基于本体论的交通异构数据集成系统①

刘文韬[1] 陈智宏[1] 许 焱[1] 李星毅[2]

(1.北京市交通信息中心 北京 100053;2.北京交通大学 先进控制系统研究所 北京 100044)

摘 要: 介绍了一种基于本体论的交通异构数据集成系统和技术核心——本体的构建方法。该系统通过全局本体和局部本体的映射实现系统数据的语义解释，使得异构数据库对于应用层而言是透明的。该方法降低了开发难度，提高了开发效率，增强了系统的可维护性和可扩展性。

关键词: 本体论；异构数据；交通信息系统；数据集成

Traffic Heterogeneous Data Integrating System Based on Ontology

LIU Wen-Tao[1], CHEN Zhi-Hong[1], XU Yan[1], LI Xing-Yi[2]

(1. Beijing Transportation Information Center Advanced Control Systems Lab, Beijing 100053, China;

2. School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: This paper develops a traffic heterogeneous data integration system based on ontology. Its key technology is ontology construction. This system is able to provide the semantic explanation of systematic data by means of mapping from global ontology to partial ontology, so that heterogeneous database is transparent for the application layer. The proposed method reduces developing difficulty, improves developing efficiency, and enhances the maintainability and expandability of the system.

Keywords: ontology; heterogeneous data; traffic information system; data integration

随着交通信息化建设开展，缺少统筹规划的数据中心所带来的问题日渐凸显。各单位建设的业务系统都是为了满足自身业务管理需要而建立的，业务应用和数据关联上缺少总体规划和设计协调，系统之间数据交换存在困难，数据不能及时提供给其他部门和单位使用。同时，由于各业务系统采用的开发形式、整体构架和数据库不同，数据的种类、类型和存储方式、数据编码、标准和接口也不统一，无法保证数据的一致性和准确性，更无法进行综合、全面、深入的数据应用，不能满足综合业务管理、公众信息服务和政府决策数据支持的需要。为此，交通信息整合成为近年来交通主管部门普遍关心的问题。交通异构数据集成是信息整合的一个技术要点。本文讨论以本体论为基础的交通异构数据集成系统，以实现交通信息资源的整合。

1 基于本体的数据集成方法

本体(ontology)是一个哲学概念[1]，是对客观存在的系统解释或说明，通过本体化约定方式近似地描述了关于现实世界的概念。一个本体就是一套关于某一领域的规范描述，它包含概念、属性和限制条件。一个完整的本体还要包含一系列与某个类相关的实例，这些实例组成了一个知识库。

基于本体的数据集成研究十分活跃，被广泛应用于知识管理、信息检索和教育等领域。SKC[2]是Standford 大学开展的项目，目标是解决信息系统中语义异构问题，实现异构自治系统间的互操作；Noy[3]等人描述了本体映射工具；Yang[4]等人讨论了基于本

① 基金项目：国家“十一五”科技支撑计划 (2006BAG01A01)
收稿时间:2009-07-07

体的综合语义的映射关系；Miao[5]等人基于本体的领域知识的语义映射；吴玲丽等人[6]讨论了语义集成方法，Malucelli 等人[7]讨论了电子商务中数据异构问题。

数据源语义异构性问题在数据集成中变得越来越突出。由于本体是对事物基本属性的概念化描述，可以使用本体通过计算机可理解的方式来描述数据源信息和全局数据模式，利用全局本体建立共享词汇库，以及待集成领域的领域知识，所有的分布数据源都利用全局本体的共享词汇和共享知识，从而最大程度的减少各个数据源数据的语义异构问题。

在基于本体的数据集成方法中，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：单一本体方法，多本体方法和混合本体方法[8]，其结构如图 1 所示。

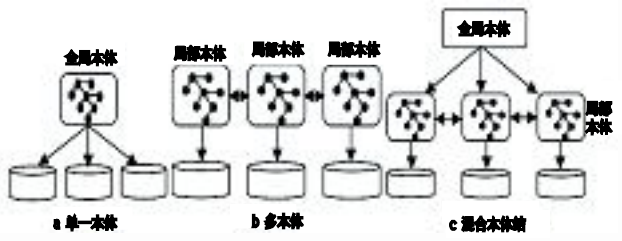

图 1　基于本体集成的三种方法

①单一本体方法：由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。当所有信息源在领域内集成，提供统一视图时，单本体方法就解决了集成问题。②多本体方法：多本体方法可以支持领域视角不同的信息源集成，不需要构建全局本体，可以支持动态性较强的信息源集成。③混合本体的方法：每个信息源由它自己的本体来描述语义。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。混合方法的优点是能够支持本体的获取和进化，便于扩展。针对交通信息系统的集成，使用混合本体的方法较为适宜。

