Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

《汉语主题词表》与英文超级科技词表概念映射构架设计*

□ 常春 曾建勋 吴雯娜 宋培彦 刘伟 邓盼盼 / 中国科学技术信息研究所 北京 100038

摘要：在网络时代，知识组织体系映射有利于实现信息的集成检索与获取。文章设计了《汉语主题词表》（工程技术版）与英文超级科技词表的概念映射模型，包括中文与英文词表的概念语义映射方法、概念映射数据描述、机器辅助和最短距离映射规则等，并讨论了中英文双语检索模型的应用前景。同时，基于英文超级科技词表的概念关系网络及其中文规范译名，提出了《汉语主题词表》扩充和完善的方法，从而更好地实现中英文概念映射，使英文超级科技词表在知识检索和知识发现中发挥更大的作用。

关键词：汉语主题词表，英文超级科技词表，映射
DOI：10.3772/j.issn.1673—2286.2012.12.005

在当今信息时代，知识表示、语义表达与推理等，成为知识组织的重要研究方向。叙词表作为知识组织体系的重要成员，是人类领域知识的积累。叙词表与其他知识组织体系的互操作，成为数字环境下实现知识组织与知识检索的重要途径，叙词表间的映射更是实现不同知识库的无缝融合的重要方法[1]。2011年新修订的ISO 25964叙词表国际标准分两部分颁布，第一部分已于2011年公开出版。第二部分目前处于修订的最后阶段，标题就是叙词表与其他词表的互操作，主要内容包括叙词表间的映射方法，叙词表与术语表、标题表、分类法、本体、名称规范列表、同义词环等的映射方法，以及映射数据管理、显示和应用等规范。ISO 25964叙词表国际标准第二部分全部是知识组织体系映射的相关规范，彰显映射在知识组织中的重要性[2]。国家“十二五”科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与应用示范”研究项目于2011年4月正式开始。其中，课题七的研究题目名称为“《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”，本文就是该研究课题相关的映射构架设计[3]。

1 映射材料与相关研究进展

1.1 《汉语主题词表》（工程技术版）编制进展

《汉语主题词表》（以下简称《汉表》）是我国第一部大型综合性叙词表，分自然科学部分和社会科学部分，由科学技术文献出版社于1980年公开出版[4]。1991年，原中国科学技术情报研究所对自然科学部分进行了修订，由科学技术文献出版社公开出版《汉表》自然科学增订本，通过压缩版面将原来的7个分册合为4个分册，第一和第二分册为字顺表，第三分册为词族索引、范畴索引，第四分册为英汉对照索引[5]。1996年，《汉表》第五分册轮排索引也公开出版。

近年来，随着计算机科学的发展、网络技术的应用，叙词表的编制和应用都发生了质的变化，基于大型数据库、词频信息、用户检索用词等信息，可以准确定位叙词表概念用词；基于网络协同技术，叙词表编制人员可以不分时间、地点，在网上共同构建叙词表知识体系；基于Web技术，用户可以直接参与叙词表的编制与应用工作；编制和维护工作也相应实现了网络在线适时完成的功能；叙词表的应用方式也发生了新的变化，机器辅助标引、网络信息系统嵌入式检索、词间关系可视化展示等等，以上所有特征表明网络环境下的叙词表编制与应用进入到一个新的阶段，

* 本文系国家“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”课题（编号：2011BAH10B07）研究成果之一。

预示着网络叙词表的正式诞生。在此大背景下，在图书情报领域具有重要影响力的《汉表》，从1991年起就没有进行更新维护，需要进行彻底的更新维护[6,7]。为了探索编制模式，首先在国家工程技术图书馆进行试验应用，中国科学技术信息研究所在2009年立项进行《汉表》（工程技术版）编制，全国16个来自大学和科研院所的领域信息研究单位参与了该项工作。为了更好地适应计算机的使用，《汉表》（工程技术版）计划概念数为22万左右，入口率为1:1，到2012年，编制工作已经进入后期专业融合阶段。《汉表》自然科学部分的整体修订，也在申请和策划阶段。

1.2 英文超级科技词表编制进展

国家“十二五”科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与应用示范”研究项目，其课题一研究题目为“面向外文科技文献信息的超级科技词表和本体建设”。超级科技词表预计收集科技概念规范名称80万条，覆盖理学、工学、医学和农学领域，由基础词库、规范概念库和范畴库三个部分组成。基础词库术语主要包含国际上重要知识组织体系中的术语，主要涉及叙词表、术语表等富含语义关系的术语或概念，计划规模为500万条；规范概念库主要包含概念形成过程中涉及的词型规范、词义规范的同义词和准同义词，以及规范概念间的共现关系；范畴库是按照使用需求修改或重新编制的概念体系结构，主要用于规范概念的分类归并以及文献信息的宏观分类导航[3]。

