第 23 卷 第 1 期
2008 年 3 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.23 No. 1
Mar. 2008

基于类树的 XML 与关系数据库转换方法

刘元红
（天津师范大学计算机与信息工程学院，天津 300387）

摘 要：随着不断增长的 XML 应用的出现，在关系数据库中可靠、有效地存储 XML 文档以及 XML 与关系数据库间的数据转换技术受到广泛关注. 在研究已有的 XML 与关系数据库映射模型和方法的基础上，提出了一种基于类树的数据映射方法，实现了 XML 文档与关系数据库的转换.

关键词：XML；关系数据库；数据映射；数据转换；类树
中图分类号：TP311 **文献标识码**：A **文章编号**：1672-6510（2008）01-0071-05

A Conversion Method of XML and Relational Database Based on Class Tree

LIU Yuan-hong
（Institute of Computer Science and Information Engineering, Tianjin Normal Univesity, Tianjin 300387, China）

Abstract: As an increasing amount of XML-based applications appears today, reliable and efficient XML storage system and data exchange between XML and DBMS are becoming more and more important. A class tree data mapping approach was introduced on the basis of current data mapping model and approaches.

Keywords: XML; relational databases; data mapping; data conversion; class tree

XML 是 eXtensible Markup Language（可扩展标记语言）的缩写，它是由 W3C 组织于 1998 年 2 月创建的一组规范. 近几年，XML 的各项标准日臻完善，在各个领域的应用也不断扩展和深入. 由于其具有简单性、可扩展性、易操作性、开放性等优点，使其在信息表示领域得到广泛的应用，甚至成为一种通用的数据交换格式. 目前大量的信息系统和 Web 站点，已经开始基于 XML 的方式进行数据组织和管理. 随着 XML 文档数量的增加，对 XML 文档的存储、管理和查询提出了更高的要求[1—5].

数据库是目前最主要的数据存储方式，由于关系型数据库系统在存储、管理和查询优化等许多方面都较其他系统成熟和稳定，多采用关系数据库来有效地存储 XML 数据；另外，为了信息集成和数据交换，存储在关系数据库中的信息又需要转换成 XML 形式进行数据发布. XML 文档与关系数据库的转换已成为当今的研究热点之一. 目前有许多关系型数据库管理系统都对 XML 提供支持，同时各大软件公司也开发了大量支持 XML 与数据库的数据转换的软件工具.

目前关系型数据库管理系统对 XML 提供支持的有：Microsoft SQL Server 2005，Oracle 8i、9i、10g 都将 XML 作为一种内置数据类型来对待，并提供了 XML 查询的支持. 在 Oracle 8i 中用户可以通过建立带有 XML 类型的基本表来存储XML文档，也可以将已有的关系数据通过函数产生相应 XML 视图[6]. 对于这些 XML 基本表和视图的查询，Oracle 也提供了从 XML 文档中定位和提取结点的函数，用户可以通过这些函数构造出新的 XML 结点.

上述技术和方法都是开发公司投入市场的软件产品，系统大而全，对于实验性研究来说，资源消耗较大，因此本文提出一种 XML 文档的树形结构模型，模型中将 XML 的 DTD 所包含的元素作为树中的结点，描述各元素之间的关系，通过这种基于类树

收稿日期：2007–09–14；**修回日期**：2007–10–17
作者简介：刘元红（1967—），女，天津人，副研究馆员.

的面向对象方法，实现XML与关系数据库的转换.

1　基于类树的数据映射

XML 文档与关系模式之间的数据映射，有两大类映射方法[3]：模型映射和结构映射. 利用类树实现 XML 与关系模式之间的映射属于结构映射. 首先建立基于 DTD 的类树模型，之后在类树模型基础上，解析 XML 文档的 DTD 建立 DOM 树，对类树进行修剪，同时定义类树与关系数据库的映射规则，并建立树中结点与数据库表字段的对应关系，从而实现 XML 与关系数据库的映射转换.

