

面向医学学科领域的检索词库构建^{*}

李艳超 王 艳 金新建

(安徽医科大学 合肥 230032)

〔摘要〕 分析专业数据库和搜索引擎在进行文献检索时存在的问题及构建医学学科领域检索词库的原因,以教育部科技查新站 2014 年以来的科技查新报告为数据源,建立经过科技查新和学科专家共同审验通过的检索词库。该词库可应用于科技查新、文献检索教学等领域,还可推广到其他领域。

〔关键词〕 科技查新; 学科服务; 文献检索; 关键词; 高校图书馆

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.05.018

Construction of the Medical Field Oriented Retrieval Thesaurus LI Yan - chao, WANG Yan, JIN Xin - jian, Anhui Medical University, Hefei 230032, China

〔Abstract〕 The paper analyzes the problems of the specialized databases and search engines during document retrieval and the reasons for constructing retrieval thesaurus in the field of medical field, and establishes a retrieval thesaurus jointly reviewed by sci - tech novelty assessment and subject specialists by taking sci - tech novelty report of the sci - tech novelty station of the Ministry of Education since 2014 as the data sources. The thesaurus applies to sci - tech novelty assessment, document retrieval teaching and other fields, which can also be promoted in other fields.

〔Keywords〕 Sci - tech novelty assessment; Subject service; Literature retrieval; Keywords; University library

1 引言

科技文献记录着科学技术研究活动的丰硕成果,反映了社会科技水平的发展状况,是人类智慧与科技水平的集中体现。对科技文献的开发利用直接影响着国民经济的发展历程和科技事业的发展水平,对全面提升科技创新能力,进一步促进科技成果转化起着非常重要的作用^[1]。在大数据时代,科

技文献信息以不同形式广泛存在于各种搜索引擎和专业数据库中^[2]。但在专业数据库和搜索引擎中,检索结果是基于检索人员提供的关键词进行相关运算、匹配得出的,检索结果的全面性与检索词的全面程度呈正相关^[3],而且,同一检索词的不同表述也会因检索人员的检索经验和学科背景不同而千差万别^[4-5],该问题也同样存在于学科服务过程中,若学科馆员对课题的理解与检索式的构建合理性不足,则会影响其检索结果的全面性和准确性,进而影响其服务的课题,因此构建面向学科服务的检索词库具有重要意义。本文以教育部科技查新站 2014 年以来的科技查新报告为数据源,构建经过科技查新及学科专家共同审核通过的检索词库,该词库不仅可以应用于学科服务、科技查新等领域,还可应

〔收稿日期〕 2017-03-20

〔作者简介〕 李艳超,馆员,发表论文 12 篇。

〔基金项目〕 安徽省高校图工委重点项目“基于科技查新的高校图书馆专利信息服务模式研究”(项目编号: TGW16A07)。

用于文献检索教学,也推广到其他领域。

2 构建医学学科领域检索词库的原因

2.1 医学学术用语的多样性

医学学术用语具有多样性,如主题词、自由词、全称、简称、学名、俗名、药品名、商品名等不同表述,因此一部分科技论文作者在发表文章时经常会根据自己的理解和使用习惯选用较口语化的关键词进行标引,如再障、人流等;还有一部分作者将英文缩写作为关键词,专指性不强,如 CHD,不仅可以代表充血性心力衰竭、先天性心脏病,还可以代表先天性髋关节疾病。此外,科技文献中普遍存在一物多名的情况,尤其是医学学科的学术用语,如一些药品会有不同的商品名,200 种常用药品中,20% 有 4 个药名,25% 有 5 个药名,有的甚至多达 10 个药名,例如消化系统药物“法莫替丁”的商品名就有胃舒达、保胃健、卡玛特、愈疡宁等不同名称^[6-7]。在检索时如果只使用其中的一种或几种表述,就会造成漏检,遗漏的文献中可能包含该学科领域内的重要文献或经典文献,从而影响对文献的研究与判断。而且医学文献中还存在全称和简称的情况,如白细胞介素-17 也称白细胞介素 17、白介素-17、白介素 17、白介-17、白介 17、IL-17、IL17,简称在信息检索过程中具有举足轻重的作用,也很容易造成漏检^[8]。如“Bcap31 基因的检测”,如果只使用检索内容的概念名称检索中国知网,Bcap31=0 篇。Bcap31 存在其他名称表述,如“B 细胞受体相关蛋白 31”、“BAP31”,再次检索可得:B 细胞受体相关蛋白 31=4 篇,BAP31=14 篇。由此可见,只是基于科研工作者提供的检索词进行检索,与经过同义词组配检索的结果相差甚远。因此必须要对检索概念进行全面详细的了解,全面考虑学科术语的不同表述,这样不仅可以提高检索结果的全面性,也可以在检索文献过程中拓宽科研工作者查找文献的思路,避免遗漏关键文献。

