

MeSH在医学信息检索中的应用研究

高金玉¹, 朱小梅²

(1. 中国核科技信息与经济研究院, 北京 100048; 2. 中国人民大学信息资源管理学院, 北京 100872)

摘要: 分析MeSH在PubMed、中国生物医学文献数据库、中国医院知识资源总库和万方医学网中的应用, 总结出MeSH在医学信息检索中的三种应用机制, 即直接使用主题词检索、实现自然语言或分类号向主题词的转换以及利用主题词等级范畴表进行知识导航, 进而提出加强中文主题词表和中文一体化医学语言系统的构建与研究的建议。

关键词: 主题词表; MeSH; 检索; 应用

中图分类号: G254.2

DOI: 10.3772/j.issn.1673—2286.2014.10.006

1 MeSH概述

21世纪生物医学和信息技术的快速发展使得医学信息资源类型多样, 数量庞杂, 更新迅速。医学文献数量占科技文献总量的20%至25%, 居学科文献之首。和其他学科文献相比, 医学文献的寿命最短, 半衰期仅为3.5年^[1]。因此, 医学领域知识需要有高效的工具来组织和获取。医学主题词表(Medical Subject Headings, MeSH)正是当今世界上使用最广泛、最权威的医学信息组织工具, 在标引、编目、生物医学文献数据库和相关网络检索系统中得到了广泛应用。

MeSH由美国国立医学图书馆(National Library of Medicine, NLM)编制, 是一部规范化、可扩充、动态性的叙词表, 具有收词丰富、注释详尽、准确性和专指度高等优点。2013版MeSH收主题词达到26 853个, 入口词超过213 000个^[2]。NLM每年都对MeSH进行版本升级和优化改进, 如对主题词做适当的增删, 力求所收的主题词既有灵活性又有稳定性。

目前关于MeSH的研究成果颇丰, 通过文献分析可知, 主要有两个方向: 一是对MeSH的微观和宏观结构

进行研究, 主要侧重词表的编制, 包括MeSH的介绍和评述^[3]、MeSH的修订和更新^[4]、MeSH主题词的语义关系和语义相似度^[5,6]、MeSH专指度研究^[7]; 二是MeSH的应用研究, 将MeSH置于具体的应用环境和条件之中, 考虑用户需求和应用的方式方法, 主要包括MeSH对其他词表的借鉴作用^[8]、MeSH在PubMed中的应用^[9,10]。后者的研究很多, 但MeSH在索引、编目和其他数据库中的应用研究不多^[11-13], 未能形成体系。因此, 本文以MeSH在数据库中的应用为切入点, 探究MeSH词表在国内外数据库中检索的应用方式, 丰富MeSH应用研究, 以达到提高医学文献检索水平的目的。

2 MeSH在医学信息检索中的具体应用

NLM以MeSH词表作为生物医学文献标引的依据, 编制《医学索引》(Index Medicus)及建立Medline数据库。许多国家都将MeSH翻译为本国语言版本进行实际应用。中国医学科学院医学信息研究所对MeSHAAL(Medical Subject Headings-Annotated Alphabetic List, 主题字顺表)不定期地进行翻译, 出版了1979、1984、1992版的《英汉对照医学主题词注释

字顺表》。我国医学图书情报单位利用和借鉴MeSH词表编制了《军用医学主题词表》、《中文医学主题词表》和《中国中医药学主题词表》等。中国生物医学文献数据库、万方医学网和中国医院知识资源总库等数字医学信息平台均参考使用了MeSH词表对生物医学文献进行处理。

2.1 MeSH在PubMed中的应用

PubMed提供生物医学方面的论文检索以及摘要，数据库来源为Medline。PubMed建立了多种适应不同用户需求的检索方式，其中通过MeSH进行主题词检索，实现方式有自动词语转换和MeSH Database辅助检索。

(1) 自动词语转换

在PubMed主页的检索框内直接输入关键词，系统会将该词自动转换成主题词进行检索，即自动词语转换，具体机制是将检索词按以下列表顺序进行匹配：

MeSH转换表 (MeSH Translation Table)

