中文生物医学关键词-主题词映射表计算机辅助构建与维护机制研究*

孙海霞[1,2]，吴英杰[1]，李丹亚[1]，李军莲[1]
（1. 中国医学科学院医学信息研究所，北京 100020；2. 南京大学信息管理学院，南京 210093）

摘要：将自然语言应用到信息组织、标引、检索和分析所需的各种词表中，实现自然语言与受控语言间的互操作，是新一代知识组织系统构建模式。文章围绕“关键词识别与扩充、关键词-主题词映射关系构建、关键词-主题词映射关系更新”3个方面，介绍了中文生物医学关键词-主题词映射表的计算机辅助构建与维护机制，以及系统的功能架构。

关键词：生物医学；关键词-主题词映射；知识组织系统；词表更新；计算机辅助
中图分类号：G350
DOI：10.3772/j.issn.1673—2286.2014.12.003

自然语言检索已成为信息检索的必然趋势之一，但长期以来，为实现信息的有效组织与利用，在信息检索领域，广泛用于文献资源的标引、组织和检索的是受控语言，如主题词表、叙词表、分类表等。在此背景下，无论是在情报检索语言理论研究，还是实际知识组织系统的编制研究，学者们开始探索新一代的知识组织系统构建模式[1-3]，主张人工语言与自然语言相结合，将自然语言应用到情报检索所需的各种词表中，实现自然语言与受控语言间的互操作。如南京农业大学侯汉清及其博士生、硕士生先后持续开展的自然语言叙词表构建、情报检索语言的兼容转换、面向信息检索的汉语同义词自动识别等系列研究[4-8]。

在生物医学信息领域，为实现领域自然语言与规范语言的结合，国内外医学信息工作者们开展了系列研究。在国外，美国国立医学图书馆一直走在前列，他们先后开发出了一些映射自由文本到领域权威词表MeSH（Medical Subject Heading）或UMLS（Unified Medical Language System）的自由文本与概念映射工具，如MicroMeSH、Chartline、Clarit、Saphire及MetaMap，并且这些工具已在词表的自动更新与维护、信息组织、处理与利用等各个领域实践中得到不同程度应用[9-15]。在国内，中国医学科学院医学信息研究所也已在九十年代开始基于中文医学主题词表CMeSH（Chinese Medical Subject Heading）[16]编制中文生物医学关键词—主题词映射表，并在中国生物医学文献“自然语言—主题语言—分类语言”一体化计算机辅助标引系统中得以应用，很大程度上提升了中文生物医学文献组织、标引和检索系统性能[17]。

但随着中文生物医学文献的飞速增长、新领域的不断兴起与发展，中文生物医学关键词—主题词映射表因编制效率低，开始面临内容更新缓慢、无法及时揭示新兴研究成果和满足中文生物医学文献组织、自动标引需求的挑战，进而在一定程度影响了中文医学文献检索系统的性能。就其问题和原因所在，可主要归纳为两点：（1）映射表中关键词文献覆盖率不高；（2）关键词和主题词映射关系的建立还主要靠人工进行，缺乏

* 本研究得到中国医学科学院医学信息研究所基本科研业务专项“中国生物医学文献服务系统发展关键问题研究”（编号：13R0103）资助。

计算辅助映射支持。

1 CMeSH和中文生物医学关键词-主题词映射表

1.1 《中文医学主题词表》CMeSH

《中文医学主题词表》CMeSH[16]由中国医学科学院医学信息研究所编制，是国内外第一部在生物医学领域广泛使用的权威的中文专业叙词表。与所有受控叙词表一样，CMeSH中一个概念只能用一个语词来表达，一个语词只能表达一个概念。若一个概念有多种表达形式，则只选其中一个词作为主题词，其他的只作为该词的入口词。CMeSH对其收录的所有主题词不仅进行了严格的词义规范、词类规范和词形规范，还通过清晰的树状结构及简明的参照系统来揭示主题词之间的属分、用代、相关、也须等语义关系，以保证作者、标引者和用户之间用词的一致性。

