基于 XML的 Web数据集成技术的研究

冯 少 荣
(厦门大学计算机科学系 福建 厦门 361005)

摘 要 本文首先介绍了 Web环境下异构数据集成技术产生的背景和研究目的以及相关的概念、技术和方法,然后针对 XML作为合适的数据交换格式的特点介绍了基于 XML的信息集成的关键因素,基于此,提出了一种基于 XML的 Web数据集成操作模型,并讨论了该模型在 Web数据集成时数据交换和共享过程。最后,给出了集成构架的结构及组成。

关键词 XML 操作模型 数据集成 集成构架

RESEARCH ON WEB DATA INTEGRATION TECHNOLOGY BASED ON XML

Feng Shaorong
(Department of Computer Science, Xiamen University, Xiamen Fujian 361005, China)

Abstract Firstly, this paper discusses the origination background and research object of technology of heterogeneous data integration under Web environment, as well as the related concepts, technologies and method, the key factors XML-based information integration are presented for the characteristics of XML as the data interchange format well A Web data integration interoperation model of XML based is proposed according to it, and discusses the process of exchanging or sharing data under the Web data integration using this model At last the structure and constitution of the integration framework are given.

Keywords XML Interoperation model Data integration Integration framework

1 引 言

随着 WWW的蓬勃发展,网上可共享的数据资源急剧增加,Web数据成为网络数据库应用的主流。如何很好地集成 Web上日益广泛的半结构化数据和非结构化数据,成为当前异构数据集成研究的一个热点。

异构数据不仅指不同的数据库系统之间是异构的,而且还包括不同结构数据之间的异构。数据集成是对各种异构数据提供统一的表示、存储和管理。它屏蔽了各种异构数据间的差异,通过异构数据集成系统进行统一操作。集成后的异构数据对用户来说是统一的和无差异的。

数据集成的主要内容是基于网络环境下的不同硬件、操作系统、数据库管理系统、应用软件组成的异构数据处理环境下的数据模型、数据库的模式、查询语言、事务处理、并发性控制与数据库状态的一致性维护等一系列问题的集成;目的是提供一个统一的、集中的视图,确保应用之间的互操作性,为企业解决多平台、多结构数据的集成问题提供一条解决途径,并为很好地整合企业内外各种相关的数据资源提供了可能。

而 XML对于数据集成过程中所面临的异构的数据结构和数据操纵语言、异构的属性表示和语义,异构的数据模式、对象标识和数据融合等问题,在一定程度上都有相应的处理对策,因而,它对数据集成提供了一种新的处理手段。

2 信息集成技术及方法

2.1 异构数据集成技术

异构数据集成涉及多种计算机技术,如分布式对象技术,XML,面向对象技术及数据库技术等。

(1) 分布式对象技术 分布式对象技术主要包括: Microsoft的 COM/DCOM标准, Sun公司的 Java RMI(Java Remote Method Invocation)标准, OMG(Object Management Group)标准。

(2) XML[1] 由于 XML(Extensible Markup Language)的可扩展,自描述,结构、内容和表现分开,以及具有以下特点,使得 XML在数据存储、数据表示、数据交换、数据集成等领域具有十分广泛地应用。

①可以表达数据内容,也可以表达数据的结构;

②对特定的应用,创建特定的数据类型;

③以 XML为中介,在不同系统之间交换异构的结构化数据;

④有助于结构化数据、半结构化数据和非结构化数据的集成。

(3) 面向对象技术及数据库技术 借助于面向对象技术可以把异构环境下的数据和对数据的操作融为一体。数据库技术包括数据模型技术,数据的查询及优化技术,数据的表示和描述等。

2.2 数据集成方法

(1) 联邦数据库 数据源是独立的,通过数据源之间的数据交换格式进行一一映射,一个数据源可以访问任何其它数据源提供的信息。这种方法工作量大,扩展性差。

收稿日期: 2004 - 02 - 19。冯少荣,副教授,主研领域: XML与半结构化数据库系统,数据挖掘与 KDD,数据仓库与 OLAP技术,基于 WEB的数据库技术,多数据库信息集成技术。

