基于OWL的成语典故本体构建研究

冉 婕[1,2],孙 瑜[1],昌 霞[1],章秀君[1],李 静[1]
(1.云南师范大学 计算机科学与信息学院,云南 昆明 650092;
2.云南昭通师范高等专科学校 计算机科学系,云南 昭通 657000)

摘 要:本体是共享概念模型的明确的形式化规范说明,作为知识表示和知识共享的一种方法,本体是目前信息处理领域研究的热点。基于本体论的思想,利用骨架法构建了成语典故本体,并用OWL语言对成语典故本体进行形式化描述。详细介绍了成语典故本体的目的和使用范围、知识采集提炼及OWL描述,OWL描述分别从类、子类、属性、个体及关系几个方面进行了详细分析,为成语典故相关知识的查询奠定基础。通过成语典故本体的构建可有效对成语典故进行智能检索,是本体技术在中国传统文化中应用的尝试。

关键词:本体;OWL;成语典故;骨架法;语义检索;Protégé
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)05-0063-04

Research of Construction of the Idiom Story Ontology Based on OWL

RAN Jie[1,2], SUN Yu[1], CHANG Xia[1], ZHANG Xiu-jun[1], LI Jing[1]
(1. Institute of Computer Science and Information Technology, Yunnan Normal University, Kunming 650092, China;
2. Department of Computer Science, Zhaotong Teacher's College, Zhaotong 657000, China)

Abstract: Ontology is an explicit conceptualization of concepts in a shared domain. As a kind of means of knowledge representation and sharing, ontology is very hot in the research of information processing. Based on ontology, the Idiom Story Ontology is built through skeleton method, and OWL is used to describe the ontology formally. It introduces the Idiom Story Ontology's purpose, knowledge acquisition and OWL description, OWL description respectively analyze from several aspects including classes, subclasses, properties, individuals and relationships in detail. This methodology of building ontology establishes the foundation for the querying of the Idiom Story. By the building of the Idiom Story Ontology, the intelligent retrieval of the Idiom Story can be applied effectively. It is an attempt of ontology technology in China's traditional culture.

Key words: ontology; OWL; idiom story; skeleton method; semantic retrieval; Protégé

0 引 言

在源远流长的中国文化中,成语典故作为古人智慧的结晶、汉语言中的精华,有着言近旨远、形象生动的独有特点。它集深厚的历史底蕴与强烈的文学色彩于一体,以其内涵深刻,妙趣无穷,真实地再现了一段段传奇故事和历史遗痕,是中华民族文化的一种特殊的表现形式。

文中以成语典故语义检索为目的,对成语典故本体的构建方法进行研究,并用本体描述语言OWL(Web Ontology Language)构建出一个原型本体库,向基于本体的语义检索迈出第一步。

1 本体的相关概念

在本节中,介绍本体的定义、知识表示元素和本体构建方法。

本体是描述概念及概念之间关系的概念模型,其应用经历了从哲学到人工智能领域再到信息领域的发展,目前已被广泛应用到计算机科学的众多领域中。当前计算机领域采用的是Studer等在1998年对本体的定义,认为本体是共享概念模型的明确的形式化规范说明。

收稿日期:2009-09-14;**修回日期**:2009-12-13
基金项目:国家自然科学基金项目(60903131);云南省社会发展科技计划应用基础研究项目(2009ZC052M);云南省教育厅重点项目(07Z10661)
作者简介:冉 婕(1975-),女,四川宣汉人,讲师,硕士研究生,研究方向为本体构建及语义检索;孙 瑜,博士,教授,硕士生导师,研究方向为智能信息处理。

这个定义包含四层含义：

(1)概念化(conceptualization)：指通过抽象出客观世界中一些现象的相关概念而得到的模型，其表示的含义独立于具体的环境状态。

(2)明确(explicit)：指所使用的概念及使用这些概念的约束都有明确的定义。

(3)形式化(formal)：指本体是计算机可读的。

(4)共享(share)：指本体中体现的是共同认可的知识，反映的是相关领域中公认的概念集，它所针对的是团体而不是个体[1]。

本体通过多种知识表示元素表现领域实体的本质及实体间的关联。这些知识表示元素主要包括：类(Classes)或概念(Concepts)、属性(Properties)、关系(Relations)、函数(Functions)、公理(Axioms)、实例(Instances)[2]。

