文章编号: 1000-5013(2010)03-0288-04

抽取 XML 模式到关系模式

洪欣, 陈维斌, 杜吉祥

(华侨大学 计算机科学与技术学院, 福建 泉州 362021)

摘要: 提出一种 XML 建模技术,对多个 XML 文档构造共享模型. 通过该模型构造的 XML 共享模式到关系模式的映射,实现将同源异构的 XML 片断抽取到相同的关系表中. 试验表明,算法在同源异构 XML 片断的映射上具有比以往算法更好的映射结果.

关键词: XML 模式; XML 模型; 关系模式;数据抽取

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

许多商品数据以 XML 数据的形式被放置在网页中,如何抽取这类数据到关系数据库中,以提高检索效率,是一项值得研究的课题. 商品 A 与商品 B 的结构并不完全相同,经过结构抽取,两种商品的数据应该存入相同的表格当中,但当前的算法无法实现这一类的数据映射. 目前,流行的大型关系数据库——SQL 数据库系统并不具备 XML(eXtensible Markup Language)模式抽取的能力,必须在关系数据库中根据 XML 模式手动建立好对应的关系模式,才能够将 XML 数据映射到关系数据库中. 当前有一些相关研究,如文[1]是基于 XML 文档已经存在模式的情况下,实现 XML 模式到关系模式的直接映射,但并未考虑 XML 数据在可能不包含模式文件的情况下文档映射的不稳定性;又如文[2]采用模式分类技术建立 XML 模式,但丢失了文档结构的完备性,可能出现同一 XML 文档文件数据抽取到不同的数据库中. 针对以上关键问题,本文采用 XML 模型(XML model)[3]为多个 XML 文档建立共享模型,通过 XML 模式到关系模式的映射,实现将同源异构 XML 片断数据抽取到相同的关系表中,并保证映射的稳定性.

1 XML 文档的建模

在当前的模型中,ER(Entity Relation)模型能够表示关系和数据,但不能描述控制信息[4-5];ODMG(Object Data Management Group)模型能够表示控制信息,但是对关系的描述不充分且不直观[6-8]. 因此,采用 XML 模型(XML model)描述 XML 数据、XML 的模式及 XML 模式的模式[3].

XML 文件的结构十分灵活,即使元素声明是一致的,元素的出现次序及元素间的层次关系在两个 XML 文档中也可能不同. 其建模有以下 3 个主要的算法.

(1) 单 XML 文档的模型构造算法. 输入为 XML 文档,输出为 XML 文档对应的 XML model 模型. 主要有如下 5 个步骤: ①读取 XML 文档,建立 XML model 根节点; ②读取 XML 文档的元素节点,添加 XML model 的元素节点; ③如果元素有属性,则添加属性节点; ④如果元素有子节点,则读取子节点,转到步骤 ②;继续分析子节点,在 XML model 中添加子节点对应的元素及属性节点; ⑤如果还有同层节点,转到步骤 ②或步骤 ④,为该节点建立元素及属性节点;如果没有同层节点,则输出 XML model 模型.

(2) 多 XML 文档的模型构造算法. 输入为多个 XML 文档,而输出为包含所有 XML 文档结构的 XML model 模型. 主要有如下 3 个步骤: ①读取 XML 文档 X_i,建立初始模型 XML model_X; ②读取

收稿日期: 2009-11-26

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

基金项目: 国家自然科学基金资助项目(60805021); 华侨大学科研基金资助项目(08HZR18)

XML 文档 X_2,并进行节点匹配. 若出现新的元素,则新建元素节点;若出现新的属性,则新建属性节点; ③读取文档 X_n,转到步骤 ②,直到读取所有文档,完成模型建立.

