文章编号:1673-0062(2016)04-0088-06

一种由叙词表向本体 OWL(Ontology Web Language)快速转换的转换算法

何 伟[1,2],李 波[1],李 霜[1]

(1.怀化学院 数学与计算科学学院,湖南 怀化 418008;
2.中国科学技术信息研究所 信息资源中心,北京 100038)

摘 要:随着信息环境的根本性转变,叙词表向本体的转换成为叙词表在网络环境下应用的研究热点.在分析了叙词表向本体的映射规则基础上,根据叙词表的用、代、属、分、参等关系符号设计了一种由叙词表向本体 OWL(Ontology Web Language)快速转换的转换算法.实证结果表明该算法可以有效的大幅度的提高本体构建速度,进一步促进叙词表在网络环境下的应用.

关键词:叙词表;本体构建;OWL

中图分类号:G254.24 **文献标志码**:B

A Quick Conversion Algorithm from Thesaurus to Ontology OWL

HE Wei[1,2], LI Bo[1], LI Shuang[1]

(1. College of Mathematica and Computing Science, Huaihua University, Huaihua, Hunan 418008, China;
2. Information Resource Center, Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: With the change of information environment, it becomes a focal point for many researchers to conversion between thesaurus and ontology. In this paper, the mapping rules from thesaurus to ontology were discussed, and then a quick conversion algorithm from thesaurus to ontology OWL was proposed according to the symbol of "Y", "D", "S", "F" and "C" in thesaurus. Experimental results show that our proposed algorithm can improve the efficiency of ontology construction and further promote the application of thesaurus in network environment.

key words: thesaurus; ontology construction; OWL

收稿日期:2016-06-16

基金项目:中国博士后科学基金项目(2014M550791);国家级大学生创新创业训练计划项目资助(201510548002)

作者简介:何 伟(1978-),男,湖南江永人,怀化学院数学与计算科学学院讲师,博士.主要研究方向:本体、信息组织和信息检索.

0　引　言

自从以 Tim Berners-Lee 为首的研究人员提出了语义网(semantic network)的概念后,语义技术有了快速的发展,本体逐渐成为语义网的核心,Web2.0 的基石.虽然本体在许多领域取得了巨大的成功,但是手工构建本体还是一项极其繁重的任务,耗费大量的人力、物力和财力.而叙词表这类传统的知识组织系统自上世纪六十年代以来得到了广泛的应用,大量的叙词表随着应用领域被编制出来.随着时间的推移,在信息环境发生根本性转变情况下,这些叙词表慢慢的淡出了大家的视野.然而叙词表的分类分级结构,与本体存在着许多共同的特征,且它由领域专家多年有序组织而成,概念及概念间关系表达更加规范.如何将现有的叙词表转换为本体以适应当前语义网建设和语义检索的需要,成为叙词表在语义网环境下新的发展趋势.使用叙词表构建本体,既省时省力,又能使叙词表在网络环境下焕发新的生命,促进叙词表新的应用发展.因此,叙词表向本体的转换研究成为网络环境下叙词表应用的研究热点.

近年来,国内外许多研究学者致力于叙词表向本体的转换方法研究,提出了许多根据自身问题域的叙词表转换本体的方法.Soergel 和 Lauser 等人手工从 AGROVOC 叙词表中提取出一个基本的 OWL(Web 本体语言)模型,并将 AGROVOC 叙词表进行 OWL 表示[1];唐爱民等人提出了一种基于叙词表的领域本体构建方法,此方法分为 5 个步骤,并通过《国防科学技术叙词表》构建军用飞机领域本体原型[2];杨建伟和张志平则研究了叙词表的 OWL 表示方法,建立了叙词表与 OWL 的映射规则[3];而曾新红则提出了中文叙词表本体概念,并采用 OWL 方法实现[4];李雷和厍睿则在比较了叙词表和本体的基础上,探讨了两者转换间的相关问题[5];段瑞龙和宋文则从叙词表的特征、叙词表向本体的转换过程以及转换后本体特征总结了国内外的叙词表向本体转换的方法[6];而纪珊珊等人重点讨论了叙词表向本体转换的关键技术[7];叶光辉提出了一种基于 UML 的叙词表转换成本体的方法,该方法以 UML 为媒介,设计了元素之间的映射规则[8].衣治安等人提出了基于叙词表的油田领域本体构建方法,重点解决了概念选择的不确定性等问题,并用 OWL 方法进行表示[9].