(2) 数据仓库　把来自多个数据源的数据副本,按照集中、统一的视图要求,进行预处理和转换,形成统一的模式,存储到数据仓库中,以便于进行联机分析处理 (OLAP)和数据挖掘 (DM)。这种方法数据重复存储,及时更新困难。

(3) Mediator　Mediator是一种软件组件,Mediator不存储任何自己的数据,而是将用户的查询翻译成一个或多个对数据源的查询。Mediator将那些数据源对用户的查询响应进行综合处理,把结果返回给用户。

(4) 基于知识的信息集成　对底层关系数据库进行数据抽象,利用面向属性的泛化构造知识库,派生出来的知识库可以利用关系模型来实现和存储,底层的关系数据库和知识库可通过关系查询语言,采用同一种形式进行处理,从而可降低信息集成的难度,提高信息集成的效率。但这种方法不通用,难以扩展。

3　数据集成中基于 XML 的互操作

3.1　基于 XML 互操作的特点

XML 与 HTML 的最大区别是 XML 具有丰富的结构信息和明确的语义,而 HTML 只对表现形式作了约定。目前,Web 上的数据信息都能用 XML 表示,并且可以表示以往 HTML 无法表示的内容。文档类型声明将文档类型定义 (DTD)附加到 XML 文档或提供有关的文档结构声明,使 DTD 与文档分离,便于复用和保证一致性。

对于频繁使用和重用的 DTD、模板和 Agent等,可以存储在公共资源库中,可被开发人员引用和下载。开发人员只需创建当前公共资源库中没有的对象。但要解决创建公共资源库时对象的增加、删除、更新等管理问题,以及对象的同步更新问题。

在数据集成过程中,Agent可起下列作用:

(1) 结合 DTD进行的 XML 文档的有效性检验。

(2) 结合 XSL将 XML 转换成 HTML 或 WML 文档格式。

(3) 支持 XML 文档的数据库中间件。

(4) 公共资源库中所引用 DTD或模板的检索。

(5) 与已有数据的交互。

XML 通过制定标准元数据,解决应用系统内部的数据/信息的语义问题。

XML 语义特征表现为:

(1) 信息提供者可以根据自己的需要定义新的标记和特性名称。

(2) 文档的层次结构可以任意复杂。

(3) 可以建立 DTD作为定义 XML 特定类型的一组规则。

因此,XML 可以作为中间介质语言,定义相关的请求 DTD 和操作结果 DTD。另一方面,XML 作为一种文本标记语言,利用现有网络的底层协议,可用 XML 作为文档传输。

由于,XML 将作为下一代浏览器支持的语言,可根据 XML 的语义信息和 XSL (XML 样式表语言)格式信息完成信息的显示。并且未来会有更多的网络软件产品全面支持 XML 操作,所以,基于 XML 的互操作将有广泛的应用前景。

3.2　基于 XML 的异构数据操作模型

基于 XML 的异构数据操作模型由使用方平台和数据提供方平台构成。结构模型如图 1所示。

3.2.1 数据使用方平台

它由以下四个模块组成:

用户界面模块　作为用户指定操作对象和操作要求的入口,以及操作结果返回的出口。必须支持操作对象和操作要求的多样性。因此使用方界面接口请求操作如下:

Request(用户信息参数,操作要求参数,操作对象参数,操作条件参数,操作目标值参数);

请求 XML 文档生成模块　按照一定的规则,如使用 DTD,将用户请求生成查询 XML 文档。

结构如下:

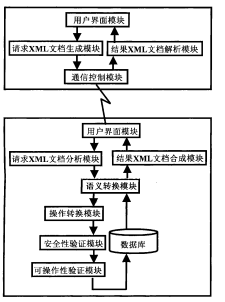

图 1　基于 XML的异构数据操作模型

```
<? xml version =“1. 0 ”>
<! DOCTYPE 请求文档 >
<! ELEMENT请求文档 (用户信息,操作要求,操作对象 +,操作条件 *,设定值 *) + >
<! ELEMENT用户信息 (用户名,用户密码) >
<! ELEMENT用户名 (#PCDATA) >
<! ELEMENT用户密码 (#PCDATA) >
<! ELEMENT操作要求 (#PCDATA) >
<! ELEMENT操作对象 (领域范畴,数据对象或表达式) >
<! ELEMENT领域范畴 (#PCDATA) >
<! ELEMENT数据对象或表达式 (#PCDATA) >
<! ELEMENT操作条件 (域名或表达式,符号,值,关系) >
<! ELEMENT域名或表达式 (#PCDATA) >
<! ELEMENT符号 (#PCDATA) >
<! ELEMENT值 (#PCDATA) >
<! ELEMENT关系 (#PCDATA) >
<! ELEMENT设定值 (#PCDATA) >
```