(3) 模型构造模式算法. 输入为 XML model 模型,输出为 XML 模式(XDR). 主要有如下 5 个步骤: ①读取 XML model,建立 XDR 文件头; ②读取节点,生成元素声明〈Element Type〉; ③如果元素有属性,则建立属性声明〈Attribute〉; ④如果元素有子节点,则读取子节点生成子元素〈Element〉,转到步骤 ③,继续分析子元素,建立子元素的元素及属性声明; ⑤如果还有同层节点,转到步骤 ④,为该节点建立元素及属性声明;如果没有同层节点,则输出 XDR 模式文件.

单个 XML 文档可以通过算法 1,3 构造模型. 多个文档片段可以采用以下 3 个步骤实现 XML 文档建模. (1) 通过算法 1 为第 1 个文档建模; (2) 根据算法 2 读取剩余文档片段,修改算法 1 建立的初始模型,最终为所有文档片段建立统一的共享模型; (3) 在该算法的基础上,根据算法 3 将共享模型映射为 XDR 模式.

2　XML 模式映射为关系模式

通过以上算法建立 XML 模式后,由于半结构化的 XML 模式与结构化的关系模式存在差异,需要将 XML 模式中非结构化的部分进行结构化处理,再将 XML 模式映射为关系模式,如图 1 所示. XML 模式映射为关系模式的算法有如下 4 个主要步骤: (1) 消除 XML 模式的重复元素; (2) 消除 XML 模式的嵌套; (3) 消除 XML 模式的递归; (4) 结构化后的 XML 模式映射为关系模式.

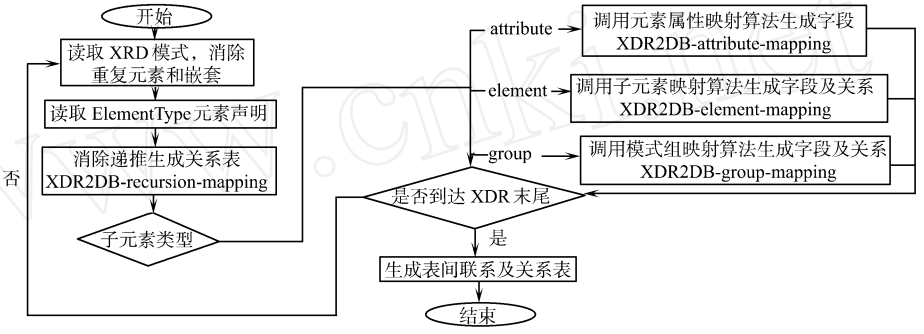


图 1　XDR 模式到关系模式的映射流程图

Fig. 1　Mapping flow chart from XDR to relation schema

3　测试结果与分析

3.1　XML 文档建模

XML 文档片断 X_1, X_2 是同源 XML 数据,如图 2 所示. 对 XML 文档片断分别建模,对应的 XML model 如图 3 所示.

```
⋮
〈 CnGame 〉
〈 Opponent 〉
  〈 TeamName 〉巴西〈 /TeamName 〉
  〈 Place 〉韩国西归埔〈 /Place 〉
  〈 Date 〉 2002 年 6 月 8 日〈 /Date 〉
〈 Time 〉 19:30〈 /Time 〉
  〈 Star 〉
    〈 StarName 〉罗纳尔多
〈 /StarName 〉
    〈 StarName 〉里瓦尔多
〈 /StarName 〉
  〈 /Star 〉
〈 Coach 〉
    〈 CoachName 〉斯科拉里
  〈 /CoachName 〉
〈 /Coach 〉
〈 /Opponent 〉
〈 /CnGame 〉
⋮
```