2　基于本体的应用系统架构

各级交通部门已建了大量的业务信息系统，这涉及交通的各种业务领域。以路政为例，包含了公路建设、管理、养护、征稽和道路运输管理等业务系统。这些系统体系结构不同，开发商和开发时期不一致，系统的编码和数据字典存在较大的差异性，数据因其宿主系统的数据字典不一致造成严重的数据间异构异义现象。因此,整合各分布的、异构的数据库资源是实现交通信息集成核心问题。

集成系统需要联合访问多个分布异构数据库系统，通过利用集成层的混合本体映射规则，实现应用层的数据请求到异构数据的映射，实现异构数据的管理和访问。全局本体实现应用层的统一视图。系统架构由数据层、集成层、应用层、客户层等 4 层构成，如图 2 所示。

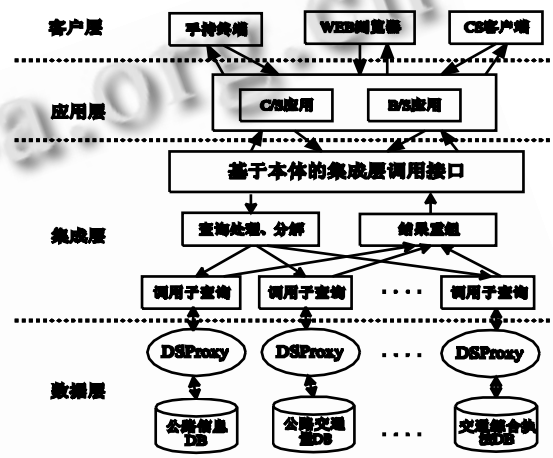

图 2　应用系统的架构图

各层主要功能：①客户层是整个应用系统对外的接口。可以通过浏览器、CS 客户端、手持设备访问系统。②应用层是系统功能实现层，不需考虑底层数据库操作，而是通过全局本体术语描述所需数据，由集成层负责解析执行。③集成层是解决异构数据库集成的核心层。提供了 Web 服务查询接口给应用层访问，并且以松耦合的方式集成各成员数据库，保持它们自身的自治性。④数据层是原始数据的存储层。它包括多个部门现有的数据库，以及在它们之上建立的数据源代理器 DSProxy。所有这些成员库都以松散耦合的方式由集成层集成，以全局统一虚拟视图为应用层提供服务。

3　集成系统的本体构建

异构数据集成系统通过建立数据源到局部本体的映射关系和局部本体到全局本体的映射关系，建立数据源间的统一语义，完成异构数据源的逻辑集中。本体构建是系统核心。

3.1 局部本体构建

局部本体对应成员数据库，从成员数据库中抽取

数据字典并构建相应的局部本体。构建流程如下：

S1：部署混合本体映射规则，实现应用层的数据请求到异构数据的映射；

S2：配置代理数据库连接信息；

S3：连接到数据源，抽取关系表的模式信息；

S4：对关系表的模式进行全局语义释义；

S5：生成局部本体和映射信息；

S6： 注册数据源的 Web Service 服务，并形成局部本体映射关系表。

以某市车管所的数据库 S1 为例来说明，该数据库主要车辆和驾驶员等信息。数据库中的部分表模式如下：

Driver (driverID , name , sex , birthday)

Drive (dirverID , carNO)

Car (carNO ,carType , buyDate, colour)

得到图 3 所示本体描述 O1，生成映射信息如表 1–表 3。

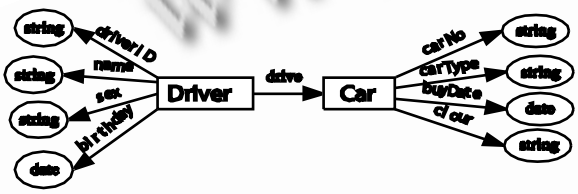

图 3　S1 对应的局部本体 O1

同样，对存放于另一个城市的交通执法数据库 S2，该库存储有司机、车辆违章与处罚等信息，部分表模式如下：

Driver (ID , name , sex)

Peccancy (peID , ID , site , cause , date)

Punish (puID , peID , money , isDeal)

本体描述 O2 如图 4,关系映射表建立方法同 S1。

通过上述步骤可以建立成员数据库的局部本体结构和映射信息供集成系统使用。

表 1　O1 中的类到 S1 中关系表的映射

ID	class	table	ID	class	table
1	Driver	Driver	2	Car	Car

表 2　O1 中的数据类型属性到 S1 中字段的映射

ID	property	column	oid	ID	property	column	oid
1	driverID	driverID	1	5	carNO	carNO	2
2	name	name	1	6	carType	carType	2
3	sex	sex	1	7	buyDate	buyDate	2
4	birthday	birthday	1	8	colour	colour	2

表 3　O1 角色到 S1 中主外键列的映射

ID	role	pfkey
1	drive	(driverID,Drive,carNO)

