叙词本体演化系统的结构与技术方案

邓仲华 杨明松 钱文静
（武汉大学信息管理学院，湖北武汉 430072）

摘 要：文章在研究叙词本体理论以及演化机制的基础上，构建出一个叙词本体演化的原型系统，并对系统的实验结果进行了详细的分析。本系统实现了叙词本体演化的过程，能够完成叙词本体的自学习和自动更新。

关键词：叙词本体；本体演化；知识组织
中图分类号：G350 文献标识码：A DOI：10.3772/j.issn.1674-1544.2012.01.008

Structure and Technical Solution of Thesauri-Ontology Evolution System

Deng Zhonghua, Yang Mingsong, Qian Wenjing
(College of Information Management, Wuhan University, Wuhan 430072)

Abstract: Based on the thesauri-ontology theory and the evolution mechanism, this paper constructs a prototype system of thesauri-ontology evolution. Detailed analysis of the system's experimental results is made indicating that the system is capable to accomplish the process of thesauri-ontology evolution and achieve the self-learning and automated updating functions.

Keywords: Thesauri-ontology, ontology evolution, knowledge organization

1 引 言

本体是共享概念模型明确的形式化规范说明[1]。本体以概念为核心，它对所用的概念类型以及对其用法的约束都有明确的定义，具有高度的形式化和共享性。引入本体的形式化方法来表示叙词表，可以提高叙词表的科学性，使叙词表能够被计算机理解和实现自动推理[1]。叙词表关于词间关系的定义不能扩展，而本体的关系是开放的，可扩展的。本体在演化方面表现出的优势，也能服务于叙词表的演化更新。因此，利用本体技术构建叙词表，可以提高叙词表概念语义描述的精确性，扩展词间关系，实现叙词表的自动演化更新。

目前，国内有不少针对本体演化的研究。刘伯嵩早在 2004 年就分析了本体演化的原因和本体演化的关键技术，提出基于 Web 的本体标识和本体演变机制[2]；其后又讨论了在知识管理领域的本体演化，设计出本体演化管理的系统框架和应用实例[3]。钱乐秋等提出了一种基于本体描述构建库中的本体演化理论框架[4]。何克清等提出一种支持可靠语义互操作的本体演化管理框架 MFI-3，通过基于该框架的本体演化可以得到可靠的本体映射[5]。董慧主持的国家自然科学基金资助项目“基于数字图书馆的本体演化和知识管理研究”的研究成果则代表国内本体演化研究的前沿，他提出了本体分子理论，分析了动态知识演化过程和控制管理[6-8]。目前国内的研究主要集中在本体演化构建[9-10]、演化管理[11-12]和演化维护模型[13-14]等方面。总体来说，理论研究较多，实际应用较少，但本体演化已经成为学界研究的重点和前沿。

本文采用本体理论构建叙词表，利用本体在概念描述，学习演化方面的技术，研究叙词本体的演化模型。在此基础上，设计和实现了叙词本体演化系统，实现叙词本体的演化，使叙词表更自动化，

第一作者简介：邓仲华（1957- ），男，武汉大学教授，研究方向：知识组织与处理、信息系统。
基金项目：教育部人文社会科学研究基金项目“多语种叙词本体的构建理论与自动维护模型研究”（07JA870013）。
收稿日期：2010 年 8 月 16 日。

有效降低叙词表构建和更新的成本。

2 叙词本体及其演化

2.1 叙词本体

叙词本体是利用本体的组织模型与管理技术构建的叙词表，具有本体技术功能，同时具有叙词表和本体两者的优点[15]。在构建理论上，叙词本体采用本体模型，以提高概念语义描述的精确性；在功能上，利用本体的智能技术，使静态的叙词表发展到动态的叙词本体，使原来被动的维护状态发展成具有主动学习演化功能的状态。叙词本体的形式化定义可表示为实体与关系的集合。

令 TO={C,I, HC,RI,O}。式中 TO(Thesauri Ontology) 为叙词本体，它由一系列分类概念集 C、从属于概念集下的实例（叙词与非叙词）I、概念的类分关系 HC、实例之间的关系 RI 以及一些公理 O 组成。这其中的分类概念集是指表示学科范畴的概念，即学科分类。I 是术语词集，对应于叙词表中的语词，包括规范化的叙词以及非规范化的非叙词。

