

远程教育技术规范 XML 文档关系数据库存储技术

吴丽萍, 颜仁泉, 朱新华, 寇丽娟

(广西师范大学 计算机科学与信息工程学院, 广西 桂林 541004)

摘要:针对在远程教育技术规范 XML 绑定文档中, 经常使用元素嵌套, 强调数据的层次性、顺序性等特点, 提出了一套基于 Schema 模式的将规范的 XML 绑定文档存储到关系数据库中的解决方案。该方案在存储文档时, 通过建立元素表保存模式的信息, 确保了元素的完整性, 并在嵌套元素的数据表中添加 Layer 属性列, 以表明元素的嵌套关系, 保证了数据的有序性, 避免数据在存储和恢复时产生混乱。对数据的存储做了详细的描述。

关键词:远程教育技术规范; XML 模式; 映射

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2009)07-0130-03

Storing Distance Learning Technical Specifications XML Binding Documents Use RDBMS

WU Li-ping, YAN Ren-quan, ZHU Xin-hua, KOU Li-juan

(Computer Science and Information

Engineering College, Guangxi Normal University, Guilin 541004, China)

Abstract: Distance learning technical specifications' XML binding documents often use nested elements, and emphasized the level and order of data, in allusion to the characteristics, put forwards a kind of XML binding document store in RDB method based on XML Schema. In this plan, establish an element table to save the schema information when store the documents. Then ensure the integrity of the element. And add layer property to the table of nested elements to show the relationship between the nested elements, and then ensure that the data ate order to avoid the data storage and recovery giving rise confusion. Has done a detailed description on data storage.

Key words: distance learning standard; XML schema; mapping

0 引言

随着远程教育日益普及, 远程教育技术规范的研究也越来越受到重视。目前, 远程教育技术规范是采用可扩展标记语言 XML 来进行定义与推广的^[1], 其基本思想为: 首先根据领域专家给出规范的数据模型, 使用 XML 模式定义规范中的词汇(元素)及词汇间的关系; 然后再按照规范的 XML 模式生成各种应用规范的 XML 实例文档。因此, 在规范的推广与应用过程中, 会产生大量的 XML 实例文档, 如何有效存储和管理这些 XML 文档数据, 成为了规范研究中的一个焦点。

目前, XML 存储方案主要有三种形式^[2]: 1) 文本文件存储方式; 2) Native-XML 数据库存储方式; 3) 支

持数据库的存储方式, 主要是关系数据库、面向对象数据库存储方式。在诸多存储方案中, 关系数据库是结构化的数据库, 是当前发展和应用的主流, 相对于其他的数据库, 它具有数据结构化、冗余度低、较高的程序与数据独立性、易于扩充、便于数据管理等优点。基于以上优点, 针对在远程教育技术规范 XML 绑定文档中, 经常使用元素嵌套, 强调数据的层次性、顺序性等特点, 把规范的 XML 绑定文档存储到关系数据库中, 可以有效地实现数据存储的管理。

1 远程教育技术规范简介

远程教育技术规范就是在远程教育领域中制定的一系列标准框架, 学习资源供应商、学习工具开发者和学习平台开发者的研究和开发都遵循这些标准, 使得教育资源得到更好的共享利用。

1.1 国内外研究现状

国际组织对于远程教育标准的研究起步于 20 世纪 90 年代末, 影响比较大的主要有 IEEE LTCS(国际

收稿日期: 2008-11-13; 修回日期: 2009-01-05

基金项目: 广西自然科学基金资助项目(0447034)

作者简介: 吴丽萍(1982-), 女, 广西陆川人, 硕士研究生, 研究方向为关系数据库、XML; 朱新华, 教授, 硕士研究生导师, 研究方向为远程教育技术规范、XML。