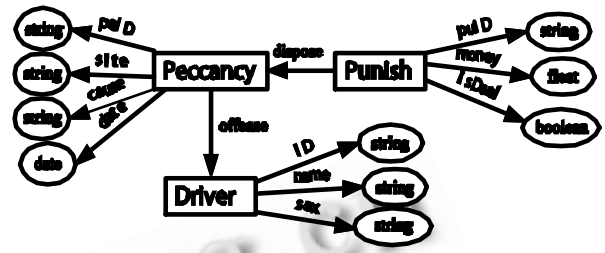

图 4　S2 对应的局部本体 O2

3.2 全局本体构建

全局本体对应系统数据库的逻辑结构，从局部本体集成为全局本体，并形成相关的映射信息。图 5 给出了由局部本体 O1 和 O2 构建的全局本体。可以使用可视化本体编辑工具(如：Protégé)对生成的全局本体进行调整和修改。同时全局本体到局部本体的映射信息，其中类如表 4 所示(其他映射表，如属性映射表、角色映射表等略去)：

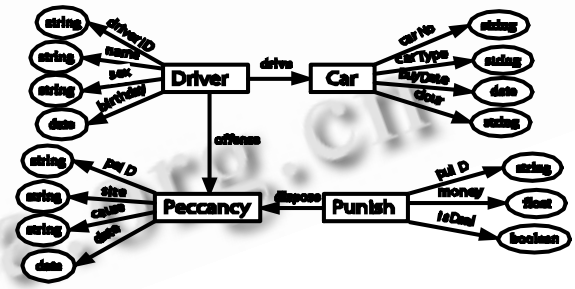

图 5　集成数据源 S1、S2 后的全局本体

表 4　全局本体中的类到各局部本体类的映射

ID	gclass	lclass	source	ID	gclass	lclass	source
1	Driver	Driver	S1	4	Car	Car	S1
2	Peccancy	Peccancy	S2	5	Driver	Driver	S2
3	Punish	Punish	S2	4	Car	Car	S1

4　本体集成映射

本体集成映射是集成系统本体构建的关键，涉及到概念归并和本体映射关系表建立两级技术，前者通过对概念及其属性的相似性计算完成，即语义释义；

后者通过合并映射计算得到。本节介绍集成映射方法。

4.1 概念和属性的义原相似度计算

概念和属性可以由多个义原表示，如图 6 表示了一个义原树片段，相似度可以通过义原之间的相似度获得。义原的相似度计算可以通过公式(1)计算。

$$Sim(p_1,p_2)=\frac{2\times Spd(p_1,p_2)}{2\times Spd(p_1,p_2)+Len(p_1,p_2)+2\times Bal(p_1,p_2)}=\frac{2\times Spd(p_1,p_2)}{Depth(p_1)+Depth(p_1)+2\times Bal(p_1,p_2)} \quad (1)$$

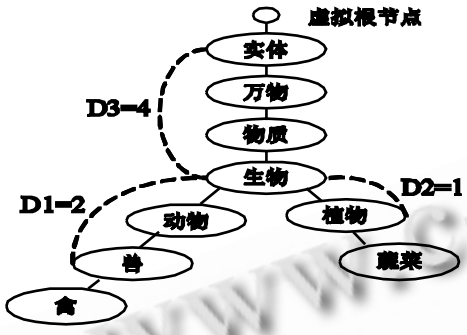

图 6　义原树的片段

其中 Depth(p)为义原的深度，指义原 p 在整体义原层次体系中所处的层数位置，并规定根节点的义原深度为 0，它的子节点深度为 1，其它依次类推；Spd(p1,p2) 为义原的重合度，指两个义原 p1 和 p2 在义原层次体系中所拥有的相同父节点的路径长度；Len(p1,p2) (Length)为义原的距离，指 2 个义原 p1 和 p2 在义原层次体系中的距离；Bal(p1,p2) 为义原的平衡度，指 2 个义原 p1 和 p2 在义原层次体系中高度的差值。例如：Sim(“兽”,”植物”)=(2×4)/(6+5+2)=8/13。使用公式 1 可以计算概念和属性的相似。

4.2 集成映射算法

全局本体映射需要对同等概念进行归并，归并过程是对语义相同和相近的概念进行映射。主要包括以下几个过程。

① 提取出全局本体和局部本体中的概念；

② 使用公式 1 进行相似概念匹配；

③ 在全局概念表中补充局部存在全局不存在的概念；

④ 根据概念匹配映射，使用公式 1 进行属性相似匹配和全局本体概念表的属性修正；

⑤ 交互确认；

⑥ 建立局部本体到全局本体的映射集。

概念匹配采用义原相似度匹配的方法获得，即相似度大于一个给定的阈值，则认为两个概念是匹配的。对于匹配概念计算属性相似，当这两个概念的相似属数大于给定阈值时，认为两概念表达同一概念，则在同一概念间建立概念映射关系和属性映射关系。本体集成映射算法如下。

