网络环境下新型《汉语主题词表》的构建*

曾建勋　常　春　吴雯娜　宋培彦

摘　要　网络环境下《汉语主题词表》的表现形态、编制维护方式和功能定位发生了深刻变化，其构建方法也需要随着时代的发展而创新。构造“基础词库—范畴体系—概念关系网络”三级联动机制，形成了新型《汉语主题词表》的构建方式。基础词库以元数据为框架，吸收各类词汇，为《汉语主题词表》编制提供丰富和可靠的素材；范畴体系将词语纳入统一的框架结构，使词语各入其类；通过概念的遴选和属性描述，以概念为中心，建立并丰富概念间关系。图1。参考文献11。

关键词　汉语主题词表　网络环境　叙词表　基础词库

分类号　G254.2

ABSTRACT　The representation and means of compilation, maintenance and functional orientation of the new *Chinese Thesaurus* in Web environment have changed profoundly and its means of construction also needs to be innovated along with the progress of the times. As a new approach to the construction of *Chinese Thesaurus*, the three level linkage mechanism is proposed, which includes the fundamental lexicon, category system and concept relationship network. Based on uniform metadata, the fundamental lexicon provides abundant and rich material. The category system incorporates a framework for each word and the relationships among concepts are built according to selected concepts and their properties. 1 fig. 11 refs.

KEY WORDS　*Chinese Thesaurus*.　Web environment.　Thesaurus.　Fundamental lexicon.

CLASS NUMBER　G254.2

1　引言

《汉语主题词表》(以下简称《汉表》)是我国第一部大型综合叙词表，1980年出版第一版[1]，1991年出版《汉表》自然科学增订本[2]。作为我国图书情报界集体智慧的结晶，30年来《汉表》在我国图书情报事业中发挥了重要作用，成为实现知识组织的有效基础工具[3]。随着人类步入网络时代，数字信息资源呈指数级增长，网络技术飞速发展，计算机处理能力日益增强，用户的需求也从海量信息检索向有效知识获取转变[4]。网络环境下从信息服务向知识服务转型过程中，《汉表》的表现形态、编制维护方式和功能定位都将发生深刻的变化，《汉表》的构建方法也需要随着时代的发展而创新。

2　新型《汉表》的建设方案

2.1　网络环境下新型《汉表》的形态特征

网络环境下，新型《汉表》是由基础词库、核心词库、叙词词库等构成的知识组织系统，将充分考虑用户检索用词和文献主题的准确表达，使叙词表词库与自然语言尽量一致。整个概念体系是机器可读和可理解的，采用RDF、OWL或SKOS机器语言表达概念关系，并以立体方式展现分布在多个树状结构中的叙词，为每个叙词设置超链接，揭示立体网状结构中的不同节点之间的关联关系[5]，构成由简到繁的知识地图和初级本体级别的语义关系。此外，新型《汉表》的编制与维护将充分发挥用户的积极性，采用在线叙词表编制平台，提供基于知识组织的术语服务，加强与用户的交互。采用智能化和

*　本文系国家社会科学基金资助项目“网络环境下叙词表的编制模式与应用方式研究”(编号：10BTQ048)研究成果之一。

可视化技术,提供更多人性化的应用方式,并建立动态变化的专业知识体系更新机制。

2.2 总体建设思路

构建新型《汉表》需要以自然语言词汇为基本单元,以概念为核心,实现词汇术语的整合。在充分利用多年对科技文献数据库建设的成果,借鉴传统汉语叙词表的词间关系的同时,开发概念关系构建工具,通过大规模语义计算展现概念间的共现语义关系,并联合专业人员进行主题概念的遴选、概念关系的构建审核,进行语义关联,并在统一的范畴体系下,对概念进行范畴归类,构建以概念为核心的“概念关系网络”,建设高度整合的新型《汉表》。

新型《汉表》编制的技术路线是调研和采集已有知识组织体系及其相关元数据集,与从文献数据库中抽取的关键词和用户检索词等一起构成来源素材;通过词形规范、词义规范等遴选规范形成概念;在借鉴综合性词表和专业词表概念语义关系的基础上,借助词共现,建立概念间相关属性关系;同时建立涵盖全学科的范畴体系,并对概念进行相应范畴体系归类,最终形成新型《汉表》,其构建框架如图1所示。