在叙词本体定义中，概念的类分关系 HC 表示叙词表中学科分类的类分关系，实例关系 RI 表示叙词与非叙词之间的用、代、属等关系。学科分类关系 HC 在确定后变更较少，应由该领域专家对其进行更新，叙词本体中主要研究实例之间的关系 RI。叙词本体要反映叙词间的等同关系、属分关系和相关关系，因此RI应包括“用（Y）”、“代（D）”、“属（S）”、“分（F）”、“参（C）”、“族（Z）”。叙词本体可以将叙词表中所有的内容包括进来，并结合学科分类，是分类法和主题法相结合的一种方案。

与传统叙词表不断更新版本不同，叙词本体在更新和修正叙词时，在原有语料库基础上进行自动化的抽取和维护。为了实现叙词本体的演化，需要利用本体的学习技术和演化技术。本体学习是指利用机器学习和统计等技术自动地从已有的数据资源中获取期望本体的过程[16]；本体演化则是指本体根据出现的变化和由这些变化引起的本体一致性问题的自适应变更[17]。所以叙词本体的演化既包括了从已有数据源中获取期望本体的过程，也包括了本体自适应变更的过程。

2.2 叙词本体演化流程设计

叙词本体的演化是科学技术发展的实际需要。在信息技术革命与网络发展的带动下，新的学术成果及新的术语与技术词汇层出不穷。现有的叙词语料库若不及时更新就很难满足应用的要求。本项目的基本思路是从网络 (Web) 的大量文献中抽取出新涌现的术语，或是检测出术语的变化，并以此对叙词表的语料库进行合理的调整（修改或更新）。因此，叙词本体的数据源来自网络，主要为 html，中文的叙词本体学习主要是基于半结构化数据源和非结构化数据源的本体学习技术。

本体作为一种概念模型，在组成上主要包含概念和概念之间的关系。其中概念之间的关系是用来描述概念的语义[18]。因此本体学习的内容也包括两个方面：概念获取和关系获取。在叙词本体演化系统的运行过程设计中，采取的运行流程是：首先从网络中获取数据资源，然后再从数据资源中获取新概念和概念关系，最后将概念和概念关系更新到叙词表中。

由于叙词表是一种将标引人员或用户使用的自然语言转换为规范化的系统语言的术语控制工具[19]，为保证叙词的规范性和权威性，且在演化过程中要严格消除冲突，保证叙词的一致性，我们在演化过程中引入了专家评审环节。即在叙词本体演化过程中，需要领域专家评审、监督和控制（图 1）。因此从已有数据源中获取候选本体后，再由领域专家对候选本体进行评审，才可以完成本体的学习。同样，从已有数据源中检测了本体的变化后，也应由领域专家对该变化进行评审，才能进行本体的更新演化。

叙词本体系统在获取候选概念以及概念关系后，将这些概念以及关系提交给该领域专家进行评审，以免发生冲突，保证叙词本体的一致性和正确性。因此叙词本体演化的整个过程包括：从网络上获取数据源；解析数据，抽取概念；分析概念之间的关系，抽取关系；专家参与，本体演化管理。

3 系统结构设计与技术方案

3.1 系统功能分析

叙词本体演化系统实现叙词本体演化的过程，自动从网络信息资源中抽取概念和概念关系，在专家评审后由工作人员根据评审结果更新本体库。因此，叙词本体系统具有本体学习和演化的功能。同时，叙词本体演化系统也是网络版的叙词表，能为用户提供叙词的 Web 查询功能。所以系统的功能可分为前台和后台两部分，前台为用户提供方便的叙词查询功能，后台实现本体库的自动演化更新。