MeSH主题词

副主题词

出版类型

入口词 (同义词)

来自UMLS (Unified Medical Language System) 的映射

补充概念名称表及补充概念同义词

刊名转换表 (Journals Translation Table)

作者和研究者名字表 (Author and Investigator names)

作者姓名全称转换表

作者索引

研究者姓名全称转换表

研究者索引

只要在表中找到匹配词后，就会停止转换。即PubMed在MeSH转换表中找到匹配的主题词，就不再在刊名转换表、作者和研究者名字表中进行查找。如果仍然找不到匹配词，PubMed就会把该词语断开后重复上述自动词汇转换过程，直到找到与键入的词语相匹配词语为止。若仍然没有匹配词，单个词会被联在一起 (用AND) 在全部字段中检索。如：“single cell”，系统就会自动将其分成两个词：“single”和“cell”检索，其检索表达式为：“single AND cell”。在图1中的“Translations”中可见

“cell”和“single”两个词的匹配过程，最后形成的检索式在“Query Translation”中列出。



图1 “single cell” 转换匹配运算图^[15]

(2) MeSH Database辅助检索

PubMed提供了MeSH Database进行辅助检索，其功能是确定主题词、副主题词；查看词义注释；根据树状结构表需要选择上位或下位主题词进行缩检或扩检。点击进入MeSH Database，在检索框内输入主题词，检索结果会自动匹配出主题词相关文献；如果输入关键词则提供同义或近义MeSH词相关文献给予选择。选择需要的文献，可显示详细结果，进一步修饰检索词包括：选择副主题词、主要主题词限定 (Restrict Search to Major Topic headings only)、限制下位词扩检 (Do Not Explode this term)、观看其他入口词等其他修饰检索、浏览树状结构表^[16]。

2.2 MeSH在中国生物医学文献数据库中的应用

中国生物医学文献数据库 (China Biology Medicine, CBM) 收录1978以来1800余种中国生物医学期刊以及汇编、会议论文的文献题录800余万篇^[17]。

中国医学科学院医学信息研究所学习和借鉴美国UMLS开发出了中文一体化医学语言系统 (China Unified Medical Language system, CUMLS)。CUMLS构建了面向计算机应用的，集自然语言、MeSH词表及《中国中医药学主题词表》等主题语言、《中国图书馆·分类法医学专业分类表》分类语言于一体的知识组织体系，相继实现了医学文本主题、分类自动标注、主题导航、智能检索等知识服务^[18]。CBM基于

CUMLS提供了主题词检索功能,具体如下:①可用中文主题词、英文主题词及同义词查找对应主题词,可浏览主题词注释信息和树形结构,帮助确定恰当的主题词。②通过选择合适的副主题词、设置是否加权、是否扩展,可使检索结果更符合用户需求。③支持多个主题词的同时检索,可使用逻辑运算符“AND”、“OR”和“NOT”进行组配^[19]。

2.3 MeSH在中国医院知识资源总库的应用

中国医院知识资源总库(CHKD)是中国知网CNKI旗下的生物医学知识资源库。CHKD将MeSH与《中国中医药学主题词表》、《中国图书馆分类法·医学专业分类表》结合,研发出主题词、分类号智能检索系统医学专业检索系统(Medical Subject Headings of Classification Code Intelligent Retrieval System, MCI)。MCI采用统一标准主题词或分类号表达,对10类医学文献进行统一的主题词标引,配备知识导航(主题词导航)、分类导航(中图法导航)两个导航系统,实现了统一的跨库主题词/分类检索和主题词/分类号智能转换,并且支持同范畴主题词、跨范畴主题词组配、中英对照主题词检索,使检索规范、标准和高效^[20]。MCI保持与MeSH同步更新,2013版主题词总数增至33 004条,词汇总量达45万余条。