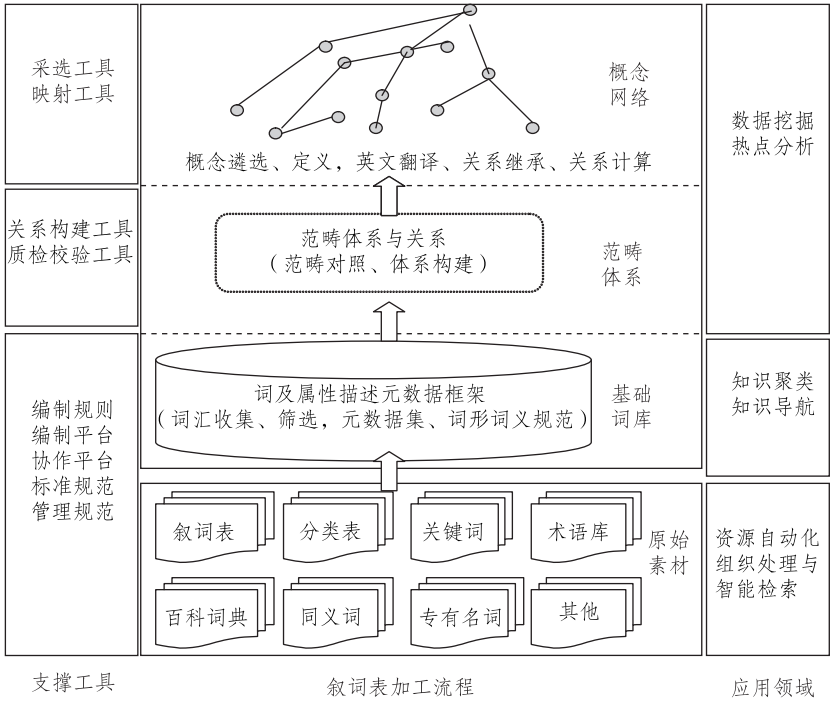

图1 新型《汉表》构建框架

3 基础词库建设

3.1 词汇词源获取途径

按照词汇的来源渠道,基础词库的词汇主要分为继承词汇来源和扩展词汇来源。继承词汇来源主要指现有的各种类型的传统叙词表、分类法、同义词集、术语列表等,是规范化词汇和概念的重要来源。扩展词汇来源主要包括各专业数据库关键词、各类领域词典、专有名词术语、专业网站专业词汇、领域用户检索词汇、百科全书等。因此,除了将传统《汉表》中的10万

多条词语作为基础词库的基本来源,还要广泛调研国家编制的综合性词表和各个学科领域的各类专业词表等知识组织体系,全面收集规范化的词汇和术语及其相关属性描述信息,同时,从数据库和专业网站采集相关关键词等科技词汇,作为构建基础词库的原料和素材。

3.2 基础词库元数据框架

多来源词汇及其描述属性信息可能有着不同的存储格式,如数据库格式、纯文本格式、XML 格式或者网页格式,元数据项也各有区别。基础词库需要集成、整合和管理这些异构的来源数据,一方面要尽可能以规范的格式进行存储,如关系型数据库或 XML 格式;另一方面,还要保留这些描述信息与词源的关联,需要对基础词库建立统一的元数据框架,准确定位相关词汇和概念,并描述其属性信息。

为了客观、真实地描述词汇的所有来源信息,必须详细分析经过遴选确定的各种来源素材及其元数据结构,在此基础上构建统一规范的元数据框架结构。首先,对各种来源词表及其元数据结构进行分析,参照都柏林核心元数据 DC 进行规范化转换;然后,把收集遴选的关键词和检索词集按照元数据标准结构进行存储。这样,按照统一的元数据框架结构标准,由来源叙词表、术语表、作者关键词和用户检索词集等构成基础词库,由人名、地名、机构等来源素材构成科学名词的规范文档。它们均是以词汇术语为中心,并包括来源素材的各种属性和来源信息。