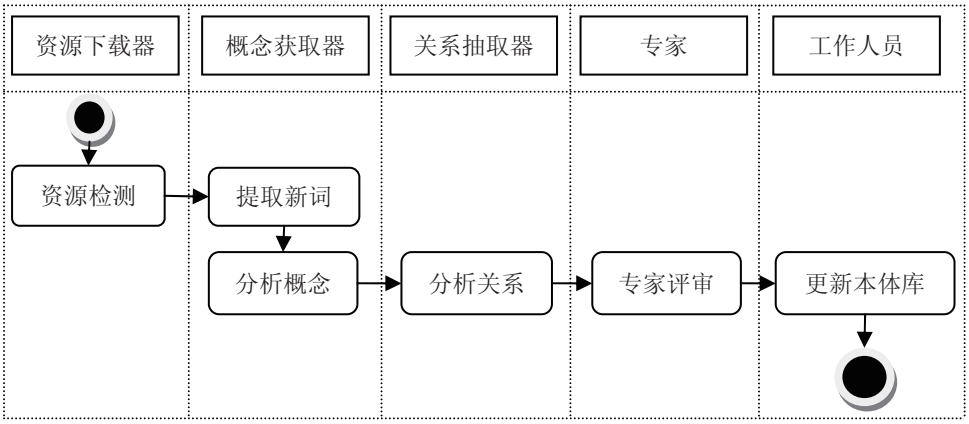

图 1　叙词本体演化总体过程

（1）叙词查询功能。用户进入系统的叙词查询页面后，可按自己的需要查询本体库中的叙词及相关信息。叙词本体采用分类法和主题法相结合的方案，因此系统提供了分类主题对应的功能，使用户能通过学科分类查找叙词，或通过叙词查找所在分类，还可直接搜索叙词及相关信息。

（2）本体演化更新。演化系统通过本体演化的过程实现本体库的丰富和更新。叙词本体演化主要由信息抽取、专家评审和本体库更新 3 部分组成。在信息抽取部分，系统按一定周期自动检测网络信息资源，从中抽取概念和概念关系，以供专家评审。在专家评审部分，系统为专家提供评审的平台，维护评审规则，并记录评审结果。在本体库更新部分，系统为工作人员提供添加新叙词和关系、更新本体库的人机交互界面，完成本体库中叙词和关系的更新演化。

3.2　系统结构与技术方案

叙词本体系统要实时检测网络数据的变化，下载相关网络资源，抽取概念和概念关系，最后更新本体库，实现叙词本体的演化更新。资源下载器下载指定数据源的资源信息，再通过概念获取器将收集的词与本体库中的词汇进行比较，提取新词，分析概念，通过关系抽取器分析关系，最后由专家进行专家评审，并以评审结果更新本体库（图 1）。系统主要包括查询、信息抽取和专家评审 3 个模块（图 2）。信息抽取模块又分为数据检测子系统、概念获取子系统和关系抽取子系统 3 个子系统。

（1）查询模块：查询模块实现的功能是提供多种检索方式检索叙词，并展示与叙词有联系的所有其他叙词。这里的联系指的是叙词的用、代、属、分、参、族几种关系。

（2）信息抽取模块：这是本系统的核心模块，包括数据检测子系统、概念获取子系统和关系抽取子系统。数据检测子系统是用来检测网络上词汇的变化，利用网络信息资源抓取技术和指定信息源的网络资源，并对该网络资源进行整理，提取网络资源中的相关中文文本信息，为新词分析和关系分析作准备。概念获取子系统是从抓取到的网络信息资源中提取词汇，并将词汇交由专家评审模块进行专家评审。关系抽取子系统要从抓取的网络信息资源中提取词与词的关系，作为候选关系交由专家进行评审。

（3）专家评审模块：专家评审模块使人加入人机协作系统中，主要为专家评定提供友好的平台，并统计专家评审的结果，确定将新叙词及其关系添加到叙词表中，从而完成叙词本体演化的全过程。专家评审是由多位专家共同参与为自动提取出的新术语和新关系评分的过程，只要评定的结果达到指定的标准，就可以确认完成叙词本体的更新。

叙词本体演化系统是一个基于 B/S 模式的 Web 应用系统，系统基于 Windows2000 Server+tomcat+mysql 架构，采用 Java 语言的 JSP/Servlet 技术来实现。以 Web 页面的形式实现叙词本体的可视化，从而让用户通过桌面平台与叙词本体底层进行交互。针对系统的 3 个模块，相应的技术方案如下。

