文章编号： 1000-5013(2011)01-0039-04

XML数据到关系数据的映射

洪欣[1]，陈维斌[1]，蹇崇军[2]

(1. 华侨大学 计算机科学与技术学院，福建 泉州 362021；
2. 华侨大学 机电及自动化学院，福建 泉州 362021)

摘要： 为实现XML数据到关系数据的数据映射，提出一种模式抽取算法，通过XML2XDR模块抽取XML模式，依据模式对XML数据分类. 分析XML模式与关系模式的差异性，通过XMLdata2DB模块建立XML数据到关系数据的映射规则，从而实现将XML数据映射到关系数据中.

关键词： 关系数据；XML；映射规则；模式抽取

中图分类号： TP 311.132.3；TP 311.12 **文献标识码：** A

企业之间进行商务数据交换时，存在数据结构不同的问题，往往通过XML格式的数据进行数据交换，并将其导入企业原有的关系数据库中[1]. 随着大量XML文档数据的出现，如何高效地存储、管理和查询这些XML数据，成为目前值得研究的重要课题. 对XML数据库系统的研究主要有两个基本途径：一种是纯XML数据库系统，优点是充分考虑XML数据的特点，以一种自然的方式来处理XML数据，能较好地支持XML数据的存储和查询，但目前技术仍不成熟；另一种是在原关系数据库系统或面向对象数据库基础之上扩充功能，实现XML数据的处理[2-5]. 本文提出一种XML模式抽取技术，抽取XML文档的模式，将XML模式映射到关系模式中，从而实现企业间的电子商务数据交换[6].

1 抽取XML模式的必要性

在XML文档中，模式不是必须的. 对于带模式的XML文档，可以将XML模式直接映射为关系模式. 对于不带模式的XML文档，如果直接映射为关系，即使XML文档来自于同一模式，也可能映射为不同结果. 如以下两个XML文档片断A和B.

(1) XML文档片断A.

```
⋮
〈CnGame xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"〉
〈Opponent〉
    〈TeamName〉巴西〈/TeamName〉
    〈Place〉韩国西归浦〈/Place〉
    〈Date〉2002年6月8日〈/Date〉
    〈Time〉19:30〈/Time〉
    〈Star〉
        〈StarName language="english"〉罗纳尔多〈/StarName〉
        〈StarName language="english"〉里瓦尔多〈/StarName〉
    〈/Star〉
    〈Coach〉
```

收稿日期： 2010-04-23

通信作者： 洪欣(1977-)，女，讲师，主要从事数据库领域的研究. E-mail：xinhong@hqu.edu.cn.

基金项目： 福建省青年人才项目资助(2007F3062)；福建省泉州市科技计划项目(2010G4)；华侨大学高层次人才科研启动项目(07BS403)；华侨大学科研基金资助项目(08HZR18)

〈CoachName〉斯科拉里〈/CoachName〉
〈/Coach〉
〈/Opponent〉
〈/CnGame〉
⋮

文档片断A直接映射生成的关系表如下：

CnGame(CnGame_PK)
Opponent(Opponent_PK,TeamName,Place,Date,Time,CnGame_FK)
Star(Star_PK,StarName,Atrribute_language,Opponent_FK)
Coach(Coach_PK,CoachName, Opponent_FK)

XML文档映射得到的关系表中带"_PK"后缀的为主码,带"_FK"后缀的为外码,带"Atrribute_"前缀的为属性映射的字段.

(2) XML文档片断B.

⋮
〈CnGame xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"〉
〈Opponent〉
〈TeamName 土耳其〈/TeamName〉
〈Place〉韩国汉城〈/Place〉
〈Date〉2002年6月13日〈/Date〉
〈Time〉14:30〈/Time〉
〈Star〉哈坎-苏克〈/Star〉
〈Coach〉
〈CoachName〉吉内斯〈/CoachName〉
〈/Coach〉
〈/Opponent〉
〈/CnGame〉
⋮

文档片断B映射生成的关系表如下：

CnGame(CnGame_PK)
Opponent(Opponent_PK,TeamName,Place,Date,Time,Star,CnGame_FK)
Coach(Coach_PK,CoachName,Opponent_FK)

以上示例(1),(2)中可见,文档片断直接映射生成的关系表存在着差异.即采用不带模式的映射,可能会生成多个具有重复字段却并不完全相同的表,不仅浪费了存储空间,而且不利于数据的管理.

