基于 XML 实现异构数据源的联合使用

United Use of Heterogeneous Data Sources Based on XML

高 明 陈 昕 李 炜 宋瀚涛

(北京理工大学计算机系 北京100081)

Abstract With the rapid development of the WWW and facing Web information sources which are complex in content, multiple in shape and changed dynamically, how to shield the different structures of these heterogeneous data sources and realize the uniform using of them is a crucial issue to solve for application integration of enterprises or e-business. Using XML is a new idea to solve the problem. We introduce the requirement of united use of heterogeneous data sources, and introduce the development and characteristic of XML. We also mention a model to solve the problem using XML.

Keywords Heterogeneous data sources, Application integration, XML, E-business

一、引言

数据库技术的发展十分迅速,已经历了网状、层次、关系、面向对象等阶段,各个数据库厂商也推出了各自丰富的数据库系统,这给企业和社会管理信息带来了很大的便利。然而,随着生产和管理的发展,对单位内部甚至单位之间不同的数据库中的数据就产生了统一联合使用的信息集成要求。异构数据源联合使用技术的提出和实现就成为特别紧迫的任务。

现在互联网上产生了许多新的需求和应用,如利用 Internet 构造企业虚拟专用网 VPN(virtual private network),电子商务的发展也如火如荼。无论是 VPN 还是电子商务都要在 Internet/Intranet 上实现异构数据源的联合使用。

另一方面,随着 WWW 和 Internet/Intranet 的迅速发展,出现了许多新的数据形式,如电子邮件、HTML 文档等信息,单个页面可能包含文本、音频、图像、动画,甚至视频数据。与这些数据相比,传统数据库中的数据具有严格的存储格式,数据的各种操作需要遵循严格的规范,所以,我们可以把传统数据库中的数据称为结构化数据。而互联网上的大量的数据缺乏统一固定的模式,数据往往是不规则的并且经常变动[1]的,这些数据是先有数据后有模式[2],可以把这种数据称为半结构化数据。这样,异构数据源联合使用系统除了要集成传统的异构数据库,还要集成 Web 上这些新的数据源。如何实现异构数据源联合使用在 Internet/Intranet 上的实现,其中一个关键技术是如何以一种统一的数据模式描述各个数据源中的数据,屏蔽它们的平台、系统环境、内部数据结构等方面的异构性,把它们进行无缝连接,对它们实现统一的使用。XML 的出现给异构数据源联合使用注入了新的解决思路。以下先对 XML 作些介绍。

二、XML 的发展及技术特点

XML(Extensible Markup Language)是 SGML 的一个优化子集,它以一个统一、开放、基于文本格式的模式来描述和交换数据。XML 也是一种元标记(meta-markup)语言,它提供了一种描述数据的格式,这方便了内容和查询结果跨平台的声明。

W3C(World Wide Web Consortium)定义 XML 有以下一些目的:保证结构化数据的统一性和应用的平台无关性,提供元数据—关于信息的数据。XML 的特点如下:

(1)内容的自描述性 HTML 是面向显示的标记语言,只定义信息的显示样式,对于信息的具体意义不作说明。XML 是面向内容的标识语言,在 XML 中的语义标识一方面限定了元素的层次结构,另一方面也说明了元素的含义。在 XML 的搜寻结果中由标记就可知道内容的含义,这也使得搜寻结果更有意义。

(2)内容的独立性 由于 XML 是自描述的,使得 XML 可以脱离具体应用来描述保存在异构环境中的各种数据,其它系统应用能直接对这些自描述的 XML 文件中的数据进行操作。由于 XML 的数据语义和数据独立性,它也将成为跨平台数据交换和操作的标准模式。实际上,数据互操作性是异构数据源研究中重要的课题。XML 将成为达到这一目标的钥匙。

(3)能描述不同复杂程度的数据 XML 提供了数据的结构化表示,并且易于操作。例如可以被用来标记下列内容[4]:① 普通文档;② 结构化记录;③ 具有数据和方法的对象,比如一个 JAVA 对象;④ 数据库查询记录;⑤ 图形显示,如应用的图形用户接口;⑥ 所有 Web 上的信息之间的链接(LINK)。这使得 XML 在异构数据源联合使用中有广泛应用前景。因为新的数据源的出现是不可预测的,而 XML 可以以一种统一的数据模式描述来自不同数据源的数据,屏蔽数据源中应用环境和数据结构的异构性,以 XML 查询语言对数据源进行统一访问,可以利用基于关键字的查询。基于关键字的搜索已被证明在 Web 的搜索技术中具有很高的效率。