（1）采用 Java 技术呈现叙词本体查询界面，编写相关程序控制检索本体库的过程，为查询者提供检索叙词的服务，可视化地展现其所检索的叙词及相关信息。

（2）采用 Java 技术编写网络资源提取程序，用以监测抓取所需的网络资源；结合 html 解析的 API 编写 html 解析器，用以从提取的 html 形式的资源

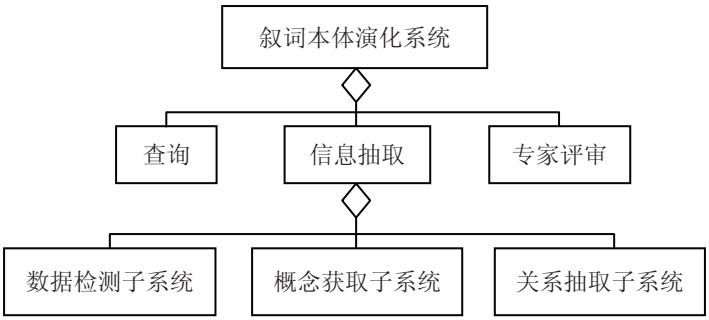

图 2　叙词本体总体结构

中解析出所需的中文文献属性信息；将中科院中文分词系统与叙词本体词汇选择的要求相结合编写中文分词器，用以对所解析出的文献信息进行分词，筛选出所需词汇；应用关联规则的方法提取词汇间的相关关系，采用模板的方法发掘词汇间的等同关系或等级关系。

（3）采用 Java 技术实现专家评审的 Web 人机交互页面。专家在该界面上能获取所需评审的叙词或关系，执行评审流程，提交评审后程序将评审数据保存。结合 jena 程序，编写本体库更新程序；建立本体库更新的 Web 页面，向执行更新任务的工作人员呈现出专家评审的统计结果，工作人员提交更新后即完成本体库的更新。

3.3　系统开发环境和开发工具

（1）开发环境平台：采用 Jave 语言进行开发。基于 Java 语言的资源非常丰富，现有很多构建本体的 API，如 Jena 等，因此采用 Java 语言的直接好处是有许多现成的资源可用，可节省大量的人力物力。

（2）数据库平台：采用 MYSQL 数据库。

（3）叙词表数据存储文件：以本体库存储叙词表数据，采用 owl 描述本体。

（4）本体库管理工具：采用 Protégé 作为本体库管理工具。

3.4　本体库构建

在综合比较 TOVE 法、骨架法、七步法等方法后，结合叙词本体构建的特点，我们在叙词本体库构建中采用了 5 个主要的步骤：确定叙词本体的应用目的、整体设计、详细设计、表示和评价。实际的叙词本体开发过程是一个反复替代的过程，需要根据实际需求反复讨论、修改、调试，最终确定本体库的原型。

3.4.1　构建步骤与内容

目前，国内外已经有了一些成型的本体构建工具，　如 Ontolingua、WebOnto、OntoEdit、Protégé，在综合考虑以上工具的优缺点后，选用 Protégé 作为叙词本体构建工具。利用本体编辑工具构建本体库。具体的构建过程包括：定义类和类的等级体系、本体的属性、语义关系和添加本体实例。

（1）定义类和类的等级体系。叙词表将叙词抽象成分类体系结构，叙词本体中以类表示叙词表中的学科分类，以类与子类的关系表示学科分类的类分关系。

（2）定义本体的属性。本体中属性用来描述概念的内部结构，通常可以分为数据类型属性和对象属性。数据类型属性又称为“内在属性”，描述类实例与 RDF 文字或 XMLschema 数据类型间的关系，这类属性通常连接一个概念和数据值。叙词本体中的概念与分类号和范围注释之间的关系可以定义为数据类型属性。对象属性又称为“外在属性”，也称之为“关系”，表示类或概念之间的关系。叙词本体中概念与术语之间的关系可以定义为对象属性。

（3）定义语义关系。叙词表中词间关系可分为：等级关系、等同关系和相关关系。为表达概念间的语义关系，叙词本体中以“Y”“D”表示概念间的等同关系，以“S”“F”“Z”表示概念间的等级关系，以“C”表示概念间的相关关系。其中“Y”和“D”、“S”和“F”互为逆反属性。

