

# 基于聚合关系的中英文词表概念 映射方法及实证\*

邓盼盼<sup>1</sup>, 常春<sup>2</sup>

(1. 中国医学科学院医学信息研究所, 北京 100020; 2. 中国科学技术信息研究所, 北京 100038)

**摘要:** 文章基于映射规则, 探索了可操作的计算机辅助概念映射方法, 选择1562个中文概念与1476个英文概念作为样本, 进行具体的映射实证, 如自动推荐候选映射、自动识别映射类型、检测矛盾映射等。关于概念映射结果的评价, 主要选择映射准确率与映射覆盖率两个指标, 根据自动推荐、人工修改或删除的映射关系数量, 分析映射方法的效率。

**关键词:** 概念映射; 叙词表; 映射类型; 精确匹配

**中图分类号:** G35

**DOI:** 10.3772/j.issn.1673—2286.2014.10.007

语义环境下, 概念映射是知识组织体系互操作的主要方式, 是实现多语种信息自动化处理的基础, 将广泛用于跨语言网络及专业资源的自动标引和聚类、知识模块抽取、智能检索、知识导航与链接等。因此, 中文和英文作为使用人数较多、范围较广泛的语言, 两种语言概念的映射对于文化认知的无缝对接更是至关重要。中英文概念映射的主要任务是实现不同词表、语言的同义概念的对应, 解决异构异质词表的语词、词间关系、元数据信息的集成、存储与管理<sup>[1,2]</sup>。

## 1 材料来源与处理

### 1.1 词表选择与分析

关于领域的选择, 根据文献调研结果, 国内外仍没有进行过水利领域中英文概念的映射研究。所以, 按范畴或词族随机抽取水利工程领域一定数量的中英文概念作为实验样本, 参与此次映射验证。根据覆盖范围及权威性, 中文选择水利领域几部比较权威

的叙词表为基础术语来源, 如专业性的《水利水电科技主题词表》<sup>[3]</sup>、《黄河水利科技主题词表》<sup>[4]</sup>及正在编制的《汉语主题词表》(工程技术版)<sup>[5]</sup>等。在2009年, 中国科学技术信息研究所立项进行《汉表》(工程技术版)编制, 新版《汉表》概念量大、关系丰富, 基本包含了所有重要的汉语概念, 且基本附有英文译文<sup>[6]</sup>。

新版汉语主题词表范畴表中“TV 水利工程”大类表示水利工程领域, 将英文词表范畴表与汉语主题词表范畴表进行映射, 此处仅需找到“TV 水利工程”对应的英文范畴即可, 进而计算词表中水利领域的概念量及比例。根据概念量及比例分布, 英文术语均选择包含水利领域概念较多的词表, 如*Transportation research thesaurus*<sup>[7]</sup>、*Ei Thesaurus*<sup>[8]</sup>、*NASA thesaurus*<sup>[9]</sup>、*Thesaurus of scientific, technical, and engineering terms*<sup>[10]</sup>、*Water resources thesaurus*<sup>[11]</sup>、*International energy: subject thesaurus*<sup>[12]</sup>、*InterWATER Thesaurus*<sup>[13]</sup>、*General Multilingual Environmental Thesaurus*<sup>[14]</sup>、*Inspec Thesaurus*<sup>[15]</sup>等24部词表。根据统计, 截至目前为止汉

\* 本研究得到国家“十二五”科技支撑计划“《汉语主题词表》(工程技术版)与英文超级科技词表的映射研究”(编号: 2011BAH10B07)资助。

语词表中约有5000个水利领域概念,非优选词约4100个,概念趋于精细化。在选择的英文词表中,初始概念共约7000个水利工程领域概念;由于来源于不同的叙词表,这些概念仍然需要进行初步的术语加工。