(4)可扩展性 通过 XML 文件中命名空间的声明,XML 标记可以在企业内部网(Intranet)中使用,并且可以通过互联网被其他组织或个人使用,这样就可以使用一种统一的数据查询和操作模式,而不必关心数据所在具体系统和应用环境。另一方面,XML 可以在不破坏现有结构和系统的情况下增加新的数据字段。应用服务器利用 XML 对所有数据建模,若改

高 明 博士生,主要研究领域为数据库和计算机网络;陈 昕 博士生,主要研究领域为计算方法和数据库;李 炜 博士生,主要研究领域为人工智能和数据库;宋瀚涛 博士生导师,教授,主要研究领域为数据库和智能搜索引擎。

变数据模型只需改变数据模式定义，如 DTD(document type definition)等，不需要重新编码现有的对象。

(5)显示的多样性　XML 一个极其鲜明的特点是把数据的显示格式和数据的表示分离。在 XML 中，可以用格式文件如 XSL(Extensible Style Language)来定义 XML 数据的显示格式，也可以利用 HTML 作为 XML 文件的显示模板，把 XML 数据以数据岛的形式内嵌到 HTML 页面中。XML 把数据的表示和操作相分离，使得用户可以利用不同的格式和应用来显示和操作数据。这种分离可以实现不同数据源数据的无缝连接。各种数据可以在中间件上转换为 XML 格式，使得数据很容易地进行在线交换和传输。

(6)粒度级的更新　XML 中的数据可以在粒度级更新，这样当数据的一部分改变时，不必重新发送全部数据，仅需要将改变的内容从服务器发送到客户端。另外，XML 允许增加数据，这些更新的信息进入用户的示图而不用重新发送新的示图。

XML 在 Internet/Intranet 环境下应用广泛，因为它提供了协同工作的功能，具有使用灵活、开放、基于标准格式的特点，这样，对于数据的访问和数据的交换提供了一种崭新的模式。可以预计 XML 将成为一种新的基于 Web 的数据浏览和互操作的标准。

三、利用 XML 实现异构数据源联合使用的系统模型

综合 XML 的特点，并结合异构数据源联合使用的特殊要求，我们提出了如图1所示的系统模型。

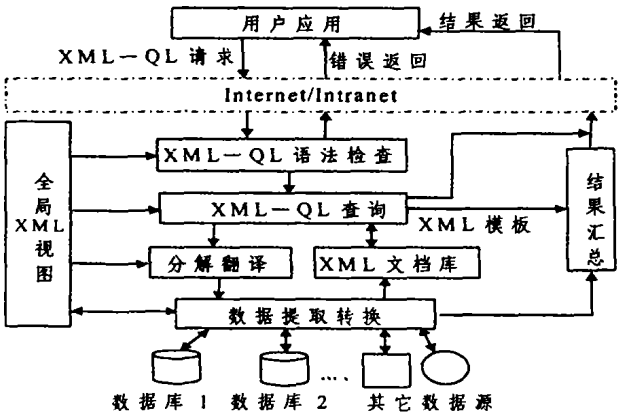

图1　系统结构模型图示

用户通过 XML-QL 进行查询。XML-QL 是 XML 的查询语言[3]。XML-QL 的结构为 where-construct 形式，where 子句由模式和过滤表达式组成，construct 子句指明了输出的 XML 的格式。

在该模式系统中，要利用 XML 的 DTD 进行数据源的数据描述，呈现给用户一个统一的 XML 文档的模式。比如在数据库1中有一个关系表 T_User 的结构为：

T_User(F_ID,F_Name,F_Description,F_Address)

它的 DTD 描述为：

```
<!ELEMENT T_User(F_ID,Name,Age,Description,Address)>
<!ELEMENT F_ID(#PCDATA)>
<!ELEMENT Name(#PCDATA)>
<!ELEMENT Age(#PCDATA)>
<!ELEMENT Description(#PCDATA)>
<!ELEMENT Address(#PCDATA)>
```

该系统模型中各个模块功能简单介绍如下：

(1)XML-QL 语法检查　对于 XML-QL 请求在这里进行语法和语义的检查，判断是否符合 XML-QL 查询语言的语法，请求的 XML 数据是否存在。

(2)XML-QL 查询　这个模块对 XML 文档库进行查询，如果 XML 文档库中没有找到符合条件的数据，该模块要把 XML-QL 查询请求传给分解翻译模块，并产生 XML 模板。如果在 XML 文档库中找到符合条件的 XML 文档，则直接返回给用户。

