《智能信息处理》课程考试

基于混合本体的异构数据集成方法研究

姚铭泽

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 20 日

基于混合本体的异构数据集成方法研究

姚铭泽

（大连海事大学信息科学技术学院 辽宁大连）

摘要: 本体以其强大的语义表达和概念推理能力，越来越多的用于解决数据集成中遇到的问题。分析局部本体和全局本体的缺点，提出一种基于混合本体的异构数据集成框架，集成查询模块，屏蔽数据源不同对于用户造成的不便，并介绍了其中的一些关键技术，并集成监控模块，生成本体的构建日志，便于系统的维护和升级。

关键字: 数据集成；混合本体；语义；映射

Heterogeneous databases integration approach based on mixed ontology

YAO MingZe

(School of Information Science and Engineering, Dalian Maritime University, Dalian China)

Abstract: With its powerful semantic expression and conceptual reasoning ability, ontology is more and more used to solve the problems encountered in data integration. This paper analyzes the shortcomings of local ontology and global ontology, puts forward a heterogeneous data integration framework based on hybrid ontology, integrates query module, shields the inconvenience caused by different data sources to users, introduces some key technologies, and integrates monitoring module to generate ontology construction log, which is convenient for system maintenance and upgrading.

.

Keyword: Data integration; Mixing body; Semantics; mapping

1 引言

随着信息技术在各个领域的广泛应用,信息量呈爆炸式增长,然而由于这些信息的存储环境、采集系统以及软硬件实施平台的差异,造成数据难以在各个平台间交流共享,给数据的有效利用造成很大的障碍,信息系统中因数据格式差异造成数据使用困难的问题称为异构问题,主要分为四个层次:系统、结构、语法和语义。对于系统和结构差异造成的数据异构问题,通常用XML作为一种公共语言构建标准化的数据模块进行解决;而对于语义的异构问题,基于本体的方法因为其独特的优越性越来越受到研究人员的重视,从搜集到的资料来看,现有的方法通常只是关注异构数据集成方法探讨,而对于数据查询系统的整合和对于本体构建过程的记录涉及较少,而查询系统直接与用户进行交互,而构建过程的记录对于本体构建的结果监控同样具有重要意义，本文基于混合本体的思想，提出了-种新的异构数据集成框架，并介绍了其中的关键技术。

本体原是哲学上的概念，原意是指世界各类具体事务具有的一般规定、一般本质、-般规律，是普遍存在于各种各类具体事务之中不可被感知但是可被人知道的相对抽象事物。在信息科学领域，本体被广泛接受的定义为“本体(Ontology)是共享概念模型的明确的形式化规范说明”，部分延续了哲学上的概念。其定义包含有四层含义，共享(Share)是指本体体现共同的知识基础，即是对于不同的参考者,其表述不会产生误解歧义;概念模型(Conceptualization)指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;明确(Explicit)指所使用的概念及使用这些概念的约束都有明确的定义；形式化(Formal)指本体的语言是计算机可读的。

由本体的定义也可以看出，本体用于数据集成系统的优势主要在于:从系统的开发者角度来看，首先，对于同一的描述对象，其本体的描述形式是唯一(这里的唯一指语义上的唯一); 其次，可将多种描述对象的概念，建立起一个"库”如果需要某一概念时，是从“库”调用出来，不需要重新定义，也就避免了因为系统开发时由于开发人员的知识背景造成语句运用的不一致;第三，本体的语言可以直接被计算机读取，避免了普通的描述语言转化为机器语言过程中出现错误。而从用户角度来看，本体作为一个用户与数据源间的中介，大量的异构的底层数据源对用户来说是透明的，即用户不需要了解数据源结构，仅需提交一个针对本体的查询，系统基于映射关系和语义定义，可以自动地将针对本体的查询转换为针对数据源的查询。这样，用户就可以仅仅提出需要什么数据，而不需要知道如何去发现数据，此外，通过本体的基础推理作用，在异构、分布环境下的数据集成中，可以提高数据的查全率和查准率。