(1) 跨库主题词检索

①支持中、英文两个语种主题词的输入检索。②支持分类检索。③支持中英文对照主题词/分类号的智能转换。当所输入检索词无法实现主题词的自动转换时,系统自动以“题名&关键词&摘要”进行检索且输出结果,同时在结果上方显示所用检索词提示。④利用主题词树形结构进行检索范畴的自选及限定(同范畴和跨范畴均可),实现同范畴和跨范畴主题词的组配。⑤基于知识导航进行主题词的自选及限定,实现与非主题词的组配。⑥基于主题词智能转换功能的外文数据库跨库检索。在主题词项输入检索词转换出主题词时,提供该主题词在PubMed、F1000、HighWire和BioMed Central四大外文数据库中检索结果的输出^[21]。

(2) 中英对照主题词注释库

CHKD整合了中英对照的主题词注释库。进入主题词注释库(中英对照)主页面,输入主题词,检索后显示含有该词素的中文主题词及其对应英文主题词,并在其右侧显示相应主题词在期刊、博硕士论文、会议

论文、报纸、年鉴等数据库中的文献检索数量,同时显示几个数据库的文献总量^[22]。

2.4 MeSH在万方医学网的应用

万方医学网利用2011版MeSH词表,在人工翻译后,对收录的医学期刊论文、学位论文、会议论文和外文期刊论文进行标引,从而实现MeSH主题词检索功能。万方医学网为用户提供主题词导航,帮助用户确定主题词,并利用主题词进行相关文献的检索。①主页面的一框式检索中包含对主题词字段的检索。输入任意检索词,会给出相关的主题词以及该检索结果相应的分类聚类分布,进而可以发现不同聚类下的检索结果。②设置MeSH主题词检索入口^[23]。③MeSH主题词导航功能。该功能需要登录后方可使用,可以引导用户使用合适的主题词进行文献检索^[24]。

3 MeSH在医学信息检索中的应用分析和建议

3.1 应用分析

对本文四个案例分析可知,不同的医学信息系统检索方式和检索策略虽然各异,但仍有共同的机制。MeSH的检索应用主要有以下三种机制:

(1) 直接主题词检索。获得主题词的捷径是借助辅助手段,如PubMed的MeSH Browser、CBM的“主题导航”、万方医学网的“MeSH检索”和CHKD的“主题词注释库”。直接使用主题词检索适用于图书或医学情报专业人员,或者掌握主题词受控语言的检索技巧的普通用户。

(2) 实现自然语言或分类号向主题词的转换。MeSH对文献中的同义词、近义词、多义词等加以严格的控制和规范,使得同一主题概念的文献相对集中在一个主题词下,提高了文献检索中的查准率和查全率,但由于自然语言或关键词检索简单方便,医学网站普通用户都喜欢自然语言的检索方式,促使医学信息网站提供自然语言向主题词的转换。同时,分类号向主题词的转换是在主题分类一体化的基础上发展起来的,两者的结合给用户提供了多样化的检索途径。

(3) 利用主题词等级范畴表进行知识导航和知识发现。如,在PubMed中使用MeSH树形结构实现扩检和缩检,万方医学网借助“分类限定”和CHKD的“知

识导航”。第2和第3种检索机制适用于缺乏检索技巧和医学专业知识的终端用户。

3.2 MeSH在我国医学信息检索中的应用建议

以CBM、CHKD以及万方医学网为代表的医学信息资源平台利用了MeSH,充分发挥了MeSH的作用,提高了医学信息检索的效率,但MeSH在我国的应用仍有很大的空间,笔者根据研究提出以下建议:

(1) 加强中文主题词表的构建和应用

中国医学科学院医学信息研究所编制的CMeSH是CBM使用的中文医学主题词表,2012版涵盖MeSH(2012中文版)、《中国中医药学主题词表》及《主题词—医学专业分类映射表》,是国内医学领域权威的主题分类一体化词表^[25]。CHKD根据自己的需要编制出了主题词、分类号智能检索系统医学专业检索系统MCI,实现了分类主题一体化。这些医学信息平台分别对MeSH进行了研究和加工,把握了主题分类一体化的发展趋势,但是由于各自为政,研究深度相对不足,导致用于标引、编目或检索的准确性不足。因此,需要协同开发和利用现有词表,及时更新词表,推进中文主题词表的构建和应用。