2.2 文献老化过程中部分学科术语改变

科学研究成果通常以科技文献形式存在,一般

来说最新发表的文献受到其所在学科专家、读者的关注和引用较多,具有较强的生命力,但其生命力会随着时间的推移而逐渐减弱。研究表明临床医学文献的半衰期为 4.515 年,肿瘤学各种期刊的平均引用半衰期为 2.9~5.14 (年)^[9-10],图书馆学、档案学、情报学的平均引用半衰期为 4.25 年^[11]。被引半衰期越长,该文献老化速度越慢,被利用时间越长,生命周期越长;反之,则生命周期越短^[12]。由此可知不同学科文献老化时间不同,其所在学科检索词汇的更新频率与变化程度也会不同,因此在检索文献时要充分考虑科技文献因不同学科文献老化情况而带来的变化。如一些专业学科术语在学科发展之初常用,但随着学科的发展该学科术语出现改变,这就使得原先的常用术语不再作为主流学科用语,如“PET 扫描”这一主题概念,1969-2004 年间,中国生物医学文献数据库(CBM)将其主题词标引为“体层摄影术,发射型计算机”,而 2004 年至今,其主题词标引为“正电子发射断层显像术”,因此在检索有关“PET 扫描”这一主题文献时,必须同时考虑这两个不同主题词表述,这样才能使检索结果较为全面,否则会造成漏检^[13]。

2.3 全文数据库分词技术的“瓶颈”

全文检索技术是当今信息检索发展的最高端、最前沿,发展至今在文本检索方面已比较成熟,用户只需以应用领域的概念即可检索出具有一定匹配度的信息,不仅改善了人机交互的效果,也提高了数据库的实用性。但由于汉语是以字为语言的基本构造单位,如何将汉语的基本语言单位转化为词,即实现无错误切分的分词系统,是目前中文信息处理技术的“瓶颈”问题^[14]。中文表述的复杂性使得数据库难以准确、彻底地对文献源的内容进行分词处理,同义词、上下位类词计算机也不能进行识别与匹配,如用“肾衰”找不到“肾功能衰竭”,用“马铃薯”检索不到“土豆”的相关内容,因此要想检索出相对全面的结果,就需要花费大量时间和精力去搜索检索词的不同表述,以使得检索概念在数据库中实现最大限度的匹配,进而使检索结果具有全面性。

3 检索词库设计与构建

3.1 设计

以教育部科技查新站（Y06）2014 年以来的科技查新报告为数据源，对这些数据源进行整理分析，将科技查新报告中的检索词抽取出来，将同一检索概念对应的主题词或关键词一一列出。使用 EPIDATA 录入软件，采用双盲双录入方法，将抽取完好的检索词进行电子化处理、统计与核查，以确保基础词库的准确性，便于数据的导出和使用。检索库的设计思路，见图 1。