(a) 片段 X_1

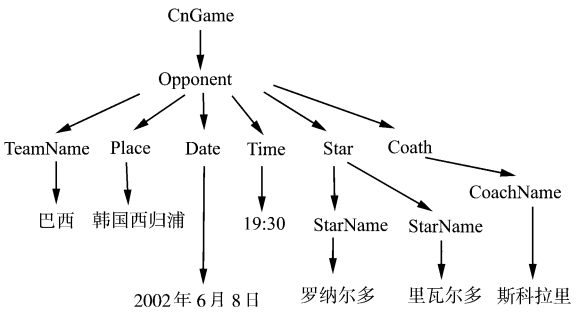
```
⋮
〈 CnGame 〉
〈 Opponent 〉
〈 TeamName 〉土耳其〈 /TeamName 〉
〈 Time 〉 14:30〈 /Time 〉
〈 Date 〉 2002 年 6 月 13 日〈 /Date 〉
〈 Place 〉韩国汉城〈 /Place 〉
〈 Coach 〉
  〈 CoachName 〉吉内斯〈 /CoachName 〉
〈 /Coach 〉
〈 /Opponent 〉
〈 /CnGame 〉
⋮
```

(b) 片段 X_2

图 2　XML 片段

Fig. 2　XML fragment

对文档片段 X_1,经算法 1 建立初始 XML 模型;然后,读取文档片段 X_2,采用算法 2 修改模型,最终



(a) X_1 模型 (b) X_2 模型

图3 XML片段模型

Fig. 3 XML fragment model

构造的文档片段 X_1 与 X_2 共享模型,如图4所示.

图4的模型经算法3构造的XDR(XML-Data Reduced)模式,如图5所示.

3.2 XML模式映射为关系模式

由图1的算法流程生成的关系模式如下,映射结果如图6所示.

3.3 与其他算法比较

3.3.1 SQL Server的直接映射方法 文档片断 X_1,X_2 经过SQL Server数据转换工具,会得到两组表及关系,

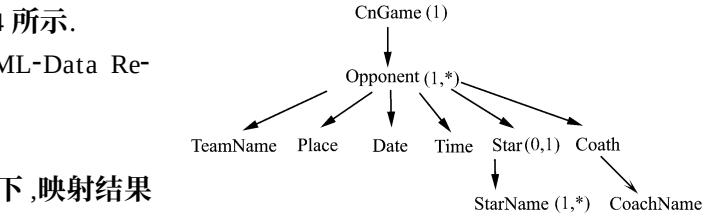


图4 文档片段的共享XML model“CnGame”

Fig. 4 Shared XML model “CnGame” of XML fragment

```
ElementType name="CnGame" content="eltOnly"〉
  〈group minOccurs="1" maxOccurs="*"〉
    〈element type="Opponent" /〉
〈/group〉
〈/ElementType〉
〈ElementType name="Opponent" content="eltOnly"〉
  〈element type="TeamName" /〉
  〈element type="Place" /〉
  〈element type="Date" /〉
  〈element type="Time" /〉
  〈group minOccurs="0" maxOccurs="1"〉
    〈element type="Star" /〉
  〈/group〉
  〈element type="Coach" /〉
〈/ElementType〉
〈ElementType name="Star" content="eltOnly"〉
  〈group minOccurs="1" maxOccurs="*"〉
    〈element type="StarName" /〉
〈/group〉
〈/ElementType〉
〈ElementType name="Coach" content="eltOnly"〉
  〈element type="CoachName" /〉
〈/ElementType〉
〈ElementType name="TeamName" content="textOnly" /〉
〈ElementType name="Place" content="textOnly" /〉
〈ElementType name="Date" content="textOnly" /〉
〈ElementType name="Time" content="textOnly" /〉
〈ElementType name="StarName" /〉
〈ElementType name="CoachName" /〉
```

图5 XML model“CnGame”生成的XDR

Fig. 5 XDR for XML model “CnGame”

即SQL Server会为文档片断 X_1,X_2 分别建表,并且无法导入关系. SQL Server为文档片断 X_1,X_2 建立的表结构分别为

CnGame(TeamName,Place,Date,Time,
StarName,CoachName);

CnGame(TeamName,Time,Date,Place,
CoachName);

