●贾君枝

《汉语主题词表》转换为本体的思考*

摘　要　叙词表具有清晰的语义结构,便于从中抽取概念和关系,目前已有十多种叙词表被用各种方法转换为本体。叙词表转换为本体的难度依赖于叙词表本身的特点。我国《汉语主题词表》有自身的一些特点和不足,转换过程中应对叙词及其存在关系明确界定,把握四个方面的基本原则。以特定应用或特定领域为核心,《汉语主题词表》在转换为本体的过程中,需要调整叙词表中概念之间的关系:一是核心概念的选择;二是概念之间关系的改造。表 1。图 5。参考文献 9。

关键词　汉语主体词表　叙词表　本体　转换原则
分类号　G354

ABSTRACT　Thesauri have clear semantic structures, and can be used to extract concepts and relations. There are more than ten thesauri which have already been used to be converted into ontologies. The difficulties of converting thesauri into ontologies depend on the characteristics of thesauri themselves. *Classified Chinese Thesaurus* has its own disadvantages, and there is a need to clarify terms and their relations before the conversion. The author proposes some principles and discusses key relations for the conversion. 1 tab. 5 figs. 9 refs.

KEY WORDS　*Classified Chinese Thesaurus*.　Thesaurus.　Ontology.　Conversion principle.
CLASS NUMBER　G354

许多学者提出对现有的词表如字典、叙词表、分类表等进行共享重用,在此基础上构建 Ontology,并在实践中进行了尝试。由于叙词表较其他词表而言,具有更清晰的语义结构,便于从中抽取概念及关系,目前已有十多种叙词表被用各种方法转换为 Ontology。从构建方法的差异性可分为两大类:一类是直接对现有的词表采用 XML/RDFS 语法进行形式化描述,不对词表作任何调整,如 Eman 提出采用本体语言 OWL 对叙词表进行描述输出[1],国内毛军提出采用 RDFS 定义叙词[2]以实现从叙词表到本体的转换;另一类是基于本体论思想,对词表的概念进行增删,调整概念之间关系,对词表进行改进以生成新的本体。如联合国粮农组织成立了农业本体论服务项目小组,采用自动本体学习系统,通过机器学习自动抽取概念间关系将 *Agrovoc* 叙词表转换为农业本体。Qin 和 Paling 采用 Ontoligua system 探索将 GEM(教育资料网关)中的受控词表转换成 Ontology[4]。阿姆斯特丹大学的 Wielinga 等运用艺术和建筑叙词表(AAT)的受控词汇表描述古代家具本体等[5]。

叙词表和本体都在基于知识理解的基础上构建,都涉及到知识的分类及语义关系的构建,二者有融合的前提。但二者构建目的不同,存在着一定的差异,叙词表作为规范化的术语词表,是为提高计算机检索效率而制定的。本体以概念和概念之间关系的建立为核心,注重计算机的形式化描述,以计算机能够理解的语义内容为前提。若使叙词表较为准确地转换为本体,必然考虑其特点,需要在进一步的数据清洗及语义关系调整的基础上进行,这样才有一定的实际意义。

1　基本转换原则

叙词表转换为本体的难度依赖于叙词表本身的特点,如果词表专业强且严格定义了语义关系,转换过程就相对容易,如 AAT 叙词表定义了严格的上下位继承关系[6],可直接使用其已有的继承关系来构建本体关系。我国《汉语主题词表》作为一部大型综合性科技检索工具,收词范围包括自然科学、医学、农业、工程技术等各学科领域的主要名词术语,是主题标引、检索和组织目录、索引的主要工具。如果有效地运用现有叙词表的成果来构建汉语领域的本体,既省时省力,又能将叙词表的作用进一步扩大,为处理中文形式的网络资源的语义问题提供可能。但考虑

* 本文系国家社会科学基金项目"汉语框架网络知识本体构建研究"(06CTQ004)成果之一。

到《汉语主题词表》本身的特点，比如词表内容专业性不强，语义关系定义欠严格，内容不很明确等，本体的转换尤其是自动转换有一定难度。需要分析叙词表中具体的叙词款目和相关的语义关系，针对性提出解决方案。在确定转换之前，应把握以下4个原则。