构建本体的方法有骨架法、IDEF5 方法、SENSUS 方法、Bernaras 方法等[3]，均采取自低向上抽取离散术语，分析概念关系，再进行形式化描述的方法。骨架法是爱丁堡大学人工智能研究所在构建企业本体过程中总结形成的方法，主要用来建立相关企业间术语和定义的集合，其流程如图 1 所示[4]。骨架法可分为两个阶段来构建本体。第一阶段是对构建对象领域的知识进行分类，首先确定本体的应用目的和范围，根据所研究的领域或任务，建立相应的领域本体或过程本体；其次是本体分析，定义本体所有术语的意义及其之间的关系。第二阶段是领域本体的表示和编码，包含本体表示、本体评价和本体建立等内容[5]。总之，本体建立是对清晰性、一致性、完善性、可扩展性进行检验。文中就是依照骨架法的基本思路来构建成语典故本体的。

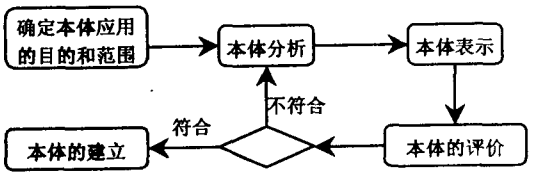

图 1 骨架法构建本体的流程

领域本体需要用形式化的本体描述语言表示，才能被计算机自动识别和处理。目前最流行的本体语言是 OWL，根据领域知识的复杂性可以选择不同描述能力的子语言(OWL Lite，OWL DL，OWL Full)[6]。

2 成语典故本体构建

在本节中，基于骨架法构建本体的方法，从两个阶段来构建本体。第一阶段是对构建对象领域的知识进行分类，首先确定本体的应用目的和范围，根据所研究的领域或任务，建立相应的领域本体或过程本体；其次是本体分析，定义本体所有术语的意义及其之间的关系。第二阶段是领域本体的表示和编码，包含本体表示、本体评价和本体建立等内容。总之，本体建立是对清晰性、一致性、完善性、可扩展性进行检验。

2.1 明确本体的目的和使用范围

该本体为成语典故相关内容的本体，属于应用本体，描述的是依赖于特定领域和任务的概念和概念之间的关系。由于成语典故较多，涉及不同的朝代，为了减小本体的规模，该本体的范围确定在楚汉相争时期。构建该本体的最终目的是为了方便用户对这一时期成语典故及相关历史知识的查询，试图在资源上提供一定程度的语义搜索，使用户的查询结构更加精确和快捷。

2.2 知识采集

这是一个获取规范名词术语的过程，在该本体中，为了搜集相关的成语典故，先后查询了《细说成语典故》、《古汉语成语典故词典》、《常用典故分类词典》、《中国成语典故》等书籍，目前共收集了楚汉战争时期的典故百余条，在以后的工作中还会进一步扩充。

2.3 分析、提炼采集到的知识

对采集到的术语名词进行细致的分析、归类、整理，确定类的特性以及类的等级等。通过对知识的采集，收集到楚汉战争时期的成语典故百余条，在整理过程中发现这些成语典故包括的内容非常丰富，涉及的领域非常广泛，有描述军事战备的，有描写社会生活中人与人关系的，也有描写人的心理情绪的，甚至还有教育学习方面的，这些在分类中都作了详细考虑。对于有的成语典故，在不同的词典中其说法不完全一致，如“智者千虑，必有一失”在黑龙江人民出版社出版的《古汉语成语典故词典》中说明其出处为《史记·淮阴候列传》，而在新世界出版社出版的《中国成语典故》一书中将其出处归为《晏子春秋·内篇杂下》，这些在本体构建中都当作属性的等价关系来考虑。

2.4 成语典故本体的 OWL 描述

成语典故本体采用 OWL 形式化编码，这个阶段主要是使用 OWL 描述 ontology，就是用 OWL 中定义好的元 ontology 对概念和关系进行形式化描述，最重要的是定义类、子类、属性和它们各自具有的特性。成语典故本体是利用 Protégé 3.2.1 编写完成的，完成后的本体以 OWL 为后缀的文件格式保存。Protégé 是由斯坦福大学的 Standford Medical Informatics 开发的一个开放源码的本体编辑器，它是用 Java 编写的。其界面风格与普通 Windows 应用程序风格一致，用户比较容易学习使用[7]。

2.4.1　文档类型及命名空间描述

一个标准的本体开始部分里包括一组 XML 命名空间(namespace)声明。这些命名空间声明提供了一种无歧义地解释标识符的方式,并使得剩余的本体表示具有更强的可读性。成语典故本体的命名空间 OWL 描述如下:

```
<? xml version="1.0"? >
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns="http://www.owl-ontologies.com/Ontology1240137142.owl#"
xml:base="http://www.owl-ontologies.com/Ontology1240137142.owl">
<owl:Ontology rdf:about=""/>
```

