本体驱动 ETL 过程的设计研究

阮文娟[1],刘勇军[1],张五生[2]
(1.武汉理工大学 管理学院,湖北 武汉 430070;
2.厦门大学 信息科学与技术学院,福建 厦门 361005)

摘 要:针对建立数据仓库时数据源存在结构多样性和语义异质性的问题,提出了本体驱动 ETL 过程的设计方法。通过元数据抽象以及语义建立本体,并运用 OWL 实现本体;再根据局部本体与全局本体之间的关系建立本体映射;最后运用本体映射和本体推理驱动 ETL 过程。该方法能有效解决数据源异构问题,并实现 ETL 过程的部分自动化。

关键词:ETL;OWL;本体
中图分类号:TP311.13 **文献标识码**:B **文章编号**:1674-7720(2010)22-0065-04

Research on design of ontology-driven ETL processes

RUAN Wen Juan[1], LIU Yong Jun[1], ZHANG Wu Sheng[2]
(1.School of Management, Wuhan University of Technology, Wuhan 430070, China;
2.School of Information Science and Technology, Xiamen University, Xiamen 361005, China)

Abstract: There exists structural diversity and semantic heterogeneity problems in the data sources when establishing data warehouse. To cope with these problems, an ontology-driven ETL process design method is proposed which establish ontology based on the metadata abstraction and semantic, and use OWL to achieve the ontology. Then establish ontology mapping across the relationship of the local body and global ontology. Finally we use ontology mapping and ontology reasoning to drive the ETL process. This method can effectively solve the problem of the heterogeneous of data sources, and achieve part of ETL process automation.

Key words: ETL; OWL; ontology

随着数据挖掘技术的不断发展,数据仓库已经能够有效地将数据集成到结构一致的数据存储环境中,从而使分散、不一致的操作数据转换为方便查询和分析所需的信息。但由于数据源具有异构性,企业需要一个能够从所有平台和环境中抽取数据,再将数据转换后加入目标数据仓库的高效处理过程,这个过程就是数据的抽取、转换、装载,即 ETL(Extract-Transform-Load)。

数据源异构问题主要表现在:(1)结构的多样性,如不同的数据库,不同的数据类型和不同的概要设计等;(2)语义异质性,这包括不同的命名定义和不同的表示格式[1]。

基于传统的 XML 元数据编码方法的 ETL 过程已经不能很好地解决数据源异构问题。首先,XML 在处理元数据语义上存在两个问题[2]:(1)同一概念有多种词汇表示;(2)同一个词有多种含义(概念)。因此 XML 无法对元数据进行准确的描述,这会直接影响 ETL 过程的效果。其次,必要的转换和内部模式映射依旧依赖手工操作,这不仅费时而且还容易出错。

为此,本文提出了一种本体驱动 ETL 过程的设计方法。

1 ETL 和本体论

1.1 ETL 概念

ETL 是负责将数据从源加载到目标数据仓库的过程,也是构建数据仓库的重要环节。ETL 包括以下三个过程[3]:(1)抽取,数据抽取是捕获数据源的过程,即将数据从各种原始的业务系统中读取出来,这是所有工作的前提。(2)转换,按照预先设计的规则将抽取得到的数据进行转换、清洗,处理一些冗余、歧义、不完整、违反业务规则的数据,统一数据的粒度,使本来异构的数据格式统一起来。(3)装载,将转换后的数据按照计划增量全部导入到数据仓库中。

ETL 作为 DW 的核心和灵魂,大约占整个 DW 项目

60%~80%的时间。在现实应用中ETL的执行效率往往成为实施DW项目的瓶颈,而ETL规则的设计和实施又是其中工作量最大的部分。

1.2 本体论和OWL

Ontology概念起源于哲学领域,即"对世界上客观存在物的系统描述"。但其明确定义是在1991年由Neches[4]等人引入人工智能领域。其后在1993年Gruber和1997年Borst也给出了Ontology的定义。直到1998年Studer[5]等人在前人基础上给出了较为广泛接受的概念,即"Ontology是共享概念模型的明确的形式化规范说明",并指出该定义包含四层含义:概念模型(conceptualization)、明确(explicit)、形式化(formal)和共享(share)。此外2001年Hendler[6]也试图作出解释。