结果 XML 文档解析模块　对数据提供方返回的结果进行有效性检查,并提取数据返回给用户界面模块。

通信控制模块　基于底层通信协议,实现连接的建立,使用方文档的发送,结果文档的接受,控制 Web数据的实际数据传输。

3.2.2 数据提供方平台

它由以下几个主要模块组成:

用户接口模块　实现与使用方通信控制模块相互连接和数据传输,接受请求文档和返回操作结果文档,接受或拒绝用户连接请求。

请求 XML 文档分析模块　按照 DTD 验证用户请求文档的有效性,提取操作对象和操作。

结果 XML 文档合成模块　按一定的规则,将语义转换后的操作结果生成 XML 文档。

语义转换模块　实现标准化操作对象与系统内部的对象相互转换。按不同的领域设计相应的转换平台,确定系统内部对象与标准对象的对照关系及其存储位置。

操作转换模块　从语义转换模块得到本地语义及其本地存储位置后,形成标准化操作。语义转换模块可设计统一的转换平台,完成操作请求到特定数据库的转换。通过统一平台,数据提供方无需另外开发转换工具,只需在提供服务前设定本地数据库系统的类型。

操作验证和安全性验证模块　对转换后的操作和语义,由数据库系统验证该操作命令的可行性,同时对请求的信息和操作要求进行安全性及合法性检查。

3.3　模型实现及相关技术研究

实现异种 Web数据的操作首先要解决相关领域的数据术语标准化,在此基础上,按标准化的术语制订特定应用领域的多

种类型的数据 DTD。在标准 DTD基础上,面向使用方设计各种特定领域的标准化用户请求平台或接口,作为用户直接加以利用或作为再次开发请求生成器的工具。面向数据提供方,提供应用平台,让数据提供方确定自己的应用范围,建立本地术语与标准化术语之间的对照关系,并指定数据在本地的位置。

为实现跨平台的数据操作问题,模型在用户方请求文档中包含了操作请求,操作请求在语义上可以定义为查询、修改、增加等操作。在形式上可借助标准的数据操作语言,达到精确表达请求的目的。

模型中的文档处理部分包含了请求文档和结果文档的生成与分析。Java提供了处理 XML文档的类和方法,SAX是 Java的 XML处理器的 API,DOM定义了 XML文档的对象模型,可以利用 Java实现文档的有效性验证、数据提取和查找等操作。利用 Java的 XML处理器来实现 XML文档分析与生成,可以避免底层文本处理细节。

4　基于 XML的信息集成[3～7]

为了摆脱对特定技术的依赖,应采用 UML作为建模标准,XML作为数据交互标准,SQL、XML-QL、XQL作为查询标准。

4.1　集成框架

逻辑上分为以下 3层:

Web数据资源层　主要指各种遗留的关系数据库及对象数据库、XML文档、HTML页面、文件系统、电子表格等。

集成模式层　指由多 Web数据资源集成产生的 XML文档。如 DTD、Schema、XML文档。

用户视图层　用以满足用户和显示属性要求的视图,提供用户端浏览操作。

4.2　重要组成部分　包括以下四个模块:

XML文档解析模块　是 XML文档索引、查询、检索和数据库存储、管理的基础。主要有基于树的 DOM和基于事件驱动的 SAX两种解析方式。

支持 XML文档的数据库中间件模块　由于数据库在事务处理、并发控制、可靠性和安全性等方面的特性,许多利用 XML做为数据交换格式的应用要求能在关系数据库和 XML文档之间传送数据。而通过支持 XML文档的数据库中间件,依据模板驱动或模型驱动方式可以实现数据库模式和 XML文档结构之间的映射。