## 1.2 抽取概念元数据

关于样本数据的选择,在5000个水利领域中文概念中选择了1562个优选词作为样本概念,非优选词共1005个,等级关系共为1676个,相关关系数量共427个。中文样本主要是根据范畴或词族进行选择,首先选定范畴“TV6 水利枢纽、水工建筑物”,然后根据词族规模、词频等确定选用词族及族首词,如“坝”、“堤”、“水库”等,再根据需要进行选择这些词族中的所有概念或部分概念;必须保证概念关系的闭合性,如“坝”参“坝基”,则需要存在另外一条记录与之对应,“坝基”参“坝”,所以在已选的词族概念参项中,也选择了一些作为样本概念。中文概念多集中在“TV5 水利工程施工”、“TV1 水利工程基础科学”、“TV6 水利枢纽、水工建筑物”、“TV7 水能利用、水电站工程”等,词族关系相当完整、系统,不需要再进行术语归并。

由于不同语种间相同学科的概念基本是相同的,所以英文概念也选择水利工程相应的子领域概念作为样本数据。首先,根据中文样本概念的范畴,找到英文词表范畴表中对应的范畴,选择这些范畴下的英文概念作为样本概念。如果没有范畴表,或范畴表中“水利工程”没有进一步细化为各子领域,则将表水利工程的全部概念均纳入样本概念。由于英语概念来源于诸多不同的词表,一般词族规模小,概念相对分散,不同词表中可能存在相同的概念,所以需要进行一定的术语归并,将相同概念的关系及属性进行归并及矛盾处理,主要借助Norm<sup>[15]</sup>将选取的7000个初始英文概念进行原型化,并基于用代关系、用代关系传导等方法对术语进行自动归并,剔除了一些使用频率不高的或不属于水利领域的概念。最终,在7000个初始英文概念中选择了1476个作为样本概念,非优选词共1197个。

在概念数据存储、处理及计算过程中,元数据主要采取款目词、属性、属性值的三元组存储形式,主要选用的是Oracle关系数据库,辅助进行英文同义术语的归并及矛盾关系处理、中英文概念的自动映射及人工判定、映射关系浏览与统计等。

## 2 基于义原特征推荐候选映射

### 2.1 确定核心义原

候选映射,主要通过英文术语与汉语术语的多个译文、英语术语的中文翻译与汉语术语进行相似匹配或用代传导,识别同义术语,然后在平台上进行推荐。

由于叙词表中的术语都具有一定的概念内涵,由多个单词或汉字组成加以限定,所以组成概念的有实际意义的每个单词或汉字都可称为该概念的概念元素(sememe),简称义原或义素。核心义原(core sememe),表示概念的本质性质或归属。如“平板坝”含有义原“平板”、“坝”,“平板”用于修饰并确定“坝”的类型;“flood discharge dam”含有义原“flood”、“discharge”和“dam”,其中“flood”、“discharge”用于指示此类“dam”的用途并与其他种类区分开。所以“坝”、“dam”分别是“平板坝”、“flood discharge dam”的核心义原。术语的核心义原示例如表1所示。

表1 术语核心义原示例

族首词	中文术语	核心义原	族首词	英文术语	核心义原
坝	尾砂坝	坝	dam	tailings dam	dam
河口	黄河河口	河口	estuary	estuary of huanghe	estuary
洪水	高含沙洪水	洪水	flood	flood with hyperconcentration	flood
水电站	坝后式水电站	水电站	power station	power station at dam toe	power station

中英文概念的核心义原一般为名词、动名词,核心义原在术语中的位置呈现一定的规律性,据此规律可以自动识别术语的核心义原,主要方法如下:

#### (1) 基于后方一致确定核心义原

位于同一个词族的概念,在词形上多数情况下存在后方一致的规律<sup>[17]</sup>,基于后方一致可以自动识别核心义原。如中文概念“坝”词族中“尾砂坝”、“泄洪坝”等词结尾均为“坝”,“河口”词族多数概念结尾词为“河口”;英文概念也存在此种规律,“leakage”