如果首先抽取XML文档片断A与XML文档片断B的模式,然后通过模式进行映射,由于两个文档片断是同源的,那么生成的关系表就是相同的.由此可见,如果XML文档带有模式,最好的方式就是依据模式来映射关系模式,这样映射生成的关系表具有通用性.如果XML文档不带模式定义,则先为其建立XDR模式,然后再调用带有模式的XML文档的映射算法进行映射.

2 系统设计

2.1 映射框架

对不带XDR模式的XML文档的映射包含如下3个步骤.(1)生成XDR模式.(2)应用含有XDR模式的映射方法,对XML文档进行映射生成关系表.(3)通过填入数据模块,将XML文档数据映射为关系数据库数据.

图1为系统流程.模块结构包括XML文档模式的抽取(XML2XDR模块)、XDR模式到关系模式的映射(XDR2DB模块)、XML数据到关系数据的映射(XMLdata2DBdata模块).

2.2　XML 文档模式的抽取

XML 文档的模式有多种，选择其中的 XDR(XML Data Reduced Language，简化 XML 数据定义语言)进行说明. 基于 XDR 模式的 XML 文档的抽取算法 XML2XDR 有如下 7 个步骤.

(1) 读取 XML 文档.

(2) 建立 XDR 文件头.

(3) 读取 XML 数据建立根节点.

(4) 生成元素声明.

(5) 如果元素有属性，则读取属性值，建立属性声明.

(6) 如果元素有子节点，则读取子节点生成子元素，转到(4)，继续分析子元素，建立子元素的元素及属性声明.

(7) 是否还有同层节点，如果有则转到(4)，为该节点建立元素及属性声明；如果没有则输出 XDR 模式文件.

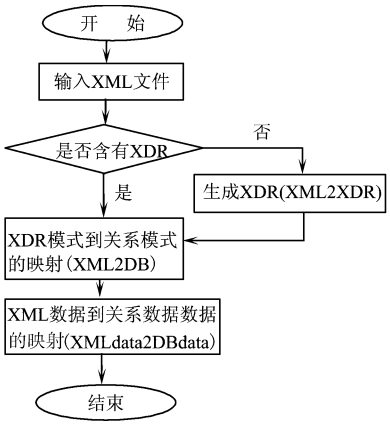

图 1　系统工作流程

Fig. 1　Working flow of the system

前例中的 A 文档与 B 文档经过模式抽取之后，生成的 XDR 模式如图 2 所示. 由于 Star 包含在 minOccurs＝“0” maxOccurs＝“＊”的 group 中，出现的次数是不定的，所以在 XDR 模式的实例中，Star 元素可能出现零次、一次或多次，与文档 A 及文档 B 的情况相符.

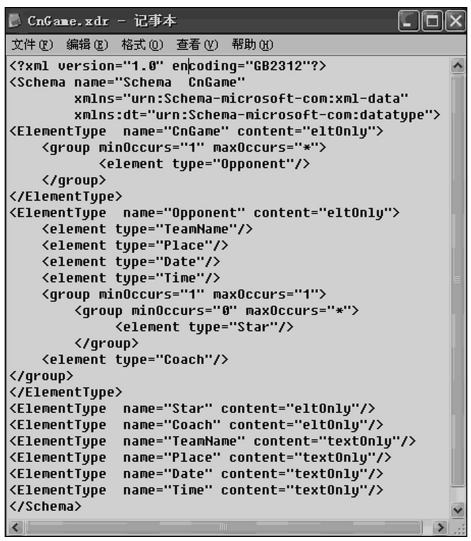

图 2　XML 文档抽取的 XDR 模式

Fig. 2　XDR schema extracted from XML documents

2.3　XML 模式到关系模式的映射算法

抽取 XML 模式之后，可以通过模式将 XML 的数据结构映射到关系数据的数据结构，建立模式映射算法 XDR2DB. 算法有如下 5 个步骤.

(1) 读取 XDR 模式.

(2) 读取 ElementType 元素声明.

(3) 消除递归，生成关系表 XDR2DB-Recursion-Mapping.

(4) 根据节点类型，进行相应的映射. 如果子元素类型为属性，调用元素属性映射算法生成字段；如果子元素类型为元素，调用子元素映射算法生成字段及关系；如果子元素为 group，调用模式组映射算法生成字段及关系.