(3)分解翻译　该模块根据全局 XML 示图把 XML-QL 分解、翻译为针对局部物理数据源的查询请求，比如产生针对数据库1的 SQL 查询语句。

(4)数据提取转换　该模块与异构数据源进行连接，完成各个物理数据源的访问操作，得到查询结果。

(5)全局 XML 示图　全局 XML 示图在本模型中占有极其重要的地位，屏蔽异构数据源的异构性的工作主要在这里实现。在系统中用 XML 的 DTD 数据模式作为全局数据模式来描述各个异构数据源中的数据，并存储在全局 XML 示图中。全局 XML 示图屏蔽了异构数据源的异构性，呈现给用户统一的数据形式，这样用户就只需理解 XML 文档形式的数据，对 XML 中的数据进行访问。在全局 XML 示图中还要有数据源的物理存储空间，如数据库表是在哪一个具体的数据库中，文本文件在哪一个数据源中。

(6)XML 文档库　存储一些已经转换为 XML 形式的数据，起到数据缓存的作用，使已经转换为 XML 形式的数据下次使用，而不用再从物理数据源中提取转换，这样提高了系统的查询效率。

(7)结果汇总　对于各个数据源的查询结果在这里进行合成，利用 XML 模板产生 XML 文档返回给用户。为了有助于对系统模型的理解，现在假设用户对 T_User 进行请求，要求查询年龄大于50岁的所有人的姓名、年龄、地址。数据查询流程如下：

首先用户产生 XML-QL 查询请求，XML-QL 如下：

```
where <T_User>
          <F_ID>$F_ID</F_ID>
          <Name>$Name</Name>
          <Age>$Age</Age>
          <Address>$Address</Address>
          <Description>$ Description</ Description>
          <Address>$Address</Address>
      </T_User> in "http://My.com/T_users.xml",
$Age > 50
construct
     <result>
       <Name>$Name</Name>
       <Age>$Age</Age>
       <Address>$Address</Address>
<result>
```

系统对于用户提出的 XML-QL 请求进行语法检查。

随后系统对 XML 文档库进行 XML 文档查询，如果查询请求的数据并不全在 XML 文档库中，则系统要产生 XML 模板：

```
<result >
       <Name>$Name</Name>
       <Age>$Age</Age>
       <Address>$Address</Address>
<result>
```

接下来，系统根据全局 XML 示图分解和翻译相应的查询请求，产生数据库1的 SQL 查询语句。在这个过程中还要进行查询优化：

SELECT Name,Age,Address FROM T_User WHERE Age>50

(下转第93页)

够及时获悉。这样，出现故障的 Agent 就能很快地被系统发觉，并做出相应的补救措施。

4）永久对象服务　提供了用于维持和管理对象的永久状态的一系列通用接口。

对象最终具有管理其状态的责任，但是它可以委派永久对象服务来完成实际的工作。Agent 要保留在查找服务里面不被剔除，就要向查找服务不断地发出租借信息。对于处于未激活状态的 Agent，可利用永久对象服务代理其向查找服务发送租借信息，以保证其处于"可用"状态。等待某一事件发生时，再将其激活。

这里的永久对象服务除了代理 Agent 发送租借信息外，还要加上 Agent 特有的状态维护。

5）安全服务　基于 RMI 的安全服务机制，扩充了关于 Agent 的安全服务机制。由于 Agent 和服务器都不能完全准确地预见到相互的行为及后果，这种不确定性带来了严重的安全问题。在 Agent 系统中，安全机制是双向的，一方面保证 Agent 本身不受破坏，另一方面也保证服务器不受 Agent 的破坏。

6）事件服务　提供了非常灵活、有力的可以配置的基本能力。Jini 通过事件代管程序，使事件能够可靠地到达目的地。

事件服务设计可以扩充且适宜于分布式环境。不需要一个中央服务器或者依赖于全局服务。事件服务接口允许应用提供不同性质的服务以满足不同的应用需求。

基本的事件服务包括 Agent 对外界环境的感知和对外界环境的响应。其中包括与其它 Agent 的联系与协作过程的管理。

7）组服务　负责对于 Agent 的组进行管理，记录组的标识以及组内的成员。

8）数据库管理服务　管理基本的数据库，对其中的数据进行查询和修改等操作。

结束语　利用 Jini 技术作为平台来构建 Agent 系统，充分利用了面向对象技术的优势，为多 Agent 系统的实现提供了一条捷径。同时也继承了 Jini 这种新的分布式系统的优势，提高了系统的安全性和可靠性。