词族中“water leakage”、“canal leakage”、“dam leakage”、“reservoir leakage”等也均以“leakage”结尾，所以可以确定该词族所有概念的核心义原为“leakage”。另外，核心义原位于术语词形末尾的特征，也可以用于确定词表中无关概念的核心义原。

### (2) 基于义原频次自动识别核心义原

位于同一个词族的概念，核心义原通常是相同的，词族的核心义原出现次数相对较多。所以，可根据该词族所有义原的出现频次，确定该词族概念的核心义原。如假设“dam”词族含有概念“dam”、“core dam”、“weir dam”、“barrier dam”、“overflow dams”五个概念，则含有义原“dam”出现5次，“core”、“weir”、“barrier”、“overflow”各1次，所以自动选择出现频次较高的“dam”作为该词族的核心义原。

### (3) 基于介词位置自动识别核心义原

术语含有介词时，一般就不再符合后方一致的规律，但核心义原一般位于介词之前，如“estuary of huanghe”的核心义原“estuary”位于介词“of”之前，“flood with hyperconcentration”介词“with”前的“flood”为该概念核心义原，“power station at dam toe”的核心义原“power station”也位于“at”之前。

## 2.2 义原相似度计算

转换后的汉语术语或英语术语在词形或语义上也不都是一致的，所以根据词形匹配及相似度计算等判定概念间的映射关联，过程及原理具体如下：

参考汉英自动翻译的方法，英语术语匹配时，可以采用最大公共字符串的比例进行相似度计算，部分相似度较高的也可以归并为同一概念。假设术语t1、t2的长度为Length(str1)、Length(str2)，相同字符个数为Com(str1, str2)，则用两倍的公共字符串长度与两术语字符串长度之和的比值表示两术语的词形相似度，公式为： $Similarity(str1, str2) = 2 * Com(str1, str2) / [Length(str1) + Length(str2)]$  [18]。此种方法在一定程度上反映了术语间词形上的关系，但由于最大公共字符串不一定是一个单词或有实际意义的术语，推荐时也会出现一些错误，如“hush”与“ash”；即使最大公共字符串是一个有实际意义的概念，相似度高的也不一定具有相同的核心理念，如“hash”与“ash”虽然相似度高达85.7%，但二者是完全不同的概念。

基于最大公共字符串的相似度计算方法进行适当

的改进，将公共义原的比例作为两术语的相似度。假设术语term1、term2，其义原分别为S(term1)、S(term2)，去除介词、助词等虚词后的两术语义原个数分别为Sno(term1)、Sno(term2)，共同的义原个数为Coms(term1, term2)，则用两倍的公共义原个数与两术语义原数量之和的比值表示两术语的语义相似度，即公式为： $Proximity(term1, term2) = 2 * Coms(term1, term2) / [Sno(term1) + Sno(term2)]$ 。如“hash”、“ash”两个概念间的相似度为0，“lake water”与“lake”相似度为66.7%。如果考虑术语中各个义原的重要性，可以根据词形进行权重的设置；一般来说，概念的核心元素为名词，则名词的权重就相对高些。

## 3 基于聚合关系自动判定映射类型

概念映射主要从中文概念到英文概念，自动映射时多个映射类型同时建立，有利于保证映射关系的覆盖范围。候选映射推荐时，应标明不同的推荐方式或方法，有助于候选推荐的准确度排序及快速判定映射关系的取舍，如“义原”、“相似度”、“组代”、“多属”、“继承”、“推理”。由于英文词表词族关系不完整，目前仅基于等同关系（包括组代关系）、等级关系（包括多重属分关系）、相关关系等聚合关系，自动识别并判定映射关系类型。

### 3.1 基于义原匹配自动识别映射类型

词形完全相同或具有用代关系的两术语可推荐为精确匹配(exact match, 符号EM)。主要存在以下几种情况，具体示例如表2。

表2 基于义原匹配自动识别映射类型示例

中文概念	译文	类型	英文概念	推荐
渠道滑坡	channel landslide	BM1	landslides	义原
太湖水	taihu lake water	BM2	water	义原
水闸	sluice	EM	floodgate	义原
闸	gate	NM1	slide gates	义原
闸	gate	NM1	spillway gates	义原
地下水	underground water	RM	water level	义原