W3C为本体的开发提供了一种网络本体语言OWL[7](Web Ontology Language),该语言包含了三种表达能力依次增强的子语言,即OWL Lite、OWL DL和OWL Full。OWL Lite支持只需要一个分类层次和简单约束的用户;OWL DL支持需要最强表达能力的推理系统的用户,且这个推理系统必须确保计算的完全性和可判定性,OWL DL包括了OWL语言的所有语言成分,但使用时必须符合一定的约束,受到一定的限制;OWL Full支持那些不需要可计算保证的,但能在完全自由的RDF上进行最强描述的用户,包含OWL的全部语言成分并取消了OWL DL中的限制。

相比于传统XML,OWL有更丰富的建模原语,能够表达语义并描述复杂逻辑关系,可以解决XML无法对元数据进行准确描述的问题。而且本体语言还可以通过建立本体映射,并运用本体推理来实现部分必要转换和内部模式映射的自动化。因此引入本体驱动ETL过程能有效地解决数据源异构问题,并实现ETL过程的部分自动化。

1.3 本体驱动ETL的一般步骤

运用本体理论指导ETL过程的一般步骤为确定领域本体、寻找本体映射以及选择适当的本体推理规则。具体过程为:

(1)用本体论的理论知识指导元数据的确立,然后运用本体语言建立本体。

(2)建立局部本体到全局本体之间的关系,即本体映射。包括一个领域基本概念之间的层次关系,同时要满足不同局部本体之间的相互查询需求。目前有很多种本体映射的方法,如GLUE[8]方法是一种利用概念的实例作为计算概念间相似度的依据,然后通过机器学习技术寻找单独分开存储的概念与自治本体之间的语义映射;SF(Similarity Flooding)[9]是一种基于相邻概念节点之间的相似传递性的算法;H-MATCH[10]是一种动态分布式本体匹配的算法。

(3)本体推理,即通过一定的规则推理出本体内部或是局部本体与全局本体之间的关系,来帮助确立映射关系。KAON2[11]是一个OWL推理机制,带SWRL子集DL-safe扩展;Pellet[12]是一个用Java构建的推理机制,专门为OWL推理设计的,这两种工具都可以用来推理。

(4)将本体推理所得到的映射转化为熟悉的ETL过程。

2 基于本体的ETL架构设计

在对本体论和ETL过程研究的基础上,本文提出了基于本体的ETL架构设计,这一架构包括三个主要阶段:(1)元数据抽象阶段:任务是从数据源中抽取元数据,然后将元数据抽象为本体。它包括局部本体和全局本体的定义。(2)本体映射阶段:目的就是找到局部本体和全局本体内部以及它们之间的语义关联,解决不同本体间的知识共享和重用使使用者更好地认识结构和语义领域的异构。本体映射的方法之一是计算两个本体的相似性。(3)基于ETL规则的本体推理阶段:即根据ETL的一般规则来制定本体推理规则,帮助从映射关系中找到隐含的与相冲突的之间的关系。如果数据源更倾向于用自动化过程开发,则对映射阶段进行描述的推理至关重要,特别是ETL过程。从数据源到目标数据仓库的ETL过程如图1所示。