2.4.2　类的详细定义及 OWL 描述

一般一个类的等级体系结构有:自顶向下法,由一个领域中最大的概念开始,而后再将这些概念细化;自底向上法,由底层这个等级体系中的细枝末节即最小概念开始,然后将这些细枝末节的类加以组织、概括也就是泛化的过程[8]。由于是依循检索的原则来构建类体系结构,因此采用自顶向下法来构建。成语典故的主要分类如图 2 所示。

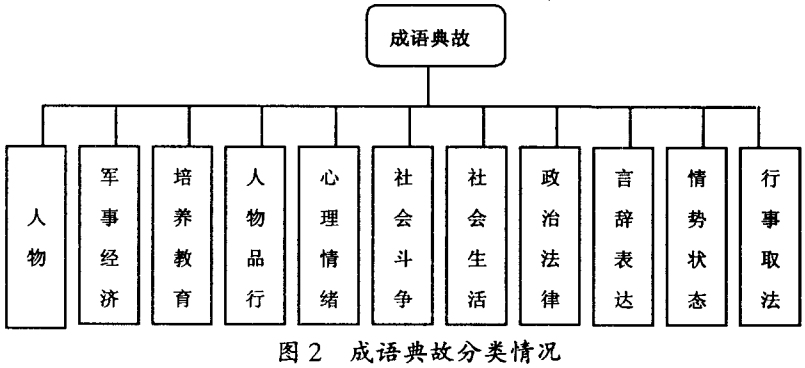

图 2　成语典故分类情况

在对成语典故进行分类时,注重全面性,既涉及相关的人物,又涉及经济军事等多方面的内容,故在这方面考虑相对周全。通过多方面资料的查询,将其分作 11 个大类 79 个小类,这种分类方式也便于以后对本体库的扩充。基于本体的知识分类需要一个准确的概念模型集,对于该本体的分类,查阅了相关的成语典故分类辞典,适应本成语典故的特点。

上述分类的部分 OWL 代码如下:

```
<owl:Class rdf:ID="行事取法">
<rdfs:subClassOf>
<owl:Class rdf:ID="成语典故"/>
</rdfs:subClassOf>
</owl:Class>
<owl:Class rdf:ID="情势状态">
<rdfs:subClassOf rdf:resource="#成语典故"/>
</owl:Class>
<owl:Class rdf:ID="政治法律">
<rdfs:subClassOf rdf:resource="#成语典故"/>
</owl:Class>
<owl:Class rdf:ID="人物">
<rdfs:subClassOf rdf:resource="#成语典故"/>
</owl:Class>
<owl:Class rdf:ID="军事经济">
<rdfs:subClassOf rdf:resource="#成语典故"/>
</owl:Class>
```

另外,楚汉之争中,涉及大量的战争,在分类时,军事经济是比较重要的一个分类,在该分类中,考虑到了战术、军备、攻防等多方面,具体分类情况如图 3 所示。

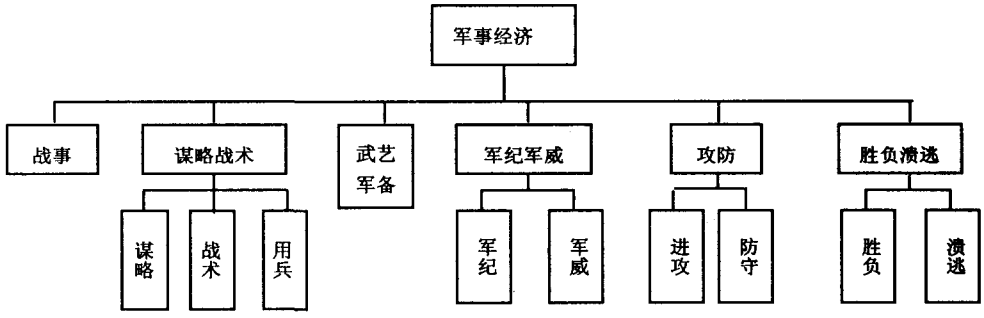

图 3　军事经济类的具体分类情况表

上面的分类,将军事经济类分为了 6 个二级类和 9 个三级类,基于该分类,所收集到的关于经济军事方面的成语都可有归属。

2.4.3　确定类之间的关系

针对查询系统的特点,在所构建的领域本体中使用了三种关系:继承关系、相关关系和同义关系,其中的继承关系可以看成是上下位关系。

(1)继承关系(is_a)。

继承表示概念之间的包含和被包含关系,也可以看成是概念之间的泛化和特化关系。如果概念 C_i 是 C_j 的一种特殊概念,那么说概念 C_i 是概念 C_j 的特化,概念 C_j 是概念 C_i 的泛化,概念 C_i 继承于概念 C_j。例

