文章编号:1008-0570(2010)09-3-0091-03

基于 XML 的关系型数据库格式转换研究

Study on Relational Database Format Conversion Based on XML

(1.福州大学;2.厦门理工学院) 涂　平[1,2]　朱晓铃[2]　满　旺[2]
TU Ping　ZHU Xiao-ling　MAN Wang

摘要: 针对现有电子政务中异构关系型数据库带来的信息共享和数据交换的难题,采用 XML 技术,实现了关系型数据库——>统一标识的 XML 文档——>关系型数据库之间的转换,即将分布在不同地点、不同存储方式的关系型数据库发布出来,并将发布的共享数据转换为统一标识的 XML 共享数据,方便共享数据的交换和使用。设计了关系型数据库转换模型的 XML 文档标准,实现了关系型数据发布、共享与下载导入。

关键词: 电子政务; 信息资源交换; 关系型数据库; 可扩展标记语言
中图分类号: TP311.13　　**文献标识码:** A

Abstract: Aiming at existing e-government information sharing and data exchange problems caused by heterogeneous relational database, conversion between relational database and unified identifier XML document are realized. Relational database in different storage forms and locations are published and transformed into unified identifier XML document to facilitate the sharing of data exchange and use. A relational database model of XML document conversion standards is designed, and the publishing sharing and importing are implemented.

Key words: E-government; Information interchanging; Relational database;Extensible markup language

随着电子政务建设快速推进, 政府部门的业务系统应用成效显著,但是仍然存在一些问题,比如各部门之间相同数据难相容、跨部门业务流程难协同、应用平台重复建设难互通、数据重复采集利用效率低等问题, 尤其是信息资源开发利用远远滞后于电子政务应用发展的需要,信息资源分散建设,政务部门间信息交换和共享困难。

为解决以上问题,必须打破政务部门之间的信息阻隔,构造一个能够根据需要快速整合和共享信息资源的体系。现有的各级政府的电子政务信息系统绝大多数都是基于关系型数据库的,因此研究关系型数据库格式转换,建设政务信息资源交换体系是实现跨部门政务信息资源共享与业务协同的一项重要举措。本文的研究目的是基于 XML 技术,实现关系型数据库——>统一标识的 XML 文档——>关系型数据库之间的转换, 即将分布在不同地点、不同存储方式的关系型数据库发布出来,并将发布的共享数据转换为统一标识的 XML 共享数据,方便共享数据的交换和使用。同时通过数据下载导入工具实现共享数据的获取,并将统一标识的 XML 共享数据导入到关系型数据库。本文第一部分简单介绍了 XML 技术,第二部分介绍了数据转换的关键技术,第三部分介绍了关系数据库转换系统的实现。

1 XML 技术

XML(Extensible Markup Language),即可扩展标记语言,它具有开放性、简单性、自我描述性、互操作性、可扩展性及跨平台、跨数据库等特性, 特别适合于网络环境下异构系统之间的信息共享和交换。

XML 的好处是数据的可交换性, 同时在数据应用方面还具有如下优点:(1)XML 文件为纯文本文件, 不受操作系统、软件平台的限制;(2)XML 具有基于 Schema 自描述语义的功能,容易描述数据的语义, 这种描述能为计算机理解和自动处理;(3)XML 可以描述结构化数据,还可有效描述半结构化,甚至非结构化数据。

XML 与关系数据库之间的转换也是当前研究的热点。XML 由于其"网络普通话"的优势,成为关系型数据库网络共享和发布的桥梁。关系型数据库转换模型使用 XML 技术作为技术支撑,具体思路是将异构、多源、跨平台的关系型数据库转换为统一标准的 XML 文档,再将该文档转换为关系型数据库,从而完成关系型数据库的网络发布与共享。

2 关键技术

2.1 关系型数据库的发布

关系型数据库的发布过程即实现各种关系型部门共享数据库的 XML 格式化过程。关系型数据发布研究的内容包括如何从 oracle、sqlserver 等关系型数据库中按一定的业务规则抽取数据, 并将抽取出来的数据转换成统一标识的 XML 文档;如何调用 XML 数据备份;如何在提供数据服务时能够直接提取以实现高效传输;如何方便用户查询发布的数据库信息。