(1) 原型化后结果完全相同或两术语存在用代关系, 均可建立精确匹配。如英文术语“floodgate”代项为“sluice”, 则可与中文术语“水闸 (sluice)”建立精确匹配。如果一个中文术语同时精确匹配到多个英文术语, 或多个中文术语同时精确匹配到一个英文术语, 则将保留由优选词推荐的精确匹配, 由代项或同义词推荐的精确匹配自动降级为近义匹配。

(2) 核心义原相同, 且一个术语的义原包含另一个术语的所有义原, 可根据义原数建立上位 (broad match, 符号BM) 或下位匹配 (narrow match, 符号NM); 义原数量可以粗略表示概念的细化程度, 所以两术语义原数的差值可用于标记两术语在词族中的距离, 建立向上匹配或向下匹配时, 类型可标记为NM1、NM2、BM1、BM2等, 也可用作人工判定时的参考。如“渠道滑坡 (channel landslide)”与“landslides”核心词相同, 中文概念义原数大于英文概念义原数, 则“渠道滑坡 (channel landslide)”可以上位匹配到“landslides”; “堤坝护岸 (levee revetment)”与“revetments”核心词均为“revetments”, 中文概念义原数小于英文概念义原数, 则建立下位匹配。核心义原及义原数均相同, 则两术语可能为同位概念或相关概念, 可能建立相关匹配 (related match, 符号RM), 如“underground water”与“water level”。相关匹配自动推荐的数量, 控制在10个以内。

(3) 如果上述两种情况均无候选推荐, 则对于英文术语推荐减少核心义原的限制, 即只需一个术语的义原包含另一个术语的所有义原, 就可以进行推荐, 映射类型也是根据义原数量判定, 如“open channel drainage”与“open channel”。

### 3.2 基于义原相似度自动识别映射类型

英语概念中的优选词和非优选词构成的同义词组, 每个术语都纳入相似度计算, 选择所有相似度中的最大值作为该同义词组与中文概念的相似度。通过计算, 根据概念词形或用代两种方式的匹配情况, 推荐相似度较高的前10位候选概念, 相似度控制在30%以上。

基于义原相似度自动识别的候选映射及类型, 示例如表3。具体情况如下:

原型化结果相同时, 义原相似度为1, 则两术语可以建立精确匹配; 如果出现同一术语精确匹配到多个术语的情况, 则转换为近义匹配。

表3 基于义原相似度自动识别映射类型示例

中文概念	译文	类型	英文概念	推荐
雨水回收系统	rain water reuse system	SM	rain water sewer system	相似度
风力输沙	sediment transport by wind	BM1	sediment transport	相似度
混凝土高坝	high concrete dam	BM1	dams concrete	相似度
国际河流	international rivers	NM1	international inland waters	相似度
国际河流	international rivers	SM	international river basin	相似度
岩体滑坡	rock landslides	RM	landslide protection works	相似度

根据实际测试义原相似度在70%以上时, 如果两术语的义原数相同, 则两概念的内涵是相近的, 可以建立近义匹配; 如果两术语的义原数不相同, 源概念义原数大于目标概念义原数, 则源概念较为专指, 可能上位匹配到目标概念; 源概念义原数小于目标概念义原数, 则源概念较为泛指, 可能下位匹配到目标概念。

在70%以下时, 多可以建立相关匹配。虽然要控制相关映射的数量, 但较多的推荐能够保证源概念能够完全匹配到最相近概念, 所以也应进行一定数量的相关映射数量推荐。关于自动匹配推荐的候选映射概念及映射关系, 可根据源概念与目标概念的内涵大小及关系, 通过人工判定进行保存或修改, 确定合适的映射关系。

