一种有效的本体创建方法——面向对象法

李宝敏[1]，张 娜[2]
（1. 西安工业大学 计算机学院，陕西 西安 710032；
2. 平顶山工学院 计算机科学与工程系，河南 平顶山 467044）

摘 要：针对目前语义网中的本体缺少有效和统一的方法，在综合传统的构建领域本体方法的基础上，将面向对象的思想引入到领域本体的构建过程中，提出了一种有效构建本体的新方法——面向对象构建法。通过在果品领域本体的构建，定义类及层次，定义类之间关系，定义类的属性，定义类之间的同义、近义关系，并将其映射为本体体系，与检索系统的应用结合起来。实践证明，该方法易学、易用，可减少领域专家的参与，减少开发系统的工作量，并且使得本体的半自动和自动构建问题成为可能。

关键词：本体；面向对象；领域；语义网
中图分类号：TP311 **文献标识码**：A **文章编号**：1673－629X(2008)10－0034－03

An Effective Method of Establishing Ontology —Object－Oriented Method

LI Bao-min[1], ZHANG Na[2]
(1. Institute of Computer Science and Engineering, Xi'an Technology University, Xi'an 710032, China;
2. Dept. of Computer Science and Engineering, Pingdingshan Institute of Tech., Pingdingshan 467044, China)

Abstract: Because there is a lack of effective and unite method of establishing ontology in the semantic Web, in the synthesis traditional construction domain main body method foundation, the object－oriented thought is introduced the domain ontology in the construction process, proposed one kind of effective construction ontology new method—object－oriented construction law. Through in the fruits domain main body's construction, the definition class and the level, between the definition class relates, the definition class attribute, the definition class between synonymy, near righteousness relations, and its mapping is the ontology system, unifies with retrieval system's application. The practice proved that this method easy to study, to use, may reduce the domain expert's participation, reduces development system's work load, and causes the ontology semiautomatic and the automatic construction question possibly becomes.

Key words: ontology; object－oriented; domain; semantic Web

0 引 言

语义网研究目前成为一个热门的话题。在语义网的七层结构模型中[1]，本体是最为关键的一层，文中在综合传统的构建本体方法的基础上，将面向对象的思想引入领域本体的构建过程中，提出一种构建本体的新方法——面向对象构建法，并将其与检索应用结合起来。实践证明，该方法易学、易用，减少了领域专家的参与、减少开发系统的工作量，并且使得本体的半自动和自动构建问题成为可能。

收稿日期：2008－01－22
基金项目：国家"星火计划"项目(2004EA850069)
作者简介：李宝敏(1949－)，男，河南巩义人，教授，研究方向为计算机系统结构、计算机网络与语义网。

1 本体的创建

1.1 领域本体的构建准则

T. R. Gruber 提出了指导本体构建的5个规则[2]：

(1) 清晰性(Clarity)。本体必须能够明确地说明所定义术语的含义。定义应该是客观的，当定义可以用逻辑公理表达时，它应该是形式化的。定义要尽可能的完整，本体中的定义应该用自然语言加以说明。

(2) 一致性(Coherence)。本体应该是一致的，也就是说，它应该支持与其它定义相一致的推理。它所定义的公理以及用自然语言进行说明的文档都应该具有一致性。

(3) 可扩展性(Extendibility)。本体应该为可预料到的任务提供概念基础。它应该可以支持在已有的概念基础上定义新的术语，以满足特殊的需求，而无需修

改已有的概念定义。

(4) 编码偏好程度最小(Minimal Encoding Bias)。概念的描述不应该依赖于某一种特殊的符号层的表示方法。因为实际的系统可能采用不同的知识表示方法。

(5) 本体约定最小(Minimal Ontology Commitment)。本体约定只要能够满足特定的知识共享需求即可。这可以通过定义约束最弱的公理以及只定义通讯所需的词汇来保证。

1.2 常用的领域本体构建方法

按照 Gruber 提出的 5 个指导原则,Gruninger 和 Fox 等人在 TOVE 项目中提出了基于场景的方法等。图 1 表示的是一种基于概念的层次常见的方法。

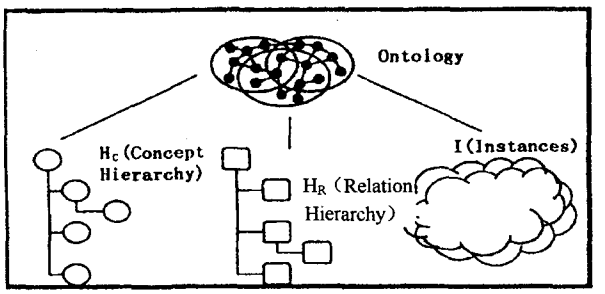

图 1　一种常见的本体组成形式

1)建立概念层次结构 HC(Concept Hierarchy)。从领域知识中抽取领域中概念之间的层次结构,并直接映射为本体中的概念层次结构 HC,而不考虑概念在层次树中的位置所表示的实际语义关系。

