

信息集成中冲突的语义解决方案

屈振新^{1,2} 唐胜群¹

(武汉大学软件工程国家重点实验室 武汉 430072)¹ (中南财经政法大学信息学院 武汉 430073)²

摘 要 在集成异质异构的信息系统时,系统间会出现大量的冲突。对可能的冲突进行了整理,列举出 5 大类冲突,并进行了深入分析。在此基础之上,提出了采用语义技术的完整解决方案,以消解各个层次的冲突。

关键词 信息集成,冲突,语义

中图法分类号 TP311 **文献标识码** A

Semantic Solution to Conflict of Information Integration

QU Zhen-xin^{1,2} TANG Sheng-qun²

(State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China)¹

(Information School, Zhongnan University of Economics and Law, Wuhan 430073, China)²

Abstract When heterogeneous information systems were integrated, a great amount of conflictions appeared. After tidying up all kinds of cases, five kinds of conflicts were found out, which were analyzed in detail. Based on the work, a full solution applying semantic technology was proposed, which can eliminate conflicts at all levels.

Keywords Information integration, Conflict, Semantic

1 概述

实际应用中存在大量的自治信息系统,它们彼此独立、分布分散,所使用的数据管理技术也存在差别、异质异构。当需要把这些信息系统进行集成,实现信息共享时,就会发现系统间存在大量的不一致,即冲突。产生冲突的原因有很多,其中很重要的原因之一就是不同的系统对相同事物理解的角度不一样,表达的方式、侧重有差别,导致彼此独立运行都很正常,一旦集成到一起,矛盾就暴露出来。

为了消解冲突,建立系统间一致的信息理解是非常重要的和必要的。源于哲学范畴的“本体”是共享概念模型明确的形式化规范说明,可以清晰地表达概念间的关系,已经成为语义技术的主流。

虽然利用语义技术进行信息系统集成并不是新话题,但是已有方法并没有对可能遇到的冲突进行系统的分析,往往解决的只是局部问题。本文提出的以语义为基础的解决方案对集成中可能遇到的冲突进行了全面分析,提出了完整的解决方案。

2 相关工作

面对异质异构的信息系统,如何集成,已经有很多方法。分布式数据库技术就是方法之一。但在实际应用中,多是同种数据库的分布式处理,其在处理异质异构数据源时仍然有局限。多数据库系统^[1]能够较好地处理异质异构的数据源,

但是与分布式数据库一样,能够消解的冲突有限。进行语义信息集成已有很多好的方法。有的将数据库 schema 映射成本体^[2,3],再将语义查询重写为 SQL 语句;有的提出语义集成时可以采用分层本体、词汇表等方法。但是这些方法都是从某个局部来看问题,解决了局部语义冲突,实际应用中遇到的冲突远不止这些。

3 信息集成面临的冲突

面对异质异构的信息系统,它们采用的数据库技术各不相同,甚至有的是文件系统,数据的语义蕴涵在程序中,语义差别可能很大,对相同事物描述的方式、结构和内容都不相同。要进行正确的集成,就必须消除系统间的冲突。要消除冲突,首先就需要对冲突有全面、正确的认识。RAM^[4]将冲突分为结构冲突和语义冲突;有的^[5]将冲突分为语法、模式和语义异构;Batini 将冲突分为命名冲突和结构冲突;Spaccapetra 将冲突分为语义冲突、描述冲突、数据模型冲突和结构冲突。他们还提出了各自的解决办法。这些分类和解决方法都是有所偏重的,没有全面分析、解决问题,没有从语义技术的角度看问题,消解的冲突有限。

从采用语义技术解决问题的角度看,按照不同的层次,本文将信息集成中可能遇到的冲突进行了归纳,如表 1 所列。

表 1 冲突分类

类型	举例
物理冲突	Oracle, FoxPro, XML, ……

到稿日期:2009-03-02 返修日期:2009-04-30 本文受国家重点基础研究发展计划 973 计划(2007CB310801),武汉市科技局科技攻关项目(200810321167)资助。