超级科技词表目前处于建设阶段，已经完成将近2000多部相关知识组织体系的原始素材收集和整理工作，涉及将近1000万条的理学、工学、医学和农学领域语词；设计了基础词库元数据元素集，包括术语的关系型元数据、描述性元数据和管理性元数据等50多个描述元素，基础词库加工平台也在研发过程中；课题组正在调研DDC、UDC等国际著名分类法的知识体系结构，以促进范畴体系结构的制定。

1.3 国内外叙词表映射研究进展

无论是用于数据库系统的知识组织体系互操作，还是网络信息系统的信息互操作，国内外已经有大量的研究报道和实践探索。欧盟Renardus项目建立了本地分类法与DDC的类目映射关系，实现了统一使用DDC进行信息资源等级结构导航[8]；OCLC术语服务通过多个受控词表间的映射信息，实现对术语资源的一站式获取[9]；英国HILT研究项目也研究了多部词表的互操作[10]；UMLS更是包含了上百部医学相关的叙词表与分类法等知识组织体系映射信息，建立了语义网络[11]；SUMO概念也实现了与WordNet术语的映射[12]。国内图书馆学、情报学机构也进行了大量的知识组织体系映射研究和探索，侯汉清等通过计算类目概念因素的相似度得到类目整体概念之间的相似度，自动确定《中国图书馆分类法》（CLC）类目与《杜威十进分类法》（DDC）类目的映射关系[13]；赖院根、曾建勋等对中图法与国际专利分类表IPC的类目映射提出了映射模型，建立分类体系之间的映射关系[14]；常春等研究并实现了将《农业科学叙词表》与粮农组织农业多语种叙词表AGROVOC在概念层次上的完整映射[15]；山西省图书馆等联合研制了《中图法》、《科图法》和《人大法》之间的映射对应系统，并且实现了以上三者与《汉表》的对应[16]；《中国分类主题词表》的编制，也是我国检索语言之间映射的重要成果。所有以上关于知识组织体系映射的研究与项目，从不同角度完善了概念映射机制与方法。在网络环境下重新构建的《汉表》与“十二五”科技支撑计划新研制的英文超级科技词表，在它们之间实现概念映射，将会是知识组织体系语义映射新的探索。

2 映射方法构架

研究项目的具体目标是将《汉表》（工程技术版）的专业概念与英文超级科技词表的规范概念（主要为工程技术部分），按照国际通用的标准规范进行映射，实现两种语言词表的语义互操作，促进中英文资源的集成揭示及跨语言检索研究和实施。同时，基于英文超级科技词表的概念关系网络及其中文规范译名，也可以对《汉表》进行扩充和完善，最终更好地通过《汉表》使英文超级科技词表在知识检索和知识发现中发挥更大的作用。

2.1 概念映射方法

参考并修订采用W3C的词表映射规则，建立映射模型，以《汉表》（工程技术版）的概念作为源（Source）概念，英文超级科技词表的概念作为目标（Target）概念，

映射采用两者概念间的以下几种主要匹配关系：

精确匹配（ExactMatch）：指两个概念含义完全相同的匹配；

向上匹配（BroadMatch）：指目标概念是源概念的上位词；

向下匹配（NarrowMatch）：指目标概念是源概念的下位词；

近义匹配（SynonymsMatch）：包括MajorMatch（主近义词）或者MinorMatch（次近义词），前者指两个概念是含义基本相同的近义词，后者指只有部分概念内涵相同的近义词。由于相近程度量化的难度较大，近似程度暂不进行区分，具体操作中也不对MajorMatch和MinorMatch两条规则进行区分，只定义为一种近义匹配；

相关匹配（RelativeMatch）：指与某一概念既不具有同义或准同义关系，亦不具有向上匹配与向下匹配的关系，但在语义上或使用中与其有密切联系的一些匹配。

建立规范映射方法标准体系，包括叙词表之间概念映射一对一关系、一对多关系。一对多关系即汉语中的一个词语对应英语中的多个词语，多对一关系即汉语中的多个词语与英语中的一个词语有映射关系，无对应关系即汉语中没有与之对应的词语等各种映射处理方法，研究构成映射关系的词语内涵的相同或相似度计算和判定算法以及语言模糊性的公允程度。

