第 25 卷 第 3 期 华侨大学学报(自然科学版) Vol. 25 No. 3
2004 年 7 月 Journal of Huaqiao University (Natural Science) Jul. 2004

文章编号 1000-5013(2004)03-0315-03

WEB 数据库 XML 数据发布及信息提取系统

赵美艳 王会进 张诗军

(暨南大学信息科学技术学院, 广东 广州 510632)

摘要 结合传统数据库和 XML 各自特点,在数据库到 XML 的数据转出、XML 数据发布、信息提取这三大模块的基础上,设计了一个系统原型. 同时,对研究过程中提出的关系数据库到 XML 数据的"相关转出"、XML 转出数据的 WEB 服务端处理、XML 数据的提取等问题和关键解决思路、技术方法做出阐述.

关键词 XML, 相关转出, 数据发布, 信息提取, WEB 服务

中图分类号 TP 311.132.3 文献标识码 A

数据库尤其是传统的关系数据库仍然是存储数据的主要手段,是 Internet 信息的一个主要来源. 数据库中的数据必须通过适当的数据载体才能在 Internet 传播,而 XML 则可以说是目前最适合在 Internet 上使用的数据载体. 因此,我们可以将两者结合起来,综合它们的优势,使 XML 成为数据库数据得以在 Internet 上传播的桥梁[1]. 目前对关系数据库到 XML 数据的转出,许多数据库应用软件都提供了支持,但提供的只是以 XML 格式表达的简单查询结果集返回支持. 从数据库中提取出来的数据只能面向某一特定显示或某一特定应用要求,往往不能被复用. 同时,只支持对某个表的数据的转出而不支持多个表关系的同时转出. 因此在本文中,除了提出了 SQL 直接查询支持外,还提出了按照应用表集实现基于表集的"相关转出"支持. 从而,充分发挥了 XML 的应用优势. 另一方面,数据库转出的 XML 格式的数据不能单纯地认为只是将 XML 数据拿来,然后原封不动的发送给请求者. 在本文中提出了 XML 数据发布和信息提取的观点. 它使得转出的 XML 数据在最大程度上得到复用,同时满足用户根据自己的需求提取相关信息[2].

1 系统的运行架构及其实现思想

1.1 系统的运行架构图

系统的运行如图 1 所示. WEB 服务器收到 WEB 请求后进行判断. 如果请求是 XML 静态文档,则从 WEB 文档发布目录中直接得到相应 XML 文档进行发布. 如果需要运行扩展服务程序,则通过扩展程序的接口来运行相应的外部扩展服务程序,得到结果后响应 WEB 请求. 扩展服务程序运行时,如果需要直接通过数据库访问得到 XML 数据或查询结果集,则通过数据库访问接口(如 JDBC)访问数据库. 如果使用"相关转出"的 XML 数据,则访问 WEB 文档发布目录中的相关 XML 文档并判断该文档是否有效. 无效则调用 WEB 文档生成功能来刷新 XML 文档,得到相关 XML 文档后可以经过相应的处理,再将处理结果返回 WEB 服务器. 如果是动态的"相关转出", XML 数据访问则由外部扩展服务程序直接调用 WEB 文档生成功能,得到动态的 XML 数据. WEB 文档生成功能从表关系保存文件目录中得到表关系保存文件,据此生成 XML 文档并放入 WEB 文档发布目录中或实时返回给 XML 数据请求程序. 表关系保存文件由表关系保存文件生成功能根据数据库中已经定义的表关系生成.

1.2 系统设计的特点

(1) 提出"相关转出"功能的支持. 提供了多个表数据及它们之间关系的转出. (2) WEB 服务端扩

收稿日期 2004-01-09

作者简介 赵美艳(1979-),女,硕士研究生,主要从事计算机网络和数据库原理的研究. E-mail: girlzmy@sohu.com

展程序的引入.(3) 关系数据库到XML数据的两种转出方法的支持.基于表集的相关转出和数据库直

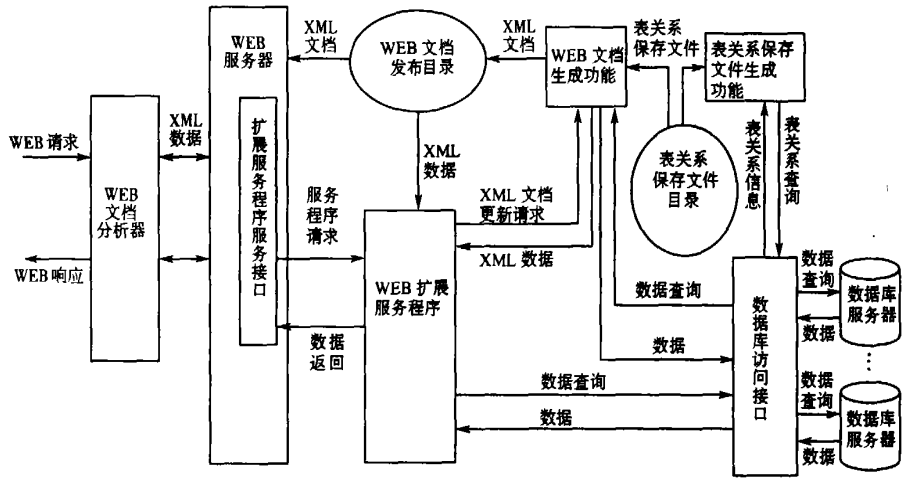

图1 XML的WEB数据库数据信息处理图

接访问转出,从而使得服务端处理程序和系统的设计更灵活,适用性更强.(4) XML数据文件的静态和动态支持.XML转出数据可以动态产生,满足一些对数据库数据更新较为敏感的应用需要.同时,转出的XML数据也可以存放为静态文件.这在一定程度上有助于WEB搜索.