元数据框架标准包含词汇的唯一 ID 标识、概念标识、描述性信息、关联信息、来源信息、使用评价信息、加工历史纪录等元数据项。

3.3 基础词库词汇的加工遴选

基础词库词汇收录需要制定统一的规范,为词汇筛选提供共同遵循的标准。词汇选择包括选择词源和在特定词源中选择词汇两个层面。需要对词源的适用性、权威性、学科性进行评价,研究基础词库选词的原则,如从词汇使用情况、词汇的知识内容关联性、词形规范、语义清晰度、词汇专业性等角度对词汇进行综合评价及筛选[6]。

在选择词汇时,按照词形规范标准,首先剔除来源数据中的非词语,继而划分为普通词和专有名词,这样把普通词中词形相似的词语集中到一起,主要解决词形变异、去重问题,可以为用户提供大量的检索入口词。同时,把专有名词划分为人名、地名、机构名、产品型号等形式类别,用于表示特定事物。对于大量的异形同义概念词,需要借助同义词词典、术语表等,把这些同义词汇聚到一起,构建同义词词群。在基础词库整合过程中,需要研究相同概念的不同表达,研究表达同一概念的多种词语(即同义词)及其词形变体的通用术语的选择,如缩略语、简称、俗称、惯用名等,以解决同形异义、同义异形的规范表示问题。

基础词库的规范化处理需要借助自动分词、词性标注、词频统计、新词发现、信息抽取、自动聚类等中文本体信息处理的最新方法和技术。

3.4 基础词库管理、维护与更新机制研究

词汇是概念的载体,随着科技进步和社会经济发展,新的学科领域和技术门类大量涌现,新词不断产生,词义不断引申,科学术语的产生、发展和演变明显加快,加强对动态词汇的研究是基础词库建设的一项重要工作,需要不断发现新术语,建立基础词库词汇维护、管理与更新机制;另一方面,基础词库的动态更新势必会增加词间关联结构的复杂度,引发多词间关联结构的变化,通过继承概念层中概念间的关联,向以概念为中心的元数据仓储中增加新词及其语义关联,进而将机器可识别的概念关系在专业领域内相对集中,在领域间互联互通,保证基础词库内容的科学性、逻辑性和及时更新。

网络技术为新型《汉表》的维护提供了良好的技术手段。例如,利用 Web 2.0 技术中的社会标注法(social tagging)和大众分类法(folksonomy)可以加强与用户的互动[7],吸收用户的修改意见;基于网络协作平台,词表编制人员可以

对词语进行在线讨论、修订和分工管理;采用可视化技术,可以直观展示词间关系,便于发现和修改错误信息,等等。

3.5 基础词库与叙词库的关系

新型《汉表》所包含的核心词汇是有限的,为了提高其与用户语言的匹配效率,需要建立自然语言与规范语言之间的关联关系。基础词库词汇数量远大于新型《汉表》的叙词库术语数量,基础词库中的词汇是新型《汉表》中叙词库词汇的母体,语义关联密切。如基础词库中的词汇是新型《汉表》叙词库词汇的近义词,即同一概念的其他表达方式;基础词库中的词汇概念颗粒度更小,表现为词汇构词中有着更多的限定,它们与新型《汉表》叙词库词汇之间可以是隶属关系,即叙词的下位词;基础词库中的词汇更为具体,是实例或专有名词,在新型《汉表》中一般不会大量收录,这类词与新型《汉表》叙词之间是概念—实例的等同关系。

除了同义关系外,基础词库中的词汇与新型《汉表》中叙词可能还存在属分关系、参照关系、反义关系等关系类型,有必要对词汇映射机制进行研究,确定一套规范化、形式化的映射关系表示方法。由于基础词库词汇量巨大,必须研究基础词与概念间自动映射方法,包括借助中间工具,即基于现有的科技词典和机读词典,对自然语言中的词汇进行语义元素分析,构造同义词、近义词、反义词、上下位词等主题词群,对自然语言词汇和叙词表中的叙词进行语义相似度计算,映射到叙词库中,形成以核心概念作为叙词、以非核心概念作为非叙词的词间映射关系网络[8]。

