基于XML的数据仓库系统

XML-based Data Warehouse System

仇丽青 赵庆祯 (济南山东师范大学信息管理学院 250014)

摘 要：本文对XML语言特点及其应用于数据仓库的优势进行了分析。并在此基础上，利用XML设计了一数据仓库体系，达到了数据整合、信息整合和服务整合的目的，从而在理论上大大简化了数据仓库的查询和操作。

关键词：XML 数据仓库

1 前言

1.1 XML概述

XML即可扩展标志语言（Extensible Markup Language），是W3C设计的一组规范。XML是SGML的一个简化的严格的子集。它是特别为Web应用设计的，是针对HTML和Internet设计的标准的、可扩展的、通用的数据格式。

XML通过文档类型(DTD)定义来定义某个领域的XML词汇和准则，然后基于这个DTD来开发XML文档，这样XML文档的数据具有实际意义，同一领域的不同组织间可以相互理解彼此的XML数据，这为不同系统间交换数据提供了实现客观条件，从而网上数据交换变得更加简单。因此在互联网世界XML的用途主要有两个，一是作为元置标语言，定义各种实例置标语言标准；二是作为标准交换语言，担负起描述交换数据的作用。

而从XML派生出来的XSLT能够将数据定义和表现形式分离。XSLT是把一种XML文档转化为另一种XML文档的翻译程序，一般用来把XML消息转化为HTML文档共Browser显示，或转换成为其他显示格式文档例如提供手机显示功能。

1.2 数据仓库

W.H.Inmon认为“数据仓库是面向主体的、继承的、稳定的、不同时间的数据集合，用于支持经营管理中决策的制定过程”。数据仓库概念的提出，不但为有效地支持企业经营管理决策提供了一个全局一致的数据环境，也为历史数据、综合数据的处理提出了一种行之有效的解决方法。数据仓库具有以下的特点：面向主题的、集成的、相对稳定的、随时变化的。

数据仓库是数据挖掘能有效连续进行的条件之一，在数据挖掘循环过程中，数据仓库是一个重要的组成部分。好的数据仓库环境是数据挖掘的催化剂，这两种技术相辅相成。

随着数据挖掘研究的深入，需解决的问题和面临的挑战也很多，例如：

* 怎样从异构数据源中挖掘信息。
* 怎样表达数据挖掘结果的不同形式。
* 怎样在不同的抽象层次上进行交互的挖掘。
* 怎样解决挖掘系统之间的封闭现象。
* 怎样进行网络阻塞的控制。

2 XML应用于数据仓库的优点

数据在进入数据仓库之前的预处理是一个比较繁琐的过程，而且数据在客户端的显示也是我们不得不考虑的问题。一般的数据交换采用的是DCOM与CORBA等分布式计算技术，但这些技术都存在着一些弊端，DCOM技术是依赖于Windows平台的，它不能满足异构环境下应用的要求；而CORBA技术的体系结构庞大而复杂，对应用系统和应用环境要求较高。而XML是一种表达结构化信息的语言，虽然每个数据库描述的数据是不尽相同的，但XML可以自己定义文件标签。因此它是不同数据库管理系统之间交换信息或为瘦客户机运行的数据库应用建立前端的理想机制。近年来出现了XML数据库。XML数据库有多种表示方法，其中最流行的一种是朴质数据库（native XML datebase），其具体做法是，XML使用带标记的数据，这种标记可以

用于标志数据结构，将这种XML数据存入数据库中，此种数据库专门用以存放XML文档数据，并使用一种类SQL语言用以对XML带标记数据进行查询及其他操作。其常用的语言有Xpath,Xquery等。