2.2 关系型数据下载导入

主要包括如何获取各个部门提供的共享数据 (即一系列 XML 文档);如何解析 XML 文档内容;如何将 XML 文档内容导入到指定的关系型数据库等。

3 关系型数据库转换系统设计与实现

在.NET 平台下, 建立了数据发布系统和数据下载导入系

涂　平: 副研究员 硕士

基金项目: 基金申请人: 吴升、涂平; 项目名称: 数字区域信息化技术服务体系关键技术研究、开发与应用; 颁发部门: 国家科技支撑计划项目资助(2007BAH16B02)

统。制定了关系型数据库同统一标识 XML 文档的交换标准。数据发布系统包括数据发布程序和数据发布服务；数据下载导入系统包括数据下载导入程序和数据下载导入服务。

3.1 数据发布程序

数据发布程序提供用户操作界面,并根据政务网项目开发规范接入单点登录系统，用户登录数据发布程序后可以进行关系型数据发布、目录结构的注册等操作。同时为关系型数据提供多种接口类型的发布，每个关系型数据项都可以以数据发布接口、数据查询接口、数据验证接口三种接口中的一种进行发布。

3.2 数据发布服务

数据发布服务是一个 Windows 服务，是为数据发布程序提供的后台服务，它根据数据发布程序生成的配置文档执行相关的后台操作。数据发布程序生成的配置文档放到附件文件 WebServicesMap.Config 中。

3.3 关系型数据库 XML 交换标准

关系型数据库数据 XML 交换标准主要适用于关系型数据库的数据交换,是关系型数据库转换的基石。所有电子政务关系型数据库都按照此标准进行发布、共享、下载导入。关系型数据库数据交换标准包含 5 个部分的 XML 文档。假设数据交换文档的主文件名为 datafile,则 5 个文档分别是:

(1)数据文档主文件(datafile.xml):XML 主架构文档,用于将以下 4 个部分的 XML 文档组织成一个完成的 XML 数据文件;

(2)数据描述文档(datafileDescription.xml):用于描述数据文档中包含交换数据的记录数、字段数、版本号、数据更新时间等信息;

(3)数据库基本信息文档(datafileInformation.xml)：用于描述共享数据库的名称、存储介质类型、发布单位等信息;

(4)数据字典文档(datafileDictioanry.xml):用于描述数据文档中包含交换数据的数据字典信息;

(5)数据片段文档(datafileSegment{i}.xml,{i} >=0):数据存储分片文档,用于记录各个分片的交换数据信息。

3.3.1 数据文档主文件(datafile.xml)

数据文档主文件通过将数据描述文档、数据库基本信息文档、数据字典文档和数据片断文档等文档以内部实体的方法进行组织为完整 XML 数据文档。数据文档主文件的文件结构如下所示:

```
<?xml version="1.0" encoding="utf-8" ?>
<! --文档类型声明-->
<! DOCTYPE database[
<! --数据库基本信息文档 -->
<! ENTITY Preface0 SYSTEM "datafile_information.xml">
<! --数据字典文档 -->
<! ENTITY Preface1 SYSTEM "datafile_dictionary.xml">
<! --数据片段文档,可以包含多个 -->
<! ENTITY Segment0 SYSTEM "datafile_segment0.xml">
]>
<database>
<! --实体引用-->
&Preface0;
&Preface1;
<tables>
<table>
&Segment0;
</table>
</tables>
</database>
```

3.3.2 数据描述文档(datafileDescription.xml)

数据描述文档是对交换数据文档本身的描述，它记录了交换数据的记录数、字段数、关键字段数、数据版本号、数据片段数、每个数据片段包含的记录数、数据发布时间、数据更新时间、数据文档包含的片段文档信息,以及记录在每个数据片段的分布情况等等。数据描述文档的文件结构如下所示:

```
<?xml version="1.0" encoding="utf-8"?>
<description>
<totalCount></totalCount><! --总记录数-->
<fieldCount></fieldCount> <! --总字段数-->
<segmentCount></segmentCount><! --数据片段数-->
<countPerSegment></countPerSegment><! --每个数据片断的记录数-->
<fileSignature></fileSignature><! --数据版本号-->
<sqlMD5></sqlMD5> <! --生成该数据的 SQL 语句的 MD5 哈希值-->
<createTime></createTime><! --数据发布时间-->
<modifyTime></modifyTime><! --数据更新时间-->
<supplement></supplement><! --数据补充说明信息-->
<keys> <! --关键字段列表 多个关键字表示组合关键字-->
<! --关键字段信息,name 为字段名称,type 为字段类型-->
<key name="" type="" /> </keys>
<sort order="ASC"><! --排序字段列表,order 指定升 (降)序-->
<! --排序字段信息,可以有多个排序字段, name 为字段名称,type 为字段类型-->
<field name= "" type= "" />
</sort>
<list><! —片段文档列表-->
<prefaces>
<preface id="0" file="datafile_information.xml">数据的基本信息</preface>
<preface id="1" file="datafile_dictionary.xml">数据字典表</preface>
</prefaces>
</segment>
</segments>
</list>
</description>
```

