网络科技资源中异构数据库访问技术的研究①

Research on Accessibility of Heterogeneous Database in Technological Resources of Network

张 岩 周明全 焦翠花 (北京师范大学 信息科学与技术学院 北京 100875)

摘 要: 为了解决网络科技资源信息共享问题,提出了一种采用 Java 数据库连接和可扩展标记语言技术解决异构数据库访问问题的方法。针对面临的异构数据库问题,首先介绍了网络科技资源的主要内容,接着介绍了如何识别异构数据库,包括异构数据库的异构表现、异构数据库的连接配置、元数据提取、数据库的信息描述等,然后介绍了异构数据库通用的问题解决策略,最后给出了网络科技资源应用环境建设项目中采取的具体实现方法,包括系统的体系结构、查询描述、结果处理等。结果证明,该方法可行。

关键词: 网络科技资源 异构数据库 JDBC XML 元数据

网络科技资源中异构数据库的访问技术包括异构数据库的识别,XML 语言及其相关技术,异构数据库访问的解决策略以及本文中给出的具体实现方法。

1 引言

1.1 文章安排

本文第 2 节介绍网络科技资源的主要内容。第 3 节介绍了异构数据库的识别及 XML 语言及其相关技术。第 4 节介绍了通用的问题解决策略。第 5 节介绍了本文所采取的具体实现方法。第 6 节给出结论以及展望。

1.2 基本介绍

随着信息化建设工作的推进,现在存在着大量的网络科技资源数据库和应用系统,由于管理体制等原因造成数据库和系统结构迥异,各数据库和应用系统彼此孤立,相互之间难以实现资源共享和业务信息传递,致使各个应用系统只能是孤岛式的运行,无法实现资源间、系统间的资源共享和信息联动。因此人们迫切需要解决这一问题,即要以最小的代价,使各种异构的数据库之间实现数据的互访及共享,并对用户实现数据的透明访问。

2 网络科技资源内容

网络科技资源应用集成环境建设项目中,科技资源主要指科技信息资源,包括通过科技活动或其它方式获取到的反映客观世界的本质、特征、变化规律等的原始基本数据,根据不同科技活动需要进行系统加工整理后得到的各类数据集,以及其它用于支撑科技活动的数据集。主要有五个方面数据内容:平台资源、领域资源、行业资源、科技业务资源、网络中分散的科技资源。

3 异构数据库识别

3.1 异构数据库的定义

异构数据库系统是相关的多个数据库系统的集合,可以实现数据的共享和透明访问,每个数据库系统在加入异构数据库系统之前本身就已经存在,拥有自己的 DBMS(Database Management System)。异构数据库的各个组成部分具有自身的自治性,实现数据共享的同时,每个数据库系统仍保有自己的应用特性、完整性控制和安全性控制。

其异构性主要体现在以下几个层次:

① 基金项目:国家科技基础条件平台建设项目(2005DKA63904)

(1)计算机体系结构的异构:数据库运行的计算机体系结构的不同,如运行在大型机、小型机、工作站、PC 或嵌入式系统中。

(2)基础操作系统的异构:数据库系统的基础操作系统不同,可以是 Unix、Windows Server、Linux 等。

(3)模式异构:即数据源在存储模式上的不同。存储模式主要包括关系模式、对象模式、对象关系模式和文档嵌套模式等几种,其中关系模式(关系数据库)为主流存储模式。同时,即便是同一类存储模式,它们的模式结构可能也存在着差异。例如不同的关系数据管理系统的数据类型等方面并不是完全一致的,如 DB2、Oracle、Sybase、SQL 等。

(4)数据的异构:定义数据库的数据结构和语义的不同。

3.2 XML 语言

XML[3-5](Extensible Markup Language,可扩展标记语言)是由 W3C(World Wide Web Consortium)组织于1998 年 2 月制定的一种通用语言规范。XML 在支持异构数据库系统方面有很多自身的优点。

