电子商务领域本体知识库的构建及应用①

王志强　任　燕　郭　宁　傅向华　(深圳大学 计算机与软件学院 广东 深圳 518060)

摘　要: 本文利用本体丰富的语义知识和语法结构及其共享性，构建了电子商务领域的本体知识库，用于解决数据的结构异构和语义异构问题。同时开发了面向电子商务领域本体知识库的汉语自动分词系统。结果表明，引入本体知识库能在一定程度上提高词语切分的正确率。

关键词: 本体; 领域本体知识库; 汉语自动分词

Construction and Application of Domain Ontology Repository for Electronic Commerce

WANG Zhi-Qiang, REN Yan, GUO Ning, FU Xiang-Hua

(College of Computer & Software Engineering, Shenzhen University, Shenzhen 518060, China)

Abstract: Ontology contains abundant semantic knowledge, syntax structure, and shared information. In this paper, a domain ontology repository for electronic commerce is firstly established to solve the problem of data structural and semantic isomerism. And then a prototype system of automatic Chinese word segmentation is developed based on the domain ontology repository. The experiment results show that the recall rate of Chinese word segmentation is improved partly.

Keywords: ontology; domain ontology repository; automatic Chinese word segmentation

1　引言

随着互联网使用人数的增加，利用互联网进行网络购物并以银行卡付款的消费方式已渐流行，电子商务所占的市场份额迅速增长，电子商务信息量激增。大部分电子商务网站采用数据库来存储信息，不同的网站和互联网用户之间信息交换需求日益增加，数据的应用价值低、共享性差的问题越来越引起关注。造成这种问题的主要原因是数据的结构异构和语义异构。

本文引入本体论知识，利用本体(ontology)丰富的语义知识和语法结构及其共享性，构建了电子商务领域本体知识库(Domain Ontology Repository, DOR)，用于解决数据的结构异构和语义异构问题。由此构建的本体知识库可以应用在汉语自动分词领域，利用词条间丰富的语义联系，弥补传统分词词条孤立的不足，能有效解决汉语自动分词的歧义切分问题。

本文第 2 部分论述了本体知识库的相关知识，第 3 部分构建了电子商务领域本体知识库，第 4 部分开发并实现基于领域本体知识库的汉语自动分词系统。研究结果表明，引入本体知识库构建的汉语自动分词系统，能在一定程度上提高词语切分的正确率。

2　本体知识库

本体论原是一个哲学概念，指关于存在及其本质和规律的学说，后被用于研究实体存在性和实体存在本质等方面的通用理论。在知识工程领域，人们普遍接受的呈现高引用率的本体定义是 Gruber[1]于 1993 年提出的：本体是对共享的概念化进行形式的显示规范说明。其中，“概念化”是现实世界中现象的抽象模型，要明确标识与现象相关的概念；“显式”是指被使用概念的类型以及概念在使用中的约束被明确地定义出来；“形式”是指本体应该是机器可读的；“共享”是反映本体中的知识是中立的一致认可的。

① 基金项目:广东省自然科学基金项目(07301329);深圳市科技计划资助项目(200741)
收稿时间:2010-03-17;收到修改稿时间:2010-04-09

2.1 知识库中本体的定义

知识本体可以用一个三元组定义[2]：KO=<KA，Rela，Rule>。其中，三元组中各字母代表的含义如下：

(1) KA(Knowledge Atom)为知识原子， 表示整个知识模型中最小的知识表示单元，可以是公理、概念及基本的操作关系等，即：$KA=\{a_i|1\le i\le n,a_i\notin\varnothing,a_i\in\Omega\}$，表示知识论域$\Omega$中的知识原子。

(2) Rela(Relation)表示知识原子之间以及由知识原子构成的知识实体之间存在的相互联系作用和影响的集合，即：$Rela=\{r_{ij}(a_i,a_j)\vee r_{kl}(b_k,b_l)|1\le i,j,k,l\le n,r_{ij},r_{kl}\notin\varnothing\}$，其中，b 表示由知识原子构成的知识实体， 即：$b=\{\sum b_ib_j\vee\prod b_ib_j|1\le i,j\le n,b_i,b_j\notin\varnothing,b_i,b_j\in\Omega\}$。

其中，$r_{ij}(a_i,a_j)$表示知识原子之间的关系，$r_{kl}(b_k,b_l)$表示知识实体之间的关系。

(3) Rule 表示知识原子或知识实体之间的关系组合生成的一些规则或操作集。

2.2 领域本体的表示

领域本体的建立对于需要交换信息、共享信息的用户或异构系统来说，将有助于消除在概念和术语上的分歧，对领域内的概念理解达成共识。Guarino[3]等人提出用词汇概念图(Lexical Conceptual Graph，LCG)表示本体的方法。LCG 是一种带标记的有向图，图中结点表示概念，有向边表示关系，结点中的词汇代表概念的名称，有向边上的词汇表示连接两个结点之间的关系。图 1 以“电子元器件”本体的子概念“电容器”和“电阻器”有向图为例表示了概念间的基本关系。

