

数据仓库元数据集成技术研究与应用

李瑞旭¹, 李 扬²

(1. 烟台大学, 山东 烟台 264005;

2. 中国人民武装警察部队学院, 河北 廊坊 065000)

摘 要:元数据集成是数据仓库元数据管理的一项重要内容。文章在目前元数据集成研究成果的基础上,提出了一种基于 SOA 架构的数据仓库元数据集成技术。该技术以 Web Service 技术为应用框架,以 CWM 为元数据模型,采用 XML 设计元数据封装器,实现了分布环境下数据仓库元数据的集成与重用。文章重点介绍了系统的体系构架,及 CWM 元数据模型的结构设计和不同 Web Service 方法设计和调用。最后,将该技术应用到消防工程领域的一个实际数据仓库项目中,验证了该技术的可行性、有效性、实用性。

关键词:数据仓库;元数据集成;SOA 架构;Web 服务

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2011)09-0175-04

Research and Application of Metadata Integration for Data Warehouses

LI Rui-xu¹, LI Yang²

(1. Yantai University, Yantai 264005, China;

2. The Chinese People Armed Police Forces Academy, Langfang 065000, China)

Abstract: Metadata integration is an important issue in metadata management. It presents a metadata integration strategy based on SOA architecture for data warehouses. In this strategy, Web Service is adopted as application frame; Metadata is modeled by CWM and metadata encapsulation is designed in XML. The metadata from distributed data warehouses has been efficiently integrated and reused. It focuses on the architecture of the system, the structure design of CWM, as well as various Web Service methods creation and call. Finally, verify the feasibility, effectiveness and practicability of this solutions by experiment.

Key words: data warehouse; metadata integration; SOA architecture; Web Service

0 引 言

元数据是数据仓库的一个重要组成部分,是联系数据仓库中各部分的纽带,它作用于数据仓库的创建、维护、管理和使用的各个方面^[1]。目前,元数据管理已成为数据仓库技术研究的重点之一。数据仓库元数据管理主要包括元数据分类、存储、维护、更新、集成和交换等内容^[2]。目前对元数据管理的研究主要集中于元数据分类、元数据管理系统的功能与系统结构等方面^[2,3],对元数据集成的研究相对较少。文献[4]讨论了元数据集成问题,认为建立一个核心的、统一的元数据模型是元数据集成的前提。文献[5]提出了一种通用数据仓库元模型(Common Warehouse Metamodel, CWM)进行了研究,剖析了具体包以及如何元数

据存储和管理。文献[6,7]研究了基于 CWM 模型的元数据集成与交换格式问题。文献[8]研究了基于 XML 的元数据封装问题。文献[9]提出了一种用于元数据集成的 FIRA(Federated Interoperable Relational Algebra)关系代数。这些研究为进一步研究元数据集成问题奠定了基础。但目前这些研究对分布环境下数据仓库元数据集成问题缺乏深入探讨。

文献[10~12]介绍了 Web Service 在数据集成方面的应用。文献[10]提出了 SOA 架构(Service Oriented Architecture, SOA)与 Web Services 的关系。SOA 架构是独立于技术的,SOA 架构没有确切地定义服务应该如何进行交互,通常情况下只是定义了服务之间应该如何进行相互理解。Web Services 则拥有具体的指导原则,通过 HTTP 协议传递的 SOAP 消息来实现 SOA 模型。从本质上讲,Web Services 可以实现 SOA 架构,而且只是 SOA 架构实现方式中的一种。由于 Web Service 具有良好的分布式结构、跨平台性、可扩展性,

收稿日期:2011-02-27;修回日期:2011-06-04

基金项目:国家自然科学基金(61070118)

作者简介:李瑞旭(1966-),男,山东莱阳人,副教授,主要研究方向为知识发现、数据仓库、计算机网络体系结构。

已被广泛地应用于各种分布式应用领域。将 Web Service 技术引入元数据集成中,采用 CWM 为元数据模型,以 XML 封装元数据,在分布环境下实现数据仓库元数据集成是文中的研究目的。

1 元数据集成系统框架

1.1 系统框架

Web Service 采用面向服务架构 (Service Oriented Architecture, SOA), 在该架构下存在三个参与者与三种基本操作。图 1 是基于 Web Service 技术设计的元数据集成系统框架图。图中, UDDI 是通用描述、发现和集成规范。服务提供者可以通过 UDDI 协议注册其要发布的服务, 服务请求者依照 UDDI 规范向 UDDI 注册表搜索相应的 Web Service, 然后得到相应的 WSDL 文档进行服务调用。SOAP 提供了标准的 RPC 方法来调用 Web Service, 并定义了 SOAP 消息的格式, 以及如何通过 HTTP 协议来使用 SOAP。WSDL 则是一种基于 XML, 用于描述 Web Service 及其操作、参数以及返回值的语言。它详细描述了服务的完整信息, 包括服务的功能、位置、访问指令以及发送和接收的信息结构, 通过它可以获得 Web Service 调用的所有信息。在此框架下, 服务提供者 (元数据所有者) 将需要共享的元数据借助 Web Service 对外提供服务, 请求者 (元数据消费者) 通过引用这种 Web Service, 获取元数据, 以实现元数据的集成和重用。