根据以上学者的已有研究成果可知,利用 OWL 描述叙词表向本体转换后的文件是合理有效的.基于此,本文尝试从叙词表中术语间语义关系出发,制定叙词表到本体转换的映射规则,建立一种快速的叙词表向本体 OWL 转换算法,并进行叙词表向本体 OWL 的批量自动转换,使叙词表在新的信息环境下焕发出更强劲的生命力,进一步降低本体构建的成本,提高本体应用效率.

1　叙词表和本体概述

1.1　叙词表

叙词表又称主题词表,根据 GB13190—91 定义为“叙词表是一种将文献、标引人员或用户的自然语言转换成规范化语言的术语控制工具,它是概括各门或某一学科领域并由语义相关、族性相关的术语组成的可以不断补充的规范化的词表.”叙词表的术语由叙词和非叙词组成,术语间规定了三种语义关系,其中叙词和非叙词间存在同义关系,叙词间有等级和相关关系,这些关系通过参照符号表示并清楚的区分术语间的语义关系,如表 1 所示.叙词表主要用于对信息进行标引时的自动或者辅助选择索引词以及进行检索时的后控制,是提高查全率、查准率、实现语义检索的重要途径[5].

表 1　叙词表的术语间语义关系符号表示

Table 1　Symbol representation of semantic relations between terms in thesaurus

关系描述	叙词表中描述	国家标准符号表示	西方标准英文简称	英文全称
同义关系	用	Y	U	Use
	代	D	UF	Use for
等级关系	属	S	BT	Boarder Term
	分	F	NT	Narrower Term
相关关系	参	C	RTRelated Term	

虽然叙词表在编制时预定义了术语间的语义关系,但在语义网环境下,无论从概念的语义及关系描述上来说,还是从推理功能上来说,叙词表都不能满足对网络信息资源进行语义标注或检索的要求.因此,在语义网建设中本体这一概念被国内外学者借用.

1.2 本体

本体来源于哲学,是哲学范畴的一个概念.一般被认为起源于古希腊,由亚里士多德和柏拉图提出的.它是对客观世界中所有事物或现象本原的一个追问,所关心的是客观现实的抽象本质.本体被引入到信息科学,引申为一种描述客观世界事物或现象的概念模型,用来组织、管理和维护信息.Gruber 给出了至今为止最为流行且被信息科学领域广泛接受的本体定义:"Ontology is an explicit specification of a conceptualization".本体是刻画某一领域内明确的知识体系结构和规范化描述,是通过大家所公认的知识对客观世界事物或现象的规范化描述.

本体被形式化地定义为五元组 O=(C,R,I,F,A),其中 C 代表概念,有时也叫做类,用来描述领域内具有相同属性的一组对象;R 用来表示关系,是指领域中概念之间以及实例之间存在的各种联系;I 代表实例,是领域内某一特定的具体对象;F 表示函数,是一种特殊的关系,在这种关系中,前 $n-1$ 个元素可惟一确定第 n 个元素;A 代表公理,是一个永真断言,用来表示对领域内一些常识性知识的描述.由此看出,概念、关系、实例是构成本体的基本要素,是本体的重要组成部分.从语义关系上分析,本体中基本的关系有 4 种:part-of、kind-of、instance-of 和 attribute-of[4].这些关系类型在进行 OWL 表示时,可转化为"IS-A"、"ispartof"等.在具体的本体构建过程中,不一定要按照本体的五元组形式化定义来完成,可省略某些成分如函数 F、公理 A,本体上概念间语义关系也不拘泥于上述 4 种,如通用本体大多数只包含同义、等级和相关关系,可根据实际的需要进行本体构建.