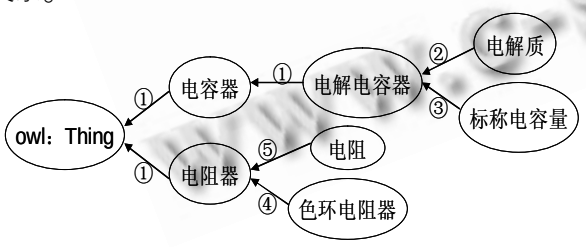

图 1 “电子元器件”本体部分有向图

有向边上的符号表示概念间的关系。符号①②③④分别表示概念之间的 4 种最基本的关系：part-of、kind-of、attribute-of 和 instance-of，其中 part-of 表达概念之间整体和局部的关系；kind-of 表达概念之间的继承关系；attribute-of 表达某个概念是另外某个概念的属性；instance-of 表达概念和概念的实例之间的关系。符号⑤表示同义词关系 synonym， 它表达了在相似数据源间的一种等价关系。如图 1 所示，“色环电阻器”是“电阻器”的一个实例，“电阻”是“电阻器”的同义词等。

3 电子商务领域本体知识库的构建

3.1 领域本体的构建方法[4]

本体构造实质就是模型化一个领域，它的构造过程必然是一个多次重复、逐步求精的过程。首先，获取领域的有关知识实体并建立知识链。通过概念以及概念间的关系语义将知识进行层次化。其次，用中间表达集合对知识链进行概念化，中间表达用类对领域知识的知识实体、属性进行描述和定义，实现本体的结构化。中间表达阶段包含 3 个模块，对现有本体的整合、编码及其概念化。最后，实现概念模型并对本体进行评估。一种结构化开发领域本体的方法如图 2 所示。

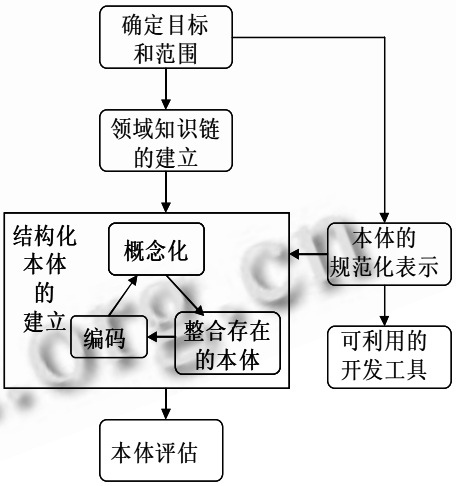

图 2 领域本体的构建方法

3.2 本体知识库的构建

借助常用的本体编辑工具，选取了电子商务领域中的电子元器件、服装、传媒、家居、机械及行业设备、照明工业、精细化学品、家用电器、通讯产品及安全防护等 10 个栏目。以电子元器件栏目为例，选取了电容器、电阻器、集成电路 IC、二极管、三极管、光电子器件等共 6 个大类。每个大类都含有常规的子类，并对子类建立电子商务领域较为常用的实例。词条的选取依据门户网站“阿里巴巴”中关于电子元器件的分类。每个类别的子类和实例的词条数量和从 30

到 80 不等，整体的词库数量约为 400 条。全部栏目的词库数目约有 4300 条，各个栏目之间不可避免的将出现一些重复词条，这是词语本身的特点所决定的。“电子元器件”本体的子概念“电容器”和“电阻器”部分有向图关系如图 1 所示。

表 1 是本文构建的部分电子商务领域本体所产生的 OWL 格式文件：

表 1　部分电子商务领域本体的 OWL 格式文件

```
</owl:Class>
<owl:Classrdf:ID=“电容器”>
<rdfs:SubClassofrdf:resouree=“# 电 子 元 器 件”/>
</owl:Class>
<owl:Classrdf:ID=“云母电容器”>
<rdfs:SubClassofrdf:resource=“#电容器”/>
</owl:Class>
<owl:Classrdf:ID=“服装”>
<rdfs:SubClassofrdf:resouree=“#电子商务”/>
</owl:Class>
<owl:DatatypePropertyulf:ID=“品牌”>
<rdfs:Domainrdf:resource=“#服装”/>
</owl:DatatypeProperty>
```

