

基于本体和 Web Services 的数据交换平台

王艳敏, 谢 强, 丁秋林

(南京航空航天大学 信息科学与技术学院, 江苏 南京 210016)

摘 要:针对目前数据交换方式在解决交换信息语义异构方面存在的不足,在 XML 技术的基础上,提出一种基于本体和 Web Services 的数据交换平台。首先给出一种数据交换平台的系统框架,对关键技术进行研究,提取各异构数据源的数据构造 XML Schema 文件,然后采用本体技术对其进行语义标记,形成带有语义信息的模式文件,最后通过对 XML Schema 文件进行模式匹配和映射,生成转换方案。实例效果表明该数据交换平台能有效地解决语义异构问题,并通过 Web 服务调用各业务系统实现数据交换和共享。

关键词:数据交换;本体;Web Services;模式匹配;语义标记

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2010)05-0112-05

Data Exchange Platform Based on Ontology and Web Services

WANG Yan-min, XIE Qiang, DING Qiu-lin

(Coll. of Information Sci. and Tech., Nanjing Univ. of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: The current data exchange model had gap in solving the problem of semantic heterogeneous of exchange information, based on the XML technology, a data-exchange platform which was based on Ontology and Web Services was proposed. First, given a system framework of data-exchange platform, researched on the key technologies, and taken the data of various heterogeneous data sources, XML Schema documents were created. Ontology technology was used for marking up their semantic so as to form pattern documents with semantic information, and then by mapping and matching of the pattern between schemas, conversion programs were produced. The instance results show that this data exchange platform can obviously solve the problem of semantic heterogeneous, which realizes data exchange and sharing by Web Services calling the operational systems.

Key words: data exchange; Ontology; Web Services; pattern matching; semantic marking

0 引言

随着信息化的高速发展,政府部门及各大企业都建立了自己的信息管理系统。这些信息系统往往是在不同时期、由不同厂商利用不同的工具、不同的平台开发的,并且运行在不同的操作系统和不同的数据库平台之上。这些业务系统由于缺少统一规划,彼此之间很难实现信息共享和数据交换。

目前常用的数据交换模式有以下几种^[1,2]:

(1) 点对点的数据交换:通常没有或者很难形成一个统一的数据交换标准,导致了相同的数据分析处理模块在很多应用中被重复地撰写。随着新系统的不断增加,需要与原有系统分别建立交换的接口,造成系统

瓶颈,效率低下;

(2) 基于 XML 的数据交换:能有效地解决系统异构、数据模式异构的问题,从而实现数据交换。但是通用性差,映射过程复杂且无法解决语义异构问题;

(3) 基于数据仓库的数据交换:中心数据仓库负责提取各个分布场地的自治系统的数据,并对各个数据具有高度的控制权。缺点是交互性与实时性较差。

为此,文中提出了一种基于本体和 Web Services 的数据交换平台,解决异构数据交换和共享问题。通过这一平台,可以把各业务系统内外部的各种相关数据资源进行整合,实现实时的数据交换和共享。

1 基于本体和 Web Services 的数据交换平台架构

1.1 数据交换平台的总体框架

基于本体和 Web Services 的数据交换平台目的是为不同的应用系统提供统一的、自动化半自动化的信息交换功能,最大限度地解决各业务系统的“信息孤

收稿日期:2009-08-03;修回日期:2009-11-07

作者简介:王艳敏(1985-),女,河南周口人,硕士研究生,研究方向为知识工程、信息系统集成、人机交互;谢 强,副教授,博士,研究方向为知识工程、信息系统、信息安全、人机交互;丁秋林,教授,博士生导师,研究方向为 CIM、DSS、MIS。

岛”问题。本数据交换平台,采用 Web Services 技术,实现跨越各种不同的系统进行数据交换;运用 XML 和本体技术解决系统异构、数据模式异构以及数据语义异构等问题,从而使不同的数据源模式之间能够准确地进行数据转换。数据交换平台总体框架如图 1 所示。