1.2 中文生物医学关键词-主题词映射表

中文生物医学关键词—主题词映射表（下简称映射表）源于CMeSH，所有主题词与CMeSH主题词保持一致，但如表1所示，它不是一个叙词表，也不是一个严格的同义词映射表，其关键词来源包括但不限于CMeSH入口词。其构建目标包含两个方面，一是要打破CMeSH“入口词规范”和“入口率适中”体制限制，提高关键词的文献覆盖率；二是为CMeSH的词汇系统和语义关系系统的更新提供基础。

表1 中文生物医学关键词-主题词映射表主要内容片段

序号	内容项	注释
(1)	主题词	同版本的CMeSH主题词
(2)	关键词	包括主题词的同义词、近义词、缩写、不同拼写形式及其他代用形式等用户较熟悉的形式
(3)	关键词来源	CMeSH入口词、CBM文献数据库、外部知识组织系统
(4)	关键词映射权重	关键词与主题词的语义相似度。CMeSH入口词和经人工确认的为1。其他为计算机计算值，取值[0-1]。
(5)	词频	术语（词汇）在CBM文献数据库中的词频

2 映射表计算机辅助构建与维护技术路线

如图1所示，映射表的计算机辅助构建与维护整体可分为三个步骤：词源扩充、映射关系自动构建与更新、人工审核。

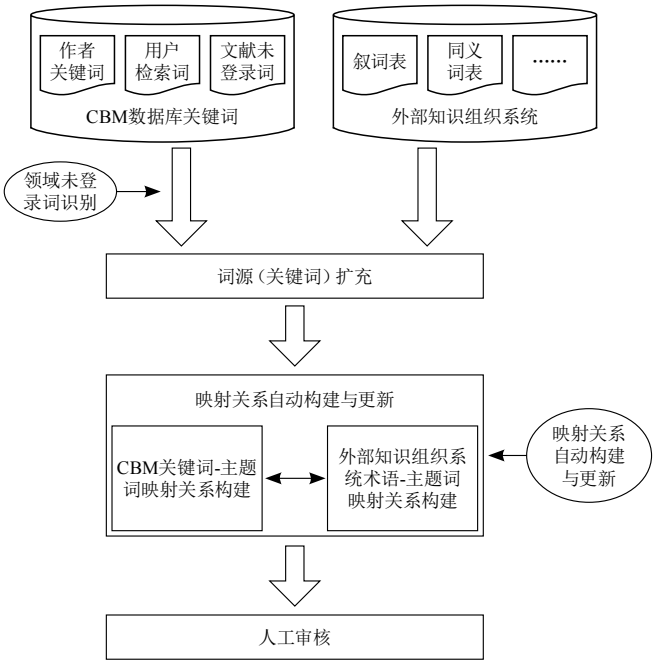

图1 映射表计算机辅助维护技术路线

2.1 词源扩充

主要指关键词的扩充，目标是提高映射表中关键词的文献覆盖率。就其实际应用角度而言，其来源可分为两大部分：中国生物医学文献数据库（Chinese Biomedical Literature Database，简称CBM）内部资源和外部资源。CBM内部资源包括CBM收录文献的作者关键词、文献中的未登录词和检索系统长期积累下的用户检索词。外部资源主要包括现有的各种用于文献处理的知识组织系统，如专业同义词典、叙词表或主题词表中的入口词，等等。

2.2 映射关系自动构建与更新

即在新扩充进来的关键词和主题词间建立同义或准同义映射关系，扩充现有主题词的入口词，包括来源于CBM文献的未登录词和外部知识组织系统的已有专业术语。当CMeSH主题词发生变更时，能够及时更新

相关关键词与主题词之间的映射关系。前者具体实现机制详见3.2和3.3；后者具体实现机制详见3.4。

2.3 人工审核

包括两方面的审核：词源审核和映射关系审核。词源审核用以判定计算机扩充进来的关键词是否适合作为文献处理。映射关系审核用以判定计算机自动构建的同义或准同义映射关系的正确性，并进行相应的调整。

3 关键问题分析与解决

上述技术路线的实现需解决4个难题：1）如何从CBM文献中获取有效未登录词；2）关键词与主题词间的映射关系的正确建立；3）如何利用外部知识组织系统中的专业术语；4）如何根据主题词的变化及时更新现有映射关系。

