

XML 存储方案评述

董泉伶, 郝春辉, 王江涛

(淮阴师范学院 计算机科学系, 江苏 淮安 223001)

摘 要: XML 近来已成为 Internet 领域数据交换、数据表示的标准。随着 XML 的广泛应用, 基于 XML 的数据交换数量呈现出指数增长的趋势。一个设计良好的 XML 存储方案是可靠、有效的存储、查询、操作 XML 数据的重要基础。为了更好地解决 XML 数据的存储问题, 文中阐述了目前 XML 文档在数据库中的各种存储技术。对于各种典型的 XML 存储方案进行了分类, 并对这些方案进行了评述。对模式映射和模型映射存储技术进行了一定的研究。

关键词: XML; CLOB/BLOB; NXD; 模式映射; 模型映射

中图分类号: TP311.13

文献标识码: A

文章编号: 1673-629X(2006)12-0150-03

Comments of XML Storage Scheme

DONG Quan-ling, HAO Chun-hui, WANG Jiang-tao

(Dept. of Computer Science, Huaiyin Teachers College, Huaian 223001, China)

Abstract: XML has recently become the standard for representing data and exchanging data among applications on the Internet. With the development of the XML, the data exchange has an upward trend in quantity. A well designed XML storage system is very important to store, query and operate these data. In order to solve the storing problem of XML data, this paper has a comment upon all kinds of storing technologies of XML documents in database. It briefly surveys some important XML storage schemes and makes comments upon them. Also studies schema mapping and model mapping.

Key words: XML; CLOB/BLOB; NXD; schema mapping; model mapping

0 引言

XML (Extensible Markup Language) 是一种可扩展标记语言。它是标准通用标记语言 (Standard Generic Markup Language, SGML) 的一个子集, 近年来已经成为数据交换、电子商务和 Internet 等领域内应用程序数据表示和交换的标准^[1]。

随着 XML 的广泛应用, 海量的数据被存储在 XML 文档中。如何可靠和有效地存储、查询、管理这些数据变得越来越重要。一个设计良好的 XML 存储方案是查询、管理 XML 数据的重要基础, 在 XML 的应用中占有相当重要的地位。

文中主要对当前 XML 存储方案的研究现状进行综述。

1 XML 文档特点

XML 本身是一种可扩展的半结构化的层次结构, 与经典的数据库模型有较大的差异, 应用经典数据库理论 (主要是关系理论) 解决 XML 的存储问题具有一定的难

度。另外, 不同类型的 XML 文档在一定程度上也会影响存储方案的设计。目前人们一般倾向于将 XML 文档分为三类^[2]:

(1) 以数据为中心的 XML 文档, 其特点是结构规范、数据颗粒度好、很少或没有混合内容, 同层次元素和 PCDATA 的出现顺序并不很重要, 如销售订单、时刻表和菜单等。这种类型的 XML 文档比较适合用分解存储方式进行存储。

(2) 以文档为中心的 XML 文档, 其特点是结构不规范、数据颗粒度大以及包含大量混合内容, 同层次元素和 PCDATA 的出现顺序非常重要, 典型的例子包括电子书籍、电子邮件等。这种类型的 XML 文档比较适合用整体存储方式进行存储。

(3) 混合型文档, 它兼有以上两者的特点。

虽然以数据为中心和以文档为中心的分类法并没有一个非常清晰的界限, 但是, 这种分类确实存在, 对于存储方案的选择和设计也相当重要。

2 XML 存储方案研究现状

XML 的存储问题是 XML 数据库的基础和焦点问题, 要想建立一个高性能的 XML 数据库必然要求一个高性能的 XML 存储方案。目前 XML 的存储方案大致可以分为以下五类: 文件系统方式、CLOB/BLOB 方式、NXD

收稿日期: 2006-03-21

基金项目: 淮阴师范学院青年教师基金项目 (05HSQN076)

作者简介: 董泉伶 (1979-), 女, 山东烟台人, 助教, 硕士研究生, 研究方向为软件工程。

数据库、模式映射方式和模型映射方式。

2.1 文件系统方式

用文件系统存储和管理 XML 数据,是最简单最方便的一种方式。但是,由于文件系统不支持诸如并发访问、访问控制、复杂查询等数据库特性,以及存在数据冗余等问题。这种方式比较适合较小规模的使用,在比较正式、复杂和大型的 XML 应用场合上不多见。

2.2 CLOB(Character Large Object)/BLOB(Binary Large Object)方式

这种存储方案是将一个 XML 文档作为关系数据库的一个字段进行存储。大多数流行的数据库(Oracle 10g^[3], Informix^[4], etc)都支持该方式。

以上两种方案具有以下特点:

(1)在逻辑上,都把单个的 XML 文档作为一个存取单位,其理论模型相当简洁,技术上也很成熟。对任意类型的 XML 文档都可以无损地存储,能够很好地完成存储要求。

