

叙词表形式化描述语言的比较研究

贾君枝 卫荣娟

(山西大学管理学院 太原 030006)

文 摘 传统叙词表的本体化构建是目前叙词表发展的重要方向。当前叙词表转换为本体的描述语言有 XML(S)、RDF(S)、SKOS、OWL等,文章通过对几种典型的描述语言和叙词表之间的语义关系进行映射分析,从描述能力、人工干预、应用性三个层面总结了描述语言在表示叙词表语义关系方面的差异性。该研究不仅有助于传统叙词表在语义环境下进行发展,而且对于本体描述语言的完善具有非常重要的作用。

关键词 叙词表 描述语言 比较

Comparative Study on the Thesaurus Form al Descriptive Languages

Jia Junzhi, Wei Rongjuan

(Management School of Shanxi University, Taiyuan 030006)

Abstract The Ontology construction of traditional thesauri has become an important developmental aspect of thesauri currently. There are several descriptive languages for converting the thesauri into Ontology such as XML(S)、RDF(S)、SKOS、OWL etc. Through the map analysis of semantic relationships among the typical descriptive languages and thesauri and from the aspects of descriptive power, manual intervention and utility, this paper summarizes the difference of these descriptive languages when expressing the semantic relationships of thesauri. The study not only promotes the development of the traditional thesauri in the semantic environment, but also has a very important role for the perfection of Ontology descriptive languages.

Key words Thesauri Descriptive languages Comparison

1 引言

国家标准 GB13190-91 将叙词表定义为:叙词表是将文献、标引人员或用户的自然语言转换成规范语言的一种术语控制工具^[1],它是概括某一学科领域并由语义相关、族性相关的术语组成,并可以不断进行补充的规范化词表。在图书情报、信息管理领域,手工或计算机进行的信息标引、检索中,叙词表已形成一种传统的比较科学的信息组织方法。而 Ontology 作为一种网络时代的信息组织技术,在知识工程、人工智能、语义网等相关领域得到了广泛关注和深入研究。Ontology 主要特点是包含一个领域的术语,以及对术语间关系的完整表达,这点和规范化的词表具有相似之处。但是,叙词表仅仅含有简单的语义关系,相比之下,Ontology 所拥有的丰富的语义关系和强大的推理能力是其所不能比拟的。可

以考虑将叙词表中现有的知识体系转化到 Ontology 中,这既是对已有知识的科学利用,也保持了知识应用的延续性和继承性,同时也是语义网环境下叙词表的发展方向之一。

2 叙词表形式化描述语言

将叙词表转换为本体不仅是必要的,同时也是可行的。传统叙词表都是基于图书情报领域的国际标准制定的,主要采用 ISO 2788、ISO 5964、ISO 5963 三个标准定义术语之间的关系^[2],通常规定叙词表中的叙词必须要以概念为基础,要做到一义一词和一词一义,其语义关系通过“用、代、属、分、参”来表达词语间的等同、等级和相关关系。本体在进行语义信息的表达时,需要将一定的语法、词汇表和语言结合起来。叙词表在转换为 Ontology 的过程中,根据描述语言的不同分为以下几种:①基于 XML 的描

述,如澳大利亚 Maria Lee和 Stewart Baillie两人开发的叙词标记语言 (TML),构建了叙词描述 Ontology 的框架^[3]。②基于 RDF的描述,如加州环境资源评估系统 (CERES)和国家生物信息基础 (NBII)联合开发的一套集成有关环境的叙词表和叙词网络工具,其采用 RDF格式进行表示^[4]。③基于 SKOS语言的描述,如英国档案叙词表等。④基于 OWL的形式化描述,如 Fischer制作的美国国家癌症研究所 (NCI)叙词表的 NCI- OWL文档等^[5]。

2 1 基于 XML的形式化描述

XML是 W3C推荐的与本体语言相关的一个关键性的技术标准。它以一种开放的自我描述方式定义了数据结构,以体现数据之间的关系。然而,XML只是定义了一种文档结构的描述语言,它并不具备描述语义信息的能力。而 XMLSchema(以下简称 XMLS)是基于 XML定义其文档语法和结构的约束,具有较强的描述能力、扩展能力和维护能力。它是用一套预先规定的 XML元素和属性创建的,这些元素和属性定义了文档的结构和内容模式。