屈振新(1972—),男,博士研究生,副教授,主要研究方向为语义网、软件工程等,E-mail:quzx72@gmail.com;唐胜群(1946—),男,教授,主要研究方向为语义网、软件工程、电子政务等。

物理冲突是指不同的信息系统所使用的数据处理技术不一致,有的系统可能采用关系型数据库,有的可能采用面向对象数据库,有的可能是 xml 文件等。即便采用相同类型的数据库,如关系型数据库,不同厂商的产品也有差别。

字段冲突指在描述同一特征时,字段间出现的不一致,包含结构冲突、命名冲突、类型冲突、长度冲突、精度冲突、计量单位冲突、表达方式冲突。结构冲突分为两种:一是用不同数量的字段表示相同的特征;二是同一特征在有些系统中用字段来表达,有些系统中用表及表间外键关系来表达。命名冲突指字段同名异义、异名同义。类型冲突指表达相同的特征在不同表中采用不同的数据类型。长度冲突指表达相同特征的不同字段的数据长度不一致。精度冲突指表达相同特征在不同表中采用不同数据精度的字段。计量单位冲突指表达相同特征的数据在不同表中具有不同的计量单位。表达方式冲突指数据表示不一致,如数据格式差异、缩写差异等。

记录冲突指描述同一个对象的不同记录间出现的不一致,包括标识冲突、内容冲突等。标识冲突指描述同一对象的记录具有不同的关键字。内容冲突指描述同一对象的不同记录数据不一致,甚至相反,比如“张三”出现了两个年龄。

表冲突指表之间的不一致,包括命名冲突、结构冲突、关系冲突、概念冲突。命名冲突包括表名重复等。结构冲突指表达相同概念的不同表具有不同的字段集数据的含义隐藏在程序中,进行信息集成时,隐藏的含义也被挖掘出来,其中就有表之间的关系。关系冲突指各系统内部的表间关系集成到一起时所呈现出的不一致,比如在 A 系统中,两张表之间表达的是父子关系,而在 B 系统中,类似的两张表却表达的是等价关系。

共享语义冲突。采用语义技术进行信息集成,一般都是将各数据的 schema 映射到本体,一旦这些本体汇集到一起,就可能出现概念定义的不一致及概念间关系的不一致。

4 语义解决框架

由于信息系统中存储的数据量巨大,若先将其静态转换为本体实例,再由推理机来处理,则现在几乎没有 DL 推理机能在海量的实例上进行推理,即使推理机的性能没问题,静态转换也会带来数据的重复存储及衍生问题,因此本方案的基本思路是将信息系统中共享数据的 schema 映射到本体,用户在本体上进行语义查询。查询经过推理后,再转换成 SQL 语句等常规方式去执行。在这一过程中,同时消解遇到的各种冲突。

考虑到信息集成时各自系统并非将自己的所有数据拿出来共享,同时出于安全性的考虑,在正常的业务系统和信息共享之间应该有一定隔离,所以在各数据源上应为共享数据建立视图,再将视图的 schema 映射到本体。在建立视图的过程中,也可以消解部分冲突。

方案中涉及的本体有 3 类:一是局部本体,根据每个信息系统对外发布的共享视图建立相应的本体。二是全局本体,包含各局部本体共享的公共词汇,各局部本体概念、属性之间

的关系,全局本体和局部本体概念、属性之间的关系;三是元本体,描述局部本体中某些属性具备的冲突因子及缺省取值,如计量单位因子、类型因子等,用于后续的冲突消解。

元本体具有以下的结构:

```
<owl:Class rdf:ID="MetaOnt"/>
<owl:ObjectProperty rdf:ID="whatConcept">
  <rdfs:domain rdf:resource="# MetaOnt"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="whatProperty">
  <rdfs:domain rdf:resource="# MetaOnt"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasFactor">
  <rdfs:domain rdf:resource="# MetaOnt"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="defaultValue">
  <rdfs:domain rdf:resource="# MetaOnt"/>
</owl:ObjectProperty>
```