(2)在实现上,技术简单,容易实现。

(3)在查询上,当需要对 XML 文档进行基于内容和结构查询的时候,系统需要对众多的 XML 文档逐个进行解析,然后判断该文档是否满足查询条件。当系统存储规模和并发访问量比较大的时候,系统的性能将非常低。解决这个问题一个可行途径是索引技术。但是,一般而言,索引技术并不能解决全部类似的问题。

(4)在实际应用中,由于其简洁易行,不易出错,虽然有性能较低的缺点,但仍然得到相当广泛的使用。

2.3 NXD(Native XML Database)数据库

NXD 是专门设计用于存储和管理 XML 文档的数据库, NXD 可以描述为:

(1)为 XML 文档而不是文档中的数据定义了一个逻辑模型,并根据该模型存取 XML 文档。该模型至少应包含元素、属性、PCDATA 和文档顺序。模型包括 XPath 数据模型、XML Infoset 以及由 DOM 和 SAX1.2 事件所蕴含的数据类型。

(2)以 XML 文档作为基本逻辑存储单位。

(3)对低层物理存储模型没有特殊要求。它既可以建立在关系型、层次型或面向对象的数据库之上,也可以使用专用的存储格式,比如索引或压缩文件。

这种定义方法相当宽泛,包涵了基于关系数据库的模式映射方案。但是,一般而言, NXD 更多地是指完全独立的数据库。

NXD 数据库是最“自然”的数据库,从理论角度而言, NXD 数据库是存储 XML 的最佳方式。但是,有很多因素阻碍了 NXD 的发展:

①一般而言,一个 NXD 数据库是一个完全独立的数据库,必须从头设计和实现,有的数据库技术很难被不变更地使用,具有相当的难度和工作量。

②必须实现一个适合于 XML 存储的物理存储系统。

但同时,却不存在一个像关系数据库那样有严格理论支持、获得广泛认可的模型。

③不能很容易地重用现有的数据库技术,而又必须达到现在的数据库的一些基本要求(如并发、访问控制、事务、数据备份等)。

④XML 的查询技术尚未定型。W3C(<http://www.w3c.org/>)的 XQuery 标准已经研究多年,目前已经几次推出草稿,但正式标准尚未出现。

由于以上原因, NXD 数据库的发展有相当的难度。因此,大多数的 NXD 数据库是基于比较简单的文件系统(存在前面提及的基于内容和结构查询缺陷)。少数方案是 Element - Based, 如 Lore system^[5]; TIMBER^[6] 或者 Subtree - Based, 如 Natix^[7]。而在这少数方案中能够支持 XML 模式(尤其是 XML Schema)的则更为少见(XML Schema 是 2001 年 5 月推出的,相当复杂)。

2.4 模式映射方式

模式映射是一种最早被研究的、用于存储 XML 的技术,有相当多的方案,现在仍被广泛地应用于数据提取^[8]、数据交换、数据发布等领域。

模式映射的过程大致如下:

(1)被存储的 XML 文档要与某个模式相匹配。

(2)通过分析 XML 模式,获取 XML 的逻辑结构,然后,利用算法将该逻辑结构映射到关系结构。用数学语言表达为: $relationSet = SchemaMapping(schema)$ 。

(3)最后,从符合该模式的 XML 文档中,提取数据,并存放到相应的关系表中去。

模式映射方案的特点:

①SchemaMapping 函数不是一个可逆映射,这意味着在分解 XML 文档的过程中,会存在某些信息丢失的情况。因此,在需要完整保存 XML 文档的场合,这种方案不是很适合。

②模式映射方式更适合于从 XML 文档中提取数据,而后将数据存储到关系数据库中,从而达到以关系的方式来管理数据的目的。这种方式更倾向于将它们处理对象,称为 XML Data,而不是 XML Document。

③在映射过程中,关系表不是预定义的,而且产生表的数目也不确定。如果 XML 的嵌套深度很深,那么可能产生大量的关系表。对于一些字段名,可能需要重新定义,以防止重复。XML 定义的元素、属性顺序信息在映射后一般不予以保证。注释、处理指令等信息则可能被遗弃。这些特点决定了模式映射方案比较适合于处理以数据为中心的 XML 文档。而对以文档为中心的 XML 文档则有些力不从心。

④由于模式映射方案将 XML 的结构进行了拆分,又可能存在重新命名字段名的情况,因此,比较难于支持基于结构的 XML 查询。

⑤目前,模式映射方案大都是基于 DTD 模式的,很少有方案支持 XML Schema(XML Schema 比 DTD 更为复杂

和正式)。这主要是因为,在大多数情况下,DTD 即可满足数据提取的需要,而无须支持更为复杂的 XML Schema。

⑥目前,模式映射方案在数据提取、异构系统之间进行数据交换的场合下具有较为广泛的应用。

2.5 模型映射方式

模型映射方式的过程大致如下:

(1)首先要定义一个类似于 DOM 的通用 XML 文档模型。一般而言,这个模型是一个树状结构。

(2)其次,需要设计一个(或者多个)能够存储这个模型的关系模式。

(3)最后,需要提供一个算法能够将具体的 XML 文档转换为该模型的实例。并将该实例分解存储到数据库相应的表中去。

该方式与文件和 CLOB/BLOB 方式有类似之处:在逻辑上,它们都将 XML 文档视为一个基本存取单位。但是模型映射中的 XML 文档是经过解析的高度结构化的数据,关系数据库可以把这些结构信息保存下来。文件和 CLOB/BLOB 方式中的 XML 文档则是未经过解析的文本块。这就决定了模型映射可以支持基于结构和类型的查询。

该方式操作过程如图 1 所示。

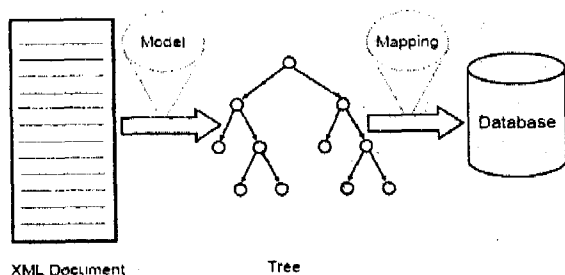


图 1 模型映射方式原理图

模型映射方式的特点:

- ①映射后,表的数目和字段都是确定的。
- ②可以保证数据的完整性,不会丢失数据。
- ③可以支持基于内容和结构的查询。

④对无模式的文档也适用。

另外,由于模型映射完全不涉及到模式;不能充分利用 XML 模式带来的优点;而且由于大多数模型抽象度比较高,存储模式相对呆板,在其上实现查询语言也有一定的难度。

3 结束语

XML 存储系统是 XML 数据库的关键。但是,由于 XML 数据本质上是一种自描述的半结构化数据,它的数据模型不同于以往的层次、关系、面向对象数据模型,已有的数据库技术和查询语言,不能完全适应于新的应用需求。目前,还没有一个能够得到广泛接受和认可的 XML 存储方案,所以有必要对该问题进行更深一步的研究。

参考文献:

- [1] 胡锡伟,陈仲委. Oracle 数据库的 XML 存储技术研究[J]. 计算机工程与设计,2005(5):179-181.
- [2] 徐德智,吴敏,赖同庆. XML 模式、查询和存储技术扫描[J]. 计算机工程与科学,2003(3):22-25.
- [3] ORACLE XML DB (An Oracle Technical White Paper [EB/OL]. 2004. <http://www.oracle.com/technology/tech/xml/xmlldb/Current/NewFeatures.pdf>.
- [4] IBM Informix Dynamic Server Getting Started Guide [EB/OL]. 2003. <http://www-306.ibm.com/software/data/informix/pubs/library/interim/ct1t1na.pdf>.
- [5] McHugh J, Abiteboul S, Goldman R, et al. Lore: A Database Management System for Semistructured Data [J]. SIGMOD Record, 1997, 26(3):54-66.
- [6] Jagadish H V, Shurug AL - Khalifa. TIMBER: A Native XML Database [R]. USA: University of Michigan, 2002.
- [7] Kanne C C, Moerkotte G. Efficient Storage of XML data [C]//In Proceedings of 16th ICDE. San Diego, California, USA: [s. n.], 2000.
- [8] 许卓明,刘琴,董逸生. 基于关系数据库的 XML 存储技术评述[J]. 计算机工程与应用,2003(21):197-201.

(上接第 149 页)

着应用环境的不断改变,这个课题也需要不断发展,如何采用合理的技术实现适应当前环境的数据集成将是一个重要的问题。

而近几年来来的发展趋势证明,XML 具有的简单性、规范性、开放性、灵活性和可扩充性等优点,能够有效地实现不同领域异构资源的集成,因此文中提出的 XML 技术在动车段(所)信息集成共享中的应用模式,能使动车段(所)内或动车段(所)间多种业务应用子系统、多种异构数据源并存,并实现异构数据源的动态及时、互访和信息的综合利用,能够满足组织低成本、阶段性、可扩展性动车段(所)信息系统建设的需要。

参考文献:

- [1] 王忠群,谢晓东. 基于 XML 的异构数据源的集成研究[J]. 安徽工程科技学院学报,2003,18(4):37-43.
- [2] 李伏欣. 铁路信息共享平台技术初探[J]. 中国铁道科学, 2002,23(5):29-35.
- [3] 孟小峰,周龙骧,王珊. 数据库技术发展趋势[J]. 软件学报,2004,15(12):1822-1836.
- [4] 李军怀,周明全,耿国华,等. XML 在异构数据集成中的应用研究[J]. 计算机应用,2002,22(9):10-12.
- [5] 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.