信息存储 Web XML 关系数据库 (13)

计算机科学2000Vol.27№.11

基于 XML 的 Web 信息存储解决方案

A XML-based Web Data Storage Solution

49-52 TP333

姚卿达 陈宁琳
(中山大学软件所 广州510275)

Abstract The emergence of XML makes much developing chance for Internet data storage and data exchange. There are two main solutions for XML data storage: semistructure database and relational database. We compare these two main solutions and present the realization of XMLDBMS, a middleware for data transformation between database and XML files. As an application of XMLDBMS, we introduce the concept of virtual XML file system and its realization.

Keywords XML, DTD, SAX, DOM, Semistructure database, Virtual XML file, Virtual XML file system

基于 Internet 数据的大量涌现，而且 Internet 上的数据不是一种规则的、有结构的数据。这种数据被称为半结构化数据。半结构化数据可能有以下特征：

·数据是不规则的，不符合严格的模式。半结构化数据是传统的数据库难以管理的。在关系模式中，处理非规则数据的方式是用空值代替。在面向对象数据库系统中，虽然复合类型、继承机制提供更多的灵活性，但设计出合适的面向对象模式来容纳非规则数据仍然是困难的；

·难以预先定义单一、合适的模式。数据结构可能常常改变，数据元素可能经常改变类型，新的数据类型不断引入。所以导致模式的阶段性变更，很难对半结构化数据预先设计一种单一的、正确的模式。同样，这种问题也是传统关系数据库难以解决的问题。

HTML 网页需要经过特殊的页面分析处理器的处理，才能被有效地查询。而且，查询仅局限于简单的关键字查询(基本由搜索引擎提供)，所有的 Web 文档被看作是字符串流，而没有理解出内在语义，

XML(eXtensible Markup Language，可扩展标记语言)的出现似乎为上述问题的解决提供转机。XML 是 Web 上新兴的文本语言，在数据表现和数据交换上越来越受欢迎，被誉为构造未来 Web 的新工具，HTML 着重描述 Web 页面的显示格式，而 XML 着重描述的是文档的内容。XML 的特长在于描述层次结构的数据，或赋予原本杂乱的信息一种清晰的结构。

对于 XML 数据的数据库存储技术，主要有两种解决方案：

1. 利用 XML 与半结构化数据的相似性，在半结构化数据的研究成果上支持 XML 的存储和查询功能。这主要有 Standford 大学 Lore 数据库研究；

2. 利用传统关系数据库技术，实现 XML 文档到关系数据库记录之间的转换，实现利用关系数据库存储、查询 XML 数据的功能。

一、XML 文档结构到数据库结构的影射

为了在数据库和 XML 文件之间相互传输数据，需要进行 XML 文档结构和数据库结构之间的相互影射。这种影射分成有两大类：模式驱动和模板驱动。

①模板驱动影射：文档结构与数据库结构之间没有预定义的影射关系，而是在模板中嵌入数据库执行指令，该指令由数据转换中间件来处理。模板影射提供很大的灵活性，例如，允许嵌套查询，也可以支持编程流程，例如 loop 循环和 if 语句。

②模型影射：为 XML 文档结构建立数据模型，该数据模型显式或隐式地影射数据库结构，因为数据从数据库到 XML 文档的转换限制到单一模型，XSL 通常被集成到模型驱动的产品，以提供模板影射所提供的灵活性。该影射通常将 XML 数据看作一个对象树，元素通常对应于对象，元素属性和 PCDATA 对应于对象属性。这种模型将 XML 直接影射为面向对象和层次数据库。如果利用传统面向对象影射技术或者 SQL3对象视图可以将 XML 影射为关系数据库，注意这个模型不是文档对象模型(DOM)；DOM 为文档本身建模型，而不是文档的数据。我们下面的讨论基于模型影射技术。

二、基于半结构化数据库的 XML 数据存储

2.1 半结构化数据库的体系结构

API 层：实现对半结构化数据库的存取。简单文本界面通常由系统开发者使用。图形界面作为用户的基本界面，提供查询、查看数据库结构等功能。

查询汇编层：分析器接收一个以文本表示的查询，将它转换为分析树，然后将分析树传送给预处理器。预处理器将半结构化数据查询转换为类似 OQL（对象查询语言）的查询。查询方案生成器从已转换的查询产生方案，再将方案传送给查询优化器。查询优化器除了对方案进行简单的转化外，还检查索引的使用是否合理，然后将已优化的查询方案提交给数据引擎。