1.3 本体 OWL 语言

OWL(ontology Web language)是由 W3C(the world wide Web consortium)在 2004 年所发布并推荐的一种用来描述语义网和本体的形式化标准语言.它是根据 DAML+OIL 发展起来的,是一种定义和具体化 Web 本体的描述语言[10].OWL 一共具有三个子语言:OWL Lite、OWL DL(OWL description logic OWL 描述逻辑)以及 OWL Full.它们所具备的表达和推理能力渐渐增强,从而可供不同需求的使用者进行选择.

OWL Lite,主要适用于只需要一个分类层次和简单约束的用户[11].该语言是表达最弱的 OWL 子语言.例如,它支持基数约束时,限定基数值只能取值为 0 和 1 两者之一.跟其它支持关系表达的本体描述语言比较,OWL Lite 相对比较简单,它为其它分类结构以及叙词表提供了一种快速的融合方法.

OWL DL,主要适用于那些需要在推理系统上进行最大表达且不失计算完整性(computational)和可决定性(decidability)的用户[11].OWL DL 提供了尽可能大的表达能力.OWL DL 的两个特性中计算完整性保证所有的结论都能被完整的计算出来,可决定性保证所有的计算都能在有限的时间内完成.该语言包含了 OWL 语言的所有约束,但是可以被仅仅置于特定的约束下.

OWL Full,主要适用于那些需要在没有计算保证的、语法自由的 RDF 上进行最大程度描述的使用者[11].它包含了 OWL 的全部语言构造成分,是一种比较自由的 Web 本体语言,它的语法可以根据使用者的具体情况随意定义.它允许在一个本体的预定义的词汇表上增加使用者适用的词汇,更进一步的满足使用者的需求.在 OWL Full 中,一个类既可以表示成个体的集合,也可以单独被说明成一个独立的个体.跟 OWL DL 最大的区别是,在 OWL Full 中一个被声明的数据类型属性(owl:DatatypeProperty)也可以被当做逆函数属性(owl:InverseFunctionalProperty)来进行定义.

用户应该根据应用的表达需求选择适合自己的 OWL 子语言.在决定是选择 OWL Full ,还是选择 OWL DL 时,使用者应考虑在应用中自动推理能力重要还是表达能力更重要,以及取决于用户对 RDF 的元模型机制的需要程度;而在选择 OWL Lite,还是选择 OWL DL 时,则依据用户对整个语言的约束的可表达性需求程度.

根据应用的需要,所转换的本体下一步主要用于智能语义检索,故本文使用表达能力更强的 OWL Full 来描述叙词表转换后的本体.

2 叙词表向本体 OWL 的转换算法

鉴于叙词表与本体间存在类似特征,将叙词表转换成本体 OWL 方式的主要思想为:首选选择叙词表中合适的术语作为本体概念;其次根据叙词表中术语间的用(Y)、代(D)、属(S)、分(F)、参(C)等符号系统来形成本体中概念间的语义关系,并同时处理与概念相关的其他问题,如范畴

号、注释等;最后根据这些映射规则,设计了一种快速的叙词表向本体 OWL 转换的算法,并用算法运行时间和空间复杂度对算法进行了性能测试.

1)本体概念的确定

叙词表中的术语由叙词和非叙词组成,这就决定了并不是叙词表中的所有术语都可作为本体概念或者类来使用.具体做法是将叙词表中的每一叙词转化为本体概念使用,并声明为一个类,而非叙词则作为对应叙词的同义属性来对待.由于叙词表中并不存在本体 OWL 中这样一个相对应的术语,该术语是所有叙词的祖先概念,所以可以人为的定义一个顶层概念 Thing 作为这些叙词的共同祖先,每个概念都是"Thing"的子类.

2)本体概念间关系的确定

通过叙词表中术语间的用(Y)、代(D)、属(S)、分(F)、参(C)来确定本体概念间的语义关系.其中用(Y)、代(D)作为叙词与非叙词间的同义属性对待,属(S)、分(F)转换为概念间的等级关系,参(C)对应概念间的相关关系.