2.2 映射技术路线

在前期充分调查研究的基础上，探索《汉表》与英文超级科技词表的映射方法，研究《汉表》与英文超级科技词表以概念为基础的完整映射。对于《汉表》（工程技术版）每个概念，首先查找精确匹配概念，有则进行映射；如果无精确匹配概念，则查找近义匹配概念（类似于叙词表的等同关系），有则映射；如果既没有精确匹配，也没有近义匹配，则查找相关匹配、向上匹配或向下匹配，相关匹配通常为3-5个概念；最终结果是将《汉表》（工程技术版）22万个中文概念通过最为贴切的映射方式，都与英文超级科技词表中的英文概念进行映射。映射完成以后，对于英文超级科技词表工程类范畴下，没有建立映射关系的英文概念，将补充到《汉表》（工程技术版）中，一方面完善《汉表》（工程技术版）的概念及其关系，另一方面修正中文概念与英语超级科技词表的英文概念的映射关系，这样持续循环，以实现《汉表》（工程技术版）中文概念与英文概念的完整映射，更好地为用户服务。总体技术路线图如图1所示。

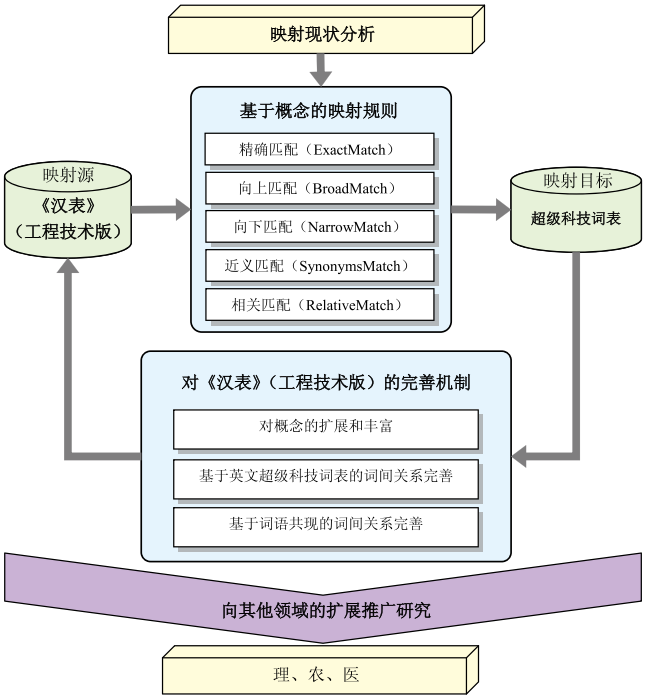

图1 课题总体技术路线图

2.3 映射规则制定

映射数据描述语言：采用XML技术进行词语的对齐和映射，用XML Schema对词语进行结构化描述和验证。表达映射的关系具体使用SKOS语言，SKOS是一种功能简便的最新知识组织体系标准语言，适合表达目前的映射关系，含有映射关系的基础数据与要开发的跨语言检索系统完全分离，有利于数据的维护和程序的升级。

计算机辅助的领域专家映射模式：使用映射工具，由计算机匹配推理运算，提出《汉表》（工程技术版）与英文超级科技词表概念间的映射关系匹配，基于《汉表》（工程技术版）概念的英文译名与英文超级科技词表的英文概念，以及《汉表》（工程技术版）的中文概念与英文超级科技词表概念的中文译

名，自动显示三大类特征的词汇，一是词汇相同、关系相同，二是词汇相同、关系不同，三是词汇不同、关系相同。系统提供关系相同与不同程度的计算功能，例如90%相同、50%相同等，具体列出相同的关系、不同的关系，然后由领域专家依据专业的不同，人工按照相同程度从高到低确定具体的映射匹配关系。

最短距离映射最近关系原则：《汉表》（工程技术版）概念具有上下位、多层次关系，英文超级科技词表概念也是网状关系，在建立概念间映射关系时，只在距离最短、关系最近的概念间建立关系，没有必要将等同的概念重复给定向上或向下匹配的关系，如果需要，只需将词表的原词间关系导入映射信息即可确定新的映射关系。

2.4 对《汉表》（工程技术版）的完善与向其他领域扩展

对概念的扩展和丰富：在《汉表》（工程技术版）与英文超级科技词表概念映射的基础上，建立中英文概念对应知识库，并且从概念出发，基于映射规则，将英文超级科技词表中没有精确匹配的英语概念及其关系，进行概念汉语翻译与组配，形成相应的汉语补充概念词汇，补充到《汉表》（工程技术版）中；对于英文超级科技词表，在完全考虑英文语言工具的基础上，同样也可以吸收《汉表》（工程技术版）的概念。最终使用户无论是用英语还是用汉语，都能达到比较一致的英语信息组织与检索效果。