3.1 如何从CBM文献中获取未登录词

本文中未登录词主要指当前所用映射表中尚未出现的词。笔者在文献18中，结合中文生物医学领域词长分布和构词特点，提出以n-gram为基础，综合利用领域词典、语料和规则的中文生物医学领域未登录词识别方案，如图1所示，并以中文生物医学文献数据库CBM中药学期刊文献的题名、摘要作为样本集进行了实验，效果表现良好[18]。

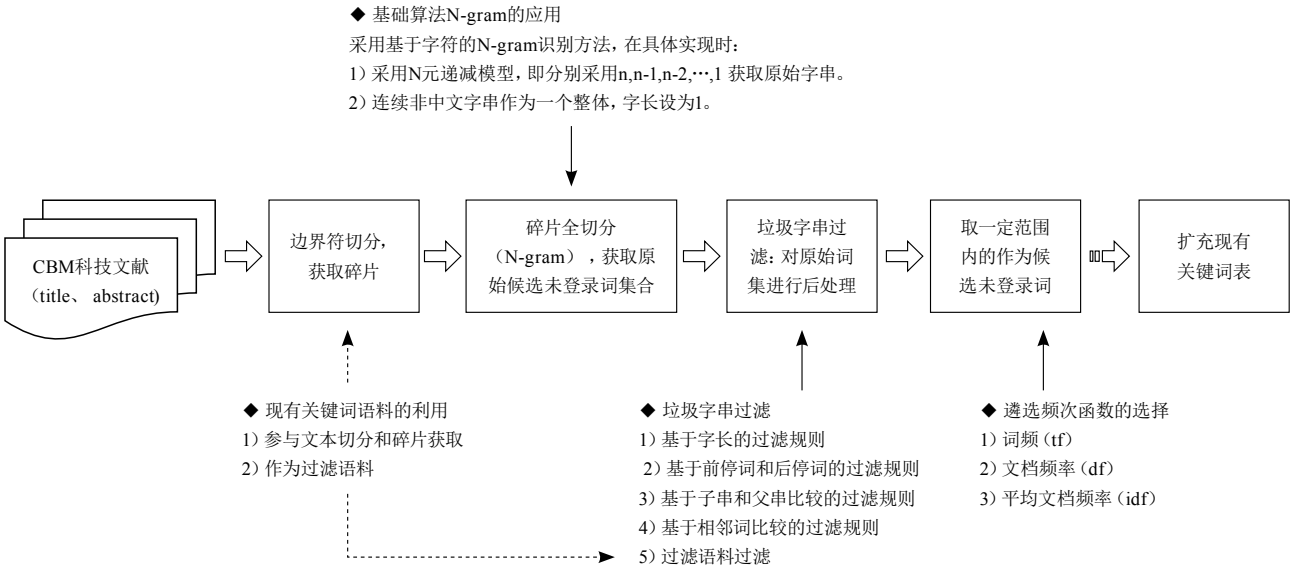

图2 基于中文生物医学期刊数据的领域未登录词识别技术方案

3.2 如何自动建立CBM关键词与主题词间的映射关系

从信息检索角度来看，关键词与主题词间的对应关系是表达同一概念的自然语言词与人工语言词间的一种等同映射关系[2]，常用的自动识别方法有基于相似度计算的方法、基于词汇共现的方法、基于规则的方法和基于链接分析的方法等[4-8]。笔者在文献19中从信息检索和文献标引角度出发，立足CMeSH体系结构，基于“在后控词表中，相邻等级的主题词中，字长较长主题词表达的概念更为专指”假设，综合考虑词汇集合相似度计算和后续过滤规则设计，提出了医学领域自由词-主题词语义映射方案和基于CMeSH的关键词-主题词语义相似度计算模型[19]。

公式1是本映射表构建与维护系统采用的CBM关键词与CMeSH主题词语义相似度计算模型。

$$Sim\_semantic(A, b)=Max\{(Sim(A, b)), Max(Sim(Aj, b))\} \quad (公式1)$$

其中A表示CMeSH中的主题词，Aj表示主题词A对应的入口词，b表示自由词。Sim(A，b)表示主题词A

本身和自由词b的字面相似度，Sim(Aj，b)表示主题词A的入口词与b的字面相似度，Max (Sim(Aj，b))表示在A的所有入口词与b的相似度中，取最大相似度值。