数据引擎层：查询操作器执行已生成的查询方案；对象管理器作为半结构对象模型和低层文件系统的翻译工具，它支持基本功能，包括返回一个对象，比较两个对象，处理简单压缩，遍历复杂对象的子对象。还有性能改善的功能，比如对经常访问的对象提供缓存。

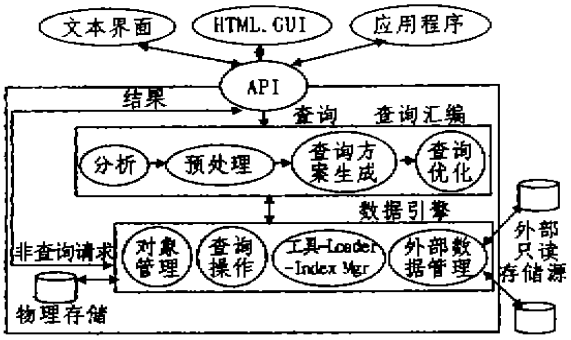

图1

2.2 XML 的半结构化数据模型

XML 模型是在半结构模型基础上建立的。在该模型中，一个 XML 元素表示为〈eid，value〉，其中 eid（Element Identifier，元素标识号）是全局唯一的。value 分为原子值和复合值。原子值为字符串，复合值包含四部分：①标识（tag），对应于 XML 文档的 Tag；②attribute-name/atomic-value 有序序列。attribute-name 是字符串，atomic-value 的类型是 integer，real，string，ID，IDREF，IDREFS 原子类型之一；③〈label，eid〉有序序列，表示结点引用关系，其中 label 是字符串，表示结点间关系，eid 是引用结点号。引用结点在 XML 文档中体现为 IDREF 或 IDREFS 属性；④〈label，id〉有序序列，表示父子结点关系，其中 label 是字符串。表示父子结点间关系。父子结点在 XML 文档中体现为元素嵌套。

XML 半结构模型有两种边：一般边和引用边。①一般边由父结点指向子结点，并标上子结点标识（tag）；②引用边由引用结点指向被引用结点，并标上引用名。

在半结构化数据库中，没有固定模式的概念。所有模式信息包括在标识（label），这些标识可以动态改变。所以，半结构化数据库在数据之上没有强加任何的规则。半结构化数据模型主要用于处理数据的不完整性，数据库结构和类型的异质性。

2.3 XML 文档到半结构化数据模型的影射

从 XML 文件到半结构模型的转换规则是：

1. XML 标识（Tag）被转换为结点标识（Tag）；

2. 标识（Tag）之间文本被转换为原子文本结点；

3. 元素属性列表转换为 attribute-name/atomic-value 序列；

4. 对于 XML 文档中 IDREF 类型的属性 i，或者 IDREFS 类型的属性部件 i，增加一个引用结点（label，eid），其中 label 是相应的属性名，eid 是结点号，其 ID 属性值为 i；

5. XML 文档中子元素转换为相应子结点、子元素标识转换为子结点 label。

三、基于关系数据库的 XML 数据存储

除了在半结构化数据的成果基础上进行 XML 存储的研究以外，利用商业关系数据库来存储 XML 文档的是另外一个发展方向。而且，这个发展方向得到更多的重视。

实现这种技术的关键在于 DTD（Document Type Definition）的存在。DTD 用于描述 XML 文件结构，如果没有 DTD，XML 将永远不能发挥它的潜力。如果应用程序之间对标识的含义没有共识的话，那么 XML 也失去了作为数据交换的意义。

由于 DTD 允许多层嵌套，这与关系模式的二维特性存在冲突。而且，由于 DTD 中存在集合型属性和递归，这些都是实现 DTD 向关系模式转化前必须解决的问题。

简化 DTD 解决 DTD 的多重嵌套问题，简化公式如下：

转化嵌套定义为非嵌套定义	简化多个操作为单一操作	同类操作归组
$(e_1,e_2)^* \rightarrow e_1^*,e_2^*$	$e_1^{**} \rightarrow e_1^*$	$\cdots,a^*,\cdots,a^*,\cdots \rightarrow a^*,\cdots$
$(e_1,e_2)? \rightarrow e_1?,e_2?$	$e_1^{*?} \rightarrow e_1^*$	$\cdots,a^*,\cdots,a?,\cdots \rightarrow a^*,\cdots$
$(e_1 \mid e_2) \rightarrow e_1?,e_2?$	$e_1^{?*} \rightarrow e_1^*$	$\cdots,a?,\cdots,a^*,\cdots \rightarrow a^*,\cdots$
	$e_1?? \rightarrow e_1?$	$\cdots,a?,\cdots,a?,\cdots \rightarrow a^*,\cdots$
		$\cdots,a,\cdots,a,\cdots \rightarrow a^+,\cdots$