4 新型《汉表》范畴体系建设

范畴是概念的一个重要属性,用来说明概念所适用的学科领域[9]。范畴体系实际上是一个学科/主题分类体系,一般以树形结构展开,用以展示范畴之间层层隶属的关系;叙词表的词族树则是把具有相同语义类型的概念按照从一般到具体的方式组织成的层级结构。这对于文献信息的主题聚类、分类组织及层级浏览具有重要意义。同时,范畴体系的建设也是通用本体建设的基础,有利于控制通用本体的维度和颗粒度,便于以后建立通用本体与新型《汉表》概念的映射关系,解决因学科交叉、表达产生的维(粒)度不同、冲突和重叠等方面的问题。

4.1 范畴体系构建原则

目前,我国分类思想大多延续《中国图书馆分类法》、《中国图书资料分类法》的类目体系,传统《汉表》的范畴表主体类目结构也与《中国图书资料分类法》相一致,其他一些专业词表范畴也基本与之吻合。所以,针对我国主要分类表的应用现状和叙词表的范畴设置情况,新型《汉表》范畴体系应以《中国图书资料分类法(第四版)》分类体系为基准,同时参考《汉语主题词表—范畴表》、《国家学科分类标准》以及《中国分类主题词表》等,面向多个学科统一构建。其范畴类目设置根据社会、经济、科技的发展现状与发展趋势以及对应学科的文献量、词汇数量、学科交叉特征等,力求达到思想性、科学性和实用性的统一。

若想既发挥叙词表词族中属分关系相对明确的优势,同时又改善其在主题聚类中的不足,就需要强化范畴表的作用,进行范畴表的扩展。根据对范畴表扩展程度的不同,可以分成两种情况:一种是将范畴体系加细加深,每个范畴中仍包含多个主题;另一种是将范畴和主题完全合一,全部范畴和概念统一形成一个等级结构,即达成分类主题一体化。如果考虑到未来词表的本体化转换和应用,就需要对叙词表的关系类型进行细化,以区分不同等级关系,达到根据不同需要(主题相关、语义类型一致)展示不同的层级结构的要求,为基于叙词表的本体化奠定基础。可以把以上两种情况看成范畴表发展的不同阶段,范畴表越细对基于主题的文献聚类越有利,当范畴表细化到极致时,就达到了分类主题一体化。

4.2 范畴体系构建与调整

根据范畴表构建原则,需要确定范畴表的

规模、层级,参照体系的设置方法、类目筛选方法、类目体系调整方法、类目命名和编码规则等。

对类目体系进行对照分析,在专业范畴表和分类法之间建立对照关系,区分不能映射的类目,结合基础词库词汇类目分布、类目文献数量,提出新型《汉表》范畴体系类目及其层级关系的调整思路。在《中国图书资料分类法》主体类目基础上,进行类目筛选和调整,对其类目体系进行细化、删减、新增、合并、移动等,调整形成范畴体系的基本框架,并考虑学科交叉特征,允许设置有主辅区分的交替类目,并建立类目间参照关系。在具体构建范畴表时,还可考虑将叙词表中的词族结构纳入范畴体系,将范畴体系进行扩展,逐步实现分类主题一体化。

为了进一步完善范畴体系,可以在基础词库中按比例随机抽取部分词汇,利用新型《汉表》范畴表对词汇进行分类试验,对分类过程进行记录,对分类结果进行统计分析,据此评价范畴表的适用性,研究范畴表类目分布与文献分布的关系。在此基础上对范畴表结构和类目进行调整优化,对类目定义和类目参照体系进行完善。

4.3 概念归类

在统一的范畴体系下,借鉴词汇原始范畴信息、来源信息、词间关系信息,解决来源词表概念的归类问题;依据在文献中的学科分布特征,或通过共现、共篇或共引等方法,借助各种计算机聚类分类手段和专家的人工判断,确定所属的范畴类目。传统《汉表》的词语映射到统一的范畴体系下,某些词语允许同时归入 2—3 个典型范畴。

5 新型《汉表》的概念建设

在统一的词表集成框架体系下,开展新型《汉表》的概念表达研究,确定新型《汉表》的概念体系。

5.1 概念归一与遴选

新型《汉表》关注的焦点是概念,而不是词汇、名称或术语。词汇的收集和组织以表达概念的涵义为目的,将具有相同概念、来源不同的词汇及其变体通过概念标识联系起来,并在新型《汉表》中通过概念定义、概念语义类型、概念内的关系(表达同一概念的各种词和词汇的关系)以及概念被使用的信息等多种方式表达概念涵义。在概念的定义方面,研究新型《汉表》中概念的定义,不同来源词表相同词汇所表达概念的定义、标示和消歧。在概念的语义类型方面,研究新型《汉表》语义类型的具体设置、类型标示、语义类型数量的确定等。在概念内关系方面,研究新型《汉表》集成过程中相同概念的不同表达、表达同一概念的多种词语(即同义词)及其词形变体的词汇选择、概念名称和其他词汇的连接、概念内词汇关系的多级表达结构模式。在概念属性方面,研究新型《汉表》如何整合和表达概念间各种属性关系。