如:军事经济分类是一种特殊的成语典故,所以概念"成语典故"是概念"军事经济"的泛化。相反,概念"军事经济"是概念"成语典故"的特化。即,"军事经济"继承于"成语典故"。如图 4 所示。

(2) 相关关系(relevant of)。

相关关系表明概念和概念由于某个主题而相互关联。在具体的结构图中相关关系可由继承关系和关系之间的联系导出,所以未做图示。

(3)同义关系(synonymy of)。

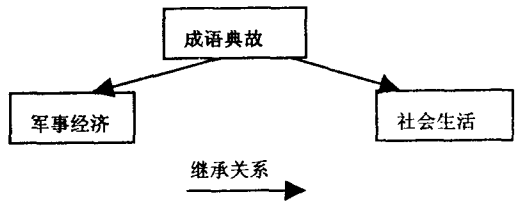

图 4 继承关系

在一个领域中,一个概念可能有几种不同的表示方式,也就是可以用不同的词语来表示一个相同的概念。例如:"刘邦"和"汉高祖"、"刘季"、"沛公"都是同一概念的几种不同的说法,所以这两个概念之间存在同义关系。

继承关系是最主要的关系,相关关系可以从继承关系中体现出来。出现同义关系的概念不是很多。

2.4.4 属性的定义

OWL 中两个主要类型的属性是对象属性(Object properties)和数据类型属性(Datatype properties),对象属性主要描述类之间的关系,数据类型属性主要定义实体的属性。成语典故本体只建立了 1 个对象属性,即 is_a 关系。14 个数据类型属性,分别是:成语名称、别名、注音、书证、典出、反义词、近义词、成语故事描述、褒贬性、释义、朋友、敌对、丈夫、妻子。其中朋友和敌对关系定义域为人物,妻子的定义域为男性人物,丈夫的定义域为女性人物,体现人物之间的关系,这种定义属性主要是便于后面检索的需要。定义近反义词主要是方便用户查询该成语的相关词汇。褒贬性的取值限定在"褒义"、"贬义"和"中性"三个词中。

对属性的定义及其一些约束对应的 OWL 的部分描述如下(只显示 Object properties 对应的 OWL 描述):

```
<owl:onProperty>
    <owl:TransitiveProperty rdf:ID="is_a"/>
</owl:onProperty> </owl:Restriction>
```

上面所定义的 is_a 属性是一个传递属性(TransitiveProperty)。

2.4.5 个体的定义

目前收集到楚汉相争时期的成语典故共有百余条,在以后的研究中可对其进行扩充。从人物分类的角度,这些成语全可分类,从其他分类角度来看,目前有 79 个成语可入类,有的分类下暂空。

对于部分成语,其名称有多种形式,但表示的都是相同的意思,如"妒贤嫉能"和"嫉贤妒能","登坛拜将"和"登台拜将","固若金汤"和"金城汤池"都是指的同一成语,这样的例子还有很多,在构建本体时,把它们当作相同的个体来处理,即 OWL 中描述的 sameAs 关系,其相应的 OWL 代码如下:

```
<韩信 rdf:ID="肝胆涂地">
    <owl:sameAs rdf:resource="#肝脑涂地"/>
</韩信>
```

3 结束语

文中定义了成语典故本体,并用 OWL 对成语本体中的类、子类、属性及其关系进行描述,为成语典故相关知识的查询奠定基础。由于主要是对楚汉相争时期的成语典故本体进行构建,因此,下一步的工作是继续构建其它时期的成语典故本体。此外,如何对所构建的本体进行评价也是下一步的研究工作。

参考文献:

[1] Fensel D. The semantic web and its languages[J]. IEEE Computer Society, 2000,7(2):75-77.

[2] Perez A G, Benjamins V R. Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem Solving Methods[C]// IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5). Stockholm, Sweden:[s.n.],1999.

[3] 韩 婕,向 阳. 本体构建研究综述[J]. 计算机应用与软件,2007,24(9):21-23.

[4] 程传业,梁春芝,杨宗霄,等. 基于骨架法的锅炉故障检测系统的领域本体构建[J]. 河南科技大学学报:自然科学版,2008,29(3):35-39.

[5] 刘光蓉. "C 程序设计"课程内容本体构建[J]. 电化教育研究,2008(12):42-45.

[6] 岳 静,张自力. 本体表示语言研究综述[J]. 计算机科学,2006,33(2):158-162.

[7] Noy N F, Sintek M, Decker S, et al. Creating Semantic Web Contents with Protégé-2000[J]. IEEE Intelligent Systems, 2001,16(2):60-71.

[8] 杜小勇. 学科领域知识本体建设方法研究[J]. 图书情报工作,2005,49(8):74-78.