文章编号:1001 - 9081(2005)02 - 0456 - 03

基于本体的信息集成技术研究

吴 昊,邢桂芬
(江苏大学 计算机科学与通信工程学院, 江苏 镇江 212013)
(littcricket@163.net)

摘 要:随着语义网技术的飞速发展,本体起了越来越重要的作用。在信息集成的过程中,本体作为一种工具解决了分布式异构信息源的语义异构问题,实现了信息源语义上的互操作。该文介绍了一种基于混合本体的信息集成方法,通过全局本体和局部本体之间的映射,向用户提供获取数据的统一接口,使用户获得语义上相关的数据。

关键词:本体;信息集成;本体集成;语义
中图分类号: TP391 **文献标识码**:A

Research on technology of information integration based on ontology

WU Hao, XING Gui-fen
(College of Computer Science & Communication, Jiangsu University, Zhenjiang Jiangsu 212013, China)

Abstract: With the rapid development of semantic Web, ontology played a prominent role on it. In the process of information integration, ontology solved the semantic heterogeneity problem of distributed, heterogeneous and autonomous data source. This paper introduces a ontology-based hybrid approach of inforamtion integration , which provides users the same interface for accessing to data corresponding to the semantic.

Key words: ontology; information integration; ontology integration; semantic

随着 Web 的迅猛发展,因特网上的资源越来越丰富,已经成为一个巨大的全球化信息仓库。Web 上的数据具有半结构化、异构性和分布性等特点。屏蔽这些特性,为用户提供统一的模式,是目前 Web 信息集成的关键问题。

信息系统的异构一般分为四种类型:结构异构、语法异构、系统异构和语义异构。结构上的异构是由于各个系统使用不同的数据模型,语法异构是由于不同的语言表示和数据表示,系统异构包括硬件和操作系统的不同,语义上的异构包括语义相等、语义相关和语义不相关等。前三种异构通过联邦数据库、虚拟数据库等方法得到了很好的解决。

造成语义异构的因素主要有:1)不同的信息源使用多种术语表示同一概念;2)同一术语在不同的信息源中表达不同的含义;3)各信息源中的概念之间存在着各种联系,但由于各信息源的分布自治性,这种隐含的联系不能体现出来(例如度量之间的不一致性)。

由于本体既准确地描述了概念含义又描述了概念之间的内在关联,能通过逻辑推理获取概念之间蕴涵的关系,具有很强的表达概念语义和获取知识的能力,因此用来解决语义异构的问题。

1 本体概述

本体(ontology)能够以一种显式、形式化的方式来表示语义,提高异构系统之间的互操作性,促进知识共享。在这个领域中,Neches 等人最早对知识本体作出定义:本体"给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延规则的定义"。但是在人工智能领域被普遍接受的却是 Studer 所作出的定义"知识本体是共享概念模型的明确形式化规范说明"。该定义包含 4 层含义:

1) 概念化:通过抽象出客观世界中一些现象的相关概念而得到的模型,其含义独立于具体的环境状态;

2) 明确:所使用的概念及使用这些概念的约束都有明确(显式)的定义;

3) 形式化:知识本体是计算机可读的;

4) 共享:知识本体中体现的是共同认可的知识,反映的是相关领域中公认的概念集,它所针对的是团体而不是个体。本体的目标是捕获相关的领域知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关系的明确定义。

为了便于讨论和研究知识本体系统的性质,Myo-Myo Naing 等人提出采用六元组方法来描述知识本体系统,采用六元组定义知识本体系统的方法如下:

本体系统包含了六个元素$\{C, A^C, R, A^R, H, X\}$。其中,C 代表概念集合;A^C 代表每一个概念的属性集合;R 代表关系集合;A^R 代表每一个关系的属性集合;H 代表概念层次体系;X 代表公理集合。

2 本体描述语言

由于机器并不能像人类一样理解蕴含在自然语言中的语义,计算机最终把所有的信息都当作 0、1 字符串进行处理。

收稿日期:2004 - 07 - 28;**修订日期**:2004 - 10 - 09 **基金项目**:江苏省信息化重点基金资助项目(1633000004)

作者简介:吴昊(1979 -),男,安徽桐城人,硕士研究生,主要研究方向:企业应用、数据库技术; 邢桂芬(1949 -),女,江苏盐城人,副教授,主要研究方向:数据库技术、人工智能.