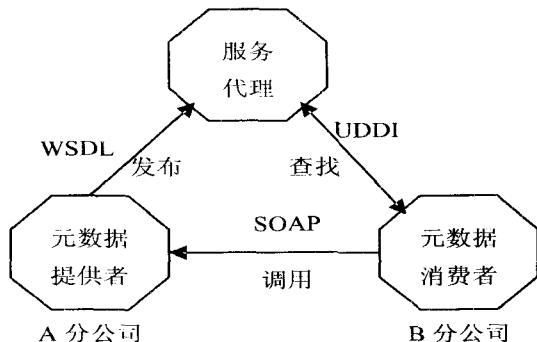


图 1 元数据集成系统框架图

对一个跨区域的集团型企业, 服务提供者可以是总公司、任何分公司, 服务请求者同样可以是总公司和任意分公司。三种基本操作流程是: 提供者将元数据集成服务发布到服务代理的一个目录上, 并基于 WSDL 描述这种服务; 当请求者需要调用该元数据集成服务时, 首先到服务代理提供的目录上搜索该服务, 取得如何调用该元数据集成服务的消息, 然后根据这些消息在 .NET 环境下构建 Web 应用或 Windows 应用去调用元数据集成服务, 实现元数据的集成和重用。

1.2 系统特点

基于 Web Service 技术的元数据集成系统具有以

下特点:

首先, 随着分布环境下数据仓库技术应用越来越广, 数据仓库元数据呈现分布范围广、管理层次多的特点, 这就要求所设计的元数据集成系统具有低耦合、易扩充、分布式管理的特点, 而 Web Service 技术的优势能够较好地解决这一问题。

其次, 采用 Web Service 技术的元数据集成系统可以实现真正意义上的异构环境下的元数据集成和重用。Web Service 建立在规范和标准的规格之上, 它允许各种异构平台下的客户端调用它提供的服务, 各种客户端都使用 SOAP 与 Web Service 进行信息交换, 而 SOAP 本身就是利用 XML 进行封装的, 并且可以利用任何的实体传输协议 (HTTP, TCP, SMTP) 来进行传输, 这就确保了在现有的 Web 平台上 Web Service 可以得到广泛支持, 并使得 Web Service 具有了很好地整合各类异构平台、构建分布式应用环境的能力, 这里的异构平台可以是不同的硬件平台、软件环境, 甚至不同的软件实现技术。而利用 CORBA 或者 DCOM 技术很难做到这一点。

再者, 采用 Web Service 技术具有更大的灵活性和通用性。基于 Web Service 的元数据集成系统已经不是采用传统意义上的单一的 C/S 或者 B/S 架构, 不管是 Web 浏览器、一般的 Windows 或者是 Java 应用甚至是手持移动设备都可以访问 Web Service 提供的服务, 这就保证了在同一软件系统架构下用户可以根据自己的需要选用不同的客户端访问 Web Service。

最后, Web Service 利用标准的 HTTP 和 SOAP 协议进行通信, 使得数据可以顺畅地通过企业间的防火墙, 解决了以往元数据集成方案无法克服的难点。

2 系统设计

基于 Web Service 技术的元数据集成系统设计分为元数据封装器设计、Web Service 方法设计、客户端代理对象的创建与 Web Service 方法调用三个部分。图 2 为数据仓库元数据集成系统结构图。其中, Web Service 方法设计主要包括数据源转化方法、数据抽取方法、数据合成方法和数据传递方法的设计。系统开发环境可选择 .NET, 开发语言可选取 C#。

2.1 元数据封装

系统采用 CWM 元数据模型实现元数据封装。每一个数据源通过一个 XML 接口作为底层数据源的包装, 以屏蔽底层的各类数据源差异, 对系统上层提供通用一致的标准数据格式。文献 [7] 较为详尽地讨论了基于 XML 的元数据封装, 文中在此基础上分别设计了不同数据源下用于元数据封装的 DTD 文档, 下面是文中所设计的 DTD 文档的格式 (部分)。使用这种格式

可对获取的元数据集进行封装。客户端在获得 SOAP 报文后,再经解析还原出元数据集,具体实现时采用文档对象模型(Document Object Model, DOM)完成。