将XML应用于数据仓库具有如下的优势：

* 容易实现数据在Web上的发布，XML数据可以不做任何修改就和HTML一样在网络中传输。

* 有利于数据集成，XML可以解决异构数据源之间的兼容问题。

* 可以使用丰富的方式显示数据，表现形式多样。

*支持本地数据处理，客户接收到数据后可以根据自己的需要解析数据，并作进一步编辑处理，减少网络流量，有利于信息共享。

* 可以实现数据的独立更新，使用XML后，一部分数据变化后，不需修改全部数据，也不影响数据表现形式。

有鉴于此，笔者设计了一个基于XML的数据仓库系统，从而大大简化了数据的处理及查询。

3 基于XML的数据仓库系统设计

3.1 系统设计

本系统采用的是流行的B/W/S结构，即客户端、Web服务器和数据库服务器三个层次。比起传统的C/S结构来说，这种结构具有的优点是显而易见的。这种结构相对独立，并行开发使客户端大大减肥，维护简单；它的应用逻辑由Web服务器提供，所以只需开发Web程序，无需开发客户端程序，大大缩短应用程序开发周期；安全性更强，应用逻辑和最终访问数据库大多由应用服务器实现，对用户来说是透明的，保证了系统的安全性。网络上的数据流量也大大减少。如图1所示，源数据直接进入一个XML格式转换器。XML格式转换器的思想就是通过统一访问接口和不同访问实现异构数据源互联，数据源的异构型从而被屏蔽，它可以免除应用开发者需熟悉各种数据源的麻烦，还可以改善应用的可移植性。这个转换器是由XML格式分析模块、XML格式转换模块、XML格式生成模块构成的。其中XML格式分析模块是对进入的数据进行分析，判断进入的数据的格式。XML格式转换模块则用来把其他的数据格式转换为XML格式，即用XML格式对数据进行封装，当然我们在XML格式转换模块中存储了相应的格式转换程序，并在这个模块中加入了智能搜索引擎，使其能够自动地进行格式匹配和格式转换，这个模块是整个转换器的中枢。XML格式生成模块即把格式转换结果进行整理，这是因为从数据库中生成的XML文档是一种规范格式，但在不同的应用中需要XML文档的不同表现形式。经过XML格式转换器处理过的数据，具有统一的格式，这样就大大简化了以后进行的数据抽取、转换、清洁操作。

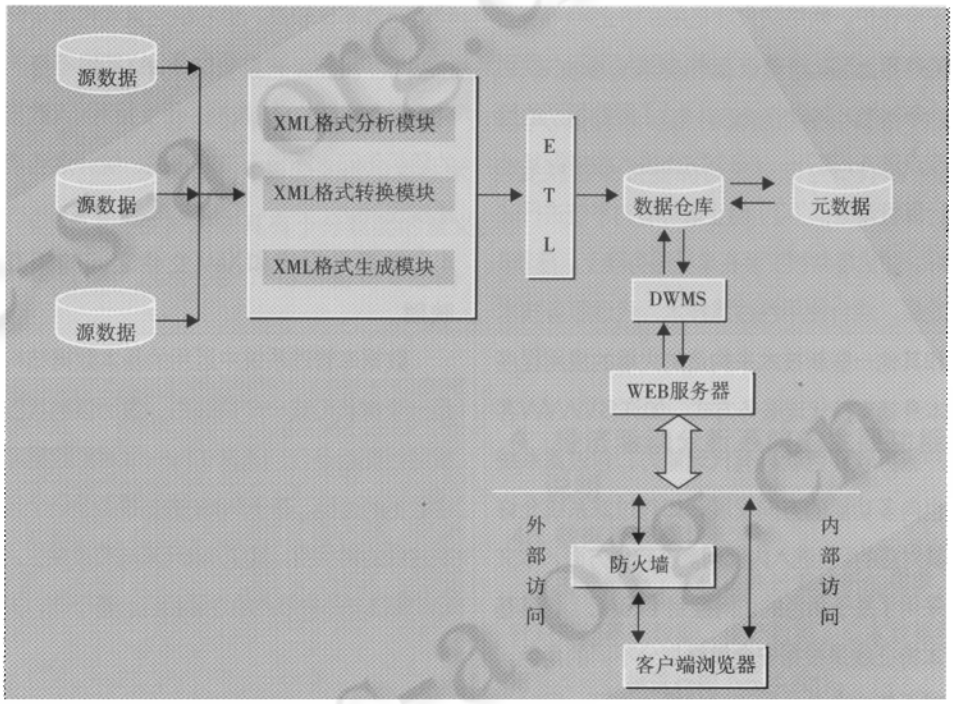

图1 基于XML的数据仓库体系设计

经过XML格式生成模块格式转换器的数据经过抽取、转换、清洁被装载入数据仓库，这时数据仓库中的数据即为统一、清洁的数据，能够被用来进行高效的数据挖掘。在客户端，采用XML接口与Web服务器进行连接（如果是外部网，则还需加上一层防火墙），用户发出的请求经过XML接口，发送给Web数据库，而Web服务器可以通过数据仓库管理系统（DWBS）直接查询数据仓库，也可以充分利用元数据库中的信息对数据仓库进行间接但更有效的查询。元数据库就像是数据仓库的地图，利用它可以大大简化操作过程。

实现Web与后台数据库动态交互的技术主要有：公共网关接口（CGI）、Internet API技术和ASP技术及JAVA和AVA Applet技术。然而XML定义的数据其显示与内容是分开的，因此允许对同一数据指定不同的显示样式，使数据更合理的表现出来。CSS和XSL为数据的显示提供了多种选择，本地数据能够以客户配置、使用者选择或其他标准决定的方式动态地表现出来。XML可以被转换为HTML、XML、WML或其他需要的格式。在这一方面XML要具有更大的优越性。