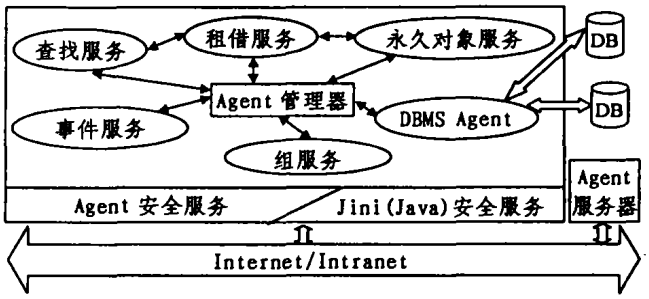

图2　Agent 服务器结构图

随着网络和信息技术的发展，Agent 技术已从人工智能领域慢慢地向其它领域渗透。借助于 Agent 技术解决本领域内的问题已成为多学科交叉研究的热点，诸多领域内已出现了 Agent 应用的先例，如在经济学领域和机械制造行业等。随着技术的发展和研究的不断深入，将 Agent 技术应用到家庭网络之中，实现个性化和智能化的家居管理和家庭信息化服务，将会成为现实。

参 考 文 献

1　史忠植. 智能主体及其应用. 北京：科学出版社，2000. 12
2　赵龙文，侯义斌. 多 Agent 系统及其组织结构. 计算机应用研究，2000，32(7)：12～14
3　Graham J R，Decker K S. Towards a Distributed，Environment-Centered Agent Framework
4　周健，吴泉源，腾猛，王怀民，孙海燕. 一种基于分布对象技术的 Agent 计算框架. 计算机研究与发展，2000，37(1)：45～49
5　Edwards W K. Jini 核心技术. 北京：机械工业出版社，2000. 7
6　高济，林东豪. 基于 Agent 技术的虚拟组织集成框架 IFVO. 计算机研究与发展，1999，36(12)：1410～1416

（上接第84页）

系统利用这个 SQL 语句对数据库1进行访问，产生的查询结果传给结果汇总模块。结果汇总模块利用 XML 模板把查询的结果转换为 XML 文件的形式：

```
〈result〉
  〈Name〉name1〈/Name〉
  〈Age〉age1〈/Age〉
  〈Address〉address1〈/Address〉
  〈Name〉name2〈/Name〉
  〈Age〉age2〈/Age〉
  〈Address〉address2〈/Address〉
  ……
〈/result〉
```

最后，结果汇总模块一方面把这些 XML 文档返回给 XML 文档库，另一方面把这些 XML 形式的数据提供给用户系统。

结论　现在，XML 在许多领域已得到应用，W3C 也在不断地修改和制定 XML 的各种标准，以期 XML 趋于更加完善。Microsoft，IBM，Oracle 等大公司已经推出了 XML 的产品。电子商务、电子政务的迅速发展，WWW 上不断出现大量的数据信息，互联网上大量的数据传输和交换需求，这些都使 XML 有着广泛的应用前景。已经有组织对 XML 的研究提出了新的观点，他们不但把 XML 看作是数据描述的标准、数据交换的标准，更把它看作是建立新的存储技术的基础。

本文结合 XML 的特点和异构数据源联合使用的特殊要求，提出了一个在企业内部和电子商务中急需解决的异构数据源联合使用问题的解决模型。由于 XML 的特点，本系统模型具有良好的可扩充性，它可以集成新的数据源，而不用对系统做大的变动，只需在全局 XML 视图中加入对新的数据源的 DTD 数据模型的描述，所以全局 XML 视图具有动态性。另外，在模型中引入 XML 文档库也增加了系统的效率。本文提出的模型还可以在一些方面进行完善，比如事物处理、一致性约束等，这些都是以后努力的方向。

参 考 文 献

1　孟小峰，曹巍，王珊. Web 查询技术研究. 计算机科学，2001，28(2)
2　王静，孟小峰. 半结构化数据的模式研究综述. 计算机科学，2001，28(2)
3　Deutsch A，Fernndez M，Florescu D. A query language for XML. In：Proc. on the Eights Intl. World Wide Web Conf.(WWW8)，Toronto，1999
4　Bourret R. XML and Databases. Available at：http://www.robourret.com/xml.
5　Wiederhold G. Mediators in the architecture of future information system. IEEE Computer. March 1992
6　刘艳梅. 基于 COM/DCOM 组件标准实现异构数据库的联合使用：[博士论文]. 北京：北京理工大学，2000(7)
7　Microsoft Platform SDK July 2000 Edition—XML. Available at http://www.microsoft.com.
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. 软件学报，2000(1)