1.1　类树结构模型的建立

由 DTD 文档生成类树模型的生成规则：

（1）DTD 中不能被其他元素内容包含的元素为根结点；

（2）DTD 中的每一个元素对应树中的每一个结点；

（3）每个结点包含以下信息：子元素列表，属性列表和其他信息；

（4）子元素列表包含该类元素的所有子元素，每个子元素对应一个子结点；

（5）属性列表包含了该元素的所有属性，每个属性为一个三元组（属性名，属性值，属性类型）；

（6）相关信息包括：父元素结点、元素内容模型、元素内容出现次数和元素文本内容等；

（7）子元素列表为空的元素的属性或活动为树的叶子结点.

根据类元素的属性分析类树的结构. 给定的某元素<!element customer（name+，tele|email，address*）>，其子元素 name，tele|email，address 按顺序出现在 XML 文档中，其中 tele 和 email 来自一个结点，具体内容需要二选一. 对于这类结点，必须考虑其子模型所定义的内容如何表示. 因此，在类树的模型中加入一种特殊结点即内容模型结点，利用该结点来描述子内容模型的选择，一个内容模型结点的子结点是同属于该子内容模型的所有元素所对应的结点.

根据定义，对于上述给定的 customer 元素，其内容模型中应该包含一个子内容模型，该子内容模型是子元素的选择列表，即 tele 和 email 二者只能出现其中之一，然后再用一个内容模型结点来体现 tele 和 email 之间的选择关系.

类树中元素结点结构如图 1 所示. 元素 customer 对应的局部类树如图 2 所示.

图 1　类树中元素结点结构

Fig. 1　Structure of class tree node

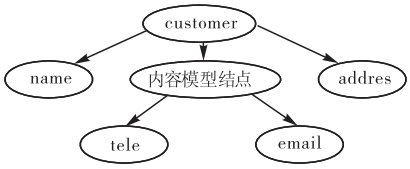

图 2　元素 customer 的类树

Fig. 2　Class tree of customer

在类树生成过程中，需要说明其中对于元素内容模型的处理：在生成一个元素的内容模型后，如果其非空则循环读取其子元素，取出一个子元素命名为 CP. 如果 CP 不是列表类型，则生成一个子结点 *S*，并找到该子元素 *S*，跳出循环；如果 CP 为子元素序列列表或选择列表，则将 CP 的每一个子元素命名为 SubCP，如果 SubCP 有名字，则 SubCP 为子元素序列列表或选择列表，生成一个内容模型的结点作为 CP 的子结点.

1.2　类树模型到关系数据库的映射

要实现 XML 和关系数据库的转换，在构造类树的基础上，需要定义 XML 类树对象中的元素与关系数据库模式之间的映射关系，映射规则如下：

（1）类树中内容模型结点不对应数据库中的对象；

（2）为类树的根结点创建数据库中的一个表及主键，称为根表；

（3）如果一个结点的属性列表或子元素列表不为空，则该结点称为表结点，为其在数据库中创建一个表，并建立主键及一个外键；

（4）除了表结点和内容模型结点以外的结点称为字段结点，它只对应其父结点所对应的表中的一个字段；

（5）如果一个表结点或子结点有父结点，那么父结点所对应的表称为该结点所对应的表的父表；

（6）数据库中，除根表以外的表与其父表通过外键相关联；

（7）如果一个结点的子元素为 PCDATA 类型，则该结点称为子结点；

（8）表结点的每个属性和子结点都对应于各自所对应的表中的一个字段.

在完成上述的构建类树模型及定义类树对象中元素与关系数据库模式的映射规则后，要进行具体的数据存储，还必须定义类树中元素到关系数据库数据

表中字段的对应关系. 对于这些映射关系,设计一个“元素——字段”映射的模板,称为映射表. 在映射表中,每一条记录都是一个(元素全名,字段全名)或(属性全名,字段全名)的二元组. 其中,元素全名是在类树中从根结点出发直接遍历到该元素的路径,不包括内容模型结点,即所有经过结点的结点名用“.”连接起来;属性全名则为元素全名再连接上“属性名”;字段全名是“表名::字段名”字符串.

通过上述映射表,对于一个给定的元素全名,可以在类树中找到一个唯一的与之对应的结点,反之对一个给定的字段全名,也可以在数据库表中找到一个唯一的字段.