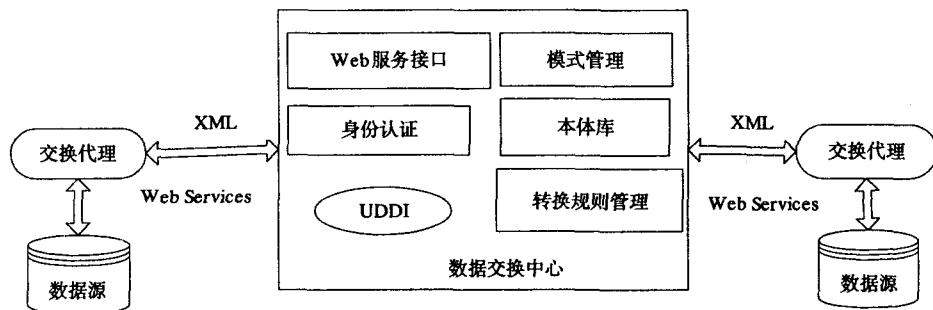


图1 数据交换平台总体框架

(1)交换代理:实现对异构数据源进行屏蔽,将数据源的数据(各种数据库、电子表格等)与 XML 文档双向转换,生成相应的 XML Schema,并提供统一的 Web 服务接口与数据交换中心进行跨平台的交互。

(2)Web 服务接口:数据交换中心的 Web Services 提供 3 类服务:数据模式访问接口、数据服务接口、数据交互接口。数据服务接口:由用户通过交换代理向数据交互中心发出请求,由数据交互中心向 UDDI 服务器查询具有特定模式的数据服务,并将查询结果返回给请求者;数据交互接口:通过此接口数据请求者向交互代理发出请求,由交换代理调用数据交换中心数据请求响应 Web Services,由其调用数据提供者的数据服务得到数据,并经过格式转换返回给数据请求者。

(3)本体库:由手工构建本体信息,存储以 OWL 描述的本体信息,对各业务系统提交的 XML Schema 文件进行语义标记。

(4)模式管理:负责保存和管理数据库中的所有数据源文件的 XML 模式,在数据交换时选择适当的模式进行匹配,并负责更新 XML Schema 文件。

(5)转换规则管理:在数据交换中心建立数据发布者共享的数据模式与数据接受者需求的数据模式之间的映射,并用 XSLT 表示转换规则,实时更新规则库。

1.2 数据交换流程

整个数据交换的流程可以描述为:应用系统 1 向数据交换中心发出数据请求,数据交换中心形成转换方案并向应用系统 2 转发数据请求,然后对系统 2 发回的数据,应用转换方案进行数据处理,并将结果经数据中心处理后发回应用系统 1。其中当应用系统第一次进入交换系统时,通过交换代理生成该数据源的

XML Schema 并将 XML Schema 提交给数据交换中心;数据交换中心利用本体对 XML Schema 进行标记,形成具有语义信息的模式文件,再经过转换方案生成器形成最终方案,并保存到转换规则库中。再次登录时,与其它应该系统交换数据时,把要交换的数据通过交换代理转换成 XML 文档并提交到数据交换中心;

数据交换中心收到 XML 文档后利用转换方案生成器根据转换规则库转换成目标系统识别的 XML 文档。数据交换流程如图 2 所示。

2 XML 文档与数据源间的数据转换

XML 文档与数据源间的数据转换是解决数据交换的一个基础性问题。由于数据库中的表与 XML 文档都具有高度规范的结构,因此关系模式与 XML 模式之间能够较容易地进行相互转换。转换时,遵循“完整映射”的原则,将关系表之间的关系完整地反应到 XML 文档的结构中,同时建立数据库表字段与相应的 XML Schema 文件元素的映射,实现二者直接的相互转换^[3]。

2.1 关系数据库到 XML 的映射

应用系统需要共享数据时,交换代理从数据库中提取数据,形成映射文件及相应的 XML Schema 文档。利用 XML 的 DOM 结构模型创建 XML 文档。首先创建一个空的 Document 对象,然后根据对应的 XML Schema 文件不断创建节点,根据数据库中的记录或者字段与 XML 文档的对应关系给节点赋值,并且把节点添加到 DOM 树中去,最终生成源 XML 文档。数据库数据转换为 XML 的转换过程如图 3 所示。