XMLS中没有类和属性的概念,只能定义元素和元素的属性^[6],并且可以表示父元素和子元素的关系,但它无法区分叙词表中的“用、代、属、分、参”关系,目前基于 XML表示的叙词表只是按照 XML简单的语法格式对词表进行再描述。叙词表中叙词及其叙词之间关系的 XML语义对照见表 1。

表 1 叙词表的 XML语义对照

叙词表	XML构成
叙词集	根元素
叙词	Term子元素
S(属)	relation子元素
F(分)	relation子元素
C(参)	relation子元素
Y(用)	relation子元素
D(代)	relation子元素
注	<! - -#PCDATA - - >

一个用 XMLSchema描述的简单本体模型如下:
<? xml version = “ 1 0” encoding = “ GB2312”?

```
>  
< Thesaums>  
< term>  
< termName> 比较法 < /termName>  
< relation>  
< relationType> S< /relationType>  
< relatedTermName> 法 (法律) < /relatedTermName>  
< /relation>
```

```
< /term>  
< term>  
< termName> 兵役法 < /termName>  
< relation>  
< relationType> S< /relationType>  
< relatedTermName> 国家法 < /relatedTermName>  
< relationType> Z< /relationType>  
< relatedTermName> 法 (法律) < /relatedTermName>  
< relationType> C< /relationType>  
< relatedTermName> 兵役 < /relatedTermName>  
< relatedTermName> 服役条例 < /relatedTermName>  
< /relation>  
< /term>  
< /Thesaums>  
.....
```

2 2 基于 RDF的形式化描述

RDF(Resource Description Framework) 是 W3C在 XML的基础上推荐的一种描述万维网上信息资源的语义关系而开发的一种 Ontology定义语言。它采用一种简单的三元组模型描述了资源的语法。这个数据框架包含节点以及节点之间的连接弧,节点表示 Web上的资源、弧表示资源所具有的属性^[7]。RDFSchem a(以下简称 RDFS)是 RDF的词汇描述语言,为 RDF资源的属性和类型提供词汇表。这样,RDF和 RDFS共同构成了一种简单的对 Web资源的知识表示机制。叙词表的 RDF表示中,通常将每一个叙词作为资源来对待,将叙词间关系定义为属性,并通过 RDFSchem a对各属性的定义域、值域进行限定,进一步增强对资源语义层次的描述能力,叙词表和 RDFS之间的语义映射见表 2。

表 2 叙词表和 RDFS之间的语义映射

| 叙词表 | RDF |
|------|--|
| 叙词集 | rdfs class |
| 叙词 | rdfs subClassOf |
| Z(族) | rdfs Property ID= "Z" |
| S(属) | rdfs Property ID= "S" |
| F(分) | rdfs Property ID= "F" |
| C(参) | rdfs Property ID= "C" |
| Y(用) | rdfs Property ID= "Y" |
| D(代) | rdfs Property ID= "D" |
| 注 | rdfs historicalNote rdfs ScopeNote rdfs CategoryNote |

一个用 RDFS描述的简单本体模型如下:

```
< rdf RDF xml lang= "chi"
xm h$ rdf= "http //www. w3. org/1999/02//22
- rdf- syntax- n#"
xm h$ rdf= "http //www. w3. org/TR /1999/PR_
rdf_19990303#" >
< rdfs class rdf ID= "tem">
< rdfs subclassOf rdf resource= "http //www.
w3. org/TR/1999/PR- rdf- schen a- 19990303#兵役
法 />
< rdfs comment> This is Chinese Thesaurus
< /rdfs comment>
< /rdfs class>
< rdfs Property ID= "S">
< rdfs domain rdf resource= "# tem" />
< rdfs range rdf resource= "# tem" />
< /rdfs Property>
< rdfs Property>
< S rdf ID= "国家法" />
< F rdf ID= "兵役" />
< C rdf ID= "服役条例" />
< /rdfs Property>
.....
```

2 3 基于 SKOS语言的形式化描述

表 3 叙词表与 SKOS之间的语义映射

| 叙词表 | SKOS |
|------|---|
| 叙词 | skos Concept |
| Y(用) | skos PrefLabel |
| D(代) | skos alLabel |
| S(属) | skos broader |
| F(属) | skos narrower |
| R(属) | skos related |
| 注 | skos defination skos scopeNote skos ex-
ample, |
| Z(族) | skos Collection |