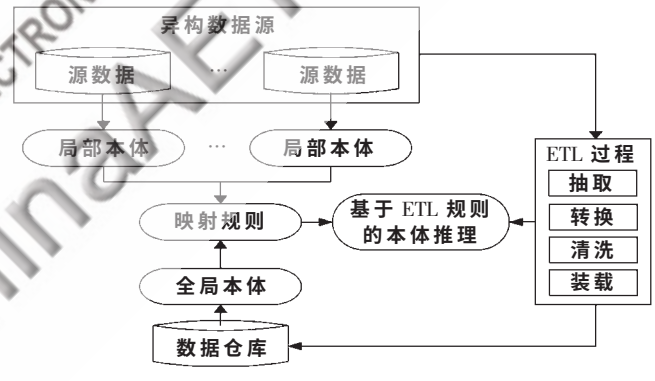

图1 基于本体的ETL架构

3 本体驱动的ETL过程

假设有两个主要实体,即客户和订单,整个设置如表1所示。一个客户有一个名字(包括他/她的姓和名)和相应的地址(其中包括他/她的国家、城市和街道)。一个订单包括一个特定的日期,其格式可以为"日/月/年"或"月/日/年"。订单的价格有美元和人民币两种货币表现形式。还有购货数量,其订货形式为"零售"或"批发"。而且有两个数据源分别是以Oracle数据库和SQL

表1 源数据和目标数据的定义

DS1	S_customer S_order	{id1,name,country,city,street} {id1,cid,date,amount,price}	Oracle
DS2	S_customer S_order	{id2,firstname,lastname,cus_address} {id2,cid,date,amount,price}	SQL Server 2005 Oracle
DW	w_customer w_order	{tid,firstname,lastname,address} {tid,cid,date,amount,price}	

Server 2005 存储客户购买商品的资料。

3.1 本体建立

3.1.1 建立局部本体

在构建局部本体前，先用本体理论指导理清字段间的关系，需明确以下关系：字段“country”、“city”和“street”是字段“address”的一部分；字段“mmddyy”和“ddmmyy”是字段“date”的不同类型；字段“retail”和“wolesale”是字段“amount”的两种售出形式。

在明确了上述关系的基础上，开始建立局部本体。以表 1 中数据源 1(DS1)为例进行说明，采用数据源的名字作为局部本体 owl 名，每个表对应一个类(概念)，实例数据源 1 的 DBMS 为 Oracle，以下是为 DS1 建立局部本体的主要 OWL 语句。

```
<owl:Class rdf:ID="#s_customer"/>     //定义客户类中的元素
<owl:DatatypeProperty rdf:ID="s_table">
    <rdfs:range rdf:resource="http://www.w3.org/2001/
                                    XMLSchema#string"/>
    <rdfs:domain rdf:resource="#s_customer"/>
</owl:DatatypeProperty>
<owl:Class rdf:ID="cid">
    <rdfs:subClassOf rdf:resource="#s_customer"/>
    </owl:Class>
```

3.1.2 建立全局本体

全局本体可以通过集成各个数据源所包括的领域信息而得到。它就像一个共享的词汇库，建立了数据源领域的知识模型，且为数据源提供了公共的语义描述，从而做到将系统的全局视图进行语义化描述，从而解决不同局部本体之间的语义异构性。

对于全局本体元素的定义基本可以借鉴对 DS1 元素的定义。建立全局本体还有一个重要的任务就是确定术语，从而对数据源有一个全局的把握。本例的术语有：客户、订单、地址、名字、数额、价格和日期。

3.2 本体映射

本体映射有多种方法，在实际应用中，可以将这些方法结合起来使用。因此发现以下关系：

```
<owl:Class rdf:ID="city">                    //表明 city 与 street 是
                                                不相交的关系。
<rdfs:subClassOf rdf:resource="# s_customer "/>
  <owl:disjointWith>
  <owl:Class rdf:ID="street">
    </owl:disjointWith>
</owl:Class>
```

以上是运用 OWL DL 描述映射关系包括层次、等价、交集、并集、互补和不相交关系。在此运用 OWL DL 提供的 rdfs：用 subClassOf 声明一个类是通过另一个或多个类的子类创建层次关系，用 equivalentProperty 以描述本体中属性间的等价关系，用 disjointWith 描述了类之间不相交关系。

3.3 基于 ETL 规则的本体推理

通过单纯的本体映射，只是找出了本体之间存在的一般关系。还需要运用基于适当规则的本体推理进一步指导 ETL 过程，即基于 ETL 规则的本体推理。

通过本体推理，可以找出每一个数据和数据仓库之间的关系，例如：等价关系、包含关系、父子关系、兄弟关系。对于本例，通过本体推理可以发现如下关系：(1)等价关系。id1 字段等价于 id2 字段，并且有相同的关系，从而推出字段 id1 与字段 tid 等价。(2)包含关系。字段 cus_address 与字段 country、city 和 street 是包含关系，因字段 cus_address 与字段 address 等价，从而可以推导出字段 address 也包含字段 country、city 和 street。(3)兄弟关系。很容易发现每个类下面的属性之间都是兄弟关系，即互不相交。