而本体的目的是使信息成为机器可理解的,因此,在计算机领域讨论本体,首先就面临着本体究竟是如何描述的,也就是概念的形式化问题。对应的研究内容就是本体的描述语言。

自 20 世纪 90 年代以来,一些基于 AI 的本体实现语言陆续被提出,如 KIF、Ontolingua、CycL、Loom、OCML、FLogic。后来,随着 Web 的发展,又出现了一系列基于 Web 的本体语言,也叫做本体标记语言,如 SHOE、XOL、RDF、RDF - S、OIL、DAML、DAML + OIL、OWL。其中 OWL 已经成为 W3C 的推荐标准语言。

OWL 是在 DAML + OIL 的基础上发展起来的,作为 RDF(S)的扩展,目的是提供更多的原语以支持更加丰富的语义表达,并更好的支持推理。针对不同的需求,OWL 有三个子语言:OWL Lite、OWL DL 和 OWL Full。OWL Lite 用于提供给那些只需要一个分类层次和简单属性约束的用户。OWL DL 支持需要在推理系统上进行最大程度表达的用户。OWL Full 支持那些需要在没有计算保证的语法自由的 RDF 上进行最大程度表达的用户。

3　基于本体的信息集成的技术

最初,本体是作为"概念的详细说明"被引进的。因此,本体能够描述数据源的语义,使得内容清晰。在信息源的集成方面,它们被用来辨识和联系相关信息源的语义信息。

在基于本体的数据集成的方法中,本体被用作信息源语义的直接描述。一般情况下,存在三种方法:单本体方法、多本体方法和混合本体的方法来对数据源进行集成,如图 1 所示。

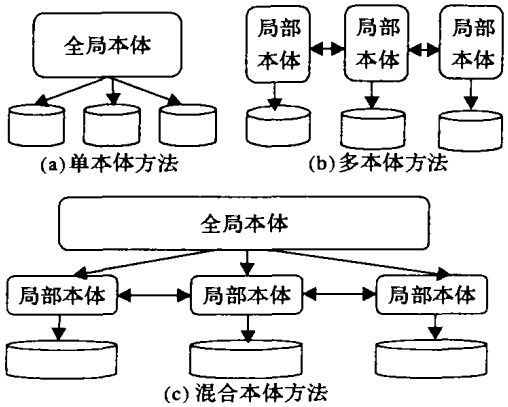

图 1　使用本体集成的三种方法

1)单本体方法

单个本体提供一个全局本体,给出共享词汇集对语义进行说明。所有的信息源都和这个全局本体相关。当所有的信息源在一个领域里被集成,提供了一个统一的视图时,单本体方法就解决了集成问题。但是如果一个信息源在领域内存在不同的视图时,也就是说提供了另一个级别的粒度,这时发现最小的本体承诺(ontology commitment)就变得很困难了。并且单本体易受信息源改变的影响,由于信息的改变,将导致全局本体的改变和对其他信息源之间映射的改变。多本体方法的产生克服了这个缺点。

2)多本体方法

在多本体方法中,每个信息源由它自己的本体来描述。多本体的优点是每个源本体不需要和其他的信息源相关而形成自己的本体,在信息源发生增改和删除时,本体的结构改动很小。缺点是由于不存在统一的全局本体,从而同其他信息源之间难以比较。

3)混合本体的方法

为了克服单本体和多本体的缺点,产生了一种混合本体方法。和多本体的方法相似,每个信息源由它自己的本体来描述语义。但是为了使每个源本体之间能够相互比较,在最上层建了一个共享的词汇集,共享的词汇集包含了领域内基本的术语。因为每一个源本体的术语是建立在原语的基础上,这样术语之间的比较就变得简单。混合方法的优点是很容易增加一个新的信息源,不需要对映射和共享的词汇集做过多的改动,并且能够支持本体进化。