### 3.3 基于概念关系自动识别映射类型

关于术语词形较长、专指度较高的概念, 如果没有直接对应的概念, 则可以推荐另一种语言中的多个优选词组合表达该概念。概念由多个义原组合而成, 对于由义原交叉组配的概念来说, 可以将概念直接拆分为多个义原组合, 查找另一语言中的相应概念, 继而实现一对多映射; 对于由义原并列组配的概念, 多是由于语言差异而引

起的分类角度不同, 导致上位概念分解为不同的下位概念组合, 可根据实例对相似度较高的组合进行推荐。如表4示例。

## 4 概念映射结果分析

自动推荐的候选映射及类型并不都是正确的, 仍然需要根据词族关系及映射关系, 识别存在的矛盾映射, 主要包括一个概念建立了多个精确匹配、两个概念建立多个不同类型的映射关系、不同领域概念或含义建立了映射等三种情况。然后加以人工判断, 采取的处理方案主要包括修改为正确的映射类型、解除映射关系等, 此类情况都判定为错误的候选映射推荐。

### 4.1 确立概念映射质量评价指标

准确率、覆盖率两个指标广泛应用于多语种自动翻译、信息检索等结果的评价<sup>[19]</sup>, 具有普适性。因此, 与人工辅助后结果相比, 自动映射的效果也采用映射准确率与映射覆盖率作为主要评价指标。

Mapping Precision (映射准确率) = CR (正确推荐的概念映射) / AR (自动推荐的所有映射关系)

Mapping Coverage (映射覆盖率) = CR (正确推荐的概念映射) / AM (可能存在的所有映射关系)

式中, CR (正确推荐的概念映射) 即为自动推荐的所有映射关系除去人工修改的映射关系数量, AM (可

能存在的所有映射关系) 即为最终建立的全部映射关系。受时间限制, 人工浏览仅判断映射关系的正确性, 没有新增可能存在的其他映射关系, 因此:

CR (正确推荐的概念映射) = AR (自动推荐的映射关系) - TR (修改为其他类型的映射关系) - DE (直接删除的映射关系)

AM (可能存在的所有映射关系) = AR (自动推荐的映射关系) - TR (修改为其他类型的映射关系) - DE (直接删除的映射关系) + CO (由其他类型转换的映射关系)

由于可能存在的所有映射关系无法准确计算, 所以可以综合当前的映射覆盖率及映射关系数量, 评价映射结果的覆盖范围。

### 4.2 基于义原匹配的映射结果分析

基于义原包含或匹配的三种情况, 将选择的中英文样本概念进行自动映射, 在1562个中文概念中共计1395个概念存在候选映射推荐, 推荐了共1116个英文概念, 自动获取了17086条映射关系。具体映射数据如表5。

在自动推荐的239条精确匹配关系中, 其中1条为推荐了一些不直接相关的概念, 采取直接删除处理, 如“井水 (well water)”与“water well (水井)”。共有29条, 推荐的英文概念和中文概念存在一定的映射关系, 但自动识别的匹配类型不恰当, 人工浏览时修改为其他匹配类型, 如“水库大坝

表4 基于概念关系自动识别映射类型

基于聚合关系识别	中文概念属性	英文概念属性	映射关系
组代关系	地下水渗流 (ground water seepage)	ground water seepage UFC ground water UFC seepage	地下水渗流 ~EQ ground water+ seepage
多重属分关系	双曲薄拱坝 (double curvature thin arch dam) 属项: 双曲拱坝 属项: 薄拱坝 双曲拱坝 (double curvature arch dam) 薄拱坝 (thin arch dam)	double curvature arch dam thin arch dam	双曲薄拱坝 ~EQ double curvature arch dam + thin arch dam
义原组配	城市地表水 (urban surface water)	surface waters urban water	城市地表水 ~EQ surface waters + urban water