由于 RDF 仅提供初级的语义关系表达, 没有统一的标签支持更具体的语义关系的描述, 因此, 需要在 RDF 基础上建立支持表达更具体的语义关系并具有灵活扩展性的统一的知识组织体系标准^[8]。SKOS (Simple Knowledge Organization System) 是在 W 3C 组织语义网下的工作组发布并维护的, 强调为知识组织系统的表达提供强有力且简单的机器理解方式。SKOS 词汇表是基于 RDFS 定义的一系列类和属性, 通过 SKOS 描述语言, 将受控词表转换为与 RDFS, OWL 兼容的概念模型, 叙词表中的叙词及叙

词关系在 SKOS 语言中都有明确的定义, 其为叙词表的 RDF 描述提供了便利条件, 并易于实现语义网中与其他 RDFS 数据的合并与融合, 为叙词表之间及其叙词表与其他词表之间的互操作提供了前提。叙词表与 SKOS 之间的语义关系映射规则见表 3

一个用 SKOS 语言描述的叙词表片段:

```
< rdf RDF
xm h$ rdf= "http //www. w3. org/1999/02/22 -
rdf- syntax- n#"
xm h$ skos = "http //www. w3. org/2004/02/
skos /cor#" >
< skos Concept rdf about= "http //www. exam-
ple. com /concept#民生主义" >
< skos narrower> 节制资本 < /skos narrower
>
< skos narrower> 平均地权 < /skos narrower
>
< skos broader> 三民主义 < /skos broader>
< skos related> 富贵在天 < /skos related>
< skos related> 经济自由论 < /skos related
>
< skos related> 经世致用 < /skos related>
< skos related> 孙中山 < /skos related>
< skos related> 利民 < /skos related>
< skos related> 小康 < /skos related>
< /skos Concept>
< /rdf RDF>
.....
```

2 4 基于 OWL的形式化描述

OWL (Web Ontology Language) 是 W 3C 开发的一种网络本体语言, 用于对本体进行语义描述, W 3C 的设计人员针对各类特征的需求制定了三种相应的 OWL 的子语言: OWL Lite, OWL DL, OWL Full 各子语言的表达能力逐步递增。OWL 中类和属性以及子类和子属性的定义和 RDFS 中的定义是一样的, 有所不同的是 OWL 在 RDFS 的基础之上添加了更多对属性和类的约束, 如类之间的不相交性、等价性以及属性的传递性、对称性, 还有基数、枚举类等。叙词表采用 OWL 描述时, 需遵循本体概念定义的严格性, 将具有明确含义的叙词作为类来对待, 叙词关系作为属性来处理, 并对每一个具体关系进行仔细分析辨认, 以建立与本体定义的四种基本关系 (继承关系、整体部分关系、实例关系和属性关系) 之间的严格对应。叙词及其各种关系与 OWL

之间的语义映射见表 4^[9]。

表 4 叙词表与 OWL 之间的语义映射

| 叙词表 | OWL |
|------|---|
| 叙词集 | <code>rdfs class</code> |
| 叙词 | <code>rdfs subClassOf</code> |
| D(代) | <code>be similar to</code> |
| S(属) | <code>kind of</code> |
| F(分) | <code>part of</code> |
| Y(用) | <code>be similar to</code> |
| C(参) | <code>< someValuesFrom > < be similar to ></code> |
| 注 | <code>comment</code> |

一个用 OWL 描述的简单本体模型如下:

```
< owl:Class rdf:ID = "商品资本" >
< rdfs:subClassOf rdf:resource = "#资本" />
< rdfs:subClassOf>
< owl:Restriction>
< owl:onProperty rdf:resource = "# broadeff-
em" />
< owl:someValuesFrom rdf:resource = "#Chronic
_Lymphocytic_Leukemia" />
< /owl:Restriction>
< /rdfs:subClassOf>
< /class>
.....
```

3 各种形式化描述语言的比较分析

基于以上几种本体表示语言对叙词表进行形式化描述, 可以得出以下结论:

(1) 从 XML(S)、RDF(S)、SKOS 到 OWL, 语言的表述能力不断提高

XML 是结构化数据的标准机制, 只具有语法性, 不能说明所定义的结构语义, 叙词之间的语义关系无法表示。RDF/RDFS 是描述资源及其关系的一种数据模型, 能够提供简单的语义表述功能, 提供了关于这些属性和类的层次结构的语义, 能够揭示叙词之间的关系, 但不同人认识不同, 可能形成不同的模式 (Schema), 易于导致不同叙词表之间映射难、互操作性差、共享性差的问题。SKOS 是 RDF(S) 的应用, 丰富和扩展了 RDF(S) 的描述能力, 定义了 SKOS 元模型和其转换的准则, 其以统一标准的模式来描述叙词表, 不仅使得叙词表的共享成为可能, 而且降低了词表共享的成本, 但由于 SKOS 遵循 RDF/RDFS 语法, 同样缺乏对属性值的限制与约束。OWL 在 RDF(S) 的基础上, 添加了更多的用于描述属性和类的词汇, 能够定义更加复杂的概念结构, 相对于前几种描述语言而言, 其具有更强大的语义表达能力和推理能力。所以采用 OWL 描述的叙

词表, 具有较强的约束和限制, 逻辑上具有严格性。

(2) 从 XML(S)、RDF(S)、SKOS 到 OWL 描述过程中, 人工干预增强

随着从 XML(S)、RDF(S)、SKOS 到 OWL 语言表述能力的提高, 无疑给描述过程增加了一定的难度, 表现为描述过程从完全的机器自动转换实现发展到人工参与的半自动转换。XML 不考虑叙词表的语义特征, 只需运用转换程序 (如 JAVA、XSLT 等) 将数据库形式的叙词表转换为 XML 文档, RDF(S)、SKOS 转换时, 则需要人工定义叙词表关系属性, 建立叙词表与转换语言之间的映射关系, 在此基础上进行自动转换。采用 OWL 形式进行描述时, 由于本体自身的复杂性和严格的建模要求, 使得原有叙词表定义的语义关系需要进一步改造, 需要专家的参与及其认真的分析确认。在此基础上完成自动转换, 实施的难度与成本进一步增大。

(3) 从 XML(S)、RDF(S)、SKOS 到 OWL, 叙词表在语义网环境下的应用性增强

叙词表形式化表示的目的在于突破传统叙词表仅仅是术语控制工具的局限性, 充分利用其已构建的专家知识体系, 在信息组织、知识表示、机器理解、人机信息间的交互匹配等方面发挥重要作用, 使其在当前语义网环境下具有新的生命力。XML(S)、RDF(S) 表示的叙词表目前在数据库检索、网络资源的元数据标引中已得到应用, 用户可以通过浏览词表选择符合检索要求的叙词, 各网站加工人员运用词表对网站的资源进行标引, 以便于用户准确地检索。SKOS 的表示方法使得各词表之间的互操作性成为可能, 方便了跨词表的检索。OWL 形式的词表使其在知识本体构建中起到重要作用, 既减少了从头建立一个 Ontology 所需的重复努力, 也可以建立一种允许不同的词表映射到 Ontology 的机制。同时其严格定义的概念及概念关系为语义检索及推理奠定了良好的基础。

4 总结

不同形式的描述语言在叙词表的不同发展阶段及应用领域发挥着重要作用, 也为叙词表在语义网环境中的发展提供了可能。就目前而言, 各类型语言正处于开发阶段, 技术还不够稳定。叙词表形式化描述方法并不成熟, 没有一套完整的统一的方法论, 我们应该根据实际应用的特点选择和借鉴具体描述语言, 使叙词表在组织和检索网络资源中发挥更大的作用。

(下转第 26 页)

容进行归类 and 关系描述,使得有用的标签词汇融入到 Taxonomy 整体结构中去,同时剔除一些恶意、错误的标注,并对原有词表中的词汇及关系进行调整。这种做法既保证了用户的参与又维持了 Taxonomy 结构的一致性,在一定程度上能够保证内容的良好组织。这种方式实现了在一个信息资源系统中自然语言与人工语言的有机结合,形成了内容建设和组织工具两者之间相互促进的良性循环,使得管理内容的 Taxonomy 总是处在变化之中,而新的内容又有地方可安置。