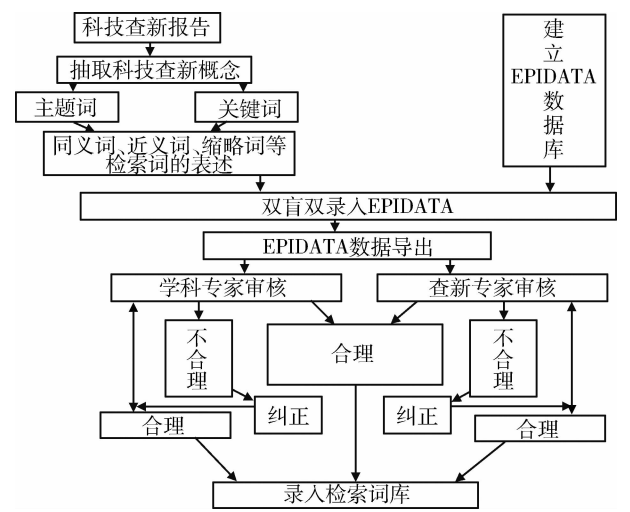


图 1 检索词库设计

3.2 构建

建立 EPIDATA 录入数据库，录入数据。对导出后的数据进行核查，邀请科技查新专家及学科专

家对归纳整理好的数据进行最终核验与指导，确定检索词的同义词、近义词及相关表述的合理性和全面性，其中，合理、全面的词汇直接进入词库进行保存，不合理的词汇进行修改和完善，再次核验成功后进入词库。最终，共收集到 2014 年以来的检索词条目 699 个，经筛选去重后共有检索词 618 个，其中少于 3 个检索词的检索概念有 207 个，拥有 3~5 个检索词的检索概念有 322 个，多于 5 个检索词的检索概念有 89 个，分别占总检索条目数的 33.5%、52.1% 和 14.4%，见表 1。由此可知，拥有 3 个及以上表述的检索词约占 66.5%，绝大多数检索词具有 3~5 个不同表述，如若不全面考虑检索词，会导致漏检。

表 1 检索词的构成比例

检索词（个）	数量（个）	百分比（%）
<3	207	33.5
3~5	322	52.1
>5	89	14.4

4 检索词库应用

4.1 概述

检索词库构建完成后，见表 2，一方面为广大教师及科研工作者在课题立项、定题跟踪时提供方便；另一方面应用于文献检索教学，为教师提供丰富的案例；此外，还可为新入职的查新员提供指导，为其在科技查新时提供参考依据，以防检索词考虑不全导致的漏检。

表 2 检索词库实体库

检索概念	检索词 1	检索词 2	检索词 3	检索词 4	检索词 5
宫颈癌	宫颈肿瘤	子宫颈鳞癌	子宫颈癌	-	-
肿瘤坏死因子 α	肿瘤坏死因子 - α	TNF α	TNF - α	-	-
循环 miRNA	循环 miRNAS	循环微小 RNA	循环微小 RNAS	-	-
益肺健脾	健脾益肺	补肺健脾	健脾补肺	培土生金	-
艾灸	艾炷	艾条	温和灸	雀啄灸	-
D-柠檬烯	苾烯	白千层烯	香芹烯	1, 8- 蒎二烯	-
血管内皮生长因子 - A	血管内皮生长因子 A	血管内皮细胞生长因子 - A	血管内皮细胞生长因子 A	VEGF - A	VEGFA
胃癌	胃肿瘤	胃腺癌	胃底癌	胃体癌	胃窦癌
白介素 - 6	白介素 6	白细胞介素 - 6	白细胞介素 6	IL - 6	IL6

4.2 为科研工作服务

学科服务是图书馆以用户为中心,学科馆员以其自身的检索技能深入院系,有针对性地利用图书馆文献资源为广大教师和科研工作者提供更全面、准确的信息源,为用户提供信息保障的同时,提升用户获取与分析利用信息的能力。目前,在开展学科服务的过程中,学科馆员根据课题组的研究方向和内容,基于检索词库中的词汇,快速检索科研工作者所需的文献,与科研工作者进行商讨,将其需要的相关文献搜索并整理出来,进而根据研究主题的文献发表情况,运用分析工具进行发文量、被引频次、引文影响力、高被引论文量、国家/地区、机构、人员等维度的学科竞争力分析,为学校的学科发展提供数据支持。“授人与鱼,不如授人以渔”,一方面,馆员在检索文献上有优势,但在筛选文献上却不如专业学科团队迅速;另一方面,学科团队对学科主题的常用表述较为熟悉,但对一些缩略词或者鲜为报道的学科词汇的掌握则不如学科馆员,所以结合二者的优势,就可以根据学科团队检索要求,学科馆员从检索词库中搜集整理相应的检索词提供给学科团队,学科团队再依据学科馆员提供的检索词到所需要的专业数据库中进行搜索、筛选文献。这样往往会快速全面、准确地搜索出学科团队所需要的文献,采集国内外选题相关成果一一进行分析,寻找热点,与科研工作者的研究方向进行关联度对比分析,从而加强高校图书馆利用数字资源进行学科服务的深度。