其中关系模式转换为 XML Schema 的转换方法如下:

(1)定义一个全局元素表示数据库(数据集)名,在全局元素内定义“complexType”类型的元素,每个元素对应一个表名,称为表元素。这样做的目的是为了在其中定义相对表中所有记录唯一的键元素。增加子元素,子元素对应表的字段,字段类型对应子元素类型,由“type”属性指定,表中字段记录的值(即表的数据)可以由该子元素的值表示,也可以以子元素的属性值形式表示。

(2)在全局元素中,也就是与表名元素同级的结构中,定义表的主键和外键。表的主键按如下方法定义:创建 XML Schema 的“key”元素,指定主键名及主键所在的域。表的外键:创建“key ref”表示,定义属性 re-

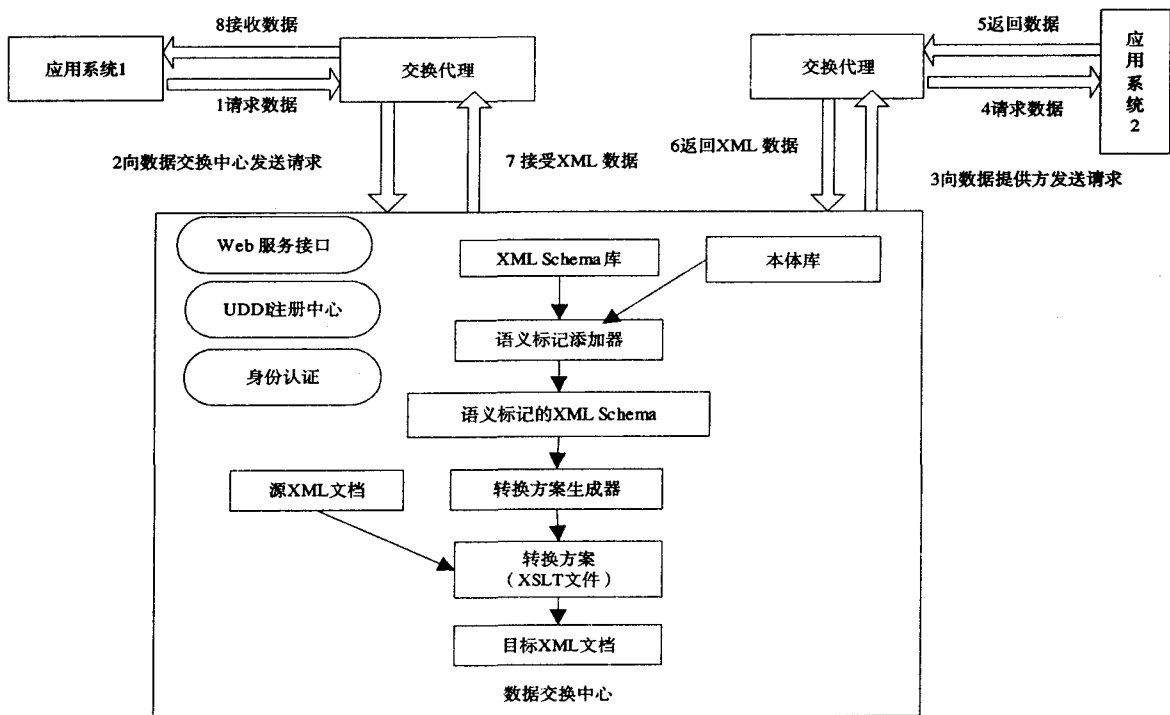


图 2 数据交换流程

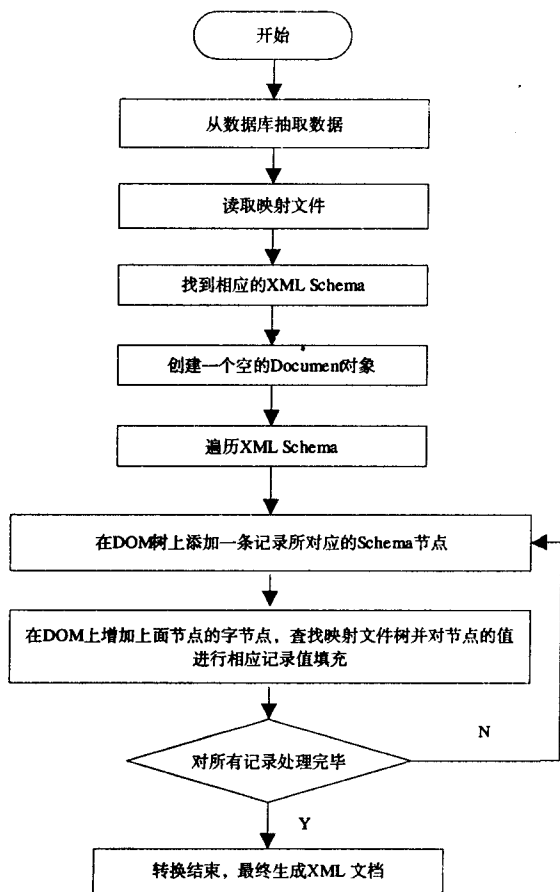


图 3 数据库到 XML 文档的转换流程

fer, 它的值对应所引用的键名, 同样还需要指定外键所在域; 关系模式中是否允许为空约束映射为 XML Schema 的 unique 约束。当表的主键多于一个时, 将主

键的集合定义成一个 complexType 类型的元素, 在元素中定义键元素。

2.2 XML 到关系数据库的映射

交换代理通过 Web Services 服务接收从数据交换中心返回的 XML 格式的消息数据, 从消息头中的 SchemaID 得到该数据属于哪个模式 (XML Schema) 并找到对应与该模式的映射文件 (定义了消息体文档元素和关系数据库表字段与 XML 格式数据之间的映射定义以及各表之间的约束关系), 接着解析 XML 格式的消息数据和映射文件, 生成对数据库的 Insert 或 Update 语句, 把 XML 数据更新到数据库中。否则返回错误信息。其处理流程如图 4 所示。

3 基于本体的 XML Schema 的匹配

两个数据源的数据格式进行转化, 首先进行模式文件 (即 XML Schema 文件) 的匹配, 使得两个模式文件的各元素之间形成映射关系。由于模式文件缺乏语义信息, 使得模式匹配的结果效率不高且常常不符合要求。因此引入本体对数据模型描述 (XML Schema) 进行标记, 使其具有语义信息, 从而提高匹配结果的准确性及效率。

3.1 本体标记 XML Schema

本体 (Ontology) 是一种新型的元数据, 其目标是捕获相关领域的知识, 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和词汇间相互关系的明确确定

义,由此实现知识重用。

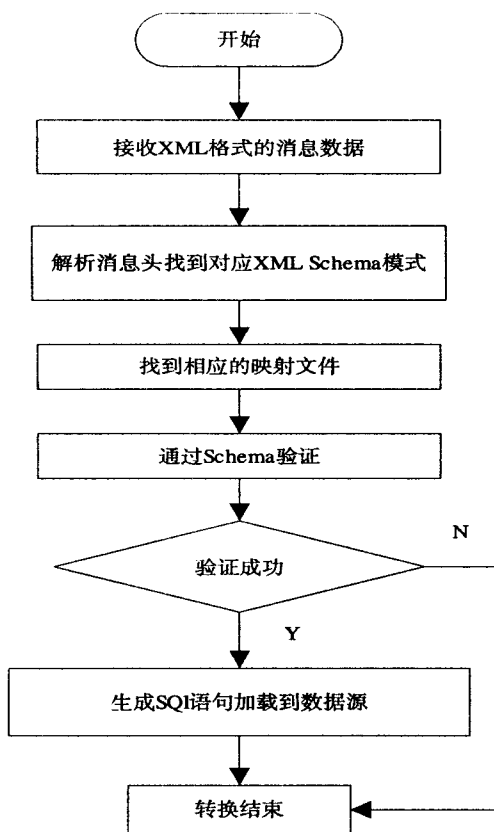


图4 XML加载到数据库的处理流程

利用本体中定义的概念和属性对 XML Schema 中的结构进行标记^[4-7]。在标记过程中,使用到辞典如 WordNet^[8]的功能实现语义标记的部分自动化处理,还用到模式匹配中已成熟的技术^[9]。利用 XML Schema 的命名空间(namespace)扩展机制添加标记到 XML Schema 中去。标记时处理的问题如下:

(1) 标记的先后顺序。首先对 XML Schema 中 ComplexType、SimpleType 类型的结构采用本体中的概念(而不是属性)进行标记,在自动标记完成之后由用户确认。在概念标记完成之后,对概念具有的属性进行标记。概念的匹配为其属性的匹配建立了一个上下文环境,在这个上下文环境中可以对其属性进行准确度比较高的自动标记。

(2) classifier 标记的使用。在结构中没有特定元素进行细粒度概念的区分时,XML Schema 可直接用粗粒度概念进行标记。反之,则用 classifier 标记这个元素。对 XML Schema 中的 Restriction、Extension 扩展机制可以使用本体中的父子类进行标记,主外键也可使用属性进行标记。

3.2 Schema 之间的匹配和映射

语义标记添加后,利用模式中的语义标记,各模式的元素和结构之间通过语义标记的匹配以及本体的参

与,实现模式之间的映射。但由于双方所对应的 XML Schema 会采用不同的术语,这样本体标记完成后,双方相应的某些概念不能很好地进行映射,为此还要考虑语义标记在本体中的关系,从而确定映射关系。当有新的数据源加入进来后,只要先对其进行语义标记,然后即可在数据转换中提取其中的数据。

在 OWL 中,类之间的包含、等价、分离关系可以用来描述 Ontology 中类的参差结构。一个 Ontology 中的类可能是另一个 Ontology 中的类的子类(OWL:sub-Class),或者等价类(OWL:equivalentClass)。例如数据源一中的 Schema 有语义标记为“general_Module”元素,数据源二的 Schema 有语义标记为“Module”元素。虽然二者语义标记不同,但从下面的 OWL 的描述可知两者具有映射关系,从而两个 Schema 相应的元素也有映射关系。general_Module 和 Module 之间的关系 OWL 表示如下:

```

<owl: Class rdf: ID= "general_Module">
  <owl: intersection rdf:parseType= "Collection">
    <owl: Class rdf: about= "Module"/>
    <owl: Restriction
      <owl: onProperty rdf: resource= "hasCategories"/>
      <owl: hasValue rdf: resource= "# general"/>
    </owl: Restriction>
  </owl: intersectionOf>
</owl: Class>
  
```

两个不同的数据模式通过匹配技术分别与本体进行匹配产生映射关系,用 XSLT 来实现。XSLT 文件是目标 XML Schema 文件和模式匹配形成的映射关系集合(目标映射节点集,源映射节点集及对应的路径集合)。首先对目标 Schema 进行遍历生成模板元素,遇到映射关系的节点。添加子节点,属性为源文件相应节点路径所对应的值。XSLT 把要处理的 XML 文档看作一个节点树,称为源树,把它转换为一个不同结构的结果树。每一个 XSLT 文件定义了源树与结果树中对应的模板规则,其中<xsl:stylesheet>表明这个文档是 XSLT 样式表。<xsl:output>元素利用其 method 属性,制定输出文本内容类型。<xsl:template>用来生成模板,定一个一个可重复使用的模板,用于特定类型和上下文的节点生成所需的输出。在转换的过程中,利用 XPath 对源树中待转换的部分进行寻址,并定义源树与一个或多个预想确定的模板相匹配的部分,找到匹配就转换为目标文档。