传统数据库是一组以数据定义语言来表达的语句集，该语句集能完整的描述数据库的结构，但不能为数据提供清晰的语义信息，数据库管理系统的检索和匹配效率不高[5]。本体知识库不仅存放数据的结构，同时还提供了清晰的语义信息和系统化知识，可以有效解决数据的结构异构和语义异构问题[6,7]。

本体构建之后，需要将其存储为数据库形式，本文成功地实现将构建的本体与 MS SQL Server 连接。本体知识库是将本体创建的 RDF、OWL 项目的本体概念内容存储到 DBMS 内的由 9 个字段构成的数据库表，部分内容如表 2 所示。

frame 字段中的数值代表本体框架中任一类、实例、槽、槽类型、槽值域、类或实例描述等要素的 ID，数值小于 10000 是系统保留数值，主要用于内部知识要素的迭代和转换。frame_type 字段代表 frame 中本体框架要素的类型， 数字 5、6、7 分别代表实例、类、槽。slot 字段代表本体槽或属性的 ID，数值从 2000 开始，基本为 4bits。short_value 字段代表本体框架中任一类、实例、槽、槽类型和槽值域等要素的描述或名称。表 2 中，“short_value = 标称电容量，frame_type = 7”表示“标称电容量”是“电容器”的一个槽，即属性。“short_value = 钽电解电容器，frame_type = 5”表示“钽电解电容器”是“电解电容器”的一个实例。表 2 中缺省的字段为 facet(值为 0)、is_template(值为 0)、value_index(值为 0)、long_value(值为 NULL)。

表 2　部分数据库表内容

frame	frame_type	slot	value_type	short_value
29938	7	2002	3	标称电容量
29944	5	2002	3	钽电解电容器
29940	6	2002	3	电解电容器
29936	6	2002	3	碳膜电阻器
29945	5	2002	3	压敏电阻器
29947	6	2002	3	水泥电阻器
29958	7	2002	3	元件尺寸
29956	7	2002	3	供货商
29960	7	2002	3	入货价格
29962	5	2002	3	低频电阻器

3.3 本体知识表示的推理[8,9]

本体中大部分知识不是显式说明的而是隐性表示的。考虑到本体过于庞大会导致知识提取效率不高，本体应用也不能把所有隐性的知识显式的表示出来。因此，需要对本体知识进行推理。知识推理的一个基本内容就是由给定的知识获得隐性的知识。推理有多方面的应用，对于知识库的建立者，推理的主要作用是检测冲突和优化表达。对于知识库的使用者，推理的作用主要在于获得知识库中的知识和运用知识库中的知识进行问题求解。

在本体语言蕴含的规则中，有这样一条规则“子类的实例也是父类的实例”。当知识库中出现“A 是 B 的子类”并且“存在一个 A 的实例 C”时，这条规则就被触发，执行的相应结果就是申明了“实例 C 也是 B 的实例”的事实，并添加到知识库中。本体推理也有一定的局限性，假设知识库中已经定义了“hasFather”和“hasGrandFather”两个属性关系，且有“D hasFather E”，“E hasFather F”，通过推理规则却无法得出“D hasGrandFather F”。因此，有必要对本体语言进行语义规则的扩充，达到对本体的

补充和完善，用来满足本体知识库不断发展更新的需求。对规则添加的研究可以结合描述逻辑和数据库触发器规则。

本文利用本体编辑工具调用推理机的 DIG 接口，对载入的电子商务领域本体知识库进行一致性检测(Consistency)、分类(Classify)、判定类型(Compute inferred types)三方面的测试。一致性检测确保本体不包含矛盾的实例；分类针对每个被命名的类，计算类与类之间的关系并产生类的层次结构；判定类型找出个体所属的特定的类，也即计算每个个体的直接类型。当用户提交概念查询时，查询被发送到推理引擎，推理引擎在知识库支持下，获得与此概念相关的概念树，同时推理出属于该概念的所有实例。将实例作为查询关键词发送至搜索程序，搜索程序检索出符合查询条件的链接，并将排序后的结果与推理结果组合成查询结果返回给用户。

4 基于本体的汉语自动分词系统

本体论作为元数据结构，提供了一个可控的语义化术语词库，其中每个术语都被清晰定义并且拥有可机器处理的语义。由此构建的本体知识库可以应用在汉语自动分词领域，利用词条间丰富的语义联系，弥补传统分词词条孤立的不足，能有效解决汉语自动分词的歧义切分问题[10]。基于电子商务领域的汉语自动分词系统的首要问题就是构筑实现分词用户的请求和分词系统间的语义化信息通信桥梁。在本体知识库的帮助下，本文构建的基于本体知识库的汉语自动分词系统(简称为 OntoSegs 系统)的基本框架如图 3 所示。