下面举例说明基于混合本体的信息集成技术。假设有系统 A 和系统 B 如图 2 所示,其结构如下。

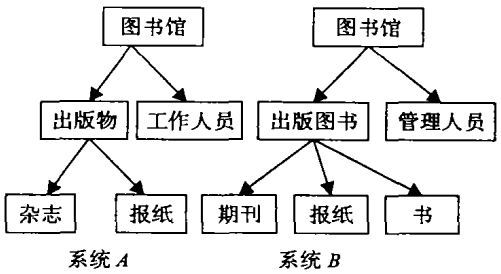

图 2　系统 A 和 B

在这里由于没有表示各个概念的个体,因此基于图 2 表示的系统 A 和 B 就可以看成本体 O_A 和本体 O_B。在集成的过程中包含几个主要步骤:分析数据源、查找原语、定义本体、本体的集成和映射。全面分析数据源,发现以前解释的语义问题,查找出数据源中重要的原语,然后用找到的原语来定义本体。例如系统 A:通过分析数据源,查找到的基本的概念有:图书馆、出版物、工作人员、杂志和报纸。

前面提到了用六元组来表示本体,由于系统 A 比较简单,只需要使用其中三个元素 C,A^C,H 来表示。其中系统 A 的三元组表示为:

C = {图书馆,出版物,工作人员,杂志,报纸};

A^C = {管理(工作人员,出版物),存放位置(出版物,图书馆),工作地点(工作人员,图书馆)};

H = {(杂志,出版物),(报纸,出版物),出版物 ……}

用 OWL 语言描述系统 A:

```
<owl: Class rdf: ID = "图书馆" >
<owl: Class rdf: ID = "出版物" >
<owl: Class rdf: ID = "工作人员" >
<owl: Class rdf: ID = "期刊" >
- <rdfs: subClassOf >
    <owl: Class rdf: about = "#出版物" / >
</rdfs: subClassOf >
</owl: Class >
<owl: Class rdf: ID = "报纸" >
  - <rdfs: subClassOf >
      <owl: Class rdf: about = "#出版物" / >
  </rdfs: subClassOf >
</owl: Class >
<owl: ObjectProperty rdf: ID = "管理" >
  <rdfs: domain rdf: resource = "#工作人员" / >
  <rdfs: range rdf: resource = "#出版物" / >
</owl: ObjectProperty >
```

```
<owl: ObjectProperty rdf: ID = "存放位置" >
    <rdfs: domain rdf: resource = "#出版物" / >
    <rdfs: range rdf: resource = "#图书馆" / >
</owl: ObjectProperty >
<owl: ObjectProperty rdf: ID = "工作地点" >
    <rdfs: domain rdf: resource = "#工作人员" / >
    <rdfs: range rdf: resource = "#图书馆" / >
</owl: ObjectProperty >
```

类似的可以得到系统 B 的本体描述。系统 A 和系统 B 存在以下问题:使用不同的元语表示同一概念的有:

管理人员 → 工作人员;

出版物 → 出版图书;

杂志 → 期刊;

以及 B 中包含了 A 中不存在的概念"图书"。

为了将各个本体之间联系起来,必须建立一个共享的词汇集即全局本体,然后在全局本体和局部本体之间建立映射。全局本体的主要目标就是提供了一个统一的视图,从而能够查询不同的局部本体表示的数据源。全局本体是通过合并局部本体即本体集成而得到。使用下面的方法从局部本体得到全局本体:

1) 类的集成:局部本体中多个概念上相等的类合并成一个全局本体中的类;

2) 属性的集成:一个类中多个概念相等的属性合并成一个属性;

3) 类之间关系的集成:从类 C_1 到类 C_2 中概念相等的关系被合并成一个关系;

