第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2007
·博士论文· 文章编号：1000—3428(2007)02—0004—02 文献标识码：A 中图分类号：TP311

本体集成的研究

王真星[1]，但唐仁[1]，叶长青[2]， 刘 岩[1]，吕 腾[3,4]，丁天怀[5]
（1. 清华大学深圳研究院，深圳 518057；2. 华东师范大学教信系，上海 200062；3. 新疆大学数学与系统科学学院，乌鲁木齐 830046；4. 炮兵学院二系，合肥 230031；5. 清华大学精仪系，北京 100080）

摘 要：本体技术在智能信息检索，电子商务，网上协作等信息技术领域的应用前景正越来越广。许多应用要求对相关信息源进行综合查询，因此必须解决这些异构数据源的语义互操作。由于同一领域的不同信息源的本体创建者采用的语义不同，将它们集成涉及许多复杂因素。本文通过对集成过程中需解决的问题进行分析，提出本体集成的框架，并给出实现方法。
关键词：本体；语义互操作；智能信息检索

Research on Ontology Integration Framework

WANG Zhenxing [1], DAN Tangren [1], YE Changeqing [2], LIU Yan [1], LV Teng [3.4], DING Tianhuai [5]
(1. Research Institute of Tsinghua University in Shenzhen, Shenzhen 518057; 2. Department of Education and Information, East China Normal University, Shanghai 200062; 3. College of Mathematics and System Science, Xinjiang University, Urumqi 830046; 4. Department 2, Artillery Academy, Hefei 230031; 5. Department of Precision Instruments, Tsinghua University, Beijing 100080)

【Abstract】 Ontology is widely used in intelligent information retrieve, e-business, and network collaboration. Meanwhile, many applications require the integrated query from related information resources, which arises the problem of semantic interoperation in heterogeneous information resources. There are many complex factors in integrating them, because different ontology creators adopt different semantics from his individual view. This paper proposes the framework of ontology integration though analyzing the problems in the process of ontology integration.
【Key words】 Ontology; Semantic interoperation; Intelligent information retrieve

随着 Internet/Intranet/Extranet 的迅速发展，网络的开放性、共享性和互联程度不断扩大，网上的信息激增。大量不同的信息源——数据库、知识库、文档集合共存于网上。许多应用需要对多个相关信息源进行联合查询，用于市场竞争分析、趋势预测和行为分析等，从单个信息源中查找数据已经远远不能满足新的需求。在信息源中查找有用数据涉及的一个重要方面便是必须知道信息源的语义。信息源的语义可以通过两种方式表达：隐含在应用程序逻辑中或用本体显式表示。过去应用程序大都采用前者处理数据查询，如果多个应用程序需要对同一信息源处理并且相互交换信息，则程序员必须在程序编制前对信息源的语义达成一致，这种方法只适合业务不发生变更的情况。现在，随着电子商务和计算机支持的协同工作环境的出现，业务需求经常发生变化，为此，不同信息源的语义开始采用本体表达。但是这些独立开发的信息源所对应的本体可能采用不同概念表达模型和本体语言，形式化程度各不相同，不同本体库之间领域知识可能重复或交叠，存在多种不一致。集成本体对网上智能信息检索、分布计算、WebService、Grid、协同工作等方面有较强的应用背景。

目前，本体集成的研究已经展开，文献[1]从本体语言的语法角度研究了本体集成,并从语义角度研究了本体集成中可能存在的语义失配，文献[2]从集成的体系结构角度研究了信息源集成的 3 种方式并分析了各种方式的优缺点，文献[3]介绍了关联本体库的概念并提出了本体代数。

本文通过对本体集成涉及的关键技术进行了分析，对全局本体库和关联本体库进行了描述，给出了相应的集成步骤。

1 本体的语义和语法

对某领域而言，可通过本体明确地表示应用程序所关心的语义。本体是对共享概念的形式化的明确说明。这里共享概念指多个主体所感兴趣的某特定领域的抽象，形式化是指可以被机器处理，明确的说明指概念、属性、关系、函数、约束、公理都有明确的定义。信息源的语义可以用适当的本体语言显式地表示出来。外界通过信息源本体的描述，可以对信息源有比较清晰的了解，从而可以对信息源进行有效地利用。本体的表示多种多样，分为不同层次和类别，将它们集成是非常困难的。