4 实例分析

根据以上设计的转换方案,举例说明主要的转换步骤。

(1)首先是一个利用 Ontology 对刀具订单表形成的 XML schema 添加语义标记的例子片段。语义注解以 semantic 前缀来声明,以 semantic: type = “”的形式表示。

```
<xs:schema id = “sword”
xmlns = “http://www.w3.org/2001/XMLSchema”
xmlns:ontology = “http://www.business.org/order.owl”>
<xs:element name = “sword”
semantic: type = “&ontology; # sword”>
<xs:complexType>
<xs:sequence>
<xs:element name = “刀具订单” type = “xs:string” minOccurs =
“0” semantic: type = “&ontology; # 刀具订单”/>
```

(2)语义标记添加器对 XML Schema 添加语义标记,产生 Schema 文件的映射图,由用户修改和确认。经手工调整后两个 XML Schema 之间的映射关系如图 5 所示。

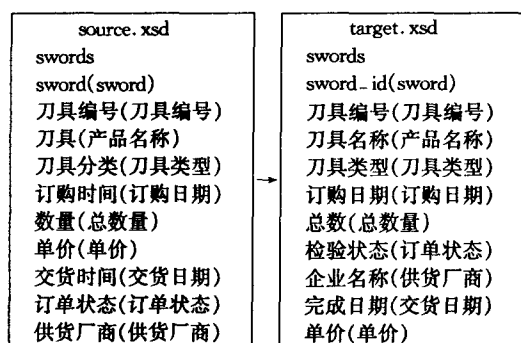


图 5 XML Schema 之间的映射关系

(3)上面举例生成的转换方案以 XSLT 表示,结果片段如下:

```
<xsl:template match = “Sword”>
< sword>
<xsl:attribute name = “ID”><xsl:value-of select = “刀具编
号”/></xsl:attribute>
<刀具分类><xsl:value-of select = “刀具类型”/></刀具分
类>
```

5 结束语

文中首先分析了目前数据交换方式存在的不足,

引入了本体和 Web Services 技术。重点介绍了关系数据库和 XML 文档之间的相互转换,采用本体技术对 XML Schema 进行语义标记,在此基础上进行了模式匹配,并由转换方案生成器完成以 XSLT 表示的数据转换方案,从而实现异构数据源间的数据转换。通过实例分析得出,本方案解决了语义异构问题,并具有良好的扩展性。

下步工作是完善本体库信息,实现模式间全自动的映射,提高匹配效率;并且针对现有的数据交换产品的不足以及使用策略的优点,预测将策略运用到数据交换领域,实现一种策略驱动的数据交换机制将是数据交换研究的一个重要方向。

参考文献:

- [1] 杨 剑,唐慧佳.基于 XML 的数据交换系统的研究与实现[J].计算机工程,2005,31(19):195-197.
- [2] 李亚楠,刘连忠,贾熾星.数据交换研究[J].计算机技术与发展,2008,18(2):5-8.
- [3] 何 忠,张申生. XML 映射器的实现[J]. 计算机工程与应用,2004(4):137-187.
- [4] Isabel F, Xiao Cruz Huiyong, Hsu Feihong. An ontology - based framework for XML semantic integration[C]// Washington: IDEAS' 04 Workshop, Proceedings of the International Database Engineering and Applications Symposium. Washington: IEEE Computer Society, 2004:217-226.
- [5] 彭 涛,张 力.基于本体和 XML 的数据交换研究[J].计算机工程,2006,32(1):90-92.
- [6] 转永光.基于本体和 Web Services 的数据交换的研究与实现[D].南京:南京航空航天大学,2008.
- [7] Huma Z, Rehman M J - U, Iftikhar N. An ontology - based framework for semi - automatic schema integration[J]. Journal of Computer Science and Technology, 2005, 20(6): 788 - 796.
- [8] Miller G A. WordNet: A Lexical Database for the English Language[M]. Cambridge, Mass: MIT Press, 1998.
- [9] Rahm E, Bernstein P A. A Survey of Approaches to Automatic Schema Matching[J]. The VLDB Journal, 2001, 10(4): 334 - 350.

2010 年中国计算机大会
(China National Computer Conference, CNCC2010)
 将于 2010 年 9 月—10 月在杭州隆重召开。