3.3.3 数据基本信息文档(datafileInformation.xml)

数据基本信息文档，描述数据的基本信息和数据发布机构信息。数据基本信息文档的文件结构如下所示:

```
<?xml version="1.0" encoding="utf-8" ?>
<basicInformation>
<DBFullName>数据库/文件全称</DBFullName>
<DBShortName>数据库/文件简称</DBShortName>
```

<memoryFormat>存储格式</memoryFormat>
<technicalParameter>主要技术参数</technicalParameter>
<contentNotes>内容说明</contentNotes>
<usage>使用方法简介</usage>
<supplement>数据库/文件补充说明</supplement>
<institution>
<name>机构全称</name>
<shortName>机构简称</shortName>
<address>单位通讯地址</address>
<contact>联系人</contact>
<telephone>联系电话</telephone>
<email>电子邮件地址</email>
<postcode>邮政编码</postcode>
<supplement>机构补充说明</supplement>
</institution>
</basicInformation>

3.3.4 数据字典文档(datafileDictioanry.xml)

数据字典文档,用来描述数据的包含的字段描述信息。数据字典文档的文件结构如下所示:

```
<?xml version="1.0" encoding="utf-8" ?>
<dictionaryTables>
<! -- 数据表,id 为表序号,name 为数据项名称,可以包含多个表,目前只有单表情况 -->
<table id="0" name="管理员列表">
<! --字段信息,id 为字段序号,可以包含多个字段-->
<column id="0">
<name></name><! --字段名称-->
<type></type><! --字段类型-->
<length></length><! --字段长度-->
<unit></unit><! --字段值对应的计量单位,比如身高字段,计量单位为厘米(cm)-->
<notes></notes><! --字段说明-->
<isKeyField></isKeyField><! --是否为关键字段{true|false} -->
<memoryMode ></memoryMode ><! —字段的存储方式{base64|accessory}-->
</column>
</table>
</dictionaryTables>
```

其中,字段存储方式中,base64 表示字段信息以 base64 编码的字符串表示存储在标准数据文档中,accessory 表示以附件形式存储,并将对应的相对路径存储在标准数据文档中,如果该不存在该节,即未指定存储格式的,则以原始内容存储。

3.3.5 数据片段文档(datafileSegment{i}.xml,{i} >=0)

数据片段文档,用于记录交换数据。对于交换数据的数据量较大的情况, 将交换数据分成多个数据片段分别保存为数据片段文档,以解决单个数据文档过大的情况。数据片段文档的文件结构如下所示:

```
<?xml version="1.0" encoding="utf-8"?>
<! --数据片段,id 为片段序号-->
<segment id="0">
 <row id="0"><! --数据记录行,id 为记录序号-->
  <col [i] format="accessory" filename="filename" >当前字段值</col[i]>
 </row>
</segment>
```

3.4 数据下载导入程序

数据下载导入程序提供用户操作页面, 用户通过该程序从另一个部门提供的有访问权限的前置系统中下载关系型数据, 还可以将下载的关系型数据导入到指定的关系型数据库。下载导入程序界面如图 1 所示。

3.5 数据下载导入服务

数据下载导入服务是一个 Windows 服务, 是为数据下载导入程序提供的后台服务, 它根据数据下载导入程序生成的配置文档执行相关的后台操作。数据下载导入服务生成的配置文档有详细的导入方案 XML 文档规范来规定。