在基于本体的数据集成方法中，主要存在三种结构，单本体结构，多本体结构，混合本体的结构。单本体结构是指:使用一个全局本体为语义规范提供一个共享词库，所有数据源关联到这个全局本体，显然存在大量数据源时，这种方法实现的难度较大;多本体方法是指:每一一个数据源由一个局部本体来描述，每一个局部本体单独开发，这种本体结构有利于数据系统的维护，但是这种方法需要一一个额外的架构来实现不同本体之间的数据交互，增加了处理的步骤和难度;混合本体整合这两种本体结构的优点，既有共享的词库，而各个数据源的语义信息由它自己的本体来描述，三种本体结构形式如图 1 所示。

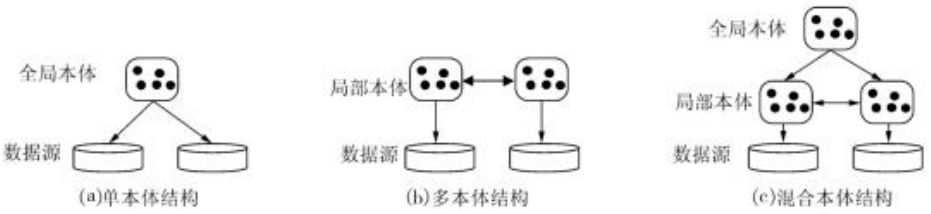

图 1 基于本体集成的三种结构

2 基于混合本体的异构数据集成框架

本文通过分析现有数据源存在的不足，综合异构数据查询问题，提出如图 2 的框架结构。

该系统分为三个模块:数据集成模块①,用户访问模块②和集成监控模块③。

(1) 数据集成模块是描述数据集成系统全过程，分为三个阶段。从下到上依此为数据源层，集成层，应用层。数据源层是由分布式异构数据源组成，数据源可以是关系数据库、Excel 表格，也可以是半结构化的 XML 文档。在这里设置包装器，可以将对于数据源的查询转化为对局部本体的查询，而开发者也可以对转化前后的数据源进行查询和数据抽取。以便于依据其差异对系统

进行维护。

集成层主要存在两个作用:对于数据用户，可以将不同地域不同形式的数据整合成同种数据，方便用户的查询使用;而从系统集成的角度来看，集成层可以整合局部本体存在的差异性，通过一系列的分解和综合，可以将数据源的信息整合为用户易于接受的数据。集成层是实现异构数据集成的核心，是实现系统用户与数据系统交互的根本。

应用层提供一个查询界面，该查询界面与用户访问模块互联，负责将用户的查询诉求输入本体模型，并通过一系列的转化，输出查询的结果。

(2) 用户访问模块与数据集成模块彼此独立，是实现系统与用户的交互作用的前提，作用包括查询类型识别、查询内容分解、查询指令的执行、查询结果的输出。可将用户的查询输入分解为计算机可以识别的形式化指令，并识别是普通用户对信息的查询，还是系统开发者对本体的查询。

(3) 集成监控模块用于生成数据集成过程中文档资料，用于记录集成的各个过程情况，作为系统集成的维护人员日后进行系统维护的依据，由于数据集成的质量好坏需要多次的验证，系统集成的日志在该系统框架下，局部本体用于描述对应数据源的语义信息，而全局本体是依据局部本体的比较，建立通用共享的语义库，在此基础上，用户只需对全局本体进行查询，即可获得数据源信息。而系统集成日志可以是集成系统开发人员日后进行维护的依据，如图 3-4 所示。