2)建立联系概念的关系层次结构 HR(Relation Hierarchy)。通过关系的定义域和值域的设置,限制概念之间的联系,从相互作用的角度隐含地体现了概念在概念层次 HC 中的实际语义。

3)用实例集 I(Instances)填充概念层次 HC 中的概念。在这种方法中,本体 Ontology 就是概念层次结构 HC、关系层次结构 HR 和实例集 I 所共同构成的一个整体。

另外一种方法是基于描述逻辑(DL, Description Logic)[3]中对 TBox 和 ABox 的定义,即:

(1) 从领域知识中找出一些最基本最一般的概念形成最初的概念集 C,一些最基本的关系形成关系集 R;

(2) 上面形成的概念集 C 中的概念和关系集 R 中的关系,通过描述逻辑中的逻辑关系运算符,定义新的复合的概念,并把新的概念添加进概念集 C;

(3) 重复(2)的过程直到认为形成的概念足够描述应用中的领域知识,这时(2)中的复合概念的定义式的全体就形成了 TBox;

(4) 用实例集 I 填充概念集 C 中的概念和关系 R 中的关系的定义域和值域,这时用于表示填充式子的全体就形成 ABox。

这种方法在 TBox 中显式地声明概念之间的语义联系,完全不同于前面的方法;这种方法还要对关系的定义域和值域进行实例集的填充。所以这种方法适合于本体工程人员较难把握需要构建本体的领域知识结构的情况。

1.3 面向对象构建法

在综合上述两种构建本体方法的基础上,提出面向对象的领域本体的构建过程中,将其与检索应用结合起来。具体方法如下:

1)定义类及层次关系,从领域中自上而下抽象出基本的类及其层次关系。这里的类对应本体体系中的 Class,对象对应本体体系中的 Instance 或 Individual。

2)定义类之间的关系,主要是聚合关系,对应到本体体系中的 Object Property。

3)定义类的属性,包括属性的名称、类型及其他约束,对应到本体体系中的原子属性(Datatype Property)。

4)把通过前 3 步定义建立的模型映射为本体体系。

5)定义同义、近义等语义扩展关系,在语义检索中来提高其查全率。

理论分析此方法将具有如下特点和意义:

(1)易学、易用,减少了领域专家的参与。面向对象的思想和方法发展至今,已比较成熟,并已在多个行业和领域中广泛应用,因此这个方法对很多从事过面向对象程序设计的软件开发人员来说易学、易用。

(2)减少开发系统的工作量。在开发基于本体论的语义检索系统时,前期的需求分析和数据库设计与此方法实现过程有不少交叉,可以减轻不少工作量。

(3)使本体的半自动和自动构建问题成为可能。面向对象的很多概念和本体体系中的术语有很强的映射关系,因此,利用面向对象建模工具和本体构建工具的 API 就有可能解决本体的半自动和自动构建问题。

2 构建方法的验证

文中采用了面向对象构建法来构建果品领域本体。便于比较,把面向对象构建法第 4 步的映射分散到前 3 步中,具体见下文。

2.1 定义类及层次关系

首先定义基本类及层次关系,从农业生物学角度定义果树的本体体系[4]。按照我国对地理位置的划分将地区分为华北、华南等,以定义地区的本体体系。基本类和层次关系如图 2 所示,映射到本体体系中的