(2) 加强中文一体化医学语言系统的构建与研究

PubMed具有强大的检索功能,MeSH的作用至关重要。除了NLM提供MeSH Browser、MeSH Database外,MeSH转换表是PubMed医学信息检索查全查准的利器,这得益于UMLS的构建和使用。MeSH具有概念和上下级、相关和等同等简单语义关系,不能机器理解,因此NLM构建了语义较丰富的UMLS,促进了自然语言向主题词的转换。UMLS由超级叙词表、语义网络、情报源图谱和专家词典组成,是人工语言和自然语言结合,概念、语义和语用等相结合的系统。超级叙词表是生物医学概念、术语、词汇及其等级范畴的广泛集成,其中MeSH是重要词汇来源。截至2014年10月,超级叙词表拥有超过150种词表源的290万个概念和1160万个专有概念名称,为自然语言和主题语言的映射和转换提供了词汇基础^[26]。中国借鉴了UMLS的经验开发出了CUMLS,CUMLS主要由医学词表和语义网构成。但是CUMLS收词量不够完整,不能进行语义类型和语义关系的自动识别,没有建立概念与语义类型、概念与概念之间语义关系的完整映射,不能及时筛选添加新词汇,词表更新存在时滞等等^[27]。这些都

需要我们研究MeSH自动词语转换,进一步学习和借鉴UMLS,加强CUMLS的研究,完善CUMLS的编制和更新,促进CUMLS的应用。

参考文献

- [1] HOLLEY E G, SCHREMSER R F. The Library Services and Construction Act: An historical overview from the viewpoint of major participants [M]. Greenwich, CT: Jai Press, 1983.
- [2] NLM. Medical Subject Headings (MeSH®) [EB/OL]. [2014-10-06]. <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>.
- [3] 严青利,张勇.医学主题词表(MeSH)评述[J].情报杂志,2001(8):64-66.
- [4] 李军莲.MeSH词表的新变化及有关标引规则[J].医学信息杂志,2007(3):285-288.
- [5] 张义煌.医学主题词的语义关系特点及其在情报检索中的应用[J].现代预防医学,2007(1):160-161.
- [6] 孙海霞,钱庆,吴英杰,等.MeSH词表的语义相似度计算研究[J].现代图书情报技术,2010(6):12-16.
- [7] 于双成,逢大欣,李占兵.医学主题词表(MeSH)专指度研究[J].情报学报,1995(6):449-452.
- [8] 范为宇,苏大明,胡艳敏,等.有关中英文版中医药学主题词表的研究[J].医学信息杂志,2007(4):411-413.
- [9] 闫微,张亚莉,邓小茹.基于PubMed数据源的生物医学文献检索工具应用特点[J].医学信息杂志,2012(1):50-53.
- [10] 王敏,许培扬.Pub Med智能检索的实现[J].医学情报工作,2005(6):431-433.
- [11] 张琳,徐平平,章锦才,等.《中国医院知识仓库》主题检索技巧及其文献主题词标引的几点建议[J].医学信息杂志,2011(7):87-90.
- [12] 孙海霞,李军莲,李丹亚,等.基于CMeSH语义系统的领域自由词—主题词语义映射研究[J].现代图书情报技术,2013(11):46-51.
- [13] 李军莲,李丹亚,阮学平,等.医学分类主题一体化系统建设[J].医学信息杂志,2011(3):69-73.
- [14] Single cell [EB/OL]. [2014-10-06]. <http://www.ncbi.nlm.nih.gov/pubmed/details?querykey=1>.
- [15] Automatic Term Mapping [EB/OL]. [2014-10-06]. http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_040.html.
- [16] MeSH Database [EB/OL]. [2014-10-06]. <http://www.ncbi.nlm.nih.gov/mesh/>.
- [17] 中国生物医学文献数据库简介[EB/OL]. [2014-10-06]. <http://www.sinomed.ac.cn/>.
- [18] 李丹亚,胡铁军,李军莲,等.中文一体化医学语言系统的构建与应用[J].情报杂志,2011(2):147-151.