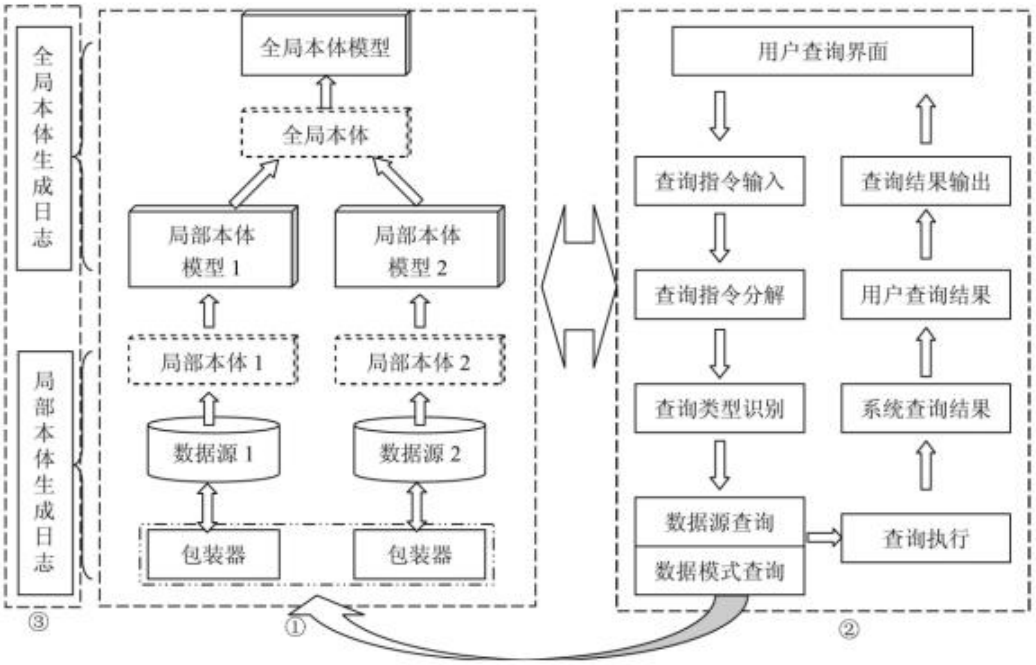

图 2 基于混合本体的异构数据集成框架

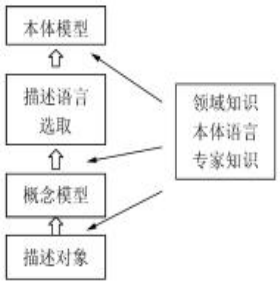

图 3 本体创建的过程

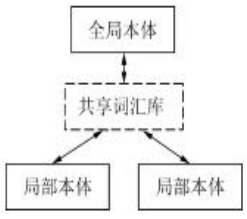

图 4 全局本体构建

3 关键技术

3.1 本体构建

本系统采用自底向上的本体集成方法，即首先提取底层各信息源的局部数据模式，其次在局部数据模式上抽取局部概念模式，最后在局部概念模式上构造全局概念模式。因此主要分为三个步骤:抽取数据库模式、构建局部本体和构建全局本体。

(1)抽取数据模式，将数据表的各个字段和属性都看成一个实体，并用相应的概念模型将其逻辑关系表达出来，以便于理解数据的格式和类别，为构建局部本体做准备。

(2)构建局部本体，在各个数据源的数据模式抽取成逻辑视图的基础上，对数据源进行全面的分析，提取感兴趣的内容，明确界定所要描述的对象，综合专家领域知识，建立数据源对应的局部本体，同时表示出从数据源到局部本体的映射，映射包括三个层次的对应:数据库与局部本体的对应，数据表和局部本体的类名的对应，表中字段和局部本体属性的对应。由于其概念模型是建立于共同的知识基础之上，因此，需要多领域专家和本体描述语言的协助完成，如图3所示。

(3)构建全 局本体，全局本体是由集成合并各个局部本体而来，它统一了各个局部本体中对同一领域内信息的描述，清理重复的数据记录，为用户提供了一个统一的视图，提供该领域的统一、一致的词汇语义，供用户查询不同的局部本体表示的数据源。全局本体作为局部本体的扩展，解决不同局部本体之间的语义异构性，同时满足不同局部本体之间的相互查询需求。

3.2 本体映射构建

全局本体和局部本体构建起来之后，还需要将各个数据源的局部本体和全局本体关联起来。具体的实施过程分为以下几个步骤:相似性提取，语义映射，映射执行。

(1)相似性提取衡量两个元素相似度可以参考多种方法，文献[6]提出一种编辑距离的方法，原理是采用一个字符串转换成另一个字符串所需要的插入、删除、替代等步骤，本文以两个数据元素x1, x2为例简要介绍给子编辑距离的相似度。

$$S(x1, x2)=\max\left(0, \frac{\min(|x1|, |x2|)-ed(x1-x2))}{\min(|x1|, |x2|)}\right)$$