由 DTD 产生关系模式要解决几个问题：关系数据库模式是由实体联系图而来，由实体联系图生成数据库模式是很直观的，因为在实体联系图中，实体和属性之间有天然的分界。然而，DTD 中的元素、属性与 ER 图中的实体、属性没有直接的对应关系，也就是说，ER 图中的属性往往在 DTD 中被表示为元素，另外，传统关系模型不支持集合值属性，例如 DTD 片段：(! ELEMENT article (title, author*, contactauthor?))，我们不能把 author 集合归为同一关系，我们遵循传统 RDBMS 存储集合的标准方法，创建一个 author 关系，并且利用外键把 author 关系和 article 连接起来，在 XMLDBMS 影射语言中详细介绍。

四、XMLDBMS 的实现

XMLDBMS 是用 Java 实现的、XML 文档和关系数据库之间数据转换中间件，XMLDBMS 实现的关键在于对象树和 MAP 对象的构造。MAP 对象初始化时设置数据库连接，获得数据库接口。然后，分析输入 DTD 文件，根据影射规则生成 DTD 文档的对象视图，将该对象视图存储到 map 文件中，以后，数据库与 XML 文件之间的转换过程中，系统将会读取 map 文件，建立 XML 文件结构与数据库模式的对应关系，进行数据转换。

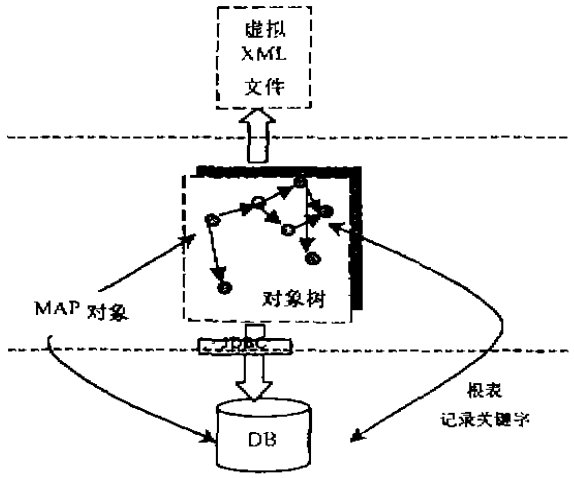

图2

4.1 XML 文档的对象视图和影射语言

XMLDBMS 把 XML 文档看作对象树，利用"对象→关系模式"影射把对象影射到关系数据库中。XMLDBMS 影射语言将描述怎样为 XML 文档创建对象视图，怎样把这数据视图影射为关系模式。在影射的过程中，引入面向对象的概念，将被影射为表的元素看作为类，将子元素或属性看作类的属性。XML 文件由以下建筑块组成：元素，标识，属性，PCDATA，CDATA。

·父元素　带有子元素的元素通常被视作类，被影射到表。

·只含 PCDATA 或 CDATA 的子元素　只带有 PCDATA、CDATA 内容的元素将被看作属性，被影射到父元素的一列。

·属性　元素属性看作类的属性，被影射到元素对应表的某列。

·最少出现一次、出现零次或多次、出现零次或一次的元素　将出现零次、一次、或多次的子元素看作子表，并用外键与主表相连。例如 DTD 片段：(! ELEMENT Para (#PCDATA | Link)*)，其影射图表示为图3。

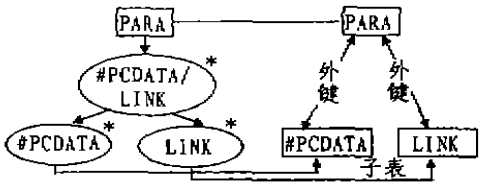

图3

·声明混合内容　混合内容包括 PCDATA（或 CDATA）和子元素，它们的出现顺序通常是重要的，所以我们通常需要保存 PCDATA（或 CDATA）和元素的顺序信息。在子表中增加一列保存系统产生顺序信息。

4.2 主要类实现

①public class DBMSToDOM extends Object

功能描述　根据给定 Map 将数据库中数据转换为 DOM 树。调用者必须提供所用 DOM 的 DocumentFactory，一个 Map 对象，以及关于如何取得数据的信息（表名和关键字，或 DocumentInfo 对象，或结果集）。