电气和电子工程师协会学习技术标准委员会)、W3C(万维网联盟)、IMS(教学管理系统全球学习联合公司)等。IEEE LTSC 具有很高的权威性,从它的系统体系结构来看,其规范是完美的。但 IEEE 的机构非常庞大、臃肿,其规范也大都面面俱到,就应用来说难以实现。W3C 致力于在 Internet 上支持资源共享和系统互操作的多种标准的开发,其主要有 XML 规范、RDF 规范、SMIL 规范、PICS 规范等。IMS 主要是制定、发布教育资源在线发布的一些标准,相对于其他标准来说,是比较完善的。它发布了大量的教育信息技术标准,为远程教育信息标准化提供了教学资源元数据规范、内容包装规范、问题与测试规范、教学管理系统规范等规范。

国家教育部于 2001 年初成立了现代远程教育技术标准委员会,于 2002 年初更名为教育部教育信息化技术标准委员会,2002 年经国家标准化管理委员会批准成为全国信息技术标准化技术委员会教育技术分技术委员会,授权承担全国教育技术相关标准的研制、认证和应用推广工作。到目前为止,该委员会已经颁布了 CELTS 系列标准,其中的 3 个标准已经发布为国家标准,即 GB/T 21364-2008 信息技术,学习、教育和培训,基于规则的 XML 绑定技术;GB/T 21365-2008 信息技术,学习、教育和培训,学习对象元数据;GB/T 21366-2008 信息技术,学习、教育和培训,参与者标识符。

1.2 远程教育技术规范的 XML 绑定

XML 作为标准交换语言^[3],起着描述交换数据的作用。远程教育信息文件格式、数据结构的标准可以通过 XML 定义数据表示的基本结构实现,XML 的简单易读性、可扩展性、保值可持续发展性等优点,为远程教育信息提供了良好的标准化环境。

以下是 XML 绑定的远程教育规范的内容包装规范中的一个小实例:

```
<organizations default="TOC1">
<organization identifier="TOC1" title="default">
  <item identifier="ITEM1" identifierref="RES1">
    <title>Lesson 1</title>
    <item identifier="ITEM2" identifierref="RES2">
      <title>Introduction 1</title>
    </item>
    <item identifier="ITEM3" identifierref="RES3">
      <title>Content 1</title>
    </item>
  </item>
  .....
</organization>
</organizations>
```

2 XML 文档到关系数据库的转换

2.1 XML Schema

Schema 其实也是一种文档,它是描述 XML 文档的文档。XML Schema 主要是用来定义 XML 文档的内容及其约束的,它是 W3C 在 DTD 定义和数据库系统建模语言的基础上提出来的。在 XML Schema 中,数据类型主要分为简单数据类型和复杂数据类型,简称为简单类型和复杂类型,数据类型可以是命名的,也可以是匿名的。以下就是关于上面的内容包装规范实例的 XML 模式定义文档:

```
<xsd:element name="organizations">
<xsd:complexType>
  <xsd:sequence>
    <xsd:element name="organization">
      <xsd:complexType>
        <xsd:sequence>
          <xsd:element name="item" maxOccurs="unbounded">
            <xsd:complexType>
              <xsd:sequence>
                <xsd:element name="title" type="xsd:string" minOccurs="0"/>
                <xsd:element ref="item"/>
              </xsd:sequence>
            </xsd:complexType>
            <xsd:attribute name="identifier"/>
            <xsd:attribute name="identifierref"/>
          </xsd:element>
        </xsd:sequence>
        <xsd:attribute name="identifier"/>
        <xsd:attribute name="title"/>
      </xsd:complexType>
    </xsd:element>
  </xsd:sequence>
  <xsd:attribute name="default"/>
</xsd:complexType>
</xsd:element>
```

从以上示例可以看出,XML Schema 的书写方式跟 XML 文档是一样的,它的定义方式和数据库系统的定义方式很相似。在 XML Schema 中不但有数据类型的定义,如 <xsd:element name="title" type="xsd:string"/>,还有数据约束的定义,如 <xsd:element name="item" maxOccurs="unbounded">。下面所讲述的 XML 文档到关系数据库的映射就是以此模式文档作为示例。