Class 及层次关系如图 3 所示。

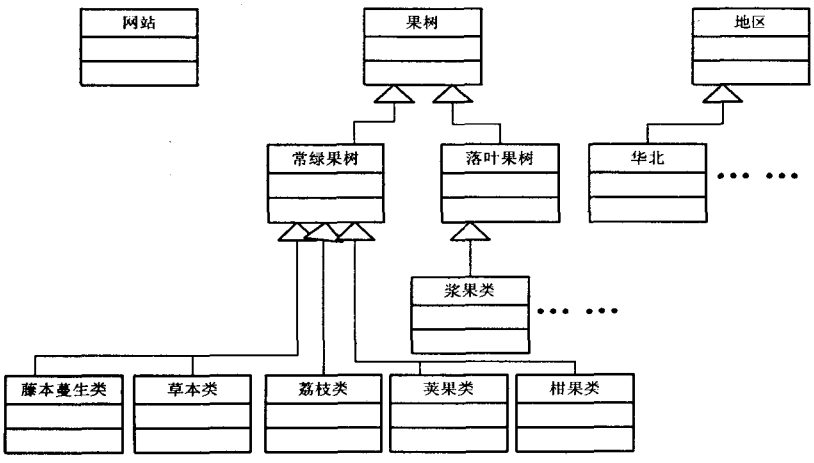

图 2 定义的类及层次关系

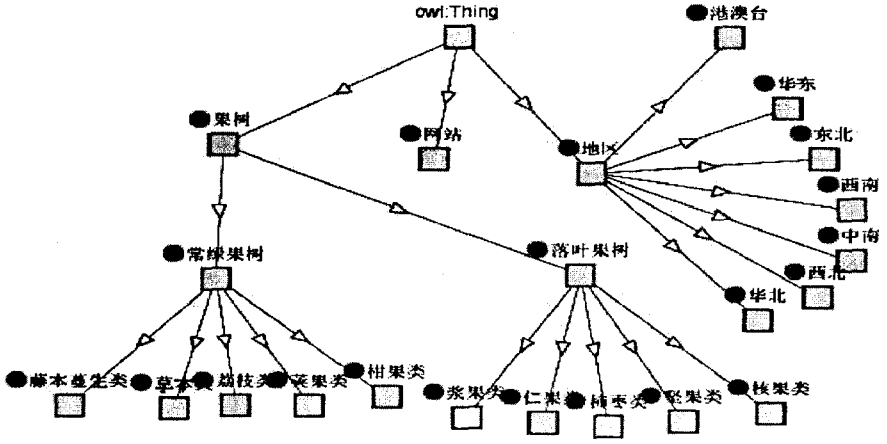

图 3 本体体系中 Class 及其层次关系

2.2 定义类之间的关系

类之间的关系如图 4 所示，映射成本体体系中的 Class 之间的关系如图 5 所示。

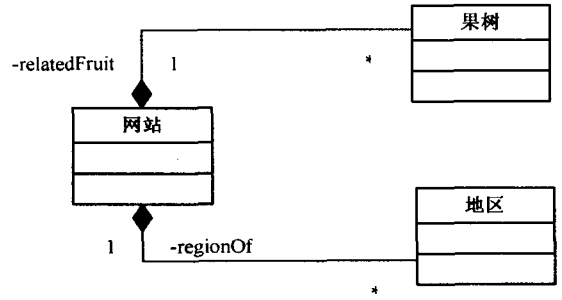

图 4 类之间的关系

2.3 定义类的属性

类的属性 Property 如表 1 所示。

表 1 Property

名称	domain	range	类别	基数
Website:WebsiteName	网站	网站	Datatype	Cardinality=1
Website:provinceOf	网站	地区	Object	Cardinality=1
Website:WebAddr	网站	网站	Datatype	Cardinality=1
Website:fullDesc	网站	网站	Datatype	Cardinality=1
Website:relatedFruit	网站	果树	Object	minCardinality=1

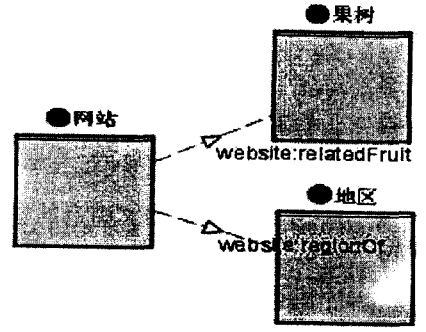

图 5 本体体系中 Class 之间的关系

2.4 定义类之间的同义、近义关系

主要是借助于 Protégé 工具来定义[5]。同义和近义表示了对象的语义关系，为的是提高检索的查全率。定义同义关系的 Protégé 界面（部分）如图 6 所示。

用本体中的概念和属性来标注网页信息，在语义智能检索中以便准确定位信息资源，也就是用户要查询的信息，以提高查准率；而图 2 中的本体体系用于上、下位关系的查询，它与图 6 所定义的同义、近义关系共同来提高系统的查全率。

3 结束语

在传统本体创建的基础上提出了新的领域本体构建方法——面向对象构建法，即：把面向对象的思想运用到领域本体的构建过程中，与传统的本体创建方法相比，该方法易学、易用，减少了系统开发量，为解决本体的自动构建问题进行了一种新的途径和思路的尝试。当然，对于不同的领域构建本体是否具有通用性答案是无需质疑的。

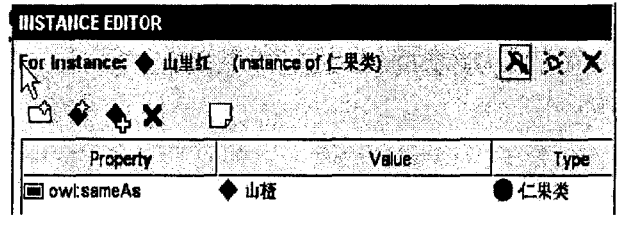

图 6 Protégé 用户界面——定义同义关系

参考文献：

[1] Harmelen F, Hendler J, Horrocks I, et al. OWL Web Ontol-

（下转第 39 页）

目,$N(w_k,di)$ 表示单词 w_k 在该类文档 di 中出现的次数,$|V|$ 表示该类中的词汇数量,$\sum_{i=1}^{|V|}\sum_{i=1}^{|D|}N(w_n,di)$ 为该类所有词的数目和。