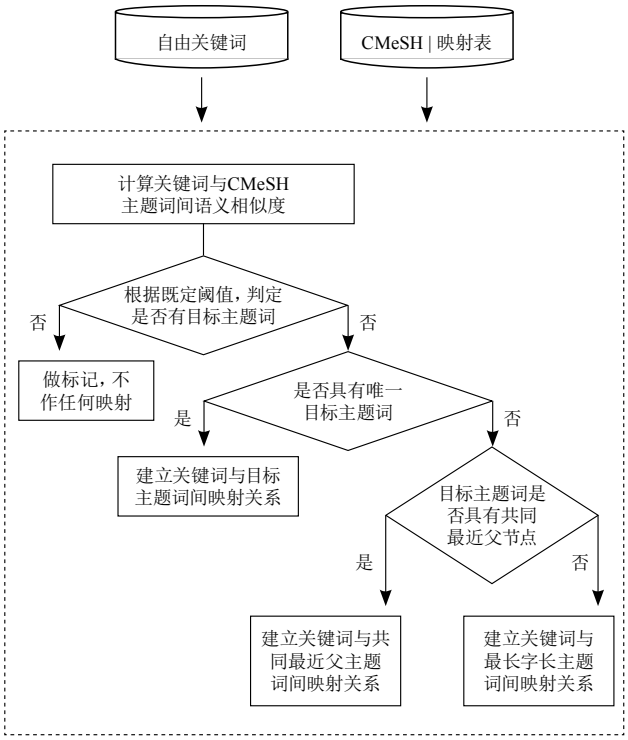

图3 CBM关键词-主题词自动映射技术方案

实际应用中，字面相似度算法模型选择的是Dice系数法。最佳相似度阈值的初始值设为0.8时，但允许用户根据具体领域的表现效果和人工干涉进度的要求进行调整。当一个关键词具有多个候选主题词时，系统推荐字长最长的主题词作为默认主题词，并且将所有满足阈值条件的候选目标主题词推荐出来供用户审核。

3.3 如何利用外部知识组织系统中的专业术语

知识组织系统是一种对内容概念及其相互关系进行描述和组织的、计算机可理解的系统，在用户信息需求与信息资源之间起着重要的桥梁作用，是实现信息组织、检索与分类、聚类、挖掘等自动化处理的重要基础之一。但由于编制目标、编制机构、编制人员、编制模式等要素的不一致，同一领域不同知识组织系统之间在收词、词间关系控制、词汇属性设置等方面存在较大差异。映射表的编制目标不是实现不同知识组织系统之间的互操作与词间关系整合，如前文所述，其目标是扩充映射表中的入口词，提高其文献覆盖率。

基于此，在利用外部知识组织系统中的专业术语进行词汇扩充时遵循“不重复收录”原则；在构建这些外来词汇与CMeSH主题词之间映射关系时，尊重词表自身的词间同义关系，但这些关系与映射表中现有同义关系存在冲突或不一致时，以映射表中的为准，或对外来词汇间的关系进行拆分，或对外来词汇间的关系进行合并。图4是以叙词表为例提出的外部知识组织系统中的专业术语在映射表构建与更新维护中的利用方案。主要包含3个步骤：