（4）添加本体实例。设计好类和属性后，就可以创建实例。叙词本体中，以某一类的实例表示该学科分类下的术语。

3.4.2　叙词本体的表示

本体的编码有基于描述的，也有基于逻辑推理的。常用的本体语言有 Ontolingua、Loom、Flogic

（Frame Logic）、XOL、RDF（S）、OWL、OIL 等。采用标准的本体描述语言，可以更好地实现本体的共享和重用，在叙词本体系统中，以 OWL 作为本体描述语言。

叙词表由叙词、非叙词及其之间的关系组成，能够反映叙词间基本语义关系。在叙词本体系统中，以本体模型构建叙词表，以 OWL 作为本体描述语言，将构建的本体库作为叙词表的数据存储平台。在本体库中，以类表示叙词表中的学科分类，以类与子类的关系表示学科分类的类分关系，以类的实例表示该分类下的叙词或非叙词，以实例的属性表示叙词或非叙词之间的“用”、“代”、“属”等关系。叙词本体中对叙词表的学科分类、叙词和关系的详细定义如表 1 所示。

3.4.3　*库结构示例*

《中国分类主题词表》是一部国家级的大型综合性分类主题一体化叙词表，结合了《中国图书馆图书分类法》和《汉语主题词表》的成果。该词表现已广泛应用于全国各类图书馆和信息机构的文献标引工作，具有一定的代表性 [20]。现从《中国分类主题词表》中截取“科学 _ 科学研究”类目及其子类目、叙词，以此为例详细阐述本体的构建过程。

（1）定义类和类的等级关系。如图 3 所示为“科学 _ 科学研究”类目下的部分分类结构，“情报学与情报工作”为“科学 _ 科学研究”的子类，“情报学”和“情报学检索”为“情报学与情报工作”的子类。在叙词本体库中，分别定义这 4 个类及其之间的父子关系。用 OWL 表示“科学 _ 科学研究”类和“情报学与情报工作”类为：

<rdfs:Class rdf:about="http://www.domain2.com# 科学 _ 科学研究 "/>

<rdfs:Class rdf:about="http://www.domain2.com# 情报学与情报工作 ">

<rdfs:subClassOf>

<rdfs:Class rdf:about="http://www.domain2.com# 科学 _ 科学研究 "/>

</rdfs:subClassOf>

</rdfs:Class>

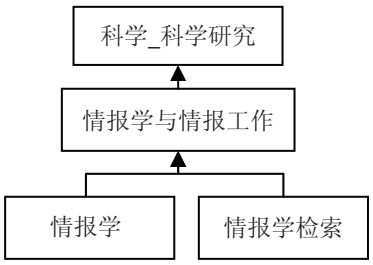

图 3　分类结构

（2）定义叙词及词间关系。类目“情报学”中具有多个叙词，在本体中都被定义为“情报学”类的实例，叙词间的语义关系在叙词的对象属性中定义。为此，要首先定义“Y”、“D”、“S”、“Z”、“F”、“C”对象属性。

科技情报

G350

D 技术情报

D 科技信息

Z 情报

S 情报

· 国防科技情报

上面被描述的叙词是“科技情报”。“G350”是“科技情报”在《中国分类主题词表》中的分类号；“D”表示“代”的关系；“Z 情报”表示族首词是“情报”；“S 情报”表示“科技情报”与“情报”具有“属”的关系；“国防科技情报”表示“科技情

表 1　叙词本体 OWL 描述

叙词表	含义	OWL定义
Concept	某一学科分类	<owl:Class rdf:Id="Concept"/>
subConcept	Concept分类下的子分类	<owl:Class rdf:Id="subConcept"> <rdfs:subClassOf rdf:resource="Concept"/></owl:Class>
Term	某一分类Catogery下的术语Term	<Category rdf:Id="Term"/>
D	关系D，与关系Y逆反	<owl:ObjectProperty rdf:ID="D"><owl:inverseOf> <owl:ObjectPropertyrdf:ID="Y"/></owl:inverseOf></owl:ObjectProperty>
…	…	…
C	关系C，对称属性	<owl:SymmetricProperty rdf:ID="C"> <owl:inverseOf rdf:resource="#C"/><rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/></owl:SymmetricProperty>

