高查准率。应用规则为:①如果在一个检索式中有几个不同的位置算符,系统按照从左至右的顺序处理;②如果在一个检索式中既有位置算符,又有逻辑算符,系统优先执行位置算符;③使用括号可保证优先权。

位置算符在不同的检索系统中有不同的规定,如medline 光盘数据库中位置算符是“with”、“near”。用With算符表示连接的两个检索词必须出现在同一字段,如:(diabetes with diet) in ab,表示检出文献中diabetes 和 diet 两个词同时出现在摘要字段。用near算符表示连接的两个检索词必须出现在同一个句子中,位置可以互换;near后面可加阿拉伯数字,指定两个检索词相邻的程度,词序不计。如:Nasopharyngeal near3 Cancer,表示Nasopharyngeal 和 Cancer 两词相邻不超过3个单词,可检出Nasopharyngeal Cancer, Cancer of Nasopharynx, Cancer of the nasopharynx 等。

位置算符的功能与逻辑与(and)功能相似,用“and”连接两个检索词,是要求两个词出现在同一条记录中,用“with”连接,要求两个词出现在同一个字段,用“near”连接,则要求两个词出现在同一个句子中,检出的数量是递减的,因此,灵活运用逻辑算符和位置算符,可以扩大或缩小检索范围,有效避免漏检和误检。

3.3 避免地域名限制

科研课题很多是以地域名限定的,如果在构造检索式时用地域名加以限制,那多数情况下检索结果都为零,但这并不能说明课题的创新性,不符合科技查新的客观、公正、实事求是的原则^[7],因此要要尽量避免使用地域名来限制。

3.4 检索策略的调整和优化

检索策略不是一成不变的,检索中我们经常会遇到检索结果文献量过多或者检不到相关文献的情况,因此,在具体的检索过程中要根据检索结果的满意程度作适当调整和优化,或缩小检索范围或扩大检索范围。如:提高检索词的专指度;用AND联接一些进一步限定主题概念的相关检索项,增加检索词组配面;用NOT排除不相关概念;(下转第1243页)

World-Wide-Web、相关性、查询等;③有关情报学研究客体的研究,如:信息、知识、管理、交流、用户、行为等;④有关情报学前沿应用的研究:如电子商务、引文分析等。以上这些热点词汇涵括了情报学领域的方方面面,从不同的学科角度,不同的方法对情报学进行相关的研究。

对于本文的研究,作者认为还存在着不足,有关情报学领域热点词汇的研究,虽然我们尽量使用较为广泛的词汇来确定其研究热点,但是由于本研究的结论是基于对文献主题词词频统计分析而做出的,因此文献中主题词的标引数量及质量将直接影响到图2及图3中的质量和效果,同时在一定程度上也会影响到我们结论的精确性,因此我们认为无论是对于情报学还是其它领域的文献,我们要尽量完整、准确地使用统一的主题词,同时还要加强对主题词标引和使用的规范化,从而提高文献检索利用的效率和价值。

参考文献

1 北大学者.核心期刊[EB/OL]. <http://baike.baidu.com/view/>

63483.htm,2010-12-27.

2 ASIST [EB/OL]. <http://www.asis.org/jasist.html>.2010-12-28.

3 Information Processing & Management[EB/OL]. http://www.elsevier.com/wps/find/journaldescription.cws_home/244/description#description,2010-12-29.

4 SCIENTOMETRICS[EB/OL]. <http://www.springerlink.com/content/101080/>,2010-12-29.

5 侯剑华.工商管理学科演进与前沿热点的可视化分析[D].大连:大连理工大学,2009.

6 Gunel K, Asliyan R. Extracting learning concepts from educational texts in intelligent tutoring systems automatically [J].2010,37(7):5017-5022.

7 Gila Prebor. Analysis of the interdisciplinary nature of library and information science [J].2010,42(4):256-267.

8 [EB/OL].<http://baike.baidu.com/view/757.htm>2010-12-23/2010-12-28.

9 赵蓉英,王 菊.国际信息检索模型研究的可视化分析[J].图书情报工作,2010,(9):61-66.

(责任编辑:徐 波)

(上接第1237页)

利用文献的外部特征加以限制,如文献类型、出版年代、语种、作者等,或将检索词限制在“题名”字段检索等,可以缩小检索范围,提高文献的查准率,减少误检。反之,降低检索词的专指度;保留主要概念和基本概念,剔除重复概念和通用概念,减少用“AND”连接的组配面;采用分类号进行簇性检索等,则可以扩大检索范围,提高文献的查全率,避免漏检。调整检索策略应在查全的基础上再根据课题需要调整,一步步加以限制,减少误检。

综上所述,信息检索中要避免漏检误检,应注意:①根据课题主要技术要点提炼准确的检索词,要充分考虑主题词、自由词以及他们的同类词(包括全称、简称、缩写、同义词、反义词、上下位词等);②检查修改检索表达式,不断调整和优化检索策略;③使用多个数据库检索,以专业数据库为主,同时补充通用数据库;要检索多种文献类型,除了检索期刊文

献外,不忘学位论文、会议论文、专利、成果、标准等的检索,只有这样才能取得满意的检索效率。

参考文献

1 李希明,梁蜀忠,孙金立,等.谈文献检索的查全性[J].中华医院管理杂志,2005,21(10):707-708.

2 曹欢增.提高科技文献查全率的几项措施[J].科技情报开发与经济,2008,18(32):72-73.

3 沈艳红.信息检索中检索词的选择对查全率的影响[J].情报探索,2006,(11):73-74.

4 黄瑞芬.文献数据库检索中的漏检与误检原因及对策[J].情报理论与实践,2004,(6):43-44.

5 张柏秋.科技查新检索中的关键词选择[J].情报科学,2008,2(9):1345-1348.

6 胡明一,李更,赵伟.医学科技查新信息检索策略研究[J].中华医学图书情报杂志,2008,17(5):52-54.

7 张 岚,张柏秋,于 非,等.探讨科技查新中检索策略的制定[J].现代情报,2008,(10):151-152,157.

(责任编辑:徐 波)