其中，S 为两个元素的相似度，ed(x1, x2)为两个数据元素的编辑距离。其中编辑距离需要在具体的数据集成案例中，依据实际情况进行定义。由公式可知,其取值范围为[0,1],在此用户可以依据需要对其阀值进行定义,如将阀值设置为0. 7,则当s的值大于0. 7时，将两者看成相似的数据记录,进行手工或计算机合并。在这里相似度只是一种简单模型，实际的相似度提取包括多种方法，常见的如基于欧式距离、马氏距离、切比雪夫、三角余弦等。

(2) 语义映射构建本步骤是为了生成映射规则，定义源本体转换为目标本体的条件和转换函数，对应方式有一对一，一对多和多对一。语义映射是本体构建的一个关键过程，很大程度上决定着本体质量的好坏。

(3) 映射构建执行该步骤是本体映射的具体实施步骤，将局部本体相关的实体依据定义出的语义映射规则转换为目标本体，并存储转换结果。

映射的作用不仅是将所反映的实体以共享的概念表达出来，而且为数据源提供了一个公用的中介，便于对于用户和数据源之间的数据交换,因此,在本体的构建过程中这是-一个关键的步骤。

3.3 用户查询系统

查询处理是异构数据集成另一个重要的研究内 容 因为只有查询系统是为用户服务的 其质量的好坏也关系着整个系统的运行效果。集成系统的设计目的就在于屏蔽数据源的异域和异构特征对于用户的信息查询带来的影响，因此查询系统的设计也应该能够依据不同的查询类别，输出用户所需要的结果。查询的处理过程如图2的用户访问模块。

查询系统的构建的几个技术环节包括查询类型的识别和查询内容的分解，将用户输入转换为本体可以识别的计算机语句。首

先是将输入语句分解，识别其中的关键词，并引导找出对应数据源的位置，将数据提取出来。具体步骤为:首先根据用户查询指令建立查询语句,然后经过本体的转化作用将查询语句映射为数据源模式,将数据源模式搜集数据源的数据并输出本体模式，将本体模式转换为可被用户接受的输出模式,将输出结果显示给用户。

4 结语

随着语义语法处理技术的进步，基于本体的方法也越来越受到研究人员的重视，本文通过对全局本体、局部本体的优缺点分析,提出一种基于混合本体的数据集成方法，并介绍其中的关键技术。但在本文中只是提出一个理论性的框架,利用本体的方法用于数据集成的实际，还有待于进一步的研究。

参考文献

[1] Natalya F Noy Deborah L d Guinness Ontology Development101: A Guide to C eating Y our First Ontology[R]. S tan ford Knowledge Systems Laboratory Technical Report K s1-01 -05 and Stanford Medical Informatics Technical Report S12001-0880, 2001

[2][EB/oL.]http//baikebaiducom/view/2998

[3] Stefen Decker Michael Eudlmann D iete Fenseletal OntonokeyOntology based access to d istributed and Sen irstuc tured infoma-tion[M]. MeersnanR, etal Sanantic Isues in Mu ltined ia Sys-tens Pncedings of DS8 khuwer A adenic Publisher Boston1999; 351 - -369.

[4]邓志鸿,唐世渭，杨东青.面向语义的集成一本体在 W eb 信息集成中的研究进展[J].计算机应用, 2002 22 15-17.

[5]沈学利,周纪超.基于本体语义映射的数据集成机制研究[J].计算机工程与科学，2010, 32

[6] G muberT R. Toward P rine iples for the Des ign ofOntologies U sed forK ow ledge Sharing[J]. in temational Joumal of hum an can pu terStudies 1995, 43(5-6): 907 -928.

. [7]孟坚. 基于规则的交互式数据清洗技术[D].东南大学硕士学位论文，2005.

[8] Guber CTR. A translation Apoach to Portable on tologies [J].