（1）围绕特定的应用或领域进行转换，而不是对整个叙词表作转换处理。《汉语主题词表》体积大，内容多，语义复杂性强，如果通盘转换，会有以下问题：手工转换会导致工作量巨大，自动转换过程中会遇到较多特殊的无法用专家所制定的规则能够解决的问题。因此需要确定本体建设的目的、范围、用途和使用者，在本体需求分析的基础上，确定叙词表转换的范围。如阿姆斯特丹大学确定以古代家具的本体构建为核心，建立描述古代家具的模板，涉及到与生产相关的描述、物理特性描述、功能特点描述、管理内容描述等25个描述元。其中功能属性的描述取值则源于AAT叙词表，运用AAT叙词表来成功构建古代家具的本体内容[7]。这样针对性较好，省时省力，既实现了叙词表的增值，又为本体构建的合理提供了一定的依据。鉴于此，首先应根据本体的需求分析确定特定的应用或者领域范围，再依据《汉语主题词表》中的范畴索引，作为核心概念集中提取的依据。

（2）明确概念与术语的差别，应基于唯一概念，而不是自然语言术语。Dahlberg（1981）指出概念是由指示物（referent，代表任何物质或非物质物体、活动、特性、空间、主题及事件）、术语（term，指示物外在沟通形式）和特征（characteristics，指示物结构的一种陈述）三者构成[8]。概念是知识的基本单位也是思维的最小单位，词是概念的外部形式。叙词表作为规范化的术语集合，只注重词汇的控制，缺乏对词汇进行概念层次的抽象，缺乏概念的明确表达和表述的一致性。《汉语主题词表》收录的社会科学和自然科学的术语较为有限，缺乏对各个领域内部概念的系统性描述，无法揭示各个领域内的共同理解知识。比如法律领域，在术语描述中，“犯罪分子、犯罪集团、同案犯、累犯、惯犯、流氓”都作为叙词表中的术语，如果从概念层次上理解，他们具有相同属性的特点，实际上都可概括为“犯罪分子”。《汉语主题词表》部分内容描述较粗泛，缺乏对概念属性特征的描述，比如“抢劫”一词，在叙词表中仅提到该概念，缺乏对这一概念特点的具体描述。因此考虑到叙词表与本体构建思想的差异性，需要在术语到概念的转换过程中，注重术语的提炼及其术语的归纳、合并及展开性描述。

（3）建立严格的概念之间的关系。《汉语主题词表》语义关系主要揭示主题词之间的关系，各叙词款目一般根据语义分析情况，通过设置用、代、属、分、参多种参照系统和索引方式建立词间关系，形成语义联系。这种语义关系含糊、较宽泛，明显表现为主观随意性。比如等级关系揭示叙词之间的上下位层级关系，包含属种关系、整体与部分关系、集合与个体层级关系，像“诉讼程序”与“起诉、上诉、审理、执行”表现为整体与部分关系，却归类于属种关系，“人口”与“常住人口、城市人口、非农业人口、平均人口、农业人口、现有人口、暂住人口”等都归于属种关系，明显表现为分类不一致，属下各个类目之间不是并列关系，重复交叉。本体构建概念之间关系的目的在于通过对关系形式化描述，便于计算机进行推理，因此需要对叙词表中不是很严格的语义关系进行明确限定，建立明确的属种关系、实例关系、整体与部分关系，推理系统才能推导出上下位类，类与实例等。

（4）采用计算机进行形式化描述。本体作为描述概念及概念关系的知识模型，必须与特定的描述语言相关联才能发挥作用。W3C组织推出与本体相关的最新标准语言OWL，能够被用于清晰地表达概念的含义以及这些概念之间的关系，相对XML、RDF和RDF Schema拥有更多的机制来表达语义，能通过定义类及类的属性来形式化一个领域，声明和定义对象和对象的属性，以及在OWL形式化语义允许程度上对类和对象进行推理。图1以“漫画”为例，采用OWL语言描述类、子类、实例及相关属性，便于计算机理解。

```
<owl:class rdf:ID='漫画'> (类定义)
  <rdfs:subClassOf rdf:resource='& 美术'/>(子类定义)
</owl:Class>
<漫画 rdf:ID='爸爸不在家时'/> (实例定义)
<漫画 rdf:ID='勤俭持家'/>
<owl:ObjectProperty rdf:ID='价格'> (属性定义)
  <rdfs:domain rdf:resource='#漫画'/>
  <rdfs:range rdf:resource='#price'/>
</owl:ObjectProperty>
```

图1 OWL语言形式化描述

2 概念之间关系的调整

以特定应用或者特定领域为核心，《汉语主题词表》在转换为本体过程中，面临两个问题：一是核心概念的选择，一是概念之间关系的构造。核心概念的选择依赖《汉语主题词表》范畴索引的叙词款目，围