②public class DOMToDBMS extends Object

功能描述　根据给定 MAP 对象将 DOM 树的数据转换到数据库。调用者必须提供一个 Map 对象，一个 KeyGenerator 对象（如果要产生关键字），一个 NameQualifier（如果 XML 文件使用名域）。它返回一个 DocumentInfo 对象，含有要获取数据必须具备的表名和关键字。

五、虚拟 XML 文件系统

数据库和 XML 提供存储数据的互补方式。数据库存储数据有利于数据的有效检索，XML 表示数据有利于应用程序之间互操作的信息交换。如果所有的 Web 数据均存放到数据库中，整个互联网的数据实质是一个大的联邦数据库。如果实现了联邦数据库的存

取接口，可以简化存取 Web 数据过程。同时可以抽取用户相关的数据传输到网络，减低了网络负载。数据的安全性方面也得到加强。利用数据库视图，可以屏蔽敏感数据，或显示不同的用户视图。

新兴的电子商务更是受到企业的青睐。内部 MIS 系统与面向 Internet 用户的系统如何进行信息交互是一个有待解决的问题。XML 的出现成为解决这一问题的良机。如果 XML 存储到数据库，与传统 MIS 系统数据库信息交互有更大可能性。

数据库信息和 XML 信息常常需要相互转换，如果信息分别在数据库和 XML 文件都有存储，会带来数据一致性问题和数据冗余问题。为了解决这个问题，我们提出虚拟 XML 文件、虚拟 XML 文件系统两个概念。

定义1　根据预先建立的 XML 文件结构与数据库模式影射关系，利用 XMLDBMS 中间件从数据库生成 XML 文件，实现 XML 数据的完全数据库管理。存储在数据库中的 XML 数据称为虚拟 XML 文件。

定义2　使用数据库来管理虚拟 XML 文件，使用 XMLDBMS 完成对虚拟 XML 文档的访问，这种集成了数据库与 XMLDBMS 的系统称为虚拟 XML 文件系统。虚拟 XML 文件系统可以由图4表示。

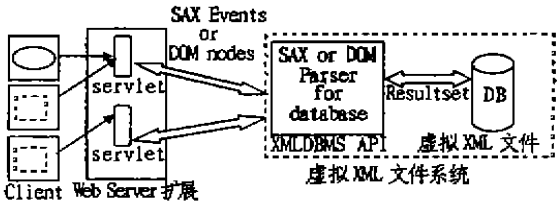

图4　虚拟 XML 文件系统

XML 存放在数据库，用户通过 XMLDBMS 直接操作于数据库，就如直接操作 XML 文档一样。从用户的角度，XML 文档和数据库之间的区别是不透明的。数据库中存储的 XML 数据可以被看作虚拟 XML 文档。对用户而言，重要的不是数据存储在何处，而是数据的操作界面。这里要改变 Web 服务器的缺省响应行为，用 Java Servlet 扩展 Web Server 的功能。扩展 Servlet 的执行流程如下：

·Server 通过 Init 函数激活 Servlet，建立数据库连接。

·初始化之后，Servlet 可以处理多个请求。每个 client 请求产生一个 service 函数调用。Service 函数创建 MAP 类和 DBMSToDOM 类的实例，调用 DOMToDBMS 类的 retrieveDocument 方法建立 XML 文档的 DOM 树。并通过 response 对象将 DOM 树传回 Server，传到客户端。

·Servlet 不断处理客户请求直到 Web 服务器调用 destroy 方法关闭 Servlet。系统回收 Servlet 资源。

结语　XMLDBMS 中间件在 SML DOM 或 SAX API 的基础上实现，是 XML 和数据库之间的数据转换接口。虚拟 XML 文件概念的提出是为了解决数据冗余和数据不一致问题。虚拟 XML 文件系统已得到实现。用户从工作站利用浏览器可以查看数据库中存放的数据。浏览器(IE5.0)显示的是从数据库中取得的 XML 格式，XMLDBMS 中间件功能还有限，只是数据库和 XML 文件之间数据转换的雏形。要真正实现实用的系统，必须改进 XMLDBMS API 的关键函数接口，以方便调用。对于如何管理 DTD 文件和 XML 对象树视图文件，本文没有涉及这方面的讨论。这也是完善本系统的一个较关键部分。

参考文献

1　陈宁琳，姚卿达. Web 信息存储的解决方案. 现代计算机，2000(2)(总85期)

2　Abiteboul S, Vianu V. Queries and Computation on the Web

3　杨冬青，裴键，唐世渭. 未来十年数据库系统研究方向. 计算机科学，1999.26(9)

4　Widom J. Data Management for XML. IEEE Data Engineering Bullentin Special Issue on XML，1999.22(3)：45～52