对于来源于现有知识组织体系的概念,在不同知识组织体系之间实现等同概念之间的关联,同时尝试等同关系的合并、相关关系的合并,为新型《汉表》概念集成奠定基础。概念是通过词汇表达的,等同概念关联需要从词汇入手。例如,一是词汇相同、关系相同;二是词汇相同、关系不同或部分相同;三是词汇不同、关系相同。等同概念关联的部分工作可以借助计算机完成,系统应该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,然后由人工按照相似度从高到低处理所有有问题、有重复的词间关系,就此找出等同的概念,实现等同概念关联。

5.2 概念属性描述

对于每个概念,通过统一的概念描述模型进行规范化。以概念为中心,尽可能准确而全面地描述词语的各个属性信息。例如拼音、英文译名、定义、同义词、知识元、概念间关联、范畴号、形式分类等。简单知识组织体系(SKOS)可以作为描述新型《汉表》概念属性的有效工

具,便于实现概念的准确描述。为了统一对概念的理解,对新型《汉表》中的每个概念,尽可能增加对应的、可参考的概念注释,可以是源词表中概念注释的继承,也可以是相关词典的定义解释或者相关的参考链接,既为同义概念认定、概念关系继承提供参考,也为上下文、同现关系确立提供支持。同时可对新型《汉表》进行基于概念的英文翻译,尽可能优选国际学术界及相关专业词表中的正式主题词作为概念的规范英文名称,以支持英汉双语检索。

6 新型《汉表》概念关系的建设

叙词表概念间关系主要包括等级关系和相关关系,可继承传统词表的概念关系,并借助新的信息技术丰富概念关系数量和类型。

6.1 对传统叙词表概念关系的继承

传统《汉表》具有相对丰富和可靠的概念关系,为编制新型《汉表》提供了良好的基础,需要最大程度地加以继承,保持《汉表》的系统性和稳定性;同时,要根据科技领域的最新进展,甄别和去除那些过时的词语或概念关系,并补充新的词语或概念,对概念的关系进行局部调整和更新。要尽可能排除由于关系集成带来的关系不一致甚至冲突的概念关系。需要通过叙词表的统一计算机化表示形式、规范和技术接口,开发词表转换的适配器、跨词表的语义分析工具、规范化主题词表的应用程序访问接口等,对概念关系集成后所形成的新的概念关系进行关系矛盾性、冲突性、一致性检查和梳理,继承重要的等级关系和相关关系。

6.2 概念关系新建及逻辑检查

概念间关系构建新方法研究。新型《汉表》概念关系的构建,需要研究知识概念关系的形成、表达和演化,研究利用计算机或通过大规模语义计算进行概念关系发现的方法。一方面,可以充分借鉴和继承已有叙词表的概念关系,对多个词表间的同一概念进行语义关系映射关联,形成跨领域、多来源的主题词表集成的概念网络体系。另一方面,使用候选词汇,统计每个词与其他词在文献中的共现频率,表现词汇概念同现关系信息;选取各专业相关的比较权威的专业文献,利用章、节各级标题间的上下位及同级结构,发现词汇概念树状结构关系;还可利用字面相似度、语义计算、关联规则等提供一些词汇概念的等级关系或相关关系的参考信息。最后,组织专业领域的专家队伍,按照叙词表编制规则和标准,对概念相互关系进行逐一思考和确认,区分并明确等级关系与相关关系,并进行相应关系的逻辑检查和修正。