(1)用代关系符号的 OWL 转换

叙词表中的术语可分为叙词和非叙词,它们之间存在着互为同义词,或近义词,或反义词等,通过"Y"和"D"符号来联系,其中用(Y)表示后接的词是叙词,而代(D)表示后接的是非叙词.由于非叙词所包含的信息在叙词款目中完全可以反映出来,它们表达的是相同或相近或相反的含义.在叙词表中叙词和非叙词间的关系统称为同义关系.因此,合并非叙词并不影响信息的损失.将所有的非叙词用叙词取代,并使非叙词成为与其对应的叙词的同义属性.

由于 OWL 没有任何预定义的数据类型,也没有提供特别的定义机制,但可以使用 XML Schema 提供的数据类型声明两个数据属性 hasDescriptor 和 hasNonDescriptor,分别指代叙词和非叙词.如:

高等教育(14F)

D 大学教育

大学教育(14F)

Y 高等教育

在 OWL 中转换为:

<owl: Class rdf: about="高等教育" >

<hasDescriptor rdf: resource="高等教育" />

< hasNonDescriptor rdf: resource = "大学教育" />

</owl: Class>

(2)分关系符号的 OWL 转换

在叙词表中,叙词间的等级关系是通过"S"和"F"符号来描述的,其中属(S)表示该符号后面的词是叙词款目词的上义词,而分(F)代表该符号后的词是叙词款目词的下义词.因此,在对这类叙词间关系进行本体 OWL 转换时,只要将它们声明成"rdfs:superClassOf"和"rdfs:subClassOf"就可描述这种关系.如:

各级教育(14F)

F 高等教育

S 教育

在 OWL 中转换成:

<owl: Class rdf: about="各级教育" >

<rdfs: subClassOf rdf: resource = "#高等教育" />

<rdfs: superClassOf rdf: resource = "#教育" />

</owl: Class>

(3)相关关系符号的 OWL 转换

叙词表中叙词间的关系除了等级关系外,其余的关系可统称为相关关系,使用符号"C"(参)表示.说明该符号后接的词是叙词款目词的语义相关词,也就是声明了该概念的相关概念.用<rdfs:seeAlso>在 OWL 中进行声明.如:上例中

高等教育(14F)

D 大学教育

S 各级教育

C 高等学校

将其进行 OWL 转换为:

<owl: Class rdf: about="高等教育">

<hasDescriptor rdf: resource="高等教育" />

< hasNonDescriptor rdf: resource = "大学教育" />

<rdfs: superClassOf rdf: resource = "#各级教育" />

<rdfs: seeAlso rdf: resource="#高等学校" />

</owl: Class>

(4)范畴号和注释的 OWL 转换

在叙词表中还有一类比较特殊的符号:范畴号和注释.其中范畴号用来表示该叙词的分类号,可声明成属性-值,通过定义数据属性 hasCLCCode 来进行表示;注释主要用于进一步解释该叙词的主要含义,可通过使用 OWL 标签 rdfs: comment 来进行描述.如

双百方针 (01B;15C) ←范畴号

(百花齐放,百家争鸣) ←注释

将其进行 OWL 转换成:

<owl：Class rdf：ID=“双百方针”>

<rdfs：comment>百花齐放,百家争鸣<rdfs：comment />

<hasCLCCode rdf：resource=“01B；15C”/>

</owl：Class>

3)快速转换算法设计及实证分析

根据上述的本体概念确定以及概念间关系的转化方式,本文设计了一种快速的算法实现叙词表向本体 OWL 的自动化转换,如算法 1 所示.