XML文档检索模块　可以使用值索引、标识索引、边索引、路径索引等多种索引技术,并且通过构建关键词表、索引条目表的多级检索体系实现 XML文档的索引、查询。

基于 Ontology的元数据读取模块　提供在查询和存取半结构化数据及非结构化数据过程中获取相应的语义解释,即元数据。

另外,在 XML的信息集成过程中要考虑,由于 XML文档的长度显著增加,在网络传输时,传输效率问题。解决的办法可使用压缩技术,以减少其大小。另外,还要考虑采用一些网络安全技术,以便提高数据集成过程中的安全性。

5　结　论

本文提出的基于 XML的 Web数据集成互操作模型和集成构架,具有以下特点:

(1) 基于术语标准化的一致性的操作平台,分布、异构环境下的信息集成方案和构架结构。

(2) 灵活的资源共享方式,可以根据自身的需要提出访问请求。半结构化数据和非结构化数据的统一、集中管理,提高了网络传输速度。

(3) 安全的验证控制措施。半结构化数据和非结构化数据的有效检索方案。

(4) 兼容不同数据库结构和平台,模型中的中间文档采用标准化的语义及操作表示,通过语义和操作的转换,兼容现有的数据库结构和平台。为信息增加语义信息,提高检索效率。

(5) 实现模块之间逻辑和技术的低耦合性。

(6) 适合 Web应用,便于用户通过 Internet访问。

(7) 一定的扩展性和可移植性。由于 XML将作为数据表示的标准,既可以表示结构化数据也可以表示非结构化数据,以 XML为数据交换中介的应用将非常普及,该模型和集成构架在 Web领域将会具有广泛的应用。

参　考　文　献

[1] Bray T., Paoli J., Sperberg-McQueen M. C., Extensible markup language (XML) 1.0 (Second Edition) [EB/OL], W3C, http://www.w3.org/TR/2000/REC-xml-20001006, 2000-10.

[2] Rohit Khare, Adam Rifkin, XML: A Door to Automated web Applications[J], IEEE Internet Computing, 1997; (7/8).

[3] R. Bourret, C. Bornhovd, A. Buchmann, A Generie Load/Extract Utility for Data Transfer Between XML Documents and Relational Databases WECWIS 2000, Milpitas, California. 2000.6.

[4] Len Seligman et al, XML's Impact on Databases and Data Sharing[J], Computer, 2001; 34(6): 59～67.

[5] McHugh J., Abiteboul S., Goldman R. et al, Lore: a database management system for semistructued data[J], AcM SIGMOD, 1997; 26(3): 54～66.

[6] Maria Arigeles et al, The Challenges That XML Faces [J], IEEE Computer, 2001; 34(10): 15～18.

[7] E. J. Lu et al, An Empirical Study of XML/EdI[J], The Journal of Systems and Software, 2001; 58(3): 271～279.

[8] 王宁、王能斌,"异构数据源集成系统查询分解和优化的实现[J]",《软件学报》, 2000, 11(2).

[9] 李冠宇、张俊、靳强勇, The Research of Heterogeneous Data Integration in Information System[C], ICMSE 2001, Harbin 2001,会议论文集, 2001.

[10] 罗伟其等,"信息大系统的信息集成结构模型设计与实现[J]",《计算机工程与应用》, 2001, 37(2): 9～12.

(上接第 38页)

[4] 孙兆林、齐占杰、李海龙,新编 SQL Server 2000图解教程,北京希望电子出版社, 2001年 6月第 1版.

[5] Sybase公司、Oracle公司技术手册.

[6] (美) Steve Bobrowski著,王焱、王磊、蒋蕊等译, Oracle 8体系结构,机械工业出版社, 2000年 1月第 1版, pp. 115～174.

[7] (美) Rama Velpuri, Anand Adkoli著,蒋蕊、王磊、王鑫等译, Oracle 8i备份与恢复手册,机械工业出版社, 2001年 9月第 1版.

[8] 胡桂香,"数据库复制的设计和管理",《电子工程师》, 2002, Vol 28, No. 5.

[9] 李青,"分布式数据库技术及其在 HNTMIS2000中的实现",《湖北大学硕士论文》, 2001年 5月.

[10] 石骁骁,"Oracle 8高级数据复制技术",互联网.