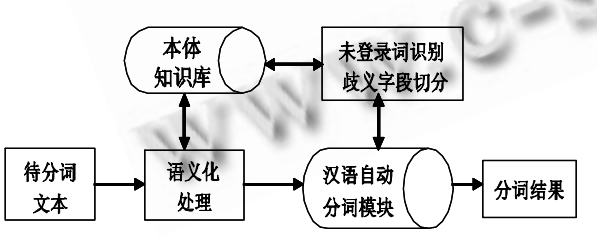

图 3 基于本体的汉语自动分词系统框架

该系统首先对输入的中文文本进行预处理，将其分割成一系列的汉语字串。系统接收到经过语义处理反馈出的词条后，将其输入汉语自动分词模块。在词语粗分阶段，得出 N 个概率最大的粗切分结果。然后，利用结构规则或统计的方法识别未登录词，并计算其概率，结合本体知识库及推理规则，动态考虑是否将其扩充为新词。利用本体概念规范、语义丰富的特点，对待分词进行词性标注排歧，优选出最大概率的文本切分标注结果。

本文的分词系统是应用于电子商务领域的，语料库中的词条按其实际所属行业存放在不同类别中。比如，“白炽灯”属于照明工业类，“晶体二极管”属于电子元器件类，“杀毒软件”属于安全、防护类等。分词系统首先对词条进行预处理，确认其所属行业类别。该系统基于隐马尔科夫模型进行词性标注并利用自动提取规则对分词结果进行优化。

汉语自动分词系统的首要功能是能够对汉语进行自动分词，衡量系统性能的主要指标是分词速度和分词的准确率。对本文构建的领域本体知识库中涵盖的 10 个栏目在 OntoSegs 分词系统上的分词速度和分词准确率进行统计，平均分词速度为 2801 个/秒，平均分词准确率为 97.94%，分词速度还有很大的提高空间，分词精度保持在较好的水平。

将 OntoSegs 分词系统与中科院的 ICTCLAS 分词系统在 3 个电子商务领域分词示例上进行分词速度、分词准确率的统计对比，结果如表 3 所示：

表 3 分词系统性能测试

示例	分词速度(个/秒)		分词准确率(%)	
	ICTCLAS	OntoSegs	ICTCLAS	OntoSegs
例 1	15000	5000	96.15	97.12
例 2	9969	2363	97.52	98.51
例 3	15750	2400	98.46	98.78

由表 3 可知，OntoSegs 分词系统的准确率高于 ICTCLAS 分词系统，体现了基于本体的汉语自动分词系统在处理本领域相关词汇时的优越性，一定程度上提高了分词正确率。在分词速度上，ICTCLAS 分词系统远远超越了 OntoSegs 分词系统，约是后者分词速度的 3—5 倍，OntoSegs 分词系统的分词速度还有待提高。

5 结束语

本文的主要贡献在于：①针对电子商务领域数据

的本体表达化，可以解决数据的结构异构和语义异构，方便数据间的传递和共享；②引入本体知识库构建汉语自动分词系统，一定程度上提高词语切分的正确率。

本系统还存在一些不足，例如基于电子商务领域的概念和术语包含的数据量非常庞大，各术语间的关系要完全的收集并加载到本体知识库中有一定难度；本体知识库中的术语提取过程效率不高，影响了分词速度等，这些将在今后研究工作中加以改进。

参考文献

1 Gruber TR. Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 1993,5(2): 199－220.

2 Uschold M, et al. The Enterprise Ontology. The Knowledge Engineering Review,1998.

3 Guarino N, Masolo C, Vetere G. OntoSeek: Content-based Access to the Web. IEEE Intelligent Systems, 1999,14(3):70－80.

4 刘文韬,陈智宏,许焱,李星毅.基于本体论的交通异构数据集成系统.计算机系统应用, 2010,19(3):7－11.

5 朱承,曹泽文,张维明.知识库系统建模框架的发展与现状.计算机工程, 2002,(8):3－5.

6 Tokosumi A, Matsumoto N, Murai H. Medical Ontologies as a Knowledge Repository.Complex Medical Engineering,2007. EEE/ICME International Conference on 23-27 May 2007.487－490.

7 Neri Mario Arrigoni, Colombetti Marco. Ontology-based Learning Objects Search and Courses Generation. Applied Artificial Intelligence, 2009, 23(3):233－260.

8 杨保明,刘晓东,姚兰,赵飞蓉. 基于本体论的农业知识的 OWL 描述.微电子学与计算机, 2007,(5):58－60,65.

9 文坤梅.基于本体知识库推理的语义搜索研究[博士学位论文].武汉:华中科技大学, 2007.

10 张春霞,郝天永.汉语自动分词的研究现状与困难.系统仿真学报, 2005,(1):138－144.