(reservoir dams)”与推荐的精确匹配概念“dams & reservoirs”，修改为上位匹配。相应地，也存在由其他类型修改为精确匹配的情况，共新增68条。自动推荐时，一个中文概念近义匹配到多个英文概念，人工判定时选择一个最贴近的概念建立精确匹配，如“地表水(surface water; open water)”近义匹配到“surface waters(地表水)”与“open water(开阔水域, 地表水)”，经过比较发现中文概念“地表水”与“surface waters”内涵相同，与“open water”概念稍有差异，所以“地表水”与“surface waters”修改为精确匹配。

这种方法，精确匹配、近义匹配、上位匹配、下位匹配的准确率都达到了60%以上，映射结果比较理想。此外也自动推荐了部分兄弟概念为相关匹配，在没有其他映射类型推荐时，能够增加映射的覆盖率，但会出现一些关系不太紧密的或错误的映射推荐，导致准确率比较低。

表5 基于义原匹配的映射结果

方法 \ 类型	精确匹配	近义匹配	上位匹配	下位匹配	相关匹配	合计
自动推荐	239	183	1861	1377	13426	17086
由其他类型转换	68	24	1211	240	777	2320
修改为其他类型	29	71	399	410	1411	2320
直接删除	1	1	16	31	6239	6288
映射准确率	87.4%	60.1%	77.7%	67.9%	43.0%	49.6%
映射覆盖率	75.5%	81.6%	54.2%	79.6%	88.1%	78.5%

### 4.3 基于义原相似度的映射结果分析

基于义原相似度的几种情况，将中英文样本概念进行自动映射，共计1508个中文概念与1028个英文概念间存在映射推荐关系，自动获取了12180条映射关系，其中上位匹配、相关匹配较多。关于自动推荐的映射关系、人工处理情况及映射结果，具体如下表6。

这种方法，精确匹配、近义匹配的准确率和覆盖率都比较高，仅依据相似度无法有效判定中英文概念的上位和下位关系，所以上位匹配、下位匹配的准确率都有所下降；借助相似度和义原数，可以识别一些相关匹配，如事物“边坡(slope)”与事物活动“slope protection”，相关匹配准确率稍显增加。

### 4.4 映射方法调整与映射结果分析

通过义原匹配或相似度的映射结果分析可知，在精确匹配或近义匹配方面，两种方法都能有效、准确地进行推荐，而且义原匹配方法关于上位匹配、下位匹配的推荐准确度较高，弥补了义原相似度的不足；义原相似度推荐的相关映射在一定程度上丰富了相关匹配的数量。综合两种自动推荐方法的优势，调整为：

(1) 原型化结果相同，原型化结果相同的两术语存在用代关系或义原相似度为1，则推荐为精确匹配，如果同一概念出现多个精确匹配现象，保留优选词的推荐为精确匹配，其他转换为近义匹配。

(2) 核心词相同，根据义原包含的方向确定上位或下位匹配。义原数量相差较小的或义原相似度较大的，一般在词族中距离最近，所以通过义原数量差标记上位或下位匹配的两概念的距离。核心词、义原数均相同时，相似度大于70%，多为相近概念或兄弟概念，推荐为近义匹配；相似度小于70%，推荐为相关匹配。核心词相同、义原数不同时，相似度在30%以上，根据义原数及方向确定匹配类型，如中文义原数大于英文义原数为上位匹配、中文义原数小于英文义原数为近义匹配。

(3) 如果无法识别概念的核心词，则直接根据义原包含的方法确定上位或下位匹配。除去(1)(2)(3)中推荐的英文概念，与剩下的样本概念进行相似度计算，相似度大于70%，推荐为近义匹配；在30%与70%之间且义原数相同，推荐为相关匹配。

1562个中文概念与1476个英文概念进行映射，基于义原或相似度等方法自动推荐候选映射关系，共1490个词获得了候选推荐，推荐的英文概念共1417个。

表6 基于义原相似度的映射结果

方法 \ 类型	精确匹配	近义匹配	上位匹配	下位匹配	相关匹配	合计
自动推荐	248	185	5457	592	5698	12180
由其他类型转换	67	41	201	143	1629	2081
修改为其他类型	41	70	1476	210	284	2081
直接删除	0	1	1725	68	1012	2806
映射准确率	83.5%	61.6%	41.3%	53.0%	77.2%	59.8%
映射覆盖率	75.5%	73.5%	91.8%	68.7%	73.0%	77.8%