(1)它的结构性强、语义性强,既能与传统关系数据库进行转换,也能与对象数据库以及多媒体数据库进行转换;

(2)它交互性好、易于处理,能方便地控制显示和浏览各种数据;

(3)它与平台无关,能在各种平台下进行处理,能用各种编程语言进行处理。所以在我们的设计中,元数据的结构是由 XML Schema 描述的,并且用 XML 在整个系统的模块间交换数据。在系统运行和数据收集期间,也用 XML 在收集数据者之间交换数据;

(4)我们还使用 XML 来解决数据库连接配置、元数据提取、数据库信息描述等技术问题。

3.3 连接配置

为用户提供进行数据源连接配置的功能,用户在此进行数据源的连接配置,所有的配置信息将以规范的格式被保存到 XML 文件中。连接配置信息为后续的统一访问提供服务。数据源的连接配置规范流程如图1所示。

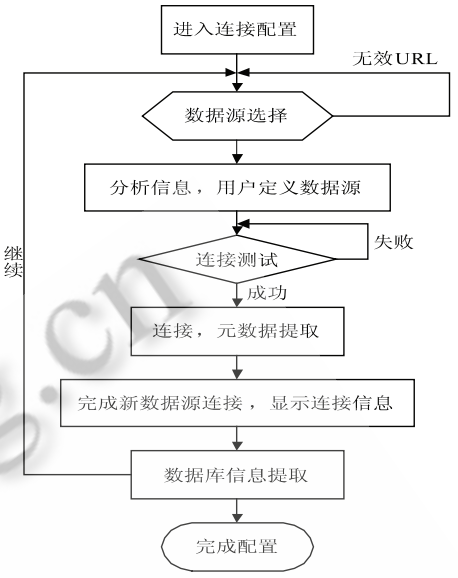

图1　数据源连接配置流程

XML 配置文件以 SQL Server 为例,DatabaseType 代表用户使用的数据库类型;DatabaseName 代表用户使用的数据库实例名称;DB_Driver 对应用户所使用的关系型数据库驱动;DB_URL 代表关系数据库的连接地址;DB_UserName 和 DB_Password 分别代表用户数据库的用户名和密码。具体格式如下所示:

```
<? xml version = "1.0" encoding = "UTF-8" ? >
<Configuration >
    <DatabaseType > MSSQL2000 </DatabaseType >
    <DatabaseName > RISTDataCenter
    </DatabaseName >
    <DB_Driver >
    com. microsoft. jdbc. sqlserver. SQLServerDriver
    </DB_Driver >
    <DB_URL >
    jdbc:microsoft:sqlserver://127.0.0.1:1433
    </DB_URL >
    <DB_UserName > SA </DB_UserName >
    <DB_Password > SA </DB_Password >
</Configuration >
```

3.4 元数据提取

元数据是描述一个具体的资源对象,并能对这个对象进行定位管理,且有助于它的发现与获取的数据,它是定义和描述其他数据的数据,用于说明数据内容、

质量、状况和其他特性的信息。

为保证整个异构数据库统一访问流程衔接良好，数据源连接配置过程中提取的元数据信息存储在相应的 XML 配置文件中，存储时必须遵循如下规则：

(1) 以 <metadata> 标示元数据信息。

(2) 以 <Title> 标示元数据标题。

(3) 以 <Creator> 标示元数据的创建者。

(4) 以 <Subject> 标示元数据主题。

(5) 以 <Description> 标示元数据的描述信息。

(6) 以 <Publisher> 标示元数据的发布者。

(7) 以 <Date> 标示元数据的创建时间。

(8) 以 <Source> 标示元数据的来源。

(9) 以 <Language> 标示元数据的语言类型描述。

示例如下：

```
<metadata>
    <Title> 标题 </Title>
    <Creator> 创始人 </Creator>
    <Subject> 主题 </Subject>
    <Description> 描述 </Description>
    <Publisher> 发布者 </Publisher>
    <Date> × × × × - × × - × × </Date>
    <Source> 来源 </Source>
    <Language> 中文 </Language>
</metadata>
```