报”与“国防科技情报”具有“分”的关系。采用OWL对该叙词详细描述如下：

```
< 情报学 rdf:ID= “科技情报” >
    <CategoryID rdf:datatype="http://
www.w3.org/2001/XMLSchema#string">
    G350</CategoryID>
    <D rdf:resource= “技术情报” />
    <D rdf:resource= “科技信息” />
    <Z rdf:resource= “情报” />
    <S rdf:resource= “情报” />
    <F rdf:resource= “国防科技情报” >
</ 情报学 >
```

4 小　结

本文根据系统的需求建立一个较完善的本体库，从而为叙词本体演化系统的实现作出基础理论的铺垫，同时完成了系统的设计准备工作，使得系统在实现过程中能够井然有序地工作。有关叙词本体演化系统的实现过程以及实验结果将另外行文详细讨论。

参考文献

[1] 曾新红.中文叙词表本体——叙词表与本体的融合[J]. 现代图书情报技术 ,2009(1): 34−43.

[2] 刘柏嵩,高济.本体演化管理研究[J].计算机科学,2004, 31(5):9−12.

[3] 贺赛龙 , 刘柏嵩 . 知识管理中本体演化研究 [J]. 情报学报 ,2004, 23(4):439−475.

[4] 杨明华 , 钱乐秋 , 赵文耘 , 等 . 基于本体描述构建库中的本体演化研究 [J]. 计算机工程 , 2007,33(9):87−89.

[5] 何扬帆 , 何克清 . 一种支持可靠语义互操作的本体演化管理框架 [J]. 计算机工程 , 2007, 33(18):26−30.

[6] 董慧 , 姜赢 , 高巾 , 等 . 基于数字图书馆的本体演化和知识管理研究 (I)——本体分子理论 [J]. 情报学报 ,2009, 28 (3):323−330.

[7] 董慧 , 姜赢 , 高巾 , 等 . 基于数字图书馆的本体演化和知识管理研究 (II)——动态知识组织 [J]. 情报学报 ,2009, 28 (4):483−491.

[8] 董慧 , 姜赢 , 高巾 , 等 . 基于数字图书馆的本体演化和知识管理研究 (III) ——动态知识描述 [J]. 情报学报 ,2009, 28 (5):643−650.

[9] 刘磊 , 范茸 , 张睿 , 等 . SetPi− 演算及其对本体演化的建模 [J]. 中国科技论文在线 ,2010,5(2):112−119.

[10] 刘晨 , 韩燕波 , 陈旺虎 , 等 . MINI——一种可减小变更影响范围的本体演化算法 [J]. 计算机学报 , 2008,31(5):712−720.

[11] 王兴 . 本体演化方法与机制研究及应用 [D]. 武汉 : 华中师范大学 ,2006.

[12] 王兴 , 何婷婷 , 庄超 . 本体演化及本体的版本管理机制研究 [J]. 计算机与数字工程 , 2006,34(7):7−10.

[13] 王进 , 陈恩红 , 林乐 . 一种网络环境中的本体演化和维护模型 [J]. 计算机科学 , 2003,30(12):126−128, 5.

[14] 马应龙 . 基于本体的信息集成、演化和动态发现研究 [D]. 北京 : 中国科学院软件研究所 , 2006.

[15] 邓仲华 , 罗玲 , 王琴 , 等 . 多语种叙词本体研究 [J]. 情报学报 , 2009, 28(5):664−671.

[16] 杜小勇 , 李曼 , 王珊 . 本体学习研究综述 [J]. 软件学报 , 2006, 17 (9):1837−1847.

[17] 刘子恒 . 本体演化研究综述 [J]. 软件导刊 , 2005, 38(20):4 −5.

[18] 邓志鸿 , 唐世渭 , 张铭 , 等 . Ontology 研究综述 [J]. 北京大学学报 : 自然科学版 , 2002, 38(5):730−738.

[19] 马张华 . 信息组织 [M]. 2 版 . 北京 : 清华大学出版社 ,2003.

[20] 国家图书馆《中国图书馆分类法》编辑委员会 . 中国分类主题词表 [M]. 北京 : 北京图书馆出版社 , 2005.