经过人工浏览判定或修改, 并进行标记, 中英文概念映射结果如表7。由计算机辅助方法自动推荐的映射关系共25801条。

表7 映射方法调整与映射结果

方法	类型	精确匹配	近义匹配	上位匹配	下位匹配	相关匹配	合计
自动推荐		249	235	4119	2261	18937	25801
由其他类型转换		82	34	569	387	2133	3205
修改为其他类型		29	93	1338	910	835	3205
直接删除		1	2	413	176	7068	7660
映射准确率		88.0%	59.6%	57.5%	52.0%	58.3%	57.9%
映射覆盖率		66.2%	52.0%	80.6%	75.2%	83.7%	82.3%

根据映射准确率计算公式, 几种匹配类型的映射准确率和覆盖率都达到了50%以上, 得出所有映射结果准确率为57.9%, 正确映射关系结果14936条。与基于义原匹配或义原相似度的单种方法相比, 两种方法的结合弥补了各自的不足, 既保证了较高的映射准确率, 也获取了较全的映射结果, 同时也证明了文中所述映射规则及方法的正确性。

## 5 结语

选择含水利领域概念较多较全的中文叙词表与英文词表作为样本概念来源, 从推荐候选映射、确定映射类型、人工审核矛盾映射、映射结果分析等四个步骤进行映射实践。关于概念映射结果的评价, 选择自动翻译与信息检索领域应用较多的准确率与覆盖率两个指标, 即映射准确率与映射覆盖率。总之, 综合义原匹配、义原相似度后的概念自动映射方法, 保证了较高映射准确率与覆盖率, 作为人工概念映射的辅助手段, 在一定程度上提高了中英文概念映射整个过程的效率, 人工参与的计算机辅助方法仍是比较理想的概念映射模式。

### 参考文献

[1] 邓盼盼, 常春. 基于精确匹配的概念映射关系规则研究[J]. 图书情报工作, 2013, 57(16): 25-29.

[2] 邓盼盼, 常春, 曾建勋. 中英文词表概念映射关系及处理方案研究[J]. 情报杂志, 2013, 32(10): 127-130, 192.

[3] 水利部信息研究所. 水利水电科技主题词表[M]. 郑州: 黄河水利出版社, 1998.

[4] 黄河水利委员会. 黄河水利科技主题词表[M]. 郑州: 黄河水利出版社, 2010.

[5] 中国科学技术情报研究所, 北京图书馆. 汉语主题词表[M]. 北京: 科学技术文献出版社, 1980.

[6] 常春, 曾建勋, 吴雯娜, 等. 《汉语主题词表》与英文超级科技词表概念映射构架设计[J]. 数字图书馆论坛, 2012, (12): 28-32.

[7] Transportation Research Board. Transportation Research Thesaurus [EB/OL]. [2013-06-28]. <http://trt.trb.org/trt.asp>.

[8] Elsevier Engineering Information. The Ei Thesaurus [EB/OL]. [2013-06-30]. <http://www.ei.org/other-publications>.

[9] NASA Scientific and Technical Information Program. NASA thesaurus [M]. NASA Scientific and Technical Information Program, 2010.

[10] Taylor & Francis Inc. Thesaurus of Scientific, Technical and Engineering Terms [M]. Taylor & Francis Inc, 1988.

[11] U.S. Department of the Interior Office of Water Research And Technology. Water resources thesaurus [EB/OL]. [2013-06-25]. <http://books.google.com.hk/books?id=An6NZOC-g9sC&printsec=frontcover&dq=Water#v=onepage&q&f=false>.

[12] International Energy Agency. International energy: subject thesaurus [EB/OL]. [2013-06-15]. <http://www.etde.org/edb/etdesuth.pdf>.