基于英文超级科技词表源词表的词间关系完善：英文超级科技词表的概念关系继承自原有叙词表，关系类型丰富；《汉表》（工程技术版）的词间关系在继承原有《汉表》及其他专业汉语词表的基础上，组织专业人员构建的知识组织体系，格式统一、规范性强。借鉴英文超级科技词表中的概念间关系，进一步丰富《汉表》（工程技术版）的词间关系，研究《汉表》（工程技术版）的词间关系是否对应于英文超级科技词表，与哪一些关系对应，语义关系是否需要归并，如何为计算机提供最短语义路径计算。

基于词语共现的词间关系完善：英文超级科技词表和《汉表》（工程技术版）都采用了词语共现技术获得词间关系，作为建立词间关系的一种有效辅助手段。英文超级科技词表所采用的词汇主要是对多个领域的80万个概念进行计算，覆盖范围更宽，而《汉表》（工程技术版）则专注于对《汉表》（工程技术版）中的叙词进行计算，专业性更强，因此二者的计算数据必定有所差别，需要进一步确定其用、代、属、分、参等关系，英文超级科技词表词间关系可为《汉表》（工程技术版）确立和完善词间关系提供参考例证，例如等同关系、相关关系、等级关系等，余下的部分则通过概念限定、词语替换等对关系网络进行补足和优化，使《汉表》（工程技术版）的词间关系趋于完善。

向其他领域扩展：海量外文科技文献由理工农医等各个不同学科领域的资源构成，在《汉表》（工程技术版）与英文超级科技词表的映射研究基础上，根据理工农医的学科特点和各领域的资源特征，研究映射的普适原理和一般性规则，探索如何将汉英双语映射和选词及词间关系构建研究，从《汉表》的工程技术领域向其他领域推广，为以后全面开展理工农医等全部科学技术领域的中英文概念映射工作、推进《汉表》的全面修订提供可操作的模式和方法。

2.5 中英文双语检索模型

在《汉表》（工程技术版）与英文超级科技词表的映射关系建立以后，基于映射数据，可以设计开发跨语言的检索系统，图2是跨语言检索系统设计图。从图2可以看出，整个系统可以分为4个区域，即用户检索入口区、检索词汇转化区、知识组织区和资源区。用户通过检索入口区，录入某一检索词，例如检索词“混凝土”，则系统在程序驱动下进行词汇匹配转化，通过《汉表》概念入口或者英文超级科技词表的中文翻译入口，转入英文超级科技词表，对应其中的相应概念，例如“Concrete”，然后进入相关海量外文数据库，检出含有概念词“Concrete”的所有外文信息，然后将检索记录返回给用户；当然，如果用户使用“Concrete”进行检索，则直接进入英文超级科技词表，然后进入外文数据库，检出含有“Concrete”的所有外文信息。这样就达到了无论用户使用的是中文还是英文，均可以得到满意的检索效果。

3 结语

鉴于我国使用外文科技文献的主要用户母语是汉语，为了方便用户和机器对英文超级科技词表的利用，增加用户的检索入口，除了英文超级科技词表概念本身的中文规范译名外，需要开展《汉表》与英文超级科技词表的映射