4) 类和属性的拷贝:当在目标本体中不存在相同或相等的类和属性时,直接拷贝其类和属性。

最后是总结相关的类为一个通用的 Superclass。Superclass 通过查找已经存在的知识库和某种推理的方法获得。

为此得到基于图 3 表示的全局本体 O_G。

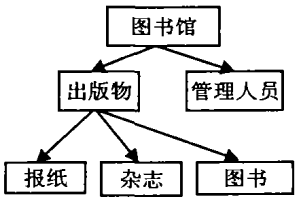

图 3 全局本体

其中全局本体中的共享词汇集为:图书馆,出版物,管理人员,图书,报纸,杂志。依照上面的原则,例如,全局本体 O_G 中的"出版物"是因为 O_A 中的"出版物"和 O_B 中的"出版图书"是相等的概念而被合并得到。全局本体 O_G 中的概念"图书"是由 O_B 中的"图书"直接拷贝得到的。其相应的三元组为:

C = {图书馆,出版物,管理人员,杂志,报纸,图书};

A^C = {管理(工作人员,出版物),存放位置(出版物,图书馆),工作地点(管理人员,图书馆)};

H = {(杂志,出版物),(报纸,出版物),(图书,出版物),出版物 ……}

除了全局本体,本体集成的过程也产生出另一个结果——映射表,它包含了全局本体和局部本体之间的映射信息。如果在全局本体中的 class, property,或者 class 之间的关系设为 M_G,这是从不同的局部本体 M_I 和 M_J 合并的结果,那么一个(M_G, M_I, M_J) 就产生了。如果全局本体中的 class 或者 property 是从局部本体中直接拷贝得到的,那么产生的映射是(M_G, M_J)。表 1 列出了 O_A 和 O_B 以及 O_G 全局本体之间部分的映射信息;

表 1 O_A 和 O_B 以及 O_G 全局本体间部分映射信息

O_G	O_A	O_B
图书馆	图书馆	图书馆
出版物	出版物	出版图书
管理人员	工作人员	管理人员
图书		书

在这里全局本体和局部本体之间只是一对一的简单映射关系,事实上,在真实的集成环境中,由于各方面的原因,存在各种不同的复杂映射关系,如一对多、多对多或者全局本体作为一个局部本体向其他层映射等。

全局本体提供给用户一个概念上的统一视图,用户通过简单的提交一个基于全局本体之上的 RDF 查询就能获取系统中所有相关数据源的数据,从而实现了概念上的互操作。

4 结语

由于语义网技术的不断发展,利用本体来解决异构信息集成中语义异构问题得到了越来越多的应用。使用本体进行数据集成存在很多优点:本体提供了丰富的、预定义的词汇作为数据库稳定概念的接口;而且是独立于数据库模式的,由本体表示的知识对于所有的相关数据源之间的转化都是可理解的;本体支持一致的管理和识别不一致性的数据。当然,在利用本体集成信息的过程中也遇到许多的问题像自动化、半自动化的获取本体,本体集成过程中的匹配问题。这是下一步研究工作的重点。

参考文献:

[1] KLEIN MCA. Interpreting XML Documents via an RDF Schema Ontology[A]. In Proceedings of the13th International Workshop on Database and ExpertSystems Applications (DEXA 2002)[C], 200.889 - 894.

[2] CALalì A, CALVANESE D, GIACOMO D, *et al.* Accessing data integration systems through conceptual schemas[A]. In Proc of the 20th Int Conf on Conceptual Modeling (ER 2001)[C], 2001.161 - 168.

[3] STUMME G, MAEDCHE A. Ontology Merging for Federated Ontologies on the Semantic Web[A]. In Proceedings of the International Workshop for Foundations of Models for Information Integration (FMII - 2001)[C], 2001.450 - 455.

[4] AMANN B, BEERI C, FUNDULAKI I, *et al.* Ontology-Based Integration of XML Web Resources[A]. In Proceedings of the 1st International Semantic Web Conference (ISWC 2002)[C], 2002.117 - 131.

[5] WACHE H, VOGELE T, VISSER U, *et al.* Ontology-Based Integration of Information - A Survey of Existing Approaches[A]. In Proc of IJCAI 2001, Workshop on Ontologies and Information Sharing[C], 2001.108 - 117, 2001.

[6] http://www.w3.org/2001/sw/[EB/OL], 2004.