在已经明确以上关系的情况下，还需要用 ETL 规则指导本体推理。为了完成 ETL 过程，需要把 DS1 中的 city、street 和 country 字段通过一定的规则加载成为 DW 中的 address 字段。因 city、street 和 country 三个字段是互不相交的，可知 DS1 中的 city、street 和 country 字段可以构成 DW 中的 address 字段。在 ETL 规则和推理规则的指导下可知 city、street 和 country 三个字段必须先组合在一起，然后再加载到目标数据库中。

同理，可以推导出从 name 字段中抽取出 firstname 和 lastname 两个字段。为了完成数据源字段 name 加载到目标数据库，根据 ETL 规则需要将字段 name 拆分成字段 firstname 和字段 lastname，然后再加载到目标数据库。

因 price 字段有两种货币形式，在把该字段从数据源加载到目标数据库的过程中需要注意将数据源的货币先转换成对应目标数据库的字段或是对应的中间字段，然后再加载到目标数据库。在实际 ETL 过程中先将该字段过滤，然后对格式与目标数据库不一致的 price 字段进行格式转换。

3.4 本体驱动的 ETL 过程

经过本体映射和推理，得到了一系列的相关关系。然后在上述基础上完成本例的数据从数据源到目标数据仓库的加载过程，具体步骤为：

(1)确定简单操作及可以直接从源到目标的操作。

(2)从源节点开始引进 ETL 数据转换操作，对相对复杂的转化过程引入中间节点；然后从中间节点出发，继续采取额外的转换直至到达目标节点。

(3)结合 ETL 规则做最后的规范，如图 2 所示。

说明：对于数据源中的“customer”、“order”、“firstname”和“lastname”等字段可以直接从源加载到目标。字段“street”、“country”和“city”三者合在一起可以构成目标字段“address”。而“date”字段需通过转换，“name”字段需要拆分，“amount”字段需要过滤。对于“price”字段，首