受元本体约束的属性称为受限属性。本体的设计过程如图 1 所示。

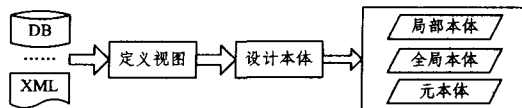


图 1 本体设计过程

有些冲突的消解可以采用纯技术手段来完成,而有些冲突的消解却要依赖用户提供的线索。如计量单位冲突,采用元本体可以描述各属性分别对应的计量单位,但用户进行信息查询时,结果数据采用什么计量单位,需要用户提供线索。

基于这些考虑,设计了语义集成框架,如图 2 所示。

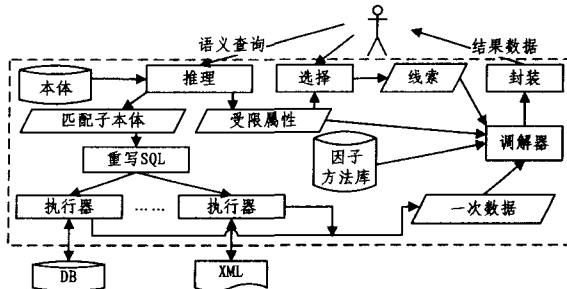


图 2 语义集成框架

用户提交语义查询后,在全局本体和局部本体上进行推理,得到匹配子本体。针对匹配子本体中的受限属性,由用户确定自己期望的冲突因子值,即线索。将匹配子本体重写为 SQL 语句,由执行器在数据源上进行具体的操作,得到一次数据。调解器进行部分冲突的消解。线索代表用户的期望,受限属性具有的冲突因子描述了一次数据中这些属性的实际取值,应用因子方法库中的方法对一次数据进行转换,得到用户期望的数据形式。最后数据封装为 RDF 或其它用户期望的格式。

在将语义查询重写为 SQL 时,本体和查询都被转换为图结构,在图上进行语义匹配,得到匹配子本体。再对匹配子本体进行深度优先搜索,转换为 SQL 语句。这一过程实现了对部分 OWL 算子的支持。算法细节另有文详述。

5 冲突的消解

在上面的语义解决框架中,已经从不同的方面对可能存

在的冲突进行了消解,其总结如表 2 所列。

表 2 冲突消解

采取的措施	消解的冲突
执行器	物理冲突
定义视图, 并映射到局部本体	字段结构冲突、字段命名冲突、记录标识冲突、 表命名冲突、表结构冲突
全局本体	字段命名冲突、表结构冲突、表关系冲突、共享语义冲突
调解器	字段类型冲突、字段长度冲突、字段精度冲突、 字段计量单位冲突、字段表达方式冲突

每个执行器对应一个独立的数据源,它接收查询请求,结合共享数据的视图定义,进行具体的查询操作。不管各信息系统采取什么样的方式存储、管理数据,无论是关系型数据库、XML 文件,或者其它任何方式,由执行器分别采用不同技术和底层数据管理系统进行交互。但执行器对外提供统一、透明的接口,这样就消解了不同数据源在数据处理上存在的物理冲突。

为共享数据单独建立视图,并将其 schema 映射到局部本体。针对字段结构冲突的第一种情况,可以在定义视图时利用自定义字段方法产生新字段,从而消除冲突。针对字段结构冲突的第二种情况,结合视图的定义,将这一特征统一用视图来表示,外键关系用本体中的对象属性来表达。在定义视图和本体时,采用重命名方法消解字段命名冲突和表命名冲突。利用 OWL 的 sameAs 构造子,指明记录间的等价关系,从而消除记录标识冲突。对于表结构冲突,即表达相同概念的不同表具有不同的字段集,可以抽象出它们的共性部分,在定义本体的时候将共性部分转换为父概念,各表对应于它的子概念。