在讨论本体的集成时，先有必要研究本体的创建过程。要创建本体，首先要对现实世界进行分析，研究应用所感兴趣部分，建立相应的概念模型。其次，根据应用选择适当的本体表达方式。最后，选用适当的本体语言来实现。其过程如图 1 所示。

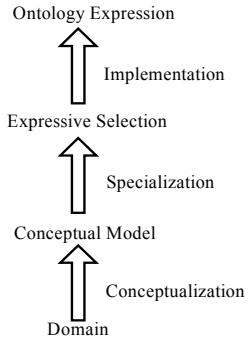

图 1 本体创建过程

2 本体集成核心问题

从上面的介绍可以发现，将不同的本体集成应分为 2 个

基金项目：国家自然科学基金资助项目(60563001)
作者简介：王真星(1970－)，男，博士后、高工，主研方向：数据库；但唐仁、叶常青，副教授；刘 岩，研究员；吕 腾，博士后；丁天怀，博导、教授
收稿日期：2006-03-26 **E-mail：**superwang2002@hotmail.com

层面：本体的概念层集成和本体的语言层集成。

2.1 本体的概念层集成

在两个或多个本体库描述的领域有交集的前提下，现要对这些本体库集成，则必须解决概念层描述的差异。不同本体库的概念层之间存在 2 种主要的差异：(1)建模时的差异，在进行领域概念化描述时，可能存在概念的范围、概念表示的粒度的差异，另外也可能采用不同的建模方法，如 ER 图和 UML 图。这种差异只能依靠领域专家以及相关知识的协助才能解决。(2)概念具体化表示的差异，可用不同的表达方式来表示同一概念，如用曲线和公式同样可以表示正弦函数，ISA 关系既可以用描述逻辑的公理表示，也可以用 OIL 表示。

2.2 本体语言层的集成

本体语言层存在的差异分为两种：语法差异和表达能力的差异。语法差异在不同的本体语言之间一定存在，如 KIF 和 OIL 之间就存在差异。而表达能力的差异和本体语言所拥有的基本原语相关，基本原语越丰富，表达就越强。但每个语言的基本原语不能太多，否则就非常庞大，解释器就很难实现。如当前的 HTML 就是基本原语太多，从而使浏览器非常复杂。较好的方式是定义一套精炼的核心原语，以后在外部扩充以满足新的需求，如 RDF/RDFS 就是这样。除了以上两种差异之外，还有词汇(Terminological)的差异，包括同义词 (Synonym) 、 同 形 异 义 词 (homonym) 、 Hyponyms 、 Hypernyms、编码的差异。表 1 给出了针对网上本体库集成过程中必须考虑的问题以及相应的对策。

表 1 本体集成过程中的各种冲突和解决方案

层次	不一致类型		存在问题	解决方案
概念层	建模	概念范围	两个类看似表达相同概念，但实际拥有不同实例	在全局本体库中创建对应的上层本体，然后以它为子树根结点生成两个兄弟结点，并建立与类的映射
		粒度	建模时采用不同的概念粒度	在全局本体库中创建对应的新概念名，并生成层次结构，然后与局部本体库中的概念名建立映射
	模型表示		同一概念采用不同模型类别表示	全部转化为采用统一格式表示
语言层	语法		不同的本体语言采用不同的语法	重写机制
	逻辑关系表达		不同的逻辑表达式表示相同的含义	提供两者之间的逻辑转换规则
	原语		同一个语言构造符字符串在不同语言中语义不同	在集成本体库中用不同的构造符字符串表示
	语言表达能力		一个语言可以表达另一个语言无法表达的内容	无
其它	同义词		不同词具有相同含义	在全局本体库中创建新的概念名，并将其和同义词建立映射
	同词异义		同一词在不同本体库具有不同含义	用不同的名字空间解决
	编码		具有数值类型的本体可用不同的编码表示	直接转换