算法 1

输入:叙词表存储文件

输出:叙词表的本体 OWL 表示方式

Step1. 声明本体的一个顶层类“Thing”;

Step2. 读取叙词表中的每一个术语,若某术语前出现符号“D”,则该术语不作为本体概念使用,只将其作为对应概念的同义属性对待,其余叙词表术语则直接当作本体概念对待;

Step3. 若某术语后直接出现“Y”,则说明该词条是非叙词,可将其忽略,因为该术语肯定会在其对应的叙词中作为同义属性出现;否则,以叙词为标识将概念声明为一个类,并根据实际情况使用 rdfs：comment 标签加上注释以及使用 hasCLCCode 说明该类范畴.根据叙词中的关系符号 Y、D、F、S、C 为类添加属性,包括此概念所包括的叙词、非叙词,此概念的上下位概念、相关概念、分类号、注释等.

Step4. 重复 Step3,直至叙词表文件的最后一行.

Step5. 输出叙词表的本体 OWL 表示文件,结束.

使用本文算法 1 对《综合电子政务主题词表》进行自动化转换本体 OWL 方式,并用 Protégé (http://protege.stanford.edu/products.php#desktop-protege)进行展示,部分截图如图 1 所示,图中显示的是有关“教育”的部分概念等级关系.所用的《综合电子政务主题词表》由中国科学技术信息研究所于 2005 年 1 月编制完成,共收录词条 20252 条,其中叙词 17421 条,非叙词 2831 条.范畴索引划分为 21 个大类,132 个二级类[12].在算法的转换实现过程中,由于“S”和“F”符号关系是完全对称的,为了压缩空间,减少冗余,提高转换效率,只声明了“S”关系符号,以进一步提高算法的效率,降低算法的空间复杂度.

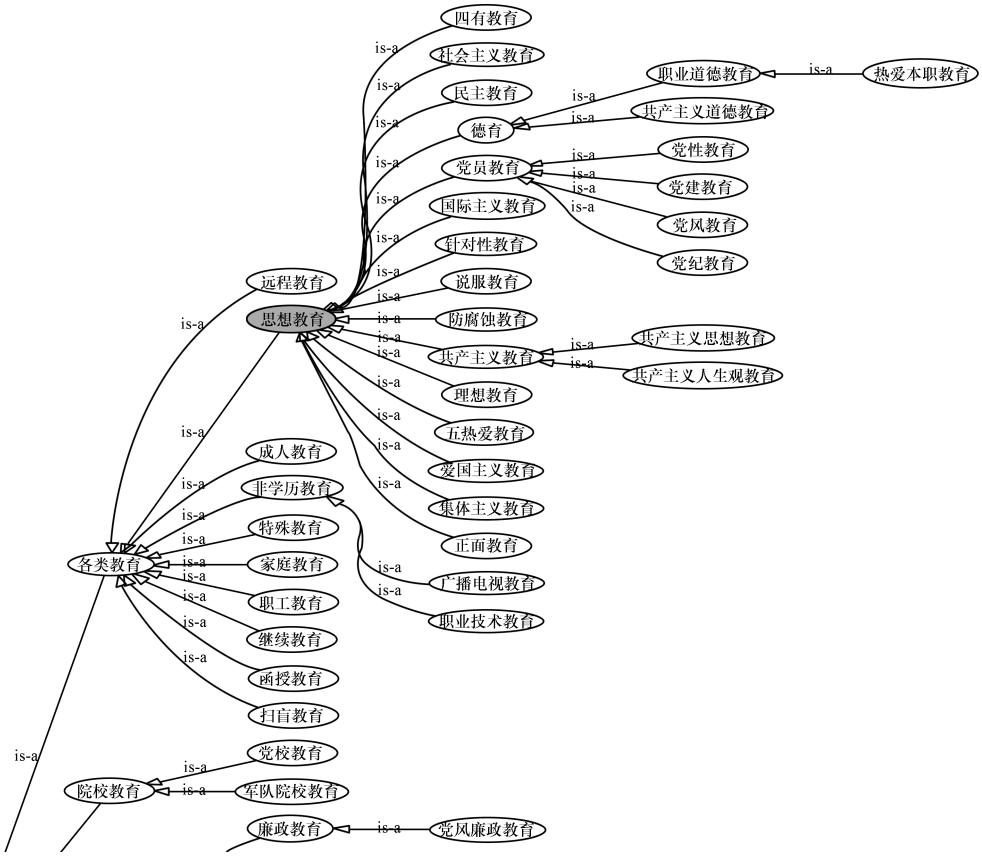

图 1 电子政务主题词表转换的本体 OWL 显示(部分)