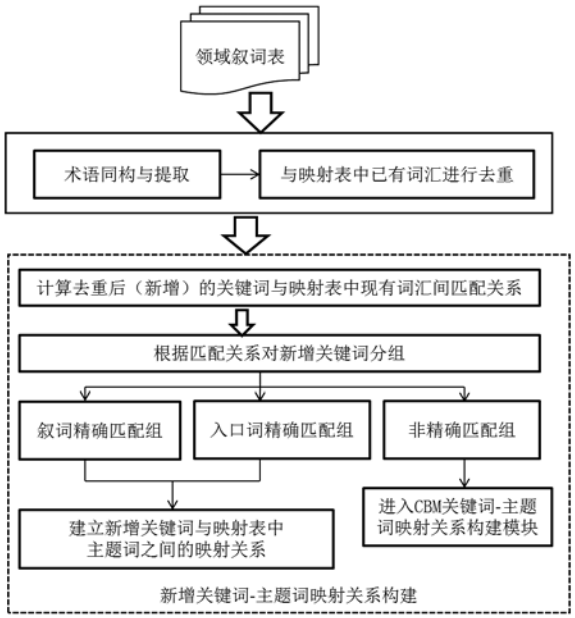

图4 外来领域叙词表中的术语在映射表中的利用方案

第一步：计算所有外来词汇与CMeSH主题词的字面匹配关系，并根据匹配关系对外来词汇进行语义分组。

（1）精确匹配组：该组中所有词汇自身在词形上与映射表中的主题词或已有关键词完全一致。

（2）叙词精确匹配组：该组中所有词汇自身在词形上与映射表中的主题词、已有关键词不完全一致，但其所在来源概念的叙词在精确匹配组中。

（3）入口词精确匹配组：该组中所有词汇自身在词形上与映射表中的主题词、已有关键词不完全一致，但其所在来源概念中的某个入口词在精确匹配组中。

（4）非精确匹配组：该组中所有词汇，无论是自己还是其所在的来源概念，没有一个词汇在词形上与映射表中的主题词、入口词完全一致。

第二步：剔除精确匹配组中的词汇，保留叙词精确匹配组和入口词精确匹配组中的词汇，并建立它们与映射表中主题词之间的映射关系。

第三步：按图3CBM关键词-主题词自动映射技术方案计算和构建非精确匹配组词汇与映射表中主题词之间的映射关系。

3.4 如何根据主题词的变化及时更新现有映射关系

以主题词A（A_1,A_2,…A_n）和B（B_1,B_2,…B_n）及其对应的关键词集合为例。一般来讲，映射表中主题词的变化情况可分为如下几类：

（1）优选词替换。即将现有主题词A变为关键词，并从现有对应关键词集合中选择A_i作为新的主题词。

（2）删除。即将整个主题词及其对应关键词集合删除。

（3）合并。即将A（A_1,A_2,…A_n）和B（B_1,B_2,…B_n）进行合并，并从中选择一个作为主题词。

（4）拆分。即将A（A_1,A_2,…A_n）拆分为两个或多个具有同位、相关或层级关系的主题概念，并分别选出主题词。

表2 映射表中主题词发生变化时的映射关系自动维护策略

序号	主题词变化情况	注释
（1）	优选词替换	在映射关系表达中替换相应的主题词和关键词的位置
（2）	删除主题词	将现有主题词及其对应关键词都作为新的自由关键词参与映射计算。从效率角度，优先与变更主题词的上位词、下位词、同位词进行计算，其次是相关词，最后才进行普遍计算
（3）	主题词合并	在映射关系表达中替换相应的主题词和对应关键词的位置
（4）	主题词拆分	重新计算入口词与拆分后的主题词间映射关系，并在映射关系表达中进行相应替换

对此，如表2所示，采取如下机制应对：

对于第1种变更情况，只需在映射关系表达中替换相应的主题词和关键词的位置。

对于第2种变更情况，将现有主题词及其对应关键词都作为新的自由关键词参与映射计算。其中，从效率角度，优先与变更主题词的上位词、下位词、同位词进行计算，其次是相关词，最后才进行普遍计算。

对于第3种变更情况，解决方案同（1）。

对于第4种变更情况，重新计算关键词与拆分后的主题词间映射关系，并在映射关系表达中进行相应替换。

（4）主题词拆分重新计算入口词与拆分后的主题词间映射关系，并在映射关系表达中进行相应替换

4 映射表计算机辅助构建与维护系统功能架构

图5是映射表计算机辅助构建与维护系统主要功能模块结构。总体分为两层，即加工层和发布应用层。其中加工层根据上述技术路线又具体分为词源扩充、映射关系自动构建和人工交互审核3层。