3.5 数据库信息描述

连接配置过程中提取的数据库信息，应包括源数据库的表信息和字段信息。每一个数据库都对应一个 XML 配置文件，存储时遵循如下规则：

(1) <db> 为根元素用于标示 XML 文档的唯一性。

(2) 以 <table> 元素表示数据表，其属性 name 用于标示数据表的表名。

(3) 以 <word> 元素标示数据表的字段名称。

示例如下：

```
<db>
    <table name = " tableName1" >
        <word > tableWord_1.1 </word >
    ......
        <word > tableWord_1. n </word >
    </table >
    ......
    <table name = " tableName_n" >
        <word > tableWord_n.1 </word >
    ......
        <word > tableWord_n. n </word >
    </table >
</db >
```

4 问题解决策略

网络环境下对于 RDMS 系统的异构数据库，目前主要有 3 种解决方法[6,7]，它们都基于客户/服务器体系结构：数据库网关(Database Gateway)、公共协议(Common Protocol)、公共编程接口(Common Programming Interface)。

公共数据库网关(Gateway)是一个转换器，客户通过它可以访问异构数据库。Oracle、Sybase 等大型数据库厂商都有自己的网关产品。

采用公共协议指对客户和服务器间通信的格式和协议以及数据库语言进行标准化，这是一种最理想的解决异种数据库系统互联的方法，目前比较典型的有 SAG(Sql Access Group)规范和 IBM 的分布式关系数据库体系结构(DRDA)等。

公共编程接口可以采用 JDBC[8] 方式。JDBC(Java DataBase Connectivity)是 Java 与数据库的接口规范，JDBC 定义了一个支持标准 SQL 功能的通用低层的应用程序编程接口(API)，它由 Java 语言编写的类和接口组成，旨在让各数据库开发商为 Java 程序员提供标准的数据库 API。JDBC API 定义了若干 Java 中的类，表示数据库连接、SQL 指令、结果集、数据库元数据等。它允许 Java 程序员发送 SQL 指令并处理结果。通过驱动程序管理器，JDBC API 可利用不同的驱动程序连接不同的数据库系统。

5 实现方法

实现方法主要包括系统的体系结构、查询描述和结果处理等。

5.1 体系结构

根据上述思想，网络科技资源应用集成环境建设项目中采用公共编程接口方法来实现异构数据库的统一访问，其体系结构如图 2 所示。

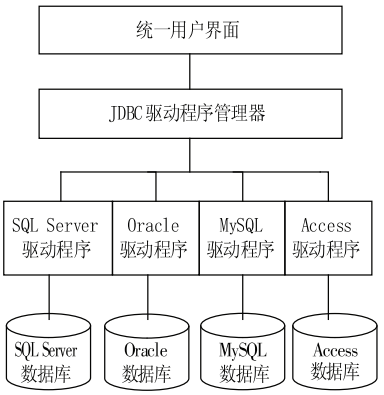

图 2　访问异构数据库的体系结构

通过统一用户界面配置连接信息，并将连接信息保存为 XML 文件，再根据配置信息通过 JDBC 驱动程序管理器选择相应的驱动程序连接数据库。根据元数据提取、数据库信息描述和查询描述等信息将访问结果写成 XML 文件，最后进行结果处理。

应注意的是，各节点上的数据库并不一定都要纳入统一访问，被统一访问的信息可能只是其中的几个数据库，几个表，甚至是表中的某几个字段。

5.2 查询描述

为了实现异构数据库的统一访问，必须实现 RDMS→XML 的转换[9-11]。通过对 XML 的数据模型与关系模型的特征的比较，笔者认为两者转换的实质是：(1) 从关系模式中字段到 XML 中的数据的映射；(2) 从关系模式中元组与元组的关系及元组与字段的关系到 XML 中数据的相互位置关系的映射。关系模式与 XML 模式的转换的核心就是按 XPDL(XML Processing Description Language) 编写的转换规则和转换规则的执行解释方法，从关系模式到 XML 模式的转换规则的组成部分和执行方法。从关系模式到 XML 模式转换的转换规则脚本由关系模式的定义和关系模式到 XML 模式转换规则的定义组成。