1.3 系统实现的基本思想

(1) 表结构的转出.主要考虑以下方法.(a) 关系表名作为XML父元素名,列名作为子元素名.(b) 关系表名作为XML元素名,列名作为该元素的属性名.(2) 表关系的转出.表关系是通过表的主、外键来实现的.可以考虑对表的主、外键进行XML的ID/IDREF对应转换,从而达到表关系转出的目的.具体可采用"包含"和"索引"两种方法.(3) WEB扩展服务程序的实现.Servlet是Sun公司开发的规范,是Java组件.它效率更高、更容易使用、功能更强大,可移植性更好.因此本系统将采用此技术.对于Servlet运行环境采用Apache Tomcat,它是免费软件中很理想的Jsp/Servlet开发和支持平台.(4) XML文档的提取分析.主要通过XML分析器读取XML文档形成数据树,然后采用遍历树的方法来取得所有节点,并根据节点的类型做相应处理[3].

2 系统设计的关键技术

2.1 表关系保存文件生成程序的主要执行步骤

(1) 连接目标数据库.(2) 选择要涉及"相关转出"的表集.(3) 选择任意两个具有转出表关系的表,通过数据库信息得到它们之间存在的表关系.(4) 选中要转出的两个表之间的表关系或删除已经选中的表关系.如果一个表的主键仅被另一个外键所依赖,则注明表关系在XML转出文档中用子元素表示.如果一个表的主键被多个外键所依赖,则注明表关系在XML文档中用ID/IDREF表示.(5) 如果还需要转出其它表关系,则转向第3步.(6) 生成整个转出表集的表关系保存文件.

2.2 表关系规范化获取

通过java.sql包中接口DatabaseMetaData提供的方法getImportedKeys(),getExportedKeys(),getCrossReference()就可以得到指定表的主外键及它们之间关联信息.

2.3 表关系保存文件结构设计

表关系保存文件对应的DTD描述为

```
<! ELEMENT relation (Database?, Table* )>
<! ELEMENT DataSource (# PCDATA)>
<! ELEMENT table(TbName?, PkColNames?, Parents* , Sons* , MapInclude* , MapIdref* )
<! ELEMENT TbName(# PCDATA)>
<! ELEMENT PkColNames(# PCDATA)>
```

<!ELEMENT Parents(#PCDATA)>

<!ELEMENT Sons(#PCDATA)>

<!ELEMENT MapInclude(#PCDATA)>

<!ELEMENT MapIdref(#PCDATA)>

其中 MapInclude 对应于被依赖主键所在表只存在当前表外键对它的依赖, 没有其它表的外键依赖它的主键的情况. 多个则需要使用 ID/IDREF 表示的表关系元素 MapIdref, 即被依赖主键所在表除了当前外键对它的依赖外, 还存在其它表的外键依赖于它的主键.

2.4 相关转出 **XML** 文件生成程序的主要执行步骤

(1) 读取表关系保存文件, 进行初始化. (a) 依据 DataSource 元素中保存的数据库连接描述信息连接数据库. (b) 针对每个表的相关描述信息生成描述对象集 A. 根据需要可以由 A 生成相关的 DTD 或 Schema. (c) 形成 XML 文件的数据框架 S, 主要包括 XML 文件头和根元素名. (2) 遍历表对象集 A, 依次搜索到未被搜索过的没有被任何外键依赖或被多个外键依赖的描述对象 Object(即 Object 的 Parents 属性无值或有多个值). (3) 根据 Object 属性值(主要是 MapInclude 和 MapIdref) 找到和 Object 有直接表关系的表, 并根据描述信息生成并执行相关的数据库查询语句. (4) 根据数据查询结果集, 生成 Object 的 XML 转出数据. (5) 由 Object 的 MapInclude 属性值得到表集 B. 遍历 B, 依次搜索到 B 中未被搜索过的表描述对象 Object, 转向(3)步处理 Object. (6) 由 S 生成 XML 转出文档.

3 结束语

本文从数据库到 XML 的数据转出和 XML 数据发布及信息提取两个方面着手. 分析讨论构建 XML 的 WEB 数据库数据发布及信息提取系统的一些主要问题及相关解决思想. 其中, 对于关系数据库到 XML 的数据转出问题的解决, 除了提供直接查询生成支持外还提出了"相关转出"的概念和实现思路, 在转出时兼顾了"整体性"和"局部性"问题. 对于 XML 数据的发布处理问题的解决, 采用 Java Servlet 技术, WEB 服务器采用 Apache+ Tomcat 技术.

参　考　文　献

1 Seligman L, Rosenthal A. XML′s Impact on DataBases and DataSharing[J]. IEEE Computer, 37(6): 59~67

2 Ferandez M, Chiew W. Trading between relations and XML[J]. Computer Networks, 2000, 33: 723~745

3 Jasnowski M 著. Java, XML 和 WEB 服务宝典[M]. 盖江南等译. 北京: 电子工业出版社, 2002. 34~63

Data Publishing and Information Extracting System of XML-Based WEB Database

Zhao Meiyan　Wang Huijin　Zhang Shijun

(College of Info. Sci. & Tech., Jinan Univ., 510632, Guangzhou, China)

Abstract Based on data roll-out from database to ZML, XML data publishing and information extracting as three modules, a system prototype is designed. The design is also integrated with the respective characteristic of traditional database and XML. Explanations are made on such problems as "correlated roll-out" from relational database to XML data, processing of XML roll-out data at WEB server, and extraction of XML data; and also on the thinking of solving and the technology.

Keywords XML, correlated roll-out, data publishing, information extracting, WEB server