4.3 为教学服务

大学生在学习过程中,遇到自己感兴趣的研究内容或者想深入了解学科时,希望通过检索相关文献并进行阅读,以快速了解研究内容的来龙去脉。研究生在课题开题时,一部分导师会根据研究内容给学生们提供少量相关文献,而大部分则需要学生自己查找。这时就会有研究生遇到困难,要么是检索不到相关文献,要么是检索到的相关文献数量非常少,不能满足需要。研究生在写作综述或者进行 Meta 分析时,有时也会感到手足无措,这时通常需要与自己的教师进行沟通,以寻求帮助和指导。面对学生们不同专业的要求,教师就要对其需要的文献进行深入了解并通过实际检索才能给予准确、全

面的指导,如果借助检索词库,就可以方便地为学生们提供快速、准确的指导,节约时间,提高效率。此外,每年医学文献检索教师需要花费大量的时间为研究生考试出题,而查新课题是科研人员的主要研究内容,有些研究可代表其所在学科的新动向,这部分研究内容如若作为研究生们的考题,则有助于其了解本学科的研究动向,进而带动学生的兴趣,调动学习积极性。但是较新的研究内容,教师们在使用过程中要注意保密原则。

4.4 应用于科技查新

只有单一学科背景的科技查新人员,如果想要准确、完整地处理好已受理的科技查新委托,就必须与课题委托人进行多次沟通以了解课题,将课题的检索概念所对应的检索词给出全面准确的表述,以确保科技查新结果的客观性和全面性,这就需要查新人员花费较多的时间和精力。此外,不同检索经验及学科背景的查新人员,遇到相同的检索概念时,会使得检索结果有所不同,即便是同一查新人员不同时间对同一检索词不同表述的选择和检索,依然需要花费大量的时间和精力。此外,科技查新检索中常用检索词主要有主题词和关键词。主题词可以从主题词库中选取,而关键词的获取比较困难。鉴于上述原因,要在信息检索过程中全面检索文献,就必须考虑各种关键词的表述,这也是科技查新过程中的主要难点。目前获取检索词的途径主要有:(1) 委托人提供。(2) 根据词表查找。(3) 根据数据库查找同义词的功能检索。(4) 根据数据库预检文献,阅读并选取关键词。这些途径除了第一种较为方便省时外,其他途径都颇为费时费力,这也正是困扰科技查新人员的主要难题所在。而经科技查新专家和学科专家共同审核校准后建构的检索词库,可以为查新人员提供准确合理的关键词。

5 结语

针对学术用语的多样性和标引不规范的情况,需要多方面的协调与统一,从检索源头进行改善。科技工作者在撰写学术论文时,应尽可能地以公认的关键词表述,不用英文缩写和一些商品名。期刊处理过程中,应尽可能地与国家标准统一,尤其是针对药品名的标引,各种期刊对同一概念的标

引尽可能有一个合理的规范,确保同一概念的表述相同。除此之外,检索词库的使用尽可能地保证检索词的全面,进而使得科技查新的结果全面、客观。在应用检索词库进行信息检索时,除选择和使用检索词库中的主题词与同义词、缩写词、简称外,还要进一步考虑不同搜索引擎中的常用符号及特点,应用检索技巧,以提高全文检索系统或因特网搜索中文信息的查全率和查准率。

参考文献

- 1 Hilbert M, López P. The World's Technological Capacity to Store, Communicate, and Compute Information [J]. Science, 2011, 332 (6025): 60-65.
- 2 李慧美. 开阔国际视野,点亮学术人生——借助 CCSI-A&SHCI 进行人文、社会科学研究 [EB/OL]. [2016-10-31]. <http://www.docin.com/p-976524759-f3.html>.
- 3 林静,伊雷,陈珊珊,等. 大数据时代高校图书馆开展学科服务研究——学科馆员工作案例解析 [J]. 现代情报, 2015, 35 (12): 65-69.
- 4 孙淑萍,孙玉坤. 主题词、同义词、缩写词、简称对科技查新影响的实例分析 [J]. 情报探索, 2013, (8): 85-87.

