异构数据源的集成与访问*)

熊海灵 伍 胜 余建桥
(西南农业大学信息学院 重庆400716)

Integrating and Accessing of Heterogeneous Data Sources

XIONG Hai-Ling WU Sheng YU Jian-Qiao
(College of Information, Southwest Agricultural University, Chongqing 400716)

Abstract As the Web comes to be viewed as a large semistructured database with XML as its model, issues related to querying semistructured data in general, and XML in particular, become more important. In this paper, based on the MIX and W4F, a system of integrating and accessing heterogeneous data sources is presented. The system accesses a set of XML sources, which are currently information systems wrapped to provide (1) an XML view of their data and (2) a set of XMAS queries that they can answer. Conceptually, it exports an integrated view of the source data. The view definition, provided by the view developer, specifies how the source data will be integrated. Especially, the wrapper for Web data sources is generated by the W4F and the system does not materialize the integrated view in advance.

Keywords Heterogeneous data sources, Semistructured data, Integrating and accessing, XMAS

1 引言

近年来,Internet/Web 技术和计算机硬件的迅速发展对数据库研究领域产生了巨大的影响,提出了新的挑战性问题:一是如何使数据库系统和技术成为 Web 的有机组成部分,而不仅仅充当 Web 体系的外围角色;二是如何实现 Web 动态信息的管理,完成日益增多的新一代 Web 应用等。人们已认识到 Web 正在逐渐成为全球性的自主分布式计算环境,Web 上的多数站点都具有丰富的数据资源。如果能够把遍及全球的 Web 数据源集成起来,Web 将成为一个全球统一的数据库,由全世界共享。然而 Web 数据源的集成并非易事,数据源的异构问题是影响 Web 数据源集成的最大障碍。Web 数据源的异构问题主要包括三个方面:第一是模式异构,表现在不同数据源具有不同的存在形式;第二是数据异构,表现在不同数据源具有不同的数据类型;第三是语义异构,表现在相同的数据形式表示不同的语义或同一个语义由不同形式的数据表示。

这些问题的解决需要我们研究和开发支持 Web 信息自动创建、管理、查询和安全控制的新数据库技术和工具。其中结构化和半结构化数据的集成是解决这一问题的基础。目前异构数据库中数据的集成已经有比较成熟的研究,将 Web 上的数据与传统数据库中的数据同时集成起来实现基于 Web 的应用还需要更深入的研究。

传统的数据库都有一定的数据模式,可以根据这个模式来具体地描述特定的数据,同时还可以很好地定义和解释相关的查询语言。而 Web 上的数据特点很复杂,没有这样特定的模式来描述。而且数据本身具有自描述性和动态可变性等一系列复杂特性,其结构也不可琢磨。在这种情况下如何来解决异构数据源的集成和数据查询问题呢?这就迫切需要有一个模式来统一地描述这些异构数据[6]。研究表明:为了适应新一代的 Web 应用,需要引入半结构化数据模式,因为它在实际的数据处理中有广泛的用途:有助于用户了解数据的结构,从而提出更精确和有效的查询;有助于查询处理器对查询计划进行优化,大大缩减查询的搜索空间;有助于设计数据的物理存储结构以及索引,从而提高查询执行的效率;有助于选择适当的集成模式和定义转换规则[5]。

2 半结构化数据及其模式的特点

2.1 半结构化数据的特点

半结构化数据是介于结构化数据(如关系数据库中的数据)和完全无结构数据(如声音、图像文件)之间的数据。它具有一定的结构,但其结构与数据混在一起,没有显示的模式定义;它具有不规则的结构,一个数据可能由异构的元素组成,同样的信息可能由不同的数据表示;它没有预先定义的模式,以及数据结构的不规则性,数据缺乏严格的约束。这种数据在网络中有很大的灵活性,但不利于数据的处理[5],显然 Web 数据就是一种典型的半结构化数据。

2.2 半结构化数据模式的特点

半结构化数据模式与传统的关系或面向对象数据模式不同,主要有以下一些特点:半结构化数据是先有数据,后有模式;半结构化数据模式是用于描述数据的结构信息,而不是对数据结构进行强制性的约束;半结构化数据模式是非精确的,它可能只描述数据的一部分结构,也可能根据数据处理的不同阶段的视角而不同;半结构化数据模式经常处于动态变化中[5]。

2.3 半结构化数据模式的选择

当前 Web 数据大多是由 HTML 表达的半结构化数据,为此美国斯坦福大学提出了 OEM(Object Exchange Model)模型,并在 LORE(Lightweight Object Repository)中应用,这种模型很适合表示结构松散或结构不固定的半结构化数据[8]。但新一代 Web 上的数据不仅是由 HTML 表达的,XML 作为标准的通用标记语言,具有适于 Web 上数据交换的特

*)重庆市教委科学技术研究项目(编号:011806)资助。熊海灵 硕士生,主要研究方向为信息集成与数据库。伍 胜 硕士生,主要研究方向为网络技术与数据库。余建桥 硕士生导师,教授,主要研究方向为数据库与人工智能。