Fig.1 Figure of ontology OWL converted from e-government thesauru (Part)

为了进一步分析算法的效率,我们分别使用《综合电子政务主题词表》整个叙词表以及该表中的“教育”子族,采用算法运行时间以及空间复杂度作为性能指标对本文算法进行实证分析.结果如表 2 所示.

表 2　算法实证分析结果

Table 2　Experimental results of our proposed algorithm

待转换叙词表	性能指标	
	算法运行时间/s	空间复杂度
电子政务主题词表	13 547(约 3.57 h)	O(1)
“教育”子族	2 169 (约 0.6 h)	O(1)

从表 2 中可看出,随着叙词表规模的增大,时间也发生较大的改变,但是对于从叙词表中分解出的一个子族来说,这是一个可接受的时间.因此,可将叙词表分解成子族,对每一子族进行本体 OWL 转换,再通过子族进行融合,可大幅度的降低转换时间.而另一个性能指标算法的空间复杂度始终维持在 O(1),这是因为在设计算法时,只定义了一个字符数组用于临时存储从叙词表文件中读取的一行叙词表术语,每转换一行,就将其写入本体文件,故在算法运行过程中并不在内存中大量保存词条信息.

由于受到 protégé 工具只能展示本体 OWL 方式的等级关系的影响,图 1 并没有显示 OWL 中的同义属性以及相关关系.

3　结　论

本体是语义网建设的基石,OWL 是 W3C 组织推荐的一种本体描述语言,能较好的对本体进行语义描述,然而构建本体 OWL 表示是一项比较繁杂的任务.叙词表在某些特征上与本体存在相似的地方,如术语及其间语义关系.因此,本文在分析了叙词表向本体的映射规则基础上,设计了一种由叙词表向本体 OWL 快速转换的转换算法,实现本体 OWL 的半自动化构建.实验结果表明该方法可进一步缩短本体构建的时间,降低构建成本.从而,加快本体的应用速度,也为叙词表在网络环境下的应用探索新的方向.

参考文献:

[1] Soergel D, Lanser B, Liang A, et al. Reengineering thesauri for new applications: the AGROVOC Example [J].Journal of Digital Information,2004,4(4):1-23.

[2] 唐爱民,真溱,樊静.基于叙词表的领域本体构建研究[J].现代图书情报技术,2005,21(4):1-5.

[3] 杨建伟,张志平.叙词表的 OWL 表示方法研究[J].情报杂志,2007,26(3):19-21.

[4] 曾新红.中文叙词表本体-叙词表和本体的融合[J].现代图书情报技术,2009,25(1):34-43.

[5] 李雷,厍睿.叙词表和本体的比较研究[J].浙江高校图书情报工作,2012(1):15-17.

[6] 段瑞龙,宋文.国内外叙词表转换本体方法研究综述[J].情报杂志,2012,31(7):66-71.

[7] 纪姗姗,刘峥,宋文.叙词表向本体重构的关键技术研究[J].图书与情报,2013(1):8-12.

[8] 叶光辉.基于 UML 的叙词表转本体方法研究[J].图书馆杂志,2014,33(7):28-33.

[9] 衣治安,王月.基于叙词表的油田领域本体构建方法[J].计算机系统应用,2015,24(9):91-96.

[10] HACHEROUF M,BAHLOUL S,CRUZ C.Transforming XML documents to OWL ontologies: a survey [J]. Journal of Information Science,2015,41(2):242-259.

[11] 杜小勇,李曼,王大治.语义 Web 与本体研究综述[J].计算机应用,2004,24(10):14-16.

[12]《电子政务主题词表》编制与应用系统课题组.综合电子政务主题词表:试用本[M].1 版.北京:科学技术文献出版社,2005.