从关系模式到 XML 模式的转换，实质上是一个将一个预先定义好的视图运行，并将当前视图内容转化成 XML 格式显示。在具体实现上，我们是将格式和视图定义融合在一起。在关系模式到 XML 模式的转换中，关系模式定义段中的 entity 可以对应一个 SQL 查询语句，也就是使用 SQL 语句代替原先在 id 中指明的表名，所有的数据准备基本在关系模式定义段已经完成。

5.3 结果处理

事实上，我们已经为异构数据库建立好了数据库信息描述。这一部分是在系统建立的前期工作中完成的(当各节点上的库表结构发生变化时，在数据库信息描述中要及时反映出来)。因此，从数据库节点上返回的数据无需再作改动，直接存储在对应的 XML 文档中，并将已有的 Schema 文档作为类型定义文档，其结果就是统一的 XML 文档格式的结果。如果同时访问多个数据库或者一个数据库中的多个表，访问结果是独立的，这种情况下必须将结果集合并，生成单一 XML 文档，最后通过解析显示在 web 页面上。

格式化后的 XML 文件如下所示：

```
<? xml version = "1.0"  encoding = " gb2312" ? >
<DataList >
    <topicNode >
        <title > 用抗洪的精神抗击雪灾 - - 安徽抗雪防冻救灾工作纪实 - - 时政 - - 人民网 </title >
        <description > 的精神抗击雪灾安徽抗雪防冻救灾工作纪实 - - 时 </description >
        <link > http://politics. people. com. cn/GB/14562/6871593. html </link >
        <time > 20080120 </time >
    </topicNode >
    <topicNode >
    ......
    </topicNode >
</DataList >
```

标签 <DataList > 表示所有资源的集合，而一个标签 <topicNode > 就代表一个网络科技资源，标签 <title >、<description >、<link >、<time > 分别代表该资源的标题、资源描述、资源连接 URL 和资源发布的时间等。

6　结束语

未来网络科技资源数据库和应用系统的发展将始终保持跨系统、跨平台的局面，能够提供一个独立于特定数据库管理系统的统一编程界面和一个基于 SQL 的通用数据库访问方法，将是未来实现异构数据库访问的目标，因此，JDBC 和 XML 必将在 Internet 上的异构数据库访问中发挥重要的作用。

(下转第 79 页)

(上接第 90 页)

参考文献

1 黄镝. 异构数据库的跨库检索技术综述. 图书情报工作,2003(6).

2 冯琪,吕汉桥,冯虹. 异构数据库的连接. 电脑与信息技术,2001(5).

3 朱跃龙,杨扬,黄玮. 基于 XML 的异构数据库间联合使用. 计算机工程与设计,2003.

4 Shyh – Kwei Chen , Ming – Ling Lo, Kun – Lung Wu, Jih – Shyr Yih and Colleen Viehrig. A practical approach to extracting DTD – conforming XMLdocuments from heterogeneous data sources. Information Sciences, 2006.

5 Collins,S R. S Navathe and Leo Mark. XML schema mappings for heterogeneous database access. Information and Software Technology, 2002.

6 张少中,王秀坤,张志勇. 基于 JDBC 的异构分布式数据库访问. 计算机工程,2002.

7 唐巍,周俊林,李晓. 异构数据库集成方法初探,计算机应用研究,1999.

8 马德云,俞时权,胡浩民. 异构数据库的集成. 计算机工程,2002.

9 卿秀华. 基于 XML 的异构数据库数据交换. 武汉科技学院学报,2003, 16(5).

10 刘晓莹. 基于XML的异构数据库查询技术研究, 中国电力教育,2007.

11 邹晓静,刘伟,卢贤玲. 基于 XML 的异构数据库系统的研究. 微计算机信息,2007.