算法 1. 基于语义的集成映射算法 Onto-IntMapping：

输入：ontology O_X，O_G

输出：M_C,M_A,O_G //分别表示类/属性集和本体

$M_C=\emptyset$，$M_A=\emptyset$ // 初始化 M_C，M_A

For i=1 to n　//n 为 O_X 中概念的个数

{$P_a=\emptyset$　// 初始化属性对集合

求出 O_G 中的 C_g，使 Sim (C_i,C_g) > Sim (C_i, C_j)

// $C_j\in O_G$，$C_i\in O_X$,j! = k

If　Sim (C_i,C_g) >α　//α 为概念名称相似性阈限值

{P_a=Attr-Mapping(Ci,Cg)//计算概念属性映射集

If　P_a ! = Ø

{M_C= M_C ∪ (C_g->C_i)　//添加概念映射

M_A= M_A ∪ P_a　//添加属性映射

If　|P_a| < C_i 中的属性个数

{将 C_i 中没有在 P_a 中出现的属性添加到 C_g，并添加它们之间映射到 M_A}}}

If　Sim(C_i,C_k)<=α !! P_a= Ø //全局本体没有同等概念

{将 C_i 及其属性添加到 O_G 中，设命名为 C_n

M_C= M_C ∪ (C_n ->C_i)

将 C_n、C_i 中每对同名属性加入到 M_A}}

算法 2. 属性映射算法 Attr-Mapping：

输入：两个本体概念 C_1 ,C_2

输出：C_1 ,C_2 的属性匹配对集合 P_a

P_a= Ø　//初始化属性对集合

Set1=C_1 的属性集合，Set2=C_2 的属性集合

minAttSum= min{|Set1|,|Set2|}　// |Seti|表属性个数

while(|Set1|>0 and |Set2|>0)

{求出两个集合中相似度最大的一组属性 $A_i\in$ Set1 和 $A_j\in$ Set2；

If　Sim(A_i,A_j)> α　// α 为属性相似性阈限值

{P_a=P_a∪ (A_I,A_J)}

Else { Break }

Set1=Set1 - {A_i}; Set2=Set2 - {A_j} }

If |P_a|<β *minAttSum //β为相似性属性匹配对系数

{ P_a = Ø }

Return P_a

至此，可以构建局部和全局本体的本体映射关系，实现异构系统的集成。本方法在北京奥运智能交通管理与服务综合系统的建设中得到了应用。

5 结论

随着交通信息化建设的深入，交通部门建立了大量的信息系统，这些系统在不同业务领域起到了重要作用。由于交通信息化建设周期长、建设单位多、业务条块分割，很多信息系统是“信息孤岛”，不利于信息共享和对交通综合管理。

本文讨论了一种以本体论为基础的交通异构数据集成系统，并讨论了本体的构建方法。通过本体建立异构数据的语义映射，实现数据交换和共享。使用本体的映射关系可以在应用层和集成层建立调用耦合，当底层的数据库发生变化时，只需通过修改局部本体、全局本体、映射信息和 UDDI 注册信息即可适应变化需求。降低了应用层与底层数据库间的耦合度，使应用层专注于业务，无须考虑底层数据库间的操作、转化，增强了系统的可维护性和可扩展性。本文讨论方法在北京奥运智能交通管理与服务综合系统的建设中得到了应用，该方法可以广泛用于异构数据的集中项目。

参考文献

1 邓志鸿,唐世渭,张铭,等.Ontology研究综述.北京大学学报(自然科学版), 2002,38(5):730－738.

2 Wiederhole G, Jannink J. Composing Diverse Ontologies. Standford University, Scalable Knowledge Composition (SKC) Porject, Technical Report, 1990.

3 Noy NF, Musen MA. The PROMPT suite: interactive tools for ontology merging&mapping. International Journal of Human-Computer Studies, 2003,59(6): 983－1024.

4 Yang XD, He N, Wu LP, Liu JQ. Ontology based approach of semantic information integration. Journal of Southeast University, 2007,23(3):338－342.

5 Miao GX, Chen XY. Domain semantic mapping of database metasearch engine. Journal of Southeast University, 2007,23(3):357－360.

6 吴玲丽,余建桥,孙荣荣.一种基于 Ontology 的异构数据库语义集成方法.计算机系统应用, 2008,17(3):31－33.

7 Malucelli A, Oliveira DP. Ontology-Based Services to help solving the heterogeneity problem in ecommerce negotiations. Electronic Commerce Research and Applications, 2006,(2):29－43.

8 Wxche H, Vogele T, Visser U. Ontology-Based integration of information: a survey of existing approaches. Proc. of IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, 2001:108－117.