通过共现聚类发现概念间关系属性。网络环境下,叙词表词间关系的建立,可以充分利用海量数据库素材和上下文语境。语料库存放了大量真实使用的语言材料,提供了词语使用的语言环境,将这些概念放在语料库中进行两两组合,采用隐马尔科夫模型 HMM 统计其在语料库中的条件概率和共现频次。共现概率高,说明词间关系比较稳定;共现概率低,说明有可能出现新的词间关系或词间关系错误。量化的数据有利于提高词间关系判定的准确性,并发现新的词间关系,使词间关系更为准确和丰富[10-11]。

概念间关系逻辑检查纠正。为了集成和构建概念间的知识关系,在梳理来源词表关系,发现关系集成可能出现的矛盾冲突的同时,研究概念关系逻辑验证方法和自动修正算法。《汉表》包含多个专业,包含大量的专业劳动和知识活动。例如专业术语的确定、专业范围内概念相关关系的确立、等级关系的设定等,必须由专业领域研究人员参与,对专业领域知识结构进行总体指导和审定;对每个概念的相互关系,必须组织领域专家逐一进行确认和构建;最后将每个专业的叙词表进行合并,通过计算机检查梳理,对所反映出来的有冲突和矛盾的关系进行人工纠正。

7 结语

网络环境下,《汉表》的构建方法发生了重大变化。首先,编制词表所需要的基础词库来

自各个专业领域和大型的文献数据库,在统一的元数据框架下建立范畴和语义关系,词汇数量多、覆盖面广、规范性强,将大大提高叙词表的代表性和客观性;其次,构造了“基础词库—范畴体系—概念关系网络”三级联动机制,扩大了词语入口,用户可以以自然语言、范畴或语义关系作为使用接口,知识检索、知识导航、知识标引等更为便利和灵活,将提高《汉表》的易用性;再次,将大规模语义相似度计算、共现聚类、可视化等自动处理技术与领域专家知识相结合,进行概念的获取和审核,语义关系更为全面和丰富,编表的效率也将有很大提高;最后,借助于网络平台,用户全程参与基础词库建设、范畴归类以及叙词表的维护等各个阶段,加强与用户的交互,改变传统上过于依靠领域专家、较少考虑用户需求所造成的局限,体现“以用户为中心”这一思想,新型《汉表》将具备良好的用户基础。新型《汉表》在吸收不同知识组织体系优点的基础上进行改进,在网络环境下对传统叙词表进行创新和发展,将拓宽《汉表》在知识揭示、知识导航、知识学习、智能检索等方面的应用。

参考文献:

[1] 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.

[2] 中国科学技术情报研究所. 汉语主题词表:自然科学[M]. 增订本. 北京: 科学技术文献出版社, 1991.

[3] 贺德方.《汉语主题词表》的回顾与展望[J]. 情报理论与实践, 2010(2):1-4.

[4] 曾建勋, 常春. 网络时代叙词表的编制与应用[J]. 图书情报工作, 2009 (8): 8-11.

[5] OWL Web Ontology Language[OL]. [2010-06-25]. http://www.w3.org/TR/owl-features/.

[6] 常春, 吴雯娜. 网络时代专业叙词表选词规则实践与讨论[G]//全国第五次情报检索语言发展方向研讨会论文集. 北京: 国家图书馆出版社, 2009: 107-113.

[7] 毛军. 元数据、自由分类法(Folksonomy)和大众的因特网[J]. 现代图书情报技术,2006(2): 1-4, 9.

[8] 赖院根, 吴文娜. 基于叙词表的概念语义相似度计算[J]. 图书情报工作, 2009 (8): 21-24.

[9] 张琪玉. 情报语言学词典[M]. 北京:北京图书馆出版社,2000.

[10] 常春, 赖院根. 基于文献标题词汇共现获取词间关系研究[J]. 图书情报工作, 2009(8): 17-20.

[11] 常春, 吴雯娜, 曾建勋. 基于后方一致获取词间关系[J]. 情报科学, 2009(7): 1085-1088.

曾建勋 中国科学技术信息研究所信息资源中心主任、研究馆员。通讯地址:北京复兴路15号。邮编:100038。

常 春 中国科学技术信息研究所信息资源中心研究馆员。通讯地址同上。

吴雯娜 中国科学技术信息研究所信息资源中心副研究馆员。通讯地址同上。

宋培彦 中国科学技术信息研究所信息资源中心助理研究员。通讯地址同上。

(收稿日期:2010-09-21;修回日期:2010-11-01)