3 本体集成模型

本体的集成涉及不同本体库中概念之间的语义关系，这是本体集成的基础。集成的本体库中包含原来各本体库概念之间的关系。原子关系是所有关系的基础，其语义也是推理的基础。参照 ONION 的研究，定义原子关系有 5 种：SubClassOf,PartOf, AttributeOf, InstanceOf, ValueOf。其中，SubClassOf、part of、AttributeOf 和 OO 设计中类似，InstanceOf 表示某对象是一概念的实例，ValueOf 指某一对象属性的具体值。概念间建立联系的本质实际是二者之间存在语义相关。

定义 1 语义相关 SemRel(C1,C2)=<Context,Mapping,(D1, D2),O>。

(1)Context 表示集成本体的上下文，也就是应用程序需求环境。

(2)Mapping 表示 C1,C2 之间的映射，此映射对应 Context。有多种 Mapping 存在，包括两个概念之间的一般/特殊关系、聚集关系、整体部分关系、函数依赖关系、部分映射关系等。

(3)(D1,D2)表示本体 C1,C2 各自所在的本体库。

(4)O 表示 Mapping 所在的本体库，全局本体库或关联本体库。

根据前面提到的本体集成的 3 种策略，可知公共本体和关联本体是两个不可缺的部分。假设存在本体库 O1,O2,…,On，它们均已转换为统一的概念层表示 O1',O2',…,On'，现有一应用 G 需要对不同本体库中同类概念集成，则这些本体库中涉及的与 G 相关的概念的交集 I 组成公共本体库。

定义 2 公共本体库。

公共本体库可用四元组描述：$CO=<C_a,C_b,C,K_{ab}>$。其中 C_a、C_b表示本体库A和B中的同类概念，C表示C_a、C_b所对应的新的概念名，K_{ab}表示C_a、C_b和C之间的联系。

当公共本体库用 DAML+OIL 表示时，可采用名字空间来区分不同本体库中的概念。名字空间的采用可以避免对原来的本体库进行修改，能十分方便地处理同义词、同形异义词；而 Rdf 的采用可以使 Web crawl 很方便地定位不同的本体库，增加了本体集成的灵活性。

定义 3 关联本体库。

关联本体库可用三元组描述：$AO=<C_a,C_b,R_{ab}>$。其中C_a、C_b表示本体库A和B中有关联的概念本体集合，R_{ab}表示依赖指派的集合。

关联本体库核心是本体的依赖，两个对象之间的关联有多种可能的形式，包括直接依赖、间接依赖、动态依赖、潜在依赖、传递依赖。直接依赖指在两个概念之间进行依赖关系的指派时，不需要任何前提条件。间接依赖指必须符合一定的前提条件依赖才能成立。间接依赖的前提条件可以是多个，当它们均满足时，两概念之间的间接依赖成立。令C_1和C_2表示两个要通过间接依赖的本体，f表示C_1和C_2之间的关系名，P表示前提条件谓词，则间接依赖地表示为$[[P(c_1)|\ \alpha]\ \zeta[P(c_2)|\ \alpha]]^* \rightarrow C_1\ fC_2$。

动态依赖指两个概念本体间可能由于满足不同的前提条件而有多种间接依赖的指派。当一个对象的成员和另一个对象的成员之间存在依赖，并且达到一定的依赖数目时，则对象之间将形成一种潜在的依赖。图 2 中的标记边以及两端的结点表示关联本体库。关联本体库最终是给推理模块使用的，推理模块根据关联本体库中的本体依赖以及前提条件，分析每一个实例数据，并最终为用户作出相关本体依赖的提示。

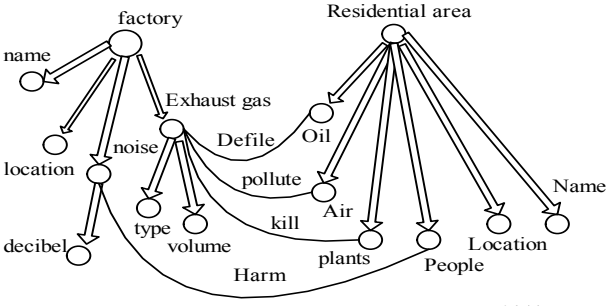

图 2 两个关联的本体库

(下转第 33 页)