绕具体的应用对其进行增、删、改而形成。概念之间关系的定义难度较大,需要在审视原有叙词表关系的基础上,结合本体对关系的界定原则,进行最终修正,需要投入的人力及时间较多。

2.1 本体的基本关系

目前本体从语义层次上,主要定义了4种基本关系:上下位类之间的关系(属种关系)、总类与分类之间的关系(整体与部分关系)、类与实例的关系(实例关系)、类与属性的关系(属性关系)。用户在构建本体时,还可根据具体应用过程构造适合特定应用的关系。

(1)属种关系(kind-of)表示概念之间的继承关系,即父类与子类的关系,它将具有某种共同属性特征的资源归入一类,大类下设置多个子类。如果两个类具有属种关系,则来自子类的实例可推理出一定属于父类的实例。如图2,叙词表中"美术"的等级层次划分可直接转换为该类型关系。

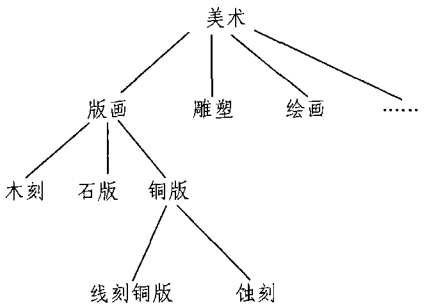

图2 属种关系

(2)整体与部分关系(part-of)表示概念之间整体与部分的关系。因为子类与总类具有部分的共同属性特征,所以子类的部分实例可能来自总类。如图3,"唐宋八大家"作为总类,其子类分别为"韩愈、柳宗元、欧阳修、苏洵……"。

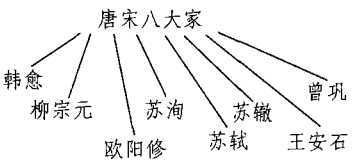

图3 整体与部分关系

(3)实例关系(instance-of)表示概念实例与概念之间的关系,即个体作为类的成员与类建立关系,其中类的共同属性特征在个体中都有体现,个体也可定义自己的特有属性特征。比如图4丰子恺漫画《爸爸不在家时》、《勤俭持家》是"漫画"的实例,同样也是"绘画"的实例。

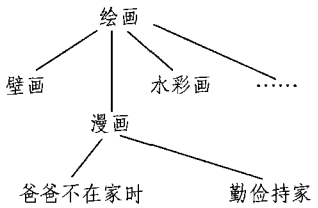

图4 实例关系

(4)属性关系(attribute-of)表示某个概念是另一个概念的属性,即某个类所具有的属性特征。比如图5"漫画"的属性有"作者、价格、年代、主题等"。

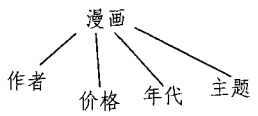

图5 属性关系

2.2 《汉语主题词表》转换成本体需要改造的关系

汉语主题词的语义关系主要有三种类型:等同关系、等级关系和相关关系。

(1)等级关系调整。等级关系揭示叙词之间的上下位层级关系,包含属种、整体与部分、集合与个体层级关系,叙词表一般采用"属、分、簇"等参照项及其词簇索引等方式进行揭示。词簇索引是将具有等级关系的叙词集成中的叙词集合,以族首词为标目,按照词族的等级展开显示,较为完整地反映具有隶属关系的款目词。因此以词族索引为处理对象,按照本体关系的定义,对叙词表中等级关系进行明确区分,主要分离出属种关系、整体与部分关系,将其准确归入对应的关系。所有族首词转换为父类,在每一类型关系下,检查各概念之间的关系是否符合本体构建中基本关系的定义,对不符合规定的概念进行删除、合并或者添加新概念。比如"人口"与"常住人口、农业人口、非农业人口、平均人口、农业人口、现有人口、暂住人口"属种关系中,我们根据具体的应用或者修改为"人口"与"农业人口、城市人口"关系,或者修改为"常住人口、暂住人口"。"诉讼程序"与"起诉、上诉、审理、执行"划分为整体与部分关系,对其内容做部分调整:"立案、侦察、起诉、审判、执行",这样较为全面地体现对犯罪行为进行一系列处理活动。