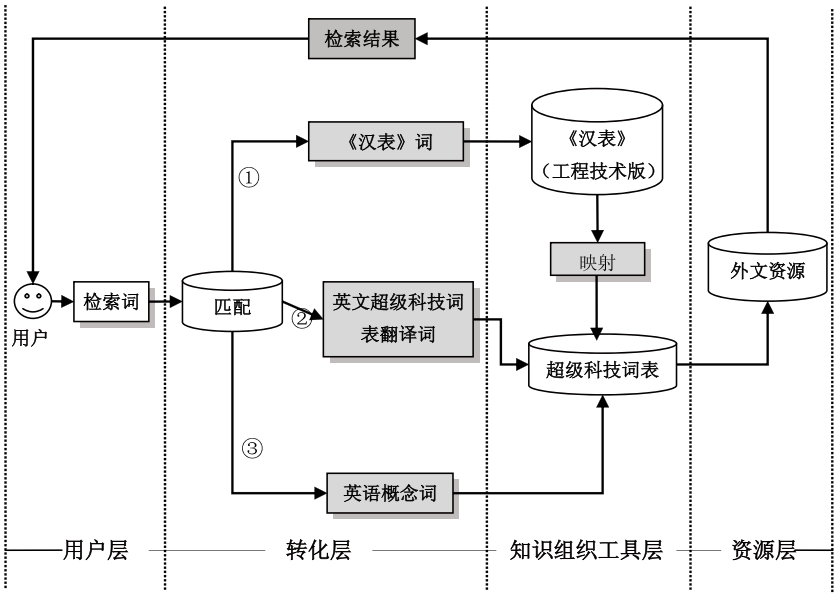

图2 跨语言检索系统设计图

研究，从而探索从《汉表》作为入口，用户同样能够检索和利用外文科技文献资源，达到对外文科技文献知识检索与知识发现的目的。

为逐步推进面向中外文文献资源的联合，同步知识组织工作，实现中英文资源的集成揭示和跨语言检索，本文设计了两种语言词表的概念语义映射方法，提出了从《汉表》（工程技术版）向英文超级科技词表的概念映射模型，设计了概念映射的数据描述、机器辅助和最短距离的映射规则，给出了中英文双语检索模型的应用前景。

“《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”课题处于开始阶段，随着《汉表》（工程技术版）与英文超级科技词表的编制完成，本项设计也将进入具体映射与实施阶段，在具体实践过程中，相信会有更多问题需要研究和探索。

参考文献

[1] 常春.信息检索系统中的映射特征[J].情报杂志,2009(3):141-143.
[2] ISO 25964-1:2011. Information and documentation -- Thesauri and interoperability with other vocabularies -- Part 1: Thesauri for information retrieval [S/OL]. [2012-08-18]. http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657.
[3] 国家“十二五”科技支撑计划项目启动会召开[EB/OL]. [20120-08-18]. http://www.nstl.gov.cn/NSTL/facade/news/newsInfo.do?act=toNewsContent&id=89383.
[4] 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.
[5] 中国科学技术情报研究所.汉语主题词表：自然科学（增订本）[M].北京:科学技术文献出版社,1991.
[6] 贺德方.《汉语主题词表》的回顾与展望[J].情报理论与实践,2010,33(2):1-4.
[7] 曾建勋,常春,吴雯娜,等.网络环境下新型《汉语主题词表》的构建[J].中国图书馆学报,2011,37(4):43-49.
[8] Renardus: Academic Subject Gateway Service Europe [EB/OL]. [2012-08-18]. http://www.ukoln.ac.uk/metadata/renardus/.
[9] Terminology Web Services [OL]. [2012-08-18]. http://tspilot.oclc.org/resources/standards.html.
[10] High-Level Thesaurus Project (HILT) [OL]. [2012-08-18]. http://hilt.cdlr.strath.ac.uk/index.html.
[11] UMLS [OL]. [2012-08-18]. http://www.nlm.nih.gov/research/umls/.
[12] SUMO Search Tool. From the SUMO ontology by means of mappings to WordNet synsets [OL]. [2012-08-18]. http://sigma.ontologyportal.org:4010/sigma/WordNet.jsp?word=cat&POS=1.
[13] 戴剑波,侯汉清.文献分类法自动映射系统的构建——以《中国图书馆分类法》和《杜威十进分类法》为例[J].情报学报,2006,25(5):594-599.
[14] 赖院根,曾建勋.期刊论文与专利文献的整合框架研究[J].图书情报工作,2010,54(4):109-112.
[15] 常春.基于叙词表映射的农业跨语言检索系统设计[J].情报学报,2008(增刊):294-296.
[16] 张琪玉.我国情报语言20年来的进步与向21世纪前进的目标[J].图书馆,1999(4):1-7.

作者简介

常春（1966-），男，博士，研究馆员。研究方向：为信息组织。E-mail: changchun@istic.ac.cn

Architecture Design of Concept Mapping from Chinese Thesaurus to ESVST

Chang Chun, Zeng Jianxun, Wu Wenna, Song Peiyan, Liu Wei, Deng Panpan / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: In the Internet era, mapping of knowledge organization systems is helpful to information retrieval and access. The article designed the concept mapping model from the Chinese Thesaurus (Engineering Edition) to English Super Vocabulary of Science & Technology (ESVST). It includes the semantic mapping method of concept between Chinese and English vocabulary, mapping data description, machine-aided mapping, and the shortest distance mapping rules. The article also discussed the application prospects in the bilingual retrieval model. Based on conceptual relationship network and Chinese translation of ESVST, it proposed the methods to expand and improve the Chinese Thesaurus, and made it easy on mapping, promoting ESVET to play a greater role in knowledge retrieval and knowledge discovery.

Keywords: The Chinese Thesaurus, English Super Vocabulary of Science & Technology (ESVST), Mapping

（收稿日期：2012-09-03）