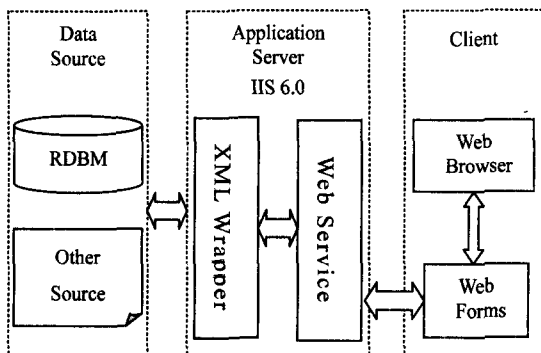


图2 基于 Web Service 技术的元数据集成系统结构

```
<? xml version="1.0"? >
<! ELEMENT Schema ( DimensionSpecification *
| CubeSpecification * )>
<! ATTLIST Schema
name CDATA #REQUIRED
datasource CDATA #REQUIRED
<! ELEMENT DimensionSpecification ( LevelSpecifi-
cation * )>
.....
<! ELEMENT LevelSpecification ( PropertySpecifi-
cation * )>
.....
<! ELEMENT PropertySpecification EMPTY>
.....
<! ELEMENT CubeSpecification ( OuterDimension-
Specification * | DimensionSpecification * | Measure-
Specification * )>
.....
<! ELEMENT OuterDimensionSpecification EMPTY>
>
.....
<! ELEMENT MeasureSpecification EMPTY>
<! ATTLIST MeasureSpecification
name CDATA #REQUIRED
column CDATA #REQUIRED !>
```

2.2 Web Service 方法设计

Web Service 调用元数据封装器,通过定义不同的 Web 方法,分别获取不同种类的元数据,如关系数据库中各类元数据、多维数据集中各类元数据。其中主要的 Web Service 方法有四个:

(1)数据源转化方法:系统采用关系模型作为公共模型,其他数据源则经过此方法先转化为关系模型表达的模式,再进行其他操作。

(2)数据抽取方法:用来连接数据源。通过该方法可以查看已连接数据源中所有表及表的结构,然后选择需要的表,将其数据传递到合成方法。下面代码为其基本框架。

[WebMethod]

```
public DataSet getMetaData ( )
```

```
{
```

//获取保存在 Web. Config 文件中的数据库连接信息;

//创建封装连接字符串连接对象;

//打开连接;

//获取关系数据库中元数据或借助 ADO MD. net 对象获取 CUBE 元数据;

//调用元数据封装器将元数据进行封装;

//由 DataSet 类实例化创建对象;

//填充数据集;

//异常处理;

//关闭连接;

//返回数据集;

```
}
```

(3)数据合成方法:用来实现多路数据的合成。一个数据合成方法可以连接多个数据源抽取数据,其输出可以通过数据传递方法传递到下游合成方法或目标库。该方法提供了多种合成方法,包括多表连接,字段名更改、过滤、分组、排序等操作,允许进行不同数据类型、不同值度量单位的变换,可以部分解决集成时语义层的异构问题。在最后一级合成中间表形成后,可以通过一对一映射操作,将其映射为需要的 XML 格式,解决结构层异构问题。

(4)数据传递方法:用来连接数据合成方法和目标库。它可以将合成方法的数据传递给下游合成方法或直接输出到目标库中。

2.3 代理对象创建与 Web Service 方法调用

Web Service 创建之后,可以通过 http 访问,此时通过 HTTP-GET 协议显示 WSDL 文件。WSDL 文件描述了创建的 Web Service 的服务名称、服务说明、在何处及如何调用。在 WSDL 文件中, <message> 与 <port-type> 用来说明提供服务内容, <binding> 说明如何调用, <service> 说明服务位置。

在客户端 Web Forms 实现 Web Service 代理对象的创建与 Web Service 方法的调用。根据 WSDL 文件提供的信息,在客户端首先创建 Web Service 的代理对象,然后调用 Web Service。下面的 02 语句显示了构建 Web Service 代理。在 02 语句中, Web Service 的名字空间为 localhost(默认值为 localhost),可根据具体情况进行设置改变。构建 Web Service 代理之后,通过 04

语句本地的 Web 应用程序和异地的应用程序可以使用 SOAP 协议进行通信,通过调用 Web Service 的方法实现将远程的元数据集成到本地。

```
01. //创建 Web Service 的 proxy 对象
02. ws35_2. Service1 ss = new DrugAdmin. ws35_2.
Service1();
03. //调用远程 Web Service 的方法
04. DataGrid1. DataSource = dataFromWebService.
getMetaData();
```

3 应用实例

将上述数据仓库元数据集成技术应用于消防工程领域的一个实际数据仓库建设项目中,取得了较好的效果:

(1) 该技术以 Web Service 技术为应用框架,以 CWM 为元数据模型,采用 XML 设计元数据封装器,实现了真正意义上的异地、异构元数据的集成与重用;