图 1 数据下载与导入界面

4 结论

本文通过具有统一标识符的 XML 标准文档做为中介,使用上传发布程序与服务、下载导入程序与服务,完成了分布在各地的异构关系型数据库的数据交换与信息共享。本文提出的关系型数据库转换模型及其提供的数据上载发布、下载导入程序在区域电子政务信息资源交换应用示范中得到应用,效果良好。这对于加快政府职能转变,提高行政质量和效率,增强政府协同监管能力,提升政府公共服务水平,将会起到重要的促进作用。

创新点及社会效益: 本文设计了关系数据库信息共享的 XML 文档标准,设计实现了关系数据库发布、下载、导入模块。并在国家科技支撑计划项目"数字区域信息化技术服务体系关键技术研究、开发与应用"进行示范应用,取得了良好效果。

参考文献

[1]穆昕, 王浣尘, 王晓华.电子政务信息共享问题研究[J].中国管理科学,2004,12(3):121-124
[2]王志平.基于 XML 的异质多数据库集成系统的设计与实现[J].河南大学学报(自然科学版),2007,37(5):530-532
[3]曹 亮, 王 茜, 卢 菁. XML 数据在关系数据库中存储和检索的研究和实现[J]. 东南大学学报(自然科学版), 2002,32(1):124-127
[4]姜莉莉,彭和平等.XML 和关系数据库在 IETM 开发中的应用[J].微计算机信息.2009,0-3:28-30

作者简介:涂平(1973-),男,汉族,福州大学,副研究员,硕士,主要研究方向:信息网络共享技术、软件工程和地理信息系统等应用开发; 朱晓铃(1979-),女,硕士;满旺(1979-),男,讲师,博士。

Biography:TU Ping (1973-), male, Han nationality, Fuzhou University, Associate Research Fellow, Master, mainly research: information sharing on Network, software engineering; ZHU Xiao-ling(1979-),female, Instructor, Master.

(350002 福建省福州市 福州大学空间信息工程研究中心) 涂 平

(下转第 154 页)

技术创新

4.集合的交集、差集运算的标准 SQL 语言的等价表示

标准 SQL 语言中没有提供集合交操作和差操作,但可用 select 语句通过其它方法间接实现。例如:查询信息系学生集合与年龄小于 20 岁的学生集合的交集,可转化为查询满足条件为"信息系且年龄小于 20 岁"的学生,即求交集运算转化成了复合条件单表查询运算。同样,查询信息系学生集合与年龄小于 20 岁的学生集合的并集,可转化为查询满足条件为"信息系或者年龄小于 20 岁"的学生,即求并集运算转化成了复合条件单表查询运算。

5 从关系代数语言到 SQL 语言上的等价关系

关系代数中的选择、连接、投影、除、并、交、差、广义笛卡尔积这 8 种运算实现的是查询操作,它们在 SQL 语言中都可用 SELECT 语句来表达。选择、连接、投影、并、广义笛卡尔积这 5 种关系代数运算与 SQL—SELECT 语句间的等价转换情况如表 2 所示。

标准 SQL 语言中没提供交集和差集操作,对于下面求交集的关系代数运算表达式:$\sigma_{F1}(R)\cap\sigma_{F2}(R)$

可用语句:Select * from R where F1 and F2 ;

实现。

而求差的关系代数运算表达式: $\sigma_{F1}(R)\cup\sigma_{F2}(S)$

可用语句:Select * from R where F1 and (not F2) ;

实现。

可见,若两个查询的源表是同一个关系,则这两个查询的差集与交集运算可分别转化成不含集合运算的 SQL—SELECT 语句来实现。

表 2 部分关系代数运算与 SQL—SELECT 语句间对应的等价关系

等价类别	关系代数表达式	Select 语句
选择等价类	$\sigma_F(S)$	Select * from S where F;
连接等价类	R ⋈S A θB	Select R.*,S.* from R,S where A θ B;
投影等价类	$\pi_{Y1,Y2,Y3,\ldots,Yn}(R)$	Select Y1,Y2,Y3,…,Yn from R;
广义笛卡尔积等价类	R ×S	Select R.*,S.* from R , S;
并等价类	R ∪S	Select * from R UNION Select * from S ;