在本方案中，XML仅仅起到中间的数据表示和消息传输的作用。也就是说在源数据端，数据可以是多种格式的，比方说COBOL程序、MVS作业控制语言（JCL）、UNIX脚本、和SQL语

句等；而在客户端客户可以在不同的平台上采用不同的语言和工具进行开发，只不过是用XML来封装各个模块和接口。虽然数据在进入数据仓库之前进行了XML格式转换，但是基于XML的数据仓库体系具有很大的扩展能力，而且具有很好的开放能力，比其它数据仓库体系来说具有更大的优越性。

3.2 系统安全性设计

为了保护数据仓库及网络的安全，本文在硬件设施方面，采用了网络防火墙来进行安全保护。网络防火墙是用来防止Internet上的病毒泛滥、资源被盗用、CRACKER入侵到内部网络，其服务目的：（1）限制特别的控制点进入/离开；（2）防止侵入者接近其他计算机设施；（3）阻止破坏者对系统进行破坏。防火墙常被安装在单独服务器上。随着IT发展，混合使用包过滤技术、代理服务技术和其他一些新技术来构造防火墙的应用程序本身就支持代理服务方式，如许多WWW客户服务软件包就具有代理能力；包过滤系统也向多功能的方向发展，如包过滤系统允许其对应的IP包进入内部网。在软件方面，本文采用了基于角色的访问控制（RBAC），其基本特征就是根据安全策略划分不同的角色，对于每个角色分配不同的操作许可，同时为用户指派不同的角色，用户通过角色间接地对信息资源进行访问。具体地在本系统中，每个用户的操作权限是由系统管理员或部门负责人往下分配，每个用户只能做自己权限范围内的各项工作。用户的权限可以根据用户在工作中的具体职责进行设置，各司其职，系统的安全性高。本系统将用户分为三级：普通用户级、操作员级和管理员级。

4 系统实现中的几个关键性技术

4.1 对 XML 文档进行访问

现有的XML文档访问有两种最常用的模式：文档树模型和回调模型。这两种模型各有利弊，在实际中应根据需要进行选择。文档树模型是最早产生的XML文档访问方法，典型就是DOM（Document Object Model）。DOM解析器在内存中建立一棵层次结构的、能被任意访问的文档树，适合需要大量频繁访问的DOM文档。但是由于DOM在内存中放置了整棵树，其中包含了许多XML语法的描述性内容，而且有很多XML结构的代码，DOM内存占用量大，效率较低。回调模型的典型实现是SAX（Simple API for XML），SAX是一种事件驱动模型。由于这种接口模型是把层次结构线性化，因此基于SAX的数据处理运行效率比较高。但它缺乏直观性而给使用增加了难度，更重要的是SAX接口不在内存中保存完整的XML数据，不支持对XML数据的随机访问，不便于类似查询、统计这些功能实现。

4.2 关系数据库和 XML 文件之间的相互映射

数据库管理系统中运用的基本数据结构是一个树状图或一个曲线图，遍历该树状图或曲线图就是我们通常访问一个对象数据库中数据的过程。基于XML的数据模型是一个带有注释的树状图，其中XML元素与节点相应，XML属性与那些节点的注释相应。基于此，关系数据库的二维表结构完全可以通过XML文档来映射。因此通过XML整合不同的数据库，首先要构造一个统一的DTD文档，来实现这种一一对应的映射关系。

4.3 XML 的查询方案

如何为普通用户提供界面友好的XML文档的非专业查询，这一问题目前研究的资料还不多。在文献[1]中对国外的研究成果进行了简单的介绍，并提出一用户界面友好的XML文档查询方案，使用户只需要输入相关的已知条件和待求信息，就能够实现Web上XML文档的统一查询。

5 小结

本文作者首先对XML语言做了简要论述，然后对数据仓库的特点及当前发展中遇到的挑战进行分析，在此基础上，作者应用XML提出了一种面向Web的数据仓库设计体系，从理论上简化了对数据仓库的操作、挖掘。系统的具体实现还有许多细节需要进一步的完善。但相信XML必然将会在数据仓库体系设计中起着举足轻重的作用。

参 考 文 献

1 路燕、张彪等，用户界面友好的XML查询方案，小型微型计算机系统，2003（10），1849—1852。

2 杨冬青，把握数据挖掘新动向，中国计算机报，1999（61）。

3 廖俊松、汤宏斌等，基于XML的电子商务应用体系构建，计算机系统应用，2002（3），9—12。