4.4 搜索与浏览

良好的内容组织架构是实现有效找寻的基石。通过系统前端的用户标签描述和系统后端的后控词表控制两种方式的结合,为用户查找内容提供了多种途径:1) 标签云图,通过依附于 Taxonomy 的 Tagadelic 模块可以实现用户标签集合的云图显示,用户通过点击感兴趣的标签多角度选择内容。2) Taxonomy 术语组合检索,Drupal 提供了一种与大众标签系统类似的组合检索功能,用户可以将不同的术语词汇以布尔逻辑方式在 URL 里构造组合检索式。3) 分类目录,通过将 Taxonomy 组织结构在界面层次上的外显化形成一个导航式(类似 DMOZ)的分类目录浏览。导引式浏览对用户信息需求模糊情况下进行探索性找寻特别有效。4) 关键词搜索,Drupal 的搜索模块提供了关键词搜索功能。在高级搜索功能中提供了从 Taxonomy 词表结构中选择特定目录进行限定搜索。

5 小结

从以上的分析和讨论中可以看出,Drupal 作为一项开源的内容管理技术平台实际上提供了一套完

整的资源描述、组织和管理解决方案,其中 Taxonomy 功能与自由标注相结合的分类组织机制是其整体技术架构中的一个亮点所在。本文的研究仅是针对 Drupal 分类组织机制展开分析和讨论,并未涉及到 Drupal 的区域设置、视图结构、主题界面、人员管理、系统配置等其他方面。这里提出的基于 Drupal 的复合信息组织模式可供信息资源库建设参考。

最后,需要指出的是,技术驱动的信息组织模式创新与变革是当前需要正视的一个问题。在数字化网络环境下,关注技术所能解决的信息组织问题是关键所在,而非单纯研究技术本身。信息组织方法的研究与应用不可避免地要结合技术手段来实现,发掘技术发展所带来的新型组织模式并纳入到信息组织体系和为传统信息组织方法寻找技术支撑实则是同一问题的两个方面。

参考文献

- 1 张琪玉. 情报语言学基础. 武汉: 武汉大学出版社, 1997: 304
- 2 John C. Dvorak. To Tag or Not to Tag That Is the Question. <http://www.pcmag.com/article2/0,2817,1819101,00.asp> [2009-3-25]
- 3 <http://www.drupal.org/> [2009-3-25]
- 4 Drupal Handbook. <http://drupal.org/handbooks> [2009-3-25]
- 5 <http://drupal.org/project/CCK> [2009-3-25]
- 6 <http://drupal.org/project/tagadelic> [2009-3-25]
- 7 http://drupal.org/project/taxonomy_manager [2009-3-25]

范 炜 博士研究生。

(收稿日期: 2009-03-27 编发: 许桂菊)

(上接第 30 页)

参考文献

- 1 张继东. 予以胜. 利用叙词表构建本体的方法研究. 图书情报知识, 2006(6): 82-85
- 2 贾君枝. 简单知识组织系统与汉语主题词表. 中国图书馆学报, 2008(1): 75-78
- 3 Maria Lee Stewart Baillie. TML thesaurus markup language. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.2101.2009.01.01>
- 4 Thesaurus RDF- The RDF Thesaurus descriptor standard. <http://ceres.ca.gov/thesaurus/RDF.html> 2009-01-02
- 5 Dietrich H. Fischer. Converting a Thesaurus to OWL: Notes on the Paper "The National Cancer Institute's Thesaurus and Ontology". Fraunhofer Institute Integrated Pub-

lication and Information Systems 2004(2)

- 6 袁梅宇. XML Schema, RDF Schema 及 DAML 比较. 计算机应用研究, 2004(10): 67-69
- 7 杨力. 从 RDF, DAML+OIL 到 OWL. 农业图书情报学刊, 2005, 17(11): 108-110
- 8 刘春艳, 陈淑萍, 伍玉成. 基于 SKOS 的叙词表到本体的转换研究. 现代图书情报技术, 2007(5): 32-35
- 9 范 炜. 语义网环境中的叙词表实例研究. 情报科学, 2006 24(7): 1073-1077

贾君枝 女, 教授, 山西大学管理学院硕士生导师。

卫荣娟 女, 山西大学管理学院在读研究生。

(收稿日期: 2009-04-17 编发: 许桂菊)