- 5 于曦. 面向嵌入式服务的科技查新服务体系优化 [J]. 图书馆工作与研究, 2016, (5): 96-100.
- 6 张柏秋,吴晓镔. 科技查新检索中的关键词选择 [J]. 情报科学, 2008, 26 (9): 1344-1348.
- 7 张岚,张柏秋,于非. 科技查新检索质量优化策略研究 [J]. 情报科学, 2011, 29 (6): 852-855.
- 8 朱康玲. 同义词的获取对医学科技查新查全率和查准率的影响 [J]. 中华医学图书情报杂志, 2012, 21 (3): 78-80.
- 9 宋玉,王筠,曲磊. 临床医学期刊论文的发表时滞与老化分析 [J]. 情报探索, 2014, (4): 15-17.
- 10 杭文文,谢靖. 我国肿瘤学期刊引用半衰期研究 [J]. 江苏科技信息, 2015, (26): 19-21.
- 11 刘伙玉. 基于 CNKI 的图书、情报学与档案学学科文献半衰期分析 [J]. 图书与情报, 2015, (1): 106-111.
- 12 罗式胜. 文献半衰期的类型及其应用 [J]. 情报学报, 1997, 16 (1): 62-67.
- 13 马磊,宋建玮. IPC 分类法在科技查新工作中的应用 [J]. 图书馆学刊, 2012, (3): 32-34.
- 14 王振风. 基于 Lucene 的分布式全文检索技术的研究与应用 [D]. 上海: 东华大学, 2015.

(上接第 64 页)

中医药治疗结直肠癌临床疗效的前提和基础。但目前临床研究对中医证型的分析缺乏多中心、大样本的流行病学研究,为辨证的客观化、标准化带来了一定的困难,从而制约了中国医临床研究水平的进一步提高。因此在今后的研究中应遵循循证医学的原则,本着科学严谨的态度,对大肠癌的症、证进行调研,开展多中心、随机对照试验,以揭示大肠癌的中医辨证论治规律,规范大肠癌中医证型。

参考文献

- 1 万德森. 结直肠癌的流行病学及其危险因素研究近况 [J]. 实用癌症杂志, 2000, 15 (2): 220-222.
- 2 陈叶,刘金涛,朱源,等. 大肠癌中医辨证及治疗概况 [J]. 中国肿瘤, 2015, 24 (4): 319-324.
- 3 许云,杨宇飞. 结直肠癌中医药研究进展与思考 [J]. 世界中医药, 2014, 9 (7): 828-832.
- 4 姚乃礼. 中医症状鉴别诊断学. 2 版 [M]. 北京: 人

民卫生出版社, 2000: 27-35.

- 5 王颖纯,白丽娜. 基于 BP 神经网络的中医脉诊体质类型判定 [J]. 中医杂志, 2014, 55 (15): 1288-1291.
- 6 连俊彦. 多因素回归生存分析探讨影响大肠癌病人术后生存期的因素 [J]. 实用医学杂志, 2002, 8 (6): 597-598.
- 7 白云静,孟庆刚,申洪波,等. 基于改进的 BP 神经网络的糖尿病肾病中医证候非线性建模研究 [J]. 北京中医药大学学报, 2008, 31 (5): 309-311.
- 8 Dayhof JE, Deleo JM. Artificial Neural Networks [J]. Cancer, 2001, 91 (8): 1615-1634.
- 9 Cross SS, Harrison RF, Kennedy RL. Introduction to Neural Networks [J]. Lancet, 1995, 346 (8982): 1075, 1079.
- 10 李建生,王至婉,余学庆,等. 基于慢性阻塞性肺病 (COPD) 急性加重期文献的多元统计方法在证候研究的应用探讨 [J]. 中医学报, 2007, 22 (6): 8-10.
- 11 孙贵香,袁肇凯. 人工神经网络在中医证候研究中的应用 [J]. 中华中医药学刊, 2007, 25 (7): 1450-1452.