点，将成为数据组织和交换的事实标准，并且大量的 XML 数据将很快出现在 Web 上。因此选用 XML 数据模型作为异构数据源集成的模式是合适的。因为 XML 图是一个非常灵活的数据模型，它能很容易地构造关系数据和面向对象的数据，并且数据与 XML 图能很好地映射，XML 图非常适合描述分布的、多态的、动态改变的 Web 数据。

3 异构数据源集成与访问的系统模型

目前，异构信息集成的系统已很多，但大多只具有某一方面的优势。本文综合考虑 California 大学提出的 MIX[1](Mediation of Information Using XML)系统的灵活性和对 XML 数据模式的完全支持，以及 Pennsylvania 大学提出的 W4F[4](World Wide Web Wrapper Factory)系统对 HTML 文档的强大处理能力，结合以上对半结构化数据的分析，提出如下一个基于 XML 的异构数据源的集成与访问系统(见图1)。图中向下的实线箭头表示 XMAS 查询的处理过程，向上的实线箭头表示查询结果的形成过程，向下的虚线箭头表示视图的定义，向上的虚线箭头表示 XML 集成视图的形成过程。

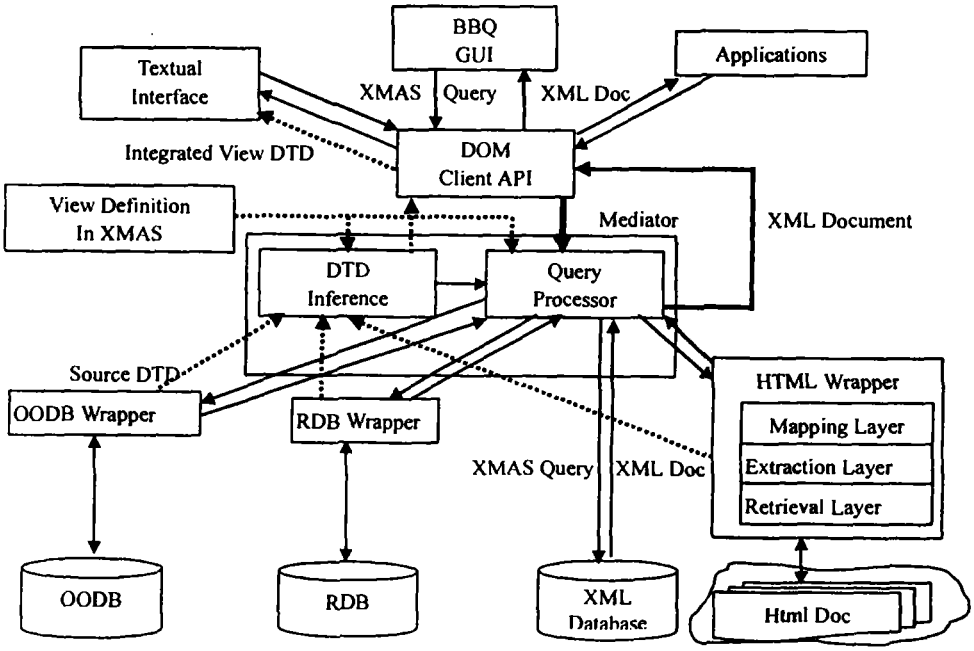

图1 异构数据源集成与访问的系统模型

该系统中数据的交换与集成完全依赖 XML，XML 查询用简洁的 XMAS[2](XML Matching And Structuring)语言表达，并且 XMAS 查询的形成完全基于友好的图形用户界面，查询结果也通过该界面清晰地表示，该系统更突出的是能将传统数据库里的数据与 Web 上的数据集成，有效地支持新一代的 Web 应用。

3.1 系统主要组件及功能

用户通过 XMAS 语言进行查询。XMAS 是在 XML-QL、YAT、MSL 和 UnQL 的基础上发展起来的一种 XML 的查询语言，也是基于规则的一种描述性语言[2]。其结构为 CONSTRUCT head WHERE body 形式。该系统模型中各个模块的功能简单介绍如下：

(1)BBQ(Blended Browsing and Querying)GUI 模块 是一个图形用户接口，帮助用户形成一个复杂的基于 XML 文档的查询要求，并且规定了直观地反映查询结果的模式。查询和浏览都基于 Mediator 提供的集成视图，这样就使查询和浏览都具有一个统一的界面[3]。

(2)DOM Client API 模块 主要是提供使用 XML 文档对象模型的应用程序接口。

(3)Mediator 模块 由几个子模块组成，是数据集成的关键模块，涉及到用 XMAS 语言定义视图，解析、分解、优化和执行 BBQ 模块所形成的 XMAS 查询，最后返回 XML 文档数据。其中 DTD Inference 模块可以自动地从经过 Wrapper 模块处理后的 XML 数据中得到 DTD 和视图的定义，这个视图可以看作是一个全局 XML 视图，它屏蔽了异构数据源的异构性。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。Query Processor 模块根据这个全局视图将 XMAS 查询分解成一系列展开的针对 XML 数据源的查询，并分配给相应的 Wrapper。