[13] IRC International Water and Sanitation Centre. InterWATER Thesaurus [DB/OL]. [2014-09-24]. <http://www.vocabularyserver.com/interwater/en/index.php>.

[14] European Environment Information and Observation Network. General Multilingual Environmental Thesaurus [DB/OL]. [2014-09-26]. <http://www.eionet.europa.eu/gemet/zh-CN/themes/>.

[15] Institution of Electrical Engineers. Inspec Thesaurus [EB/OL]. [2014-08-31]. <http://www.theiet.org/resources/inspec/about/records/ithesaurus.cfm>.

[16] the lexical systems group. Lexical web tools [EB/OL]. [2014-06-21]. <http://lexsrv3.nlm.nih.gov/LexSysGroup/Summary/lexicalTools.html>.

[17] 杜慧平. 概念等级关系自动识别研究[J]. 情报学报, 2011, 28(3): 237-243.

[18] 邓丹, 刘群, 俞鸿魁. 基于双语词典的汉英词语对齐算法研究[J]. 计算机工程, 2005, 31(16): 45-47.

[19] 王石, 曹存根. WNCT: 一种WordNet概念自动翻译方法[J]. 中文信息学报, 2009, 23(4): 63-70, 94.

## 作者简介

邓盼盼, 女, 硕士, 研究方向: 知识组织与服务, E-mail: deng.panpan@imicams.ac.cn。  
常春, 男, 博士, 研究馆员。

### Concept Mapping between Chinese and English Thesaurus Based on Paradigmatic Relationship

DENG PanPan<sup>1</sup>, CHANG Chun<sup>2</sup>

(1. Institute of Medical Information & Library of Chinese Academy of Medical Sciences, Beijing 100020, China;  
2. Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Based on previous research of mapping rules, the paper proposes a computer-aided concept mapping method. It selects 1562 Chinese concepts and 1476 English concepts as sample to make empirical research, such as recommending candidate mappings automatically, identifying mapping types and detecting mapping contradictory. It uses mapping precision and mapping coverage as the evaluating indicator of concept mapping results to analyze mapping efficiency, according to the mapping quantity of automatic recommendation, modification or deletion.

Keywords: Concept mapping; Thesaurus; Mapping type; Paradigmatic relationship

(收稿日期: 2014-09-16)

## 《数字图书馆论坛》2014年征稿启事

《数字图书馆论坛》是由科学技术部主管、中国科学技术信息研究所主办的专业性学术刊物(月刊), 国际标准刊号ISSN: 1673-2286, 国内统一刊号: CN:11-5359/G2。本刊是“中国科技核心期刊”统计源刊, 是CSSCI扩展版来源期刊。

本刊是我国唯一一本以“数字图书馆”命名的刊物, 一直关注国内外数字图书馆领域的相关研究和实践, 设有: 专家访谈、专家视点、专题研究、技术前沿、应用案例、业界动态等栏目, 报道主题涵盖信息检索、数字资源、知识组织、语义技术、开放获取、用户服务等, 侧重反映数字图书馆领域在资源建设、技术应用和产品服务等方面的新趋势、新发展和新变革。

本刊注重稿件的学术水准、研究内容和研究特色, 来稿需要满足以下基本要求: ①未发表过、未一稿多投的原创性论文; ②主题鲜明、数据可靠、文字通顺、引用规范; ③来稿应包含以下项目: 中文和英文的标题、作者姓名、单位、摘要、和关键词, 以及中图分类号、参考文献和作者联系方式; ④本刊在30天内将通知作者是否录用, 30天后作者可另行投稿。请登录本刊网站 (<http://www.DLF.net.cn>) 进行在线投稿。

本刊既厚名家、更重新人。欢迎国内外作者赐稿, 一经发表即向作者寄送稿酬和样刊。本刊开放出版(网址: <http://www.DLF.net.cn>), 也期待着相关专家在阅读或利用后提出宝贵意见和建议。