关系代数中的除运算可以用数理逻辑中的全称谓词(∀)和逻辑蕴含(->)等价表示。按照除运算的定义,关系代数式 R(X,Y)÷S(Y)用数理逻辑知识等价表示为 $\{x|\ \forall y(S(y)\rightarrow R(x,y))\}$

其中 X,Y 表示属性组,x、y 分别表示相应关系中的元组在 X,Y 属性组上的分量。

例如,查询选修了全部课程的同学学号,其关系代数式是:

$$\prod_{Sno,Cno}(SC)\div(\prod_{Cno}(Course))$$

//该除运算等价的数理逻辑表达式是{x|∀c(c 是学校开出的课程→x 选修了 c)}。而

$\forall y(S(y)\rightarrow R(x,y))\Leftrightarrow\neg\exists y(\neg(S(y)\rightarrow R(x,y)))\Leftrightarrow\neg\exists y(\neg(\neg S(y)\vee R(x,y)))\Leftrightarrow\neg\exists y(S(y)\wedge\neg R(x,y))$

因此,关系代数除运算可用含有 EXISTS/NOT EXISTS 谓词的 select 语句等价表达。

对于上面"查询选修了全部课程的同学学号"的例子,其 select 语句如下:

Select sname from student where not exists (select * from course where not exists(select * from sc where student.sno=sc.sno and course.cno=sc.cno))

按照数理逻辑知识,有 $(\forall x)(p)\Leftrightarrow\neg\exists x(\neg p)$,因此全称谓词()运算都可用带有 EXISTS/NOT EXISTS 谓词的查询实现。

由此可见,利用关系代数运算与 SQL—SELECT 语句间存在的等价关系,可以很容易地用 SQL—Select 语句解决诸如"除"运算这样的较复杂、较难的关系操作问题。

6 结束语

本文结合教学实践总结了关系代数语言和 SQL 语言中的按照实现的功能相同的定义所存在的等价关系。利用等价关系,可实现等价类内的不同表达形式之间的等价转换。按照等价关系这一主线,可以将关系代数运算与关系操作有机联系起来;此外,利用关系代数运算与 SQL—Select 语句的对应等价关系,还可用关系代数运算的数学理论知识指导应用 SQL 语言进行数据的处理,从而提高运用 SQL 语言实现数据操作的理论指导性、预见性。

本文作者创新点:利用离散数学中等价关系的理论,分析了关系代数和 SQL 语言上存在的基于语义相同关系的等价关系、给出了等价类,为以关系代数知识指导 SQL 语言的运用提供了理论依据。

参考文献

[1]邓辉文.离散数学[M] .北京:清华大学出版社,2006 年 10 月:59-62

[2]李京辉,邵温,许向众.数据库语义和词性双标注模型研究[J].微计算机信息,2009,7-3:127-128

[3]王珊,萨师煊.数据库系统概论(第四版) [M].高等教育出版社.2007 年 11 月:60,109,110

作者简介:陈龙猛(1968-),男(汉族),山东莒县人,青岛农业大学理学与信息科学学院,副教授,工学硕士。主要从事软件工程,知识工程,数据库等方面的研究。

Biography:CHEN Long-meng (1968-),Male (Han Nationality), Shandong Province,Qingdao Agricultural University, Associate Professor,Mastor. Research field: Software Engineering, Knowledge Engineering,Database etc.

(266109 山东 青岛市城阳区青岛农业大学理学与信息科学学院) 陈龙猛

(College of Science and Information Science of Qingdao Agricultural University, Qingdao, Shandong province, 266109, China) CHEN Long-meng

通讯地址:(266109 山东 青岛市城阳区青岛农业大学理学与信息科学学院) 陈龙猛

(收稿日期:2009.10.15)(修稿日期:2010.01.15)

(上接第 93 页)

(361024 福建省厦门市 厦门理工学院空间信息科学与工程系) 涂 平 朱晓铃 满 旺

(Spatial Information Engineering Research Center, Fuzhou University, Fuzhou Fujian 350002, China) TU Ping

(Department of Spatial Information Science & Engineering, Xiamen University of Technology, Xiamen, Fujian 361024, China) TU Ping ZHU Xiao-ling MAN Wang

通讯地址:(361024 福建省厦门市集美区厦门理工学院空间信息系) 涂 平

(收稿日期:2009.11.30)(修稿日期:2010.03.01)