(4)Wrapper 模块 Wrapper 是一种软件构件，负责将数据和查询请求由一种模式转换成另一种模式。一个 Wrapper 实际是一类页面到该页面所含元组集合的函数[7]。其目的一方面是以统一的格式向 Mediator 提供数据源，另一方面是将 Mediator 传送来的基于 XML 数据源的查询具体化为针对某一特定数据源的查询。一个 Wrapper 一般针对某一单一数据源中的一类页面，目前，一般是针对不同的数据源手工编制特定的 Wrapper。考虑到 W4F 是专门用于生成 Web Wrapper 的 Java 工具包，它生成的 Wrapper 可以集成于任何 Java 应用程序中，并且所有部件都是声明性的，因此本系统采用 W4F 生成针对 HTML 文件的 Wrapper，该 Wrapper 分3层，包括检索、抽取和映射，可以将 HTML 文件直接转换成 XML 文档。

3.2 系统行为分析

以查询 HTML 文档中有关电影的信息为例[4]，对于一个单独的 HTML 页面可以利用 Wrapper 直接生成 XML 文档，对于多个相关的 HTML 页面，在将它们集成一个 XML 文档前还需用下面的 Java 代码进行处理：

```
XmlDoc doc=new XmlDoc();
doc.appendDTD_multiple("List",movie_t);
MovieRef[] refs=get_top250();
```

(下转封四)

(上接第184页)

```
for(int i=0;i<refs.length;i++)
doc.appendElement(getMovie(ref[i].url));
doc.print(new PrintWriter(System.out));
```

利用XML的DTD进行数据源的数据描述,假设生成的文档的DTD如下:

```
<! DOCTYPE W4F_DOC[
<! ELEMENT W4F_DOC(List)>
<! ELEMENT List(Movie)*>
<! ELEMENT Movie(Title,Year,Directed_By,Genres,Cast)>
<! ELEMENT Title(#PCDATA)>
<! ELEMENT Year(#PCDATA)>
<! ELEMENT Directed_By(Director)*>
<! ELEMENT Director(#PCDATA)>
<! ELEMENT Genres(Genre)*>
<! ELEMENT Genre(#PCDATA)>
<! ELEMENT Cast(Actor)*>
<! ELEMENT Actor(FirstName,LastName)>
<! ELEMENT FirstName(#PCDATA)>
<! ELEMENT LastName(#PCDATA)>
]>
<W4F_DOC>
```

生成了XML文档后,我们可以利用XMAS语言产生查询请求。假设要查询在三部电影中都扮演角色的演员的姓名,则可以这样描述(IN后边的内容代表URL):

```
CONSTRUCT
  <Actor>
  <Name> $name <Name>
  <Actor>
WHERE
<*.Cast.Actor.LastName> $name </>
IN "…Star+Wars+(1977)",
<*.Cast.Actor.LastName> $name1 </>
IN "…Empire+Strikes+Back,+The+(1980)",
<*.Cast.Actor.LastName> $name2 </>
IN "…Return+of+the+Jedi+(1983)"
text($name)= text($name1),
text($name)= text($name2)
```

接着,Mediator解析、分解、优化和执行XMAS查询,最后返回XML文档数据:

```
<Actor>
   <Name> name1 <Name>
   <Name> name2 <Name>
   <Name> name3 <Name>
   ……
<Actor>
```

总结 进入20世纪90年代以来,数据库应用环境发生了巨大的变化,Internet/Web向数据库领域提出了前所未有的挑战,一大批新一代数据库应用应运而生,如支持高层决策的数据仓库、OLAP分析、数据挖掘、数字图书馆、电子出版物、电子商务、Web医院、远程教育、虚拟现实、工作流管理、移动数据库、Web上的信息管理与检索等。同时,异构数据源的集成问题也已成为数据库研究的热点。目前,这个问题虽然没有得到圆满解决,但是也取得了很多研究成果。

本文提出的系统模型集成了MIX系统和W4F系统的优势,界面友好,数据的集成采用虚拟方式,用户的查询基于中间模式,数据保存在局部数据源中,因此更适合于数据源数目多、各局部数据源的自治性很高且局部数据经常变化的Web环境。本研究只是Web数据管理技术的一个方面,未来的研究主要是如何以有结构的方式更有效地组织和访问Web及其它数据库资源。

参考文献

1 Baru C, et al. XML-Based Information Mediation with MIX. In Exhibitions Program of ACM SIGMOD 1999

2 XMAS sub-group of MIX. A Brief Introduction to XMAS. http://www.db.ucsd.edu/Projects/MIX/docs/XMAS

3 Baru C, et al. Features and Requirements for an XML View Definition Language: Lessons from XML Information Mediation. Position paper in W3C's Query Language Workshop

4 Sahuguet A, Azavant F. Looking at the Web through XML glasses. CoopIs'99, 1999

5 孟小峰. Web数据管理研究综述. 计算机研究与发展, 2001, 38(4)

6 高明, 陈昕, 李炜, 宋瀚涛. 基于XML实现异构数据源的联合使用. 计算机科学, 2002, 29(3)

7 李效东, 顾毓清. 基于DOM的Web信息提取. 计算机学报, 2002, 25