2.2 XML 文档到关系数据库的映射

2.2.1 元素表的建立

在映射过程中,首先生成一些元素表^[4],这些表被

所有的 XML 模式所共享。主要是用来存放 XML 模式的信息,比如元素间的父子关系,同级元素的顺序,是否为属性等信息。这些信息在关系模式转换到 XML 模式中的应用也十分重要,在转换过程中,元素的顺序就不会产生混乱。以前面的模式文档为例,建立的元素表见表 1。

表 1 元素表

Ele_ID	Ele_Name	Parent_ID	Sequence	IS_Arrt	In_Par	Is_Nest
1	organizations	0	1	false	false	false
2	default	1	0	true	true	false
3	organization	1	1	false	false	false
4	Identifier	3	0	true	true	false
5	title	3	0	true	true	false
6	item	3	1	false	false	true
...

其中, Ele_ID 由系统自动生成,是元素的唯一标识符; Ele_Name 是元素名称; Parent_ID 是父元素的 ID 号,没有父元素,则为 0; Sequence 是同级元素的顺序号,属性没有顺序,为 0; IS_Arrt 是布尔类型,描述是否为属性; In_Par 描述是否存放在父元素列表中, Is_Nest 描述元素是否为递归嵌套元素。

2.2.2 属性的映射

属性为简单类型,且只能在元素中出现一次,直接映射为父元素表中的同名列^[5,6]。若属性是可选的,则对应映射字段可为空,否则不能为空。

2.2.3 简单类型的映射

简单类型元素分为单值元素(在文档只出现一次)和多值元素(在文档中至少出现两次)。若为单值元素,则映射为表中的一列,否则,创建新表,指定 Ele_ID 为主键,并在新表中添加其父元素表的主键为外键。

2.2.4 复杂类型的映射

每一复杂元素映射为单独的数据表^[7~9],表名跟元素名同,同样,在新表中添加其父元素表的主键为外键,其包含的简单类型的子元素和属性映射为表中的同名列。若元素没有属性,则指定 Ele_ID 为主键,否则,表的主键采用属性列组合。

2.2.5 递归嵌套元素的映射

在远程教育技术规范的 XML 绑定文档中经常会出现同名元素的多层嵌套,若是在映射过程没处理好,很容易出现数据混乱问题。在同名元素的数据映射表中设立一个标识符来表明元素的层次嵌套情况,这样就避免了从数据库中恢复数据到 XML 文档中时产生混淆。在数据表中添加 layer 标识符,表明元素是第几层嵌套。

按照以上的映射规则,上述的内容包装规范的 XML 文档可映射为以下的数据表(见表 2~表 4):

表 2 organizations

organizations - ID	default
1	TOC1

其中, organizations_ID 是 organizations 的唯一标识符,由系统自动生成, default 是 organizations 的属性列。

表 3 organizations

organization - ID	organizations - ID	identifier	title
1	1	TOC1	default

其中, organization_ID 是 organization 的唯一标识符,由系统自动生成, organizations_ID 是 organization 的父元素的 ID 号, TOC1 和 default 是 organization 的属性列。

表 4 item

item - ID	organization - ID	ParentItem - ID	identifier	identifierref	title	layer
1	1	0	ITEM1	RES1	Lesson 1	1
2	0	1	ITEM2	RES2	Introduction 1	2
3	0	1	ITEM3	RES3	Content 1	2

这是递归元素的数据表, item_ID 是 item 的唯一标识,与前者的数据表不同,递归元素的第二层 item 元素所指向的父元素与其本身同名,不能直接用 item_ID 指向其父元素,因为数据表中不能有同名属性。所以多增加了 ParentItem_ID 指向其递归元素的父元素 ID,若元素为第一层, ParentItem_ID 为 0,否则指向上一层元素的 ID 号。在数据表中,还增加了 organization_ID 指向第一层 Item 的父元素 ID,若为 0,则表明 Item 的父元素不是 organization。layer 字段说明的是同名元素的嵌套层数,表明元素是第几层嵌套。