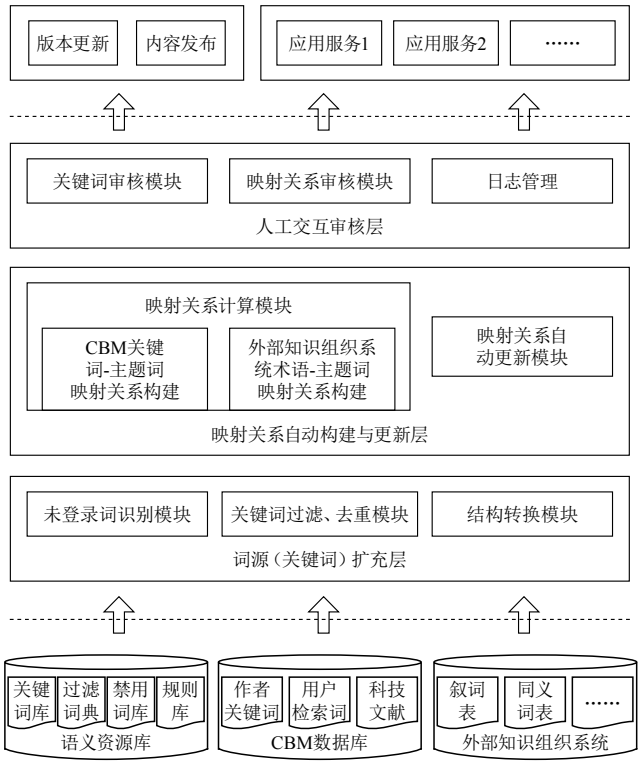

图5 映射表计算机辅助构建与维护系统主要功能架构

4.1 词源扩充层

实现关键词的扩充，由未登录词识别模块、结构转换模块和过滤、去重模块3个主要功能模块组成。未登录词识别模块主要负责从科技文献中提取现有关键

词库中尚未收录的词条；结构转换模块一方面负责将外部资源的结构转换成本系统支持的结构，另一方面负责将新收录的关键词描述方式转换成便于计算机进行映射关系计算的描述方式；过滤、去重模块按一定的规则对从不同来源获得的新词进行过滤和去重，保证关键词的唯一性。

4.2 映射关系自动构建与维护层

实现自由关键词到CMeSH主题词的自动映射。关键词词库中自由关键词可能是真正新扩充进来的全新词条，简称全新关键词，也可能是由于CMeSH主题词表更新生成的自由关键词（可以是从主题词角色转换而来的关键词，也可以是历史入口词），简称为历史关键词。全新关键词与主题词间映射关系的计算与构建主要由映射关系计算模块负责。历史关键词与主题词间映射关系的计算与更新主要由映射关系自动更新模块负责。

4.3 人工交互审核层

包括关键词审核模块、映射关系审核模块和日志管理模块。关键词审核模块用于判定计算机扩充进来的关键词是否适合作为文献处理。映射关系审核模块用于判定计算机自动构建的映射关系的正确性。日志管理模块用于记录人工审核操作情况。

4.4 发布应用层

提供内容发布、版本管理和对外应用服务。应用服务主要指能够按既定描述规范要求进行格式转换，支持外部应用系统或服务的调用。目前系统能够提供XML格式的映射表输出，Web service接口开发正在进行中。

5 小结

将自然语言应用到信息组织、标引和检索所需的各种词表中，实现自然语言与受控语言间的互操作，是新一代的知识组织系统构建模式。中文生物医学关键词-主题词映射表是基于中文医学主题词表CMeSH编制的，主要目标是用于“自然语言—主题语言—分类语言”一体化计算机辅助标引系统中，提高中文生物医学文献组织、标引和检索系统的性能与服务效果。为扩充映射表中的关键词来源，提高映射表中关键词的文献覆盖率，本文主要围绕“关键词扩充、关键词-主题词映射关系构建、关键词和关键词-主题词映射关系的更新”3个方面，介绍了中文生物医学关键词-主题词映射表的计算机辅助构建与维护机制，以及系统的功能架构，希望能够为其他领域自然语言与规范语言一体化词表建设提供参考。

截至2014年9月，中文生物医学关键词—主题词映射表已从CBM文献中新增累积2-7个字符碎片（关键词素材）40万余条，经过人工审核且扩充进映射表的关键词8万余条；从外部知识组织系统中扩充的原始专业术语累计100余万条，经去重且扩充进映射表的有40余万条。

参考文献

[1] 金燕,张玉峰.知识检索中自然语言控制机制研究[J].中国图书馆学报, 2004, 30(6):56-58.

[2] 张琪玉.积极为自然语言和情报检索语言的结合创造条件:建立大量编制自然语言词表(下)[J].图书馆杂志, 1999, 18(10):7-9.