(2) 通过 Web 方法增加与修改,可以便捷地扩充系统对各类元数据的集成能力,满足了软件扩展性的要求;

(3) 与传统元数据集成方法相比,通过元数据包装器屏蔽了不同企业数据源的差异,为在不同数据源之上构建数据仓库奠定了基础。

4 结束语

文中在对目前元数据集成技术进行综合分析的基础上,提出了一种基于 SOA 架构的数据仓库元数据集成技术,并详细讨论了实现的关键技术。采用该技术

实现了分布环境下数据仓库元数据的集成与重用。

参考文献:

- [1] 聂 茹,张 虹. 数据仓库元数据管理模式的分析与比较[J]. 计算机应用研究,2005(2):57-58.
 - [2] 张英朝,邓 苏,张维明. 数据仓库元数据管理研究[J]. 计算机工程,2003,29(1):8-10.
 - [3] 谢福成,王备战,史 亮,等. 基于银行数据仓库的元数据管理系统[J]. 计算机工程,2009,35(9):79-81.
 - [4] Sen A. Metadata management: past, present and future[J]. Decision Support Systems,2004,37(1):151-173.
 - [5] 戴超凡,刘青宝,黄宏斌,等. 数据仓库中的元数据管理[J]. 计算机工程与科学,2003,25(4):54-57.
 - [6] 王裕明,吴 忠. 商务智能中元数据管理模型研究[J]. 计算机应用与软件,2005,22(8):34-35.
 - [7] 李姗姗,宁 洪,彭绍亮. 基于 CWM 的元数据管理系统中数据交换格式的研究[J]. 计算机工程与应用,2004(14):192-195.
 - [8] Thiru T, Greggl W, Jacquell G, et al. Metadata Management: the Foundation for Enterprise Information Integration[J]. Intel Technology Journal,2004,8(4):337-344.
 - [9] Tseng F S C, Chen Chia-Wei. Integrating heterogeneous data warehouses using XML technologies[J]. Journal of Information Science,2005,31(3):209-229.
 - [10] 周爱武,李孙长. 基于 Web 服务数据集成框架的设计与应用[J]. 计算机技术与发展,2010,20(11):246-248.
 - [11] 先晓兵. 基于 Web Service 的外贸业务智能决策系统[J]. 微机发展(现更名:计算机技术与发展),2005,15(4):51-53.
 - [12] 钟 路,张开松,江 琼. Web Service 技术在数据仓库系统中的应用[J]. 武汉理工大学学报,2004,26(8):74-76.
-
- (上接第 174 页)
- 参考文献:
- [1] 王 炜,包卫东,张茂军,等. 虚拟仿真系统导论[M]. 长沙:国防科技大学出版社,2007.
 - [2] 肖 甫,王汝传,孙力娟. 多关联性虚拟现实系统的设计与实现[J]. 计算机技术与发展,2009,19(12):36-39.
 - [3] 赵 伟,穆志纯,张 玉. 基于虚拟现实的轧钢实时仿真训练系统[J]. 系统仿真学报,2006,18(4):909-912.
 - [4] 郑 坚,关正西. 虚拟现实系统中的视景仿真研究[J]. 微机发展(现更名:计算机技术与发展),2003,13(9):18-21.
 - [5] 赵金洲,张桂林. 钻井工程技术手册[M]. 北京:中国石化出版社,2005.
 - [6] 袁恩熙. 工程流体力学[M]. 北京:石油工业出版社,2006.
 - [7] 刘明昆. 三维游戏设计师宝典—Virtools 行为模块词典大全[M]. 成都:四川电子音像出版社,2005.
 - [8] 张雪鹏,陈国华,戴莺莺. 基于 3D 的虚拟运动仿真平台设计及 Virtools 功能实现[J]. 北京化工大学学报(自然科学版),2009,36(4):93-95.
 - [9] Teng Yingyan, Zheng Junsheng, Gao Zhijun. Design and Implementation of Interactive 3D Scenes Based on Virtools [C]//2009 International Forum on Computer Science-Technology and Applications. Dalian, China: [s. n.], 2009:87-89.
 - [10] Li Xunxiang, Li Anding, Chen Dingfang. Research on Distributed Multi-Screen Display Technique Based on Virtools [C]//Proceedings of 7th International Conference on Computer-Aided Industrial Design and Conceptual Design. [s. l.]:IEEE Press,2006:756-761.
 - [11] Li Jizu, Zhang Shaohong. Application of Virtual Reality Technologies to the Simulation of Coal Miners' Safety Behaviors [C]//Open-source Software for Scientific Computation(OS-SC). Guiyang, China: [s. n.], 2009:60-62.
 - [12] 王 瑶. 3ds Max 2008 完全手册—建模篇[M]. 北京:科学出版社,2008.