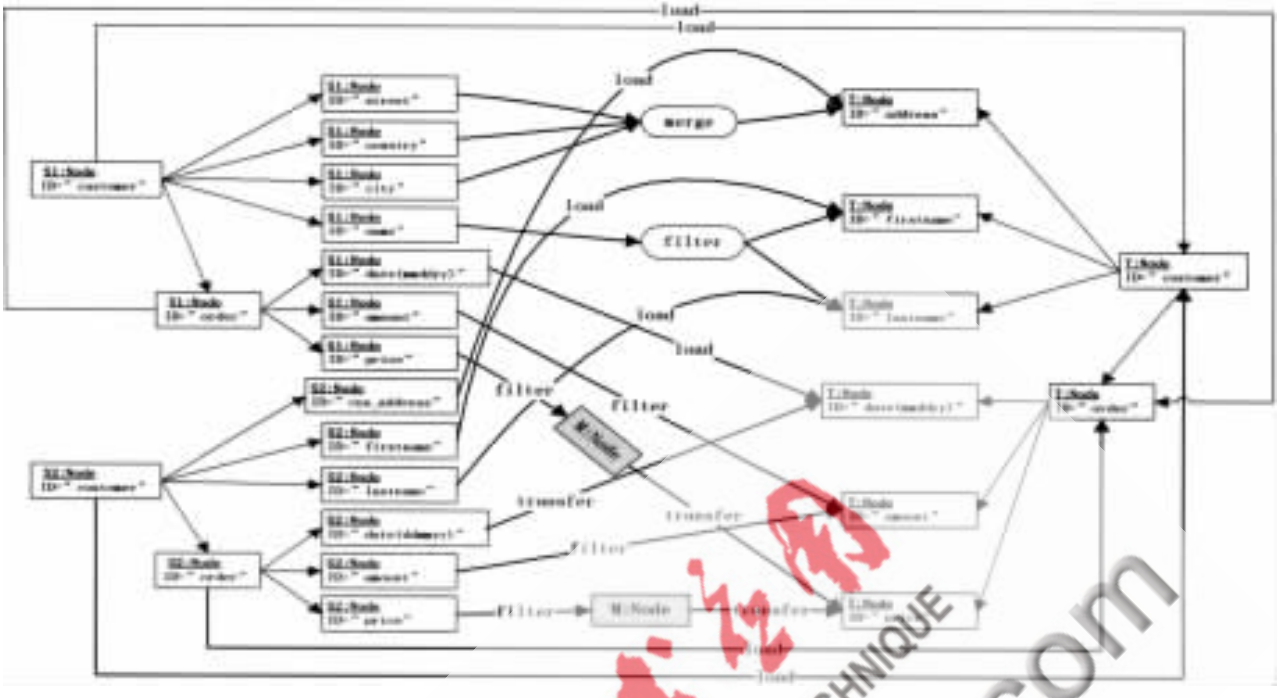

图 2 本体驱动的 ETL 过程图

先过滤，然后转换。最后用 ETL 规则规范整个过程。

本文提出了一个本体驱动的 ETL 过程的架构模型；在此基础上，建立了本体、找出了本体间的映射并在 ETL 规则的基础上进行了本体推理；最后通过实例图表的方式展现了本体驱动 ETL 过程。本体的应用使 ETL 过程更加灵活、高效，并且架构中的 ETL 过程可以部分实现自动化，从而解决了数据源结构异构和语义异构的集成问题。

参考文献

[1] ZHANG Zhuo Lun, WANG Su Fen. A framework model study for ontology-driven ETL processes[C]. IEEE International Conference on Wireless Communications, Networking and Mobile computing, 2008.

[2] 邓志鸿，唐世渭，张铭，等. Ontology 研究综述[J]. 北京大学学报(自然科学版)，2002，38(5)：730-737.

[3] 张忠平，赵瑞珍. 基于元数据驱动的 ETL 架构设计[J]. 计算机应用与软件，2009，26(6)：61-63.

[4] NECHES R, FIKES R E, GRUBER T R, et al. Enabling technology for knowledge sharing[J]. Artificial Intelligence, 1991, 12(3): 36-56.

[5] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: principles and methods[J]. Data and Knowledge Engineering, 1998, 25: 161-197.

[6] HENDLER J A. Agents and the semantic web[J]. IEEE Intelligent Systems, 2001, 16(2): 30-37.

[7] W3C Candidate Recommendation. OWL Web Ontology Language Guide[EB/OL]. [2003-08-18]. http://www.w3.org/TR/2003/CR-owl-guide-20030818/.

[8] ANHAI D, JAYANT M, PEDRO D, et al. Learning to map between ontologies on the semantic web[C]. Proceedings of the Eleventh International World Wide Web Conference, 2002.

[9] MELNIK S, GARCIA-MOLINA H, RAHM E. Similarity flooding: a versatile garph matching algorithm[C]. On Data Engineering(ICDE), 2002.

[10] CASTANO S, FERRARA A, MONTANELLI S. An algorithm for dynamically matching ontologies in peer based systems Berlin[C]. Proc of the 1st Workshop on Semantic Web and Databases, 2003, 231-250.

[11] KAON2[EB/OL]. [2008-01-14]. http://kaon2.semanticweb.org/.

[12] Pellet: OWL 2 Reasoner for Java[EB/OL]. [2008-10-27]. http://clarkparsia.com/pellet.

(收稿日期：2010-05-18)

作者简介：

阮文娟，女，1984年生，硕士研究生，主要研究方向：数据仓库与数据集成。

刘勇军，男，1975年生，副教授，硕士生导师，主要研究方向：供应链协同、语义 Web 服务系统、知识管理等。

张五生，男，1983年生，硕士研究生，主要研究方向：嵌入式。