[3] 张琪玉.分类语言、主题语言与自然语言一体化检索系统与《中国财经报刊数据库》的实践[J].现代图书情报技术,2002, (1):66-68.

[4] 戴剑波.受控词表的互操作研究[M].南京:东南大学出版社, 2009:155-158.

[5] 刘华梅.基于情报检索语言互操作技术的集成词库构建研究[D].南京:南京农业大学,2006.

[6] 仲云云,侯汉清,杜慧平.电子政务主题词表自动构建研究[J].中国图书馆学报,2008,34(175):97-102.

[7] 杜慧平,仲云云.自然语言叙词表自动构建研究[M].南京:东南大学出版社,2009.

[8] 陆勇.面向信息检索的汉语同义词自动识别[M].南京:东南大学出版社,2009.

[9] Medical Subject Headings [EB/OL]. [2014-09-20]. http://www.nlm.nih.gov/mesh/MBrowser.html.

[10] ARONSON A R. NLM Medical Text Indexer: A Tool for Automatic and Assisted Indexing [EB/OL]. [2014-08-15]. http://www.lhncbc.nlm.nih.gov/lhc/docs/reports/2008/tr2008002.pdf.

[11] RINDFLESCH T C, RAJAN J V, HUNTER L. Extracting Molecular Binding Relationships from Biomedical Text [EB/OL]. [2014-08-15]. http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2000/

pub2000016.pdf.

[12] ARONSON A R, RINDFLESCH T C, BROWNE A C. Exploiting a Large Thesaurus for Information Retrieval [EB/OL]. [2014-08-15]. http://skr.nlm.nih.gov/papers/references/riao94.final.pdf.

[13] SRINIVASAN S, RINDFLESCH T C, HOLE W T, et al. Finding UMLS Metathesaurus Concepts in MEDLINE [EB/OL]. [2014-08-15]. http://skr.nlm.nih.gov/papers/references/FindingUMLSinMEDLINE.pdf

[14] ARONSON A R. MetaMap Variant Generation [EB/OL]. [2014-08-15]. http://skr.nlm.nih.gov/papers/references/mm.variants.pdf.

[15] ARONSON A R. MetaMap Candidate Retrieval [EB/OL]. [2014-08-15]. http://skr.nlm.nih.gov/papers/references/mm.candidates.pdf.

[16] 中文医学主题词表 [EB/OL]. [2014-09-20]. http://cmesh.imicams.ac.cn/index.action?action=mainWordView&keyid=D018410&beanName=com.tbs.dictweb.bean.ZtcKmcPage.

[17] 中文生物医学文献服务系统 [EB/OL]. [2014-06-20]. http://www.sinomed.ac.cn.

[18] 孙海霞,李军莲,吴英杰,等.基于混合策略的中文生物医学领域未登录词识别研究[J].现代图书情报技术,2013,29(1):15-21.

[19] 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,2013,29(11):46-51.

作者简介

孙海霞，女，中国医学科学院医学信息研究所助理研究员，南京大学信息管理学院博士生。
李军莲，女，1972年生，中国医学科学院医学信息研究所副研究馆员，研究方向：信息组织与系统，通讯作者，E-mail：li.junlian@imicams.ac.cn。

Computer-aided Construction and Maintenance of the Chinese Biomedical Keyword - Subject Heading Mapping Vocabulary

SUN HaiXia[1,2], WU YingJie[1], LI DanYa[1], LI JunLian[1]
(1. Institute of Medical Information & Library of Chinese Academy of Medical Sciences, Beijing 100020, China; 2. School of Information Management, Nanjing University, Nanjing 210093, China)

Abstract: To realize the interoperability between natural language and controlled language, merging the natural language into vocabularies, used for information organization, indexing, retrieval and analysis, is a new mode of knowledge organization systems building. Focusing on keyword recognition and expansion, keyword - subject heading mapping building and its update, this paper introduces the computer-aided construction and maintenance mechanisms of the Chinese Biomedical keywords - subject heading mapping vocabulary, and its system's design and implementation.

Keywords: Biomedical; Keyword - Subject heading mapping; Knowledge organization system; Thesaurus maintaining; Computer-aided

（收稿日期：2014-12-04）