(4) 计算整个标题信息输入类 C_j 的几率,公式:

$$P(C_j|di,\hat{\theta}) = \frac{p(C_j|\hat{\theta})\prod_{k=1}^{n}p(w_k|C_j;\hat{\theta})^{N(w_k,di)}}{\sum_{r=1}^{|C|}p(C_r|\hat{\theta})\prod_{k=1}^{n}p(w_k|C_r;\hat{\theta})^{N(w_k;di)}}$$

其中,$P(C_j|\hat{\theta}) = \frac{\text{第 } j \text{ 个类别的训练文档数}}{\text{总训练文档数}}$,$P(C_r|\hat{\theta})$ 为相似含义,$|C|$ 为类的总数,n 为特征词总数。

(5)给出对应文本属于各类别的概率大小,将文本确定为概率最大的那个类别。

3　实验分析

中文网页数据集是实现中文网页自动分类的前提和基础,但是到目前为止还没有出现标准的中文网页数据集。为了考查提出方法的可行性,从新浪网、人民网和新华网这三大网站上采集了中文网页(共计 1000 篇)作为数据测试集。分类体系则采用门户网站的一般类别,分为政治、财经、娱乐、军事、体育、教育、饮食、卫生、环保、法治等 10 类,在此基础上,信息抽取准确率的实验结果如表 1 所示。

表 1　网页正文信息抽取结果

网页来源地	网页数目	正确抽取网页数	错误提取网页数	准确率
新浪网	426	412	14	96.7%
人民网	357	337	20	94.4%
新华网	326	291	35	89.3%
合计	1109	1040	69	93.8%

在信息抽取准确的基础上,进行了分类实验,实验结果如表 2 所示。

在观察实验结果后发现,这种方法对于少部分未含有标题信息的文本无效,同时,在部分标题中带有特征分属多个类别的特征词时,效果还有待提高。但从总的分类效果来看,分类准确率的平均数达到了 84% 以上,应该说分类结果可行,而这种方法显然比整个文本信息进行 SVM 计算简单,复杂性降低。

表 2　网页文本分类结果

具体类别	网页数目	正确分类网页	准确率
政治	216	187	86.6%
财经	225	194	86.2%
娱乐	305	286	93.8%
军事	159	127	80%
体育	247	209	84.6%
教育	168	148	88.1%
饮食	92	78	84.8%
卫生	146	117	80.1%
环保	96	79	82.3%
法治	138	106	76.8%

4　结束语

网页的正文信息的提取和文本的自动分类在信息检索领域中均占有十分重要的意义,文中结合了传统的文本文类贝叶斯算法,在 HTML 自身特点的基础上,对网页文本信息实现了正文信息提取和文本自动分类,在此基础上进行了实验,取得了较好的分类结果。

中文网页不同于普通的文本文件,它包含大量的网页标记信息,这些一般成对出现的标记对为正文的信息抽取提供了可供使用的含义,文中正是在对这些标记对信息研究的基础上实现正文信息的抽取的。而所采用的方法同一般采用的 SVM 分类算法相比,复杂度明显降低,分类效率明显提高。

参考文献:

[1] 郭庚麒.Web 文本挖掘技术[J].计算机技术与发展,2004,14(1):114－116.

[2] 冯伟华,苗长芬.基于 Web 的网页信息抽取方法的研究[J].洛阳工业高等专科学校学报,2005,15(3):30－31.

[3] 许　文,都云程,李渝勤,等.一种通用 HTML 网页主题信息提取方法[J].现代图书情报技术,2007(1):40－43.

[4] 程传鹏.中文网页分类的研究与实现[J].中原工学院学报,2007,18(1):61－64.

[5] 王晓霞,尹四清.网页分类技术的研究[J].机械工程与自动化,2007(1):75－77.

(上接第 36 页)

ogy Language Reference[EB/OL]. World Wide Web Consortium. 2004－02－10. http://www.w3.org/TR/OWL－REF,20040210/.

[2] 林鸿飞,战学刚,姚天顺. 基于概念的文本结构分析方法[J]. 计算机研究与发展,2000(3):324－328.

[3] Baader F, McGuinness D L, Nardi D, et al. The Description Logic Handbook: Theory, implementation and applications [M]. Cambridge:Cambridge University Press, 2003.

[4] 张　娜,李宝敏.语义检索及其关键技术研究[J].计算机技术与发展,2006,16(11):24－25.

[5] 李宝敏,张　娜.基于领域本体的语义智能检索研究[J].情报杂志,2007(12):125－126.