- [19] 中国生物医学文献数据库主题检索[EB/OL]. [2014-10-06]. <http://www.sinomed.ac.cn/zh/subjectSearch.html>.
- [20] 《中国医院知识资源总库(CHKD)》新版简明使用手册(2013) [EB/OL]. [2014-10-06]. <http://www.chkdcnki.net/ZSCKsc/all.htm>.
- [21] 中国医院知识资源总库[EB/OL]. [2014-10-06]. http://mci.chkdcnki.net/kns55/brief/single_default.aspx?dbPrefix=chkd.
- [22] 中国医院知识资源总库主题词注释库[EB/OL]. [2014-10-06]. http://kns.chkdcnki.net/kns55/MainWordDict/dict_list.aspx.
- [23] 万方医学网论文检索[EB/OL]. [2014-10-06]. <http://med.wanfangdata.com.cn/Paper/Index>.
- [24] 万方医学网MeSH检索[EB/OL]. [2014-10-06]. <http://old.med.wanfangdata.com.cn/Mesh/Mesh.aspx>.
- [25] CMeSH [EB/OL]. [2014-10-06]. <http://cmesh.imicams.ac.cn>.
- [26] What's new [EB/OL]. [2014-10-06]. http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html.
- [27] 李丹亚, 胡铁军, 李军莲, 等. 中文一体化医学语言系统的构建与应用[J]. 情报杂志, 2011(2):147-151.

作者简介

高金玉, 女, 硕士, 研究方向: 信息资源管理。

朱小梅, 女, 博士生, 研究方向: 信息组织。

Application Research on MeSH on Medical Information Retrieval

GAO JinYu¹, ZHU XiaoMei²

(1. China Institute of Nuclear Information & Economics, Beijing 100048, China;

2. School of Information Resource Management of Renmin University, Beijing 100872, China)

Abstract: The paper studies on the application of Medical Subject Headings (MeSH) on Medical information retrieval. The different ways of applications of MeSH are on Pubmed, CBM, CHKD, Wanfang Med Online. Three mechanisms were summarized: Using the subject headings to search, translating the natural words or classification codes into subject headings, navigating by MeSH Categories and Subcategories. Advice is proposed to promote the construction and research of MeSH and China Unified Medical Language system.

Keywords: Thesaurus; MeSH; Retrieval; Application

(收稿日期: 2014-09-22)

NSTL全国开通电子资源介绍

NSTL以“国家许可”的方式引进数字资源, 最大程度地实现了规模效益, 节约国家在文献资源经费的整体投入, 重点弥补国内文献资源的历史性缺失。

自2002年引进第一个全国开通的现刊数据库MANEY开始, 截至目前, NSTL以“国家许可”的方式共引进了40个现刊数据库及3个事实性数据库, 2014年开通期刊数量达到655种。重点保障国际知名中小学协会的期刊, 如美国植物生物学家学会、北美放射学会、美国营养学会、美国遗传学学会、美国植物学会、美国冷泉港实验室、美国医师协会、美国胸科医师学会和美国神经放射学会等期刊均极大满足了国内科研人员的需求。

自2008年至今共引进回溯期刊数据库16个, 包括Springer、OUP、IOP、RSC、LWW、BMJ和OUP等国际知名出版社、学协会, 涵盖期刊品种数量达到1500余种, 约600万篇期刊全文回溯内容。

中国大陆非盈利学术机构通过向NSTL提交开通申请(申请流程见<http://www.nstl.gov.cn/全文文献栏目>), 均可以免费使用以上资源。机构用户需通过出版社平台访问现刊数据库内容, NSTL平台只提供链接及跨库检索服务。机构用户均可通过NSTL自建的回溯平台(<http://archive.nstl.gov.cn/Archives/>)访问所有已购买的回溯资源。