类树的转换采用基于数据的映射方法,在构造类树的基础上,根据其映射关系制定一系列执行指令. 通过执行这些指令,将执行结果插入到数据模型中的相应位置,就可以得到相应 XML 文档. 同样,执行反向指令就可以把 XML 文档转换为其他格式的数据.

由于引入了内容模型结点,因此在进行 XML 和关系数据库的转换时,XML 文档结构的限制也大大放宽了. 此外,由于类树的生成是基于 DTD 的,因此对于 DTD 相同的 XML 的文档,类树可以被多次复用,从而大大提高了数据集成转换的效率.

在对类元素及其类元素属性进行数据转换后,每一个类元素所定义的操作可以在流程中借助触发,实时有效地完成相关操作. 因为数据的转换使所有的数据遵循同一标准,从而提高了操作的准确性.

2　XML 文档与关系数据库数据的转换

2.1　XML 到关系数据库的转换

数据转换过程中需要处理的数据包括 XML 文档及其 DTD、类树和映射表,从 XML 到关系数据库的转换过程按照算法 1.

算法 1 :

(1)对于给定 XML 文档的 DTD,先查找类树库以确定是否已有对应的类树,若没有则生成其类树;

(2)利用 XML 文档解析器对该 XML 文档进行解析,生成其对应的 DOM 树,并将数据内容插入到类树;

(3)查找“元素——字段”映射表,如果元素不能与数据库中的字段完全对应,根据一定规则创建新的数据库模式;

(4)遍历类树,读取数据,并根据“元素——字段”映射表,创建 SQL 语句,在数据库中插入或修改相应的记录.

通过基于模型的转换方法实现从 XML 文档到关系型数据的转换,是将 XML 文档构建成一定的对象模型,然后根据一定的映射关系映射到关系数据库的数据表和表中的字段. 那么在进行实际存储时就存在两种情况:一种情况是 XML 文档中的若干元素与数据库中的某个表的字段是完全对应的(每个字段都有元素与之对应,没有元素对应的字段是 NULL),这时只需要在这个表里插入一条记录即可;另一种是存在一些元素,没有现有的字段与之对应,或者是不能完全对应,这时要进行存储,就必须在数据库中创建新的数据库模式.

2.2　创建新数据库模式

在上述过程中的步骤(3)中,需要创建新的数据库模式,创建新数据库模式的步骤见算法 2.

算法 2 :

(1)一个表结点对应于一张表,表名为该结点名前而加上“tab_”;

(2)每个表中都有一个 ID 字段,作为该表的主键;

(3)为根表以外的所有表增加一个 ParentID 字段,对应于父表中的 ID 字段,作为该表的外键;

(4)表结点的每一个子元素映射表中的一个字段,字段的排列顺序为子结点在类树中的排列顺序;

(5)一个属性对应其所属结点所对应的表中的一个字段,字段名为“Att_属性名”;

(6)一个字段结点对应于其父表的一个字段,字段名为该结点的结点名;

(7)对于重复出现的结点不生成新表,而是在对应的表中增加一条记录;

(8)在创建了新表以后,还要在“元素——字段”映射表中增加相应的记录.

3　实　例

下面给出一个 DTD 实例(pubinfo.dtd),详细说明 XML 到关系数据库的转换的过程.

```
<!ELEMENT Pubinfo(book)+>
<!ELEMENT book(title, author, publishinfo)>
<!ATTLIST book ISBN CDATA#implied>
<!ELEMENT title(PCDATA)>
<!ELEMENT author(empty)>
```

< !ATTLIST author name CDATA#implied >

< !ATTLIST author email CDATA #implied >

< !ELEMENT publishinfo（（publisher，year）|（year，publisher））>

< !ELEMENT publisher（#PCDATA）>

< !ELEMENT year（#PCDATA）>

利用文中方法可以得到图 3 所示 DTD 对应的类树模型.