(5) 判断是否到达 XDR 模式末尾，如果不是，则转到(2)；否则，生成表间联系.

图 2 所示生成的 XDR 模式，经过 XDR2DB 映射后，其结果如图 3 所示.

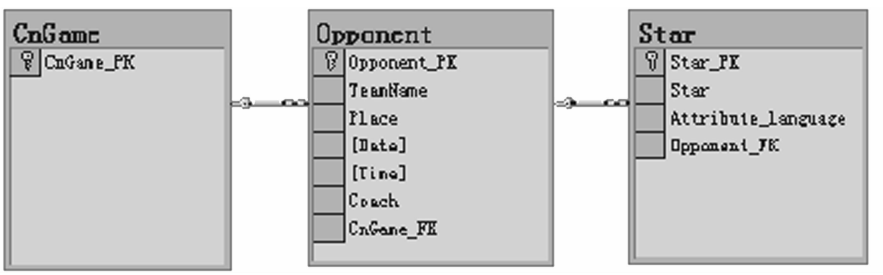

图 3　XDR 模式映射到关系数据库的结果

Fig. 3　The results of XDR schema mapping to the relational database

2.4　XML 数据到关系数据的映射算法

将 XML 模式映射为关系模式之后，需要将 XML 文档的数据填充到所建立的关系表中. XMLdata2DB 映射算法有如下 5 个步骤.

(1) 读取 XML 模式的根节点，建立表格.

(2) 读取模式的节点 element,创建子表.

(3) 如果 element 不含子节点,则转到(5);如果 element 含有子节点,则分析子节点的类型. 如果为属性节点,则为该属性节点对应的字段赋值;如果为元素节点,则判定是否还包含子节点,如果含有子节点则转到(2);否则,为该子节点对应的字段赋值.

(4) 判断子节点是否还有未访问的同层节点,如果是,则读取下一节点转到(3);否则,插入一条记录到关系表中.

(5) 查找是否还有未访问的节点,如果有转到(3);否则完成数据映射.

文档 A 与文档 B 的数据映射到关系数据库中的结果,如图 4 所示.

图 4 XML 数据映射到关系数据的结果

Fig. 4 Results of XML data mapping to relational data

3 结束语

建立了一个数据交换框架,提出一种 XML 模式的抽取技术,并给出从 XML 数据到关系数据的映射算法. 研究结果可为企业的商务数据交换提供一种可行的方法,方便企业之间在异构平台上的实现数据交换.

参考文献:

[1] 宋培义. XML 与电子商务[J]. 北京广播学院学报:自然科学版,2000,7(4):33-37.

[2] WANG Guo-ren, HAN Dong-hong, QIAO Bai-you, et al. Extending XML schema with object-oriented features [J]. Information Technology Journal,2005,4(1):44-54.

[3] WANG Guo-ren,LIU Meng-chi . Extending XML schema with nonmonotonic inheritance[C]// Proceedings of 1st International Workshop on XML Schema and Data Management. Chicago:Iustrotion,2003:402-407.

[4] LEE Dong-won,CHU W W. Comparative analysis of six XML schema language[J]. ACM SIGMOD Record,2002, 29(3):117-151.

[5] 万常选,林大海. 基于关系数据库分裂大型 XML 文档到关系存储[J]. 计算机应用研究,2004,21(8):166-167.

[6] 洪欣. 基于 XDR 模式的 XML 文档与关系数据库的映射技术研究[D]. 泉州:华侨大学, 2004.

Mapping from XML-Data to Relational-Data

HONG Xin[1], CHEN Wei-bin[1], JIAN Chong-jun[2]

(1. College of Computer Science and Technology, HuaQiao University, Quanzhou 362021, China;
2. College of Mechanical Engineering and Automation, Huaqiao University, Quanzhou 362021, China)

Abstract: In order to realize the data mapping from the XML-data to the relational-data, this paper propose a model extraction algorithm, the XML2XDR model extract the XML model, according to the model the XML-data is classified. By analysing the differences between the XML schema and relational schema, the mapping rules from XML-data to relational data was established based on XMLdata2B, and the mapping from XML-data to relational-data was realized.

Keywords: relational data; XML; mapping rules; pattern extraction

(**责任编辑**: 钱筠 **英文审校**: 吴逢铁)