(2)等同关系调整。等同关系主要揭示同义词、意义相近词或者是具有用代关系词之间的关系。叙词表一般采用"Y、D"参照项互相指代,在非叙词与叙词之间建立对应关系,这在叙词表中占有相当大的比例。通常情况下叙词对术语的表达较为标准、科学,本体构建最好选择叙词作为概念处理。许多非叙词作为不同历史时期的词汇单元,随着社会的发展已不再使用,可以不作处理,如"公粮"作为非叙词,用"农业税"代替使用。只将有必要建立等价关系的非叙词与叙词建立联系,如"罚金"与"罚款"之间建立等价关系。还有一部分意义相关但不相近的词汇则应该分别作为两个概念处理,如"版权"与"著作权",分别建立两个类。

(3)相关关系调整。相关关系揭示叙词之间除等同、等级关系以外的表示相互关联的一种关系,在主表中表现为参照关系,以"C"符号出现。它涉及的类型复杂,包含交叉关系、矛盾关系、因果关系、影响因素、事物与其性质、学科或研究领域与其研究对象、过程与所用工具等多种关系。参照关系表现类型复杂,转换过程有一定难度,需要仔细辨别这些关系。如表1,*Agrovoc* 叙词表转换为农业本体中,将相关关系细分为制作与被制作、成员关系等[9]。《汉语主题词表》相关关系中,矛盾关系居多,比如"招聘"与"解雇",我们将转换成本体对应的类时,考虑建立两个并列类,并定义两类之间的不相交性。同时将一部分参照关系转换为属种关系,如"刑事犯罪"与"放火、贿赂、交通肇事、流氓、破坏婚姻家庭罪……",一部分转换为整体与部分关系,如"唐宋八大家"与"韩愈、柳宗元、欧阳修、苏洵……",另外涉及到事物与事物所处理的对象等相关关系,如"盲文书"与"盲人图书馆","玩具"与"游戏",可作为属性关系处理。

表1 *Agrovoc* 叙词表中的相关关系处理

Relationship	Examples	Remark(More Appropriate Relationships)
RT	1. *Mutton* RT *Sheep*	*Mutton* < madeFrom > *sheep*
	2. *Rice* RT *Riceflour*	*Rice* < useToMake > *Riceflour*
	3. *Fao* RT *UN*	*Fao* < memberOf > *UN*

利用《汉语主题词表》的编制成果来构建领域本体,是一项非常有意义且有价值的研究,如果将人工研究的转换规则转换成计算机处理形式,将会大大提高转换效率。未来的研究更应注重探讨采用机器学习的方式构建人工智能系统,实现《汉语主题词表》到领域本体的自动转换。

参考文献

1 Eman Jay ven. Owl Exports From a Full Thesaurus. Bulletin of the American Society for Information Science and Technology, 2005, 32(1)

2 毛军. 基于RDF的叙词表研究. 情报学报, 2003(4)

3,9 Asanee Kawtrakul. Automatic Term Relationship Cleaning and Refinement for AGROVOC. [2006-07-14]. http://ftp.fao.org/docrep/fao/008/af240e/af240e00.pdf

4 Qin jian. Paling Stephen. Converting a controlled vocabulary into an ontology: the case of GEM. Information research, 2001, 6(2)

5,6,7 B. J. Wielinga A. Th. Schreiber J. Wielemaker J. A. C. Sandberg From Thesaurus to Ontology, http://www.cs.vu.nl/guus/papers/Wielinga01a.pdf

8 Dahlberg, I. Conceptual definitions for interconcept. International classification, 1981, 8(1)

贾君枝 山西大学管理学院副教授,博士。通讯地址:太原。邮编030006。 (来稿时间:2006-11-10)

欢迎订阅和零购《中国图书馆学报》

《中国图书馆学报》是文化部主管,中国图书馆学会和国家图书馆主办的国家级图书情报学专业期刊,被评为中国优秀图书馆学期刊、全国中文核心期刊、中国期刊方阵期刊和国家期刊奖百种重点期刊。

全国各地邮局均可订阅,国内代号2-408。中国国际图书贸易总公司负责国外发行,国外代号Q184。2007年改善封面和内文印刷用纸,但每期定价仍为13元,全年78元。

也可在本刊编辑部订阅。在编辑部直接订购的个人订户,每期优惠价10元(含寄费)。编辑部地址:北京中关村南大街33号。邮编100081。电话88545141。订阅方式可以直接汇款,在汇款单上写明订阅者名称、地址、邮编、份数。也可银行转账。开户名称:中国图书馆学会。开户银行:北京银行魏公村支行。账号:01090303200120105049050。