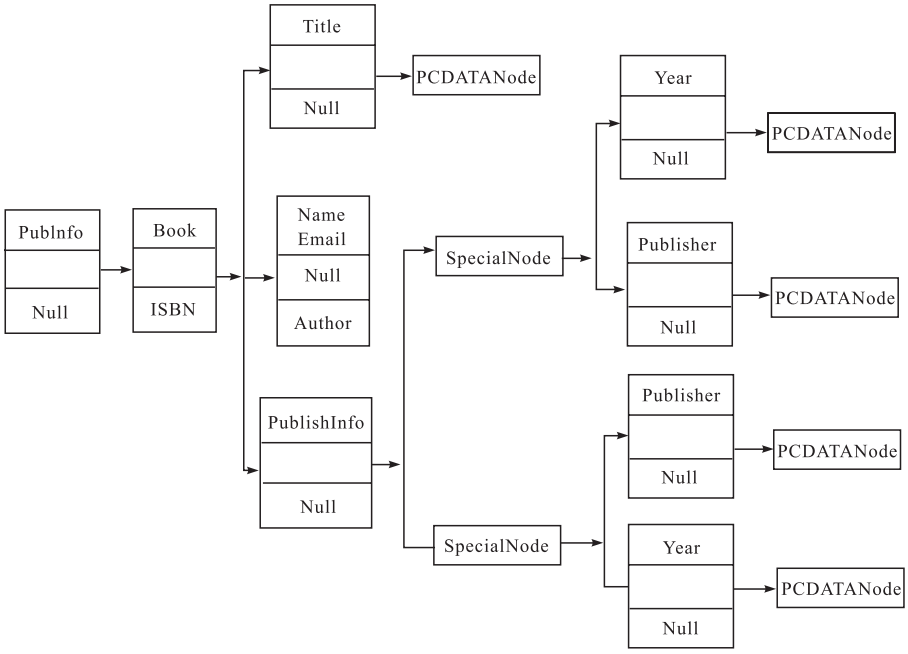

图 3　类树模型的示例

Fig. 3　Example of class tree

3.1　创建数据表

根据类树模型及数据库模式建立规则，在数据库中创建 4 张表，见表 1—表 4.

表与表之间通过 ID 和 ParentID 字段发生关联，例如 tab_book 中的 parentID 字段与 tab_pubinfo 中的 ID 字段关联. 通过这些关联，只要给出某个 XML 元素的 ID 值，就可以查找并构造出以该元素为根元素的 XML 文档片断. 而通过字段名，还可以得知该字段对应的是 XML 元素，还是 XML 属性.

表 1　数据表 tab_pubinfo

Tab. 1　Data table tab_pubinfo

ID	book
数据表主键	存储子元素 book

表 2　数据表 tab_book

Tab. 2　Data table tab_book

ID	ParentID	Att_ISBN	title	author	publishtinfo
数据表主键	关联父表的外键	存储属性 ISBN	存储子元素 title	存储子元素 author	存储子元素 publishinfo

表 3　数据表 tab_ author

Tab. 3　Data table tab_ author

ID	ParentID	Att_name	Att_email
数据表主键	关联父表的外键	存储属性 name	存储属性 email

表 4　数据表 tab_ publishinfo

Tab. 4　Data table tab_ publishinfo

ID	ParentID	Att_publisher	year
数据表主键	关联父表的外键	存储子元素 publisher	存储子元素 year

3.2　创建“元素——字段”映射表

在创建上述表的同时，系统还生成“元素——字段”映射表如下：

pubinfo. book=tab_pubinfo：：book

pubinfo. book. ISBN_ATT=tab_book：：Att_ISBN

pubinfo. book. title=tab_book：：title

pubinfo. book. author=tab_book：：author

pubinfo. book. author. name_ATT=tab_author：：ATT_name

pubinfo. book. author. email_ATT=tab_author：：

ATT_email

pubinfo. book. publishinfo=tab_book：：publishinfo

pubinfo. book. publishinfo. publisher=tab_publishinfo：：publisher

pubinfo. book. publishinfo. year=tab_publishinfo：：year

pubinfo=tab_pubinfo：：ID

book=tab_book：：ID

author=tab_author：：ID

publishinfo=tab_publishinfo：：ID *

通过上述的映射表，可以查询到一个表的所有字段及所对应的 XML 元素或属性，反之可以查询到一个 XML 元素的所有子元素和它的属性，及与它们在数据表中所对应的字段. 这样，通过转换就可以生成相应的 SQL 语句，将 XML 文档内容保存到数据库.