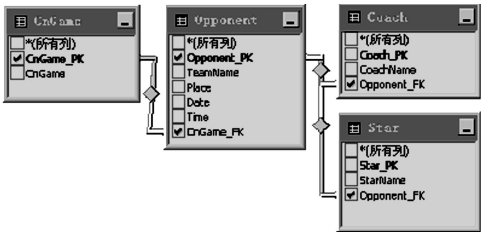


图6 XDR模式映射到关系数据库的结果

Fig. 6 Results of the mapping from XDR to DB

由此可以看到,由于结构差异导致SQL Server建立多个类似表格,出现表格冗余,并且数据分布在不同表格中不利于查询和管理.

3.3.2 特征值计算方法 若采用文[2]的方法,将文档片断 X_1,X_2 的XML数据导入时,表结构的建立是通过相似度计算得到. 在某些特殊的情况下,会有一个XML文档多个节点由于相似度差异较大而分配到不同表格中,并且不能映射关系. 文档片断 X_1,X_2 可能产生如下表结构:

CnGame(TeamName, Time,Date, Place,CoachName);

Star(StarName);

文档片断的数据会被分割到不合适的表中,使得数据产生错误.

3.3.3　多个 XML 文档共享结构的 XML 模型构造算法　采用 XML 模型构造算法,文档片断 X_1, X_2 生成的关系如下:

CnGame(CnGame_PK);

Opponent(Opponent_PK, TeamName, Place, Date, Time, Coach, CnGame_FK);

Star(Star_PK, StarName, Opponent_FK);

Coach(Coach_PK, CoachName, Opponent_FK);

由此可知,所提算法可以较好地实现 XML 同源异构的模式抽取. 映射生成的表结构包含了完整的数据结构和关系,可以保证 XML 数据被正确地映射到关系数据库中. 实验结果证明,该算法是可行并有效的.

4　结束语

提出多个 XML 文档共享结构的 XML 模型构造算法,并通过将所建立的 XML 模型依次映射为 XML 模式及关系模式,实现多个 XML 文档的模式抽取. 试验结果表明,该算法可以实现将多个的同源异构的 XML 文档映射到相同关系模式中,为电子商务的 XML 数据到关系数据的数据抽取提供一种可行的解决方案.

参考文献:

[1] 方洁,刘广钟. XML 模式到关系数据模式转换的研究[J]. 计算机工程与应用,2009,45(9):157-160.

[2] 吴扬扬,雷庆,陈锻生. 一种从 XML 数据中发现关系信息的方法[J]. 软件学报,2008,19(6):1422-1427.

[3] 洪欣,陈维斌,蹇崇军. XML 模型到关系模型的映射[J]. 华侨大学学报:自然科学版,2009,30(6):84-87.

[4] WANG G R. Extending XML schema with object-oriented features[J]. Information Technology Journal, 2005, 4(1):44-54.

[5] 王珊,萨师煊. 数据库系统概论[M]. 4 版. 北京:高等教育出版社,2006.

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

[7] LEE Dong-won, CHU W W. Comparative analysis of six XML schema language[J]. ACM SIGMOD Record, 2000, 29(3):117-151.

[8] CATTELL R G G. The object database standard: ODMG-93[S]. San Mateo: Morgan Kaufmann, 1994.

Extract Relation Schema from XML Schema

HONG Xin, CHEN Wei-Bin, DU Ji-Xiang

(College of Computer Science and Technology, Huaqiao University, Quanzhou 362021, China)

Abstract: This article put forward an XML modeling method to build a shared XML model from the muti-XML document. The shared XML schema can be created by the shared model, and be mapped into the relational Schema to realize the mapping from the isomer XML document to the same relational database. The tests show that the algorithm in this article have better mapping result than other algorithm in isomer XML document's mapping.

Keywords: XML schema; XML model; relational schema; data extraction

(责任编辑: 黄仲一　　英文审校: 吴逢铁)