全局本体包含有各局部本体共享的公共词汇,以及局部本体和公共词汇之间、局部本体词汇彼此之间的关系。利用 OWL 的 subClassOf, equivalentClass, subPropertyOf, equivalentProperty 等构造子,可以表达清楚各概念、属性之间的关系,消解冲突。在构造全局本体的过程中,对语义一致的概念、属性,可以用等价关系来描述;对语义有差异,但有共同基础的概念、属性,可以抽象出共同部分作为父概念或父属性,相关内容作为子概念或子属性。各局部本体尽可能和全局本体建立联系,利用全局本体实现局部本体间的映射。

对于可以通过简单计算、变换来消解的冲突,总结、细化其类型,得到冲突因子,并建立消解冲突的因子方法库。设计了元本体来描述局部本体中受限属性具备的冲突因子。用户

查询请求中,受限属性具有的冲突因子描述了一次数据中哪些数据的语义需要进一步明确,用户线索指明了对这些数据期望。调解器将两者进行匹配,应用因子方法库的函数对一次数据进行转换,就消解了这类冲突。

应用上述方案,可以解决绝大多数冲突,但仍有少量冲突没办法消解。比如记录的内容冲突,从两个不同的数据源可能得到同一个人“张三”的年龄分别为 30 和 40,这是由数据源的数据质量造成的,可以通过手工的方式去调整。

结束语 本文对信息集成中可能遇到的冲突进行了总结、分析,并基于语义技术,设计了解决方案,较好地消解了各个层次的冲突。在进行语义信息集成时,目前最大的困难在于如何在海量数据上进行语义处理。将业务数据静态转换为本体中的实例,现有的 DL 推理机如 Pellet, Racer 等都无法对海量数据进行推理;Oracle 11g spatial 虽然有强大的处理能力,且最大能支持 OWL Prime,但它只支持本地推理,而且要求所有待处理的数据都要进入它的 model,所以无法处理分散的数据库。如果将分散的数据静态导入一个数据库,有可能带来数据重复存储及衍生问题。所以面对异质异构的信息系统,目前常见的方法是将语义查询重写为 SQL 语句,但都只支持 OWL 的部分构造子,算法不一样,支持的数量也不相同。这对充分应用语义技术进行信息集成有一定的影响,今后可以在这方面做进一步研究。

参 考 文 献

[1] 李瑞轩,卢正鼎. 多数据库系统原理与技术[M]. 北京:电子工业出版社,2005

[2] 王洪伟,伊磊,张元元. 本体模型的逆向获取研究[J]. 计算机科学,2007,34(2):166

[3] 许卓明. 从 ER 模式到 OWL DL 本体的语义保持的翻译[J]. 计算机学报,2006,29(10):1786

[4] Sudha RAM, Jinsoo PARK. Semantic conflict resolution ontology(SCROL): an ontology for detecting and resolving data and schema-level semantic conflicts [J]. IEEE Transaction on Knowledge and Data Engineering, IEEE Transactions, 2004, 16 (2):189-202

[5] Cruz I F, Xiao Huiyong. Ontology Driven Data Integration in Heterogeneous Networks[C] // Complex Systems in Knowledge-based Environments: Theory, Models and Applications. Berlin:Springer Berlin / Heidelberg,2008:75

(上接第 113 页)

[3] Azarian K, El Gamal H, Schniter P. On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels[J]. IEEE Transactions on Information Theory, 2005, 51(12):4152-4172

[4] Dohler M. Virtual antenna arrays[D]. King's College London, Univ. London, London, U. K. , 2003

[5] Laneman J N, Wornell G W. Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks

[J]. IEEE Transactions on Information Theory, 2003, 49(10): 2415-2425

[6] Zhang Peng, Wang Furong, Tu Lai, et al. Performance Analysis and Diversity-Multiplexing Tradeoff in Virtual Antenna Array [C]//Proceedings of FGCN. Hainan, China, 2008

[7] Foschini G J. Layered space-time architecture for wireless communications in a fading environment when using multiple antennas[J]. Bell Labs Tech. J. , 1996, 1(2):41-59