4　结　语

随着 XML 技术的迅速发展，数据库领域出现了大量与 XML 相关的研究工作. 由于大量的 XML 文档需要得到有效的存储和管理；同时关系数据需要转换成 XML 在网上发布和交换. 因此，XML 和关系数据库之间的数据转换成为一个重要的研究领域. 本文在研究已有的 XML 与关系数据库映射模型和方法的基础上，提出了基于类树的数据映射方法，作为轻量级的实验产品，实现 XML 与关系型数据库的转换.

参 考 文 献：

［1］ 万常选.XML 数据库技术［M］.北京：清华大学出版社，2005：75.

［2］ 赵　毅，王浩然，庄冠华，等.一种基于 XML 的数据集成系统框架及其应用［J］.计算机工程与应用，2005（26）：181—183.

［3］ 刘科研，万丽荣，曾庆良，等. 基于 XML 的信息集成系统的研究与实现［J］.计算机应用研究，2005（4）：149—154.

［4］ 杨　剑，唐慧佳，孙林夫，等. 基于 XML 的异构数据交换系统的研究与实现［J］.计算机工程，2005（19）：195—197.

［5］ 曹宇昆，王清明，杨卫冬，等. XML 模式到关系模式的映射［J］. 计算机工程，2005（8）：37—39.

［6］ Lee D, Chu J. CPI：Constraints-preserving inlining algorithm for mapping XML DTD to relational schema［J］. Data & Knowledge Engineering，2001（39）：3—25.

(上接第 66 页)

以较小的时间复杂度获得一个闭合的路面损坏区域的边缘，进而根据放置于现场待测区域附近尺寸已知的标盘之间的像素映射比例关系进一步计算路面的损坏面积. 通过与实际测量结果的比较，本文算法可以在检测路面坑槽、凹陷、油污损坏等痕迹清晰的损坏面积时获得良好的实验效果. 并且此方法还可应用于其他大面积的图像测量的应用. 但对于龟裂、波浪拥包类路面的损坏面积测量仍有待进一步研究.

参 考 文 献：

［1］ Zhang Juan，Sha Ai-min，Gao Huai-gang，et al. Automatic pavement crack recognition and evaluation system based on digital image processing［J］. Journal of Chang-an University Natural Science Edition，2004，2（2）：18—22.

［2］ 丁爱玲，焦李成.基于支撑矢量机的路面破损识别［J］.长安大学学报，2007，27（3）：35—37.

［3］ 张　娟，沙爱民，孙朝云.数字图像处理技术在道路无损检测中的应用［J］.山西交通科技，2002（6）：10—12.

［4］ Rodriguez Tomas. Practical camera calibration and image rectification in monocular road traffic applications［J］. Machine Graphics and Vision，2006，15（1）：51—71.

［5］ 高建贞.基于图像分析的道路破损自动检测研究［D］.南京：南京理工大学，2003.

［6］ ARRB 提供资料：设备技术参数及技术服务，HAWKEYE 2000 SERIES Processing Toolkit User Manual［R/OL］.http：//www.arrb.com.au/.

［7］ 储江伟，初秀民，王荣本，等. 沥青路面破损图像特征提取方法研究［J］.中国图形图像学报，2003，8A（10）：1211—1217.

［8］ 胡小兵.基于面积特征的物体测评方法研究与实现［J］.计算技术与自动化，2002，21（4）：21—23.

［9］ 赵　鹏，浦昭邦，张田文. 基于动态轮廓线的图像面积测量研究［J］.仪器仪表学报，2006，27（9）：1150—1153.

［10］ 崔　屹. 数字图像处理技术与应用［M］.北京：电子工业出版社，1996：121—131.

［11］ 屈　彬.基于区域生长规则的快速边缘跟踪算法和改进的 Herman 插值算法的研究以及在医学图像处理中的应用［D］. 成都：四川大学，2002：24—27.

［12］ Herman T G，Robinson D F. Digital boundary tracking［J］. Pattern Analysis&Applic，1998（1）：2—17.