3 结束语

目前,关系数据库系统已相当成熟,文中提出的将远程教育技术规范的 XML 绑定文档存储到关系数据库中,可以很好地处理数据的层次性、顺序性,使得规范在推广与应用中产生的大量 XML 文档的存储得到有效管理。针对远程教育技术规范的 XML 绑定文档中经常出现同名元素的多层嵌套问题,文中提出将此同名元素映射到数据表中时添加 Layer 列,表明元素的嵌套关系,确保了数据的顺序性。XML 文档到关系数据库的存储技术在远程教育技术规范中的应用,使得远程教育技术规范更易于管理和推广。

参考文献:

- [1] 裴伟廷. XML 与远程教育信息标准化[J]. 河北广播电视大学学报, 2005, 10(1): 6-8.
- [2] 王国仁, 于戈, 杨晓春, 等. XML 数据库管理技术[M]. 北京: 电子工业出版社, 2007: 27-40.
- [3] 万常选. XML 数据库技术[M]. 北京: 清华出版社, 2005: 1

(下转第 136 页)

$\text{Core}(A) = \{d, e\}$ 。

在 S 中分别删除 a, b, c 所在的列, 均不会产生新的重复的 n 元泛值, 因此, a, b, c 是可约的。而删除属性 d 所在的列, 会出现新的重复的 n 元泛值 $(A - \{d\})(X_{14})$, 使得 $(A - \{d\})(X_{14}) \equiv (A - \{d\})(X_7) \equiv (A - \{d\})(X_{10})$ 。同理, 删除属性 e 所在的列, 也会出现新的 n 元泛值 $(A - \{d\})(X_6), (A - \{d\})(X_{13}), (A - \{d\})(X_{15})$ 使得 $(A - \{d\})(X_{13}) \equiv (A - \{d\})(X_{15})(A - \{d\})(X_6) \equiv (A - \{d\})(X_9) \equiv (A - \{d\})(X_{11})$, 所以按定义 5 及性质 2 可知 d 与 e 是不可约的或是必要的, 它们组成属性 A 的 $\text{Core}(A)$ 。结果与按原定义所求结果相同。

通过计算有 $U/\text{ind}(P) = U/\text{ind}(A - \{e\})$ $U/\text{ind}(Q) = U/\text{ind}(A - \{d\})$, 根据正区域定义有: $\text{pos}_P(Q) = \{X_1, X_3\} \cup \{X_2, X_5\} \cup \{X_4\} \cup \{X_7, X_{10}\} \cup \{X_8, X_{12}\} \cup \{X_{14}\}$ 。根据新定义可知 Q 的 P 正区域实质是 n 元泛值 $P(x)$ 与 $Q(x)$ 满足一对一映射关系的所有对象 x 的集合。 $P(X_1) \equiv P(X_3) = (1, 1, 1)$ 并且 $Q(X_1) \equiv Q(X_3) = (1, 0)$, 因此有 $X_1, X_3 \in \text{pos}_P(Q)$, 同理 $X_2, X_5, X_7, X_{10}, X_8, X_{12} \in \text{pos}_P(Q)$, $P(X_4) = (2, 3, 3) \neq P(X_i) (i \neq 4)$, 因此一定满足一对一的映射关系, 所以 $X_4 \in \text{pos}_P(Q)$ 。同理也有 $X_{14} \in \text{pos}_P(Q)$ 。由于 $P(X_6) \equiv P(X_9) \equiv P(X_{11}) = (3, 1, 1)$, 而 $Q(X_6) = (2, 0), Q(X_9) \equiv Q(X_{11}) = (2, 1)$ 。由于不满足一一映射关系, 因此有 $X_6, X_9, X_{11} \notin \text{pos}_P(Q)$ 同理 X_{13}, X_{15} 也不属于 Q 的 P 正区域。

通过以上分析知道, 求正区域的两种计算方法所得结果是一样的, 但是时间复杂度是不同的。原计算方法需要求两次划分和一次交集运算, 时间复杂度为 $O(|A| |U|^2)$; 而新的计算方法如果先对数据表进行排序, 然后再扫描一遍数据表即可, 时间复杂度为 $O(|U| \log |U|)$, 显示出了它的高效性。

由于相对核、相对约简、属性依赖、属性重要度都是建立在正区域基础上的, 所以计算正区域的时间复杂度的降低对于基于粗糙集的属性约简具有重要的意义。

义。

4 结束语

文中在深入分析粗糙集中核心概念的基础上, 从广义映射的观点对相关的概念进行另一种形式的定义, 并给出了基于这种定义形式的相关性质。从理论上证明了它们与原定义的等价性, 并给出了关于正区域及属性重要度的改进的计算方法。从理论上分析了它们的有效性, 使粗糙集能更好地应对大数据集的挑战, 这对粗糙集的应用推广无疑是一种有益的尝试。

参考文献:

- [1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Science, 1982, 11(5): 341 - 356.
- [2] Wang J, Miao D Q. Analysis on attribute reduction strategies of rough set[J]. Journal of Computer Science and Technology, 1998, 1(32): 189 - 192.
- [3] William Zhu, Wang Fei-yue. Reduction and axiomization of covering generalized rough sets [J]. Information Science, 2003, 152: 217 - 230.
- [4] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 681 - 684.
- [5] 杜晓, 刘维亭, 杜茜, 等. 基于粗糙集理论与灰色理论的属性约简算法[J]. 计算机技术与发展, 2008, 18(1): 154 - 156.
- [6] 覃伟荣, 秦亮曦. 基于粗糙集理论的条件属性动态约简算法[J]. 计算机技术与发展, 2008, 18(8): 23 - 25.
- [7] 马光志, 吴黎明. 基于粗糙集理论的一种属性约简算法[J]. 计算机工程与应用, 2006(18): 171 - 175.
- [8] 吕跃进, 刘南星, 陈磊. 一种基于并行遗传算法的粗糙集属性约简[J]. 计算机科学, 2008, 35(3): 219 - 221.
- [9] 刘少辉, 盛秋骥, 吴斌, 等. Rough 集高效算法的研究[J]. 计算机学报, 2003, 26(5): 524 - 529.
- [10] 张文修, 吴伟志, 梁吉业. Rough 集理论与方法[M]. 北京: 科学出版社, 2001.
- [11] 左孝凌, 李为鉴, 刘永才. 离散数学[M]. 上海: 科学技术出版社, 1982.

(上接第 132 页)

- 4.

- [4] 谈子敬, 陈宇达, 施伯乐. 基于模式的 XML 文档关系数据库存储[J]. 小型微型计算机系统, 2003, 24(7): 1231 - 1234.
- [5] 余贞斌, 王新伟. XML 数据在关系数据库中的存储[J]. 微机发展(现更名: 计算机技术与发展), 2005, 15(11): 120 - 122.
- [6] 吴永春. XML 数据存储方法研究及应用[J]. 计算机技术

与发展, 2006, 16(2): 139 - 141.

- [7] Florescu D, Kossmann D. Storing and querying XML data using an RDBMS[J]. IEEE Data Engineering Bulletin, 1999, 22(3): 27 - 34.
- [8] Manolescu I, Florescu D, Kossmann D. Pushing XML queries inside relational databases[S]. NRIA. 2001.
- [9] Bonifati A, Ceri S. Comparative analysis of five XML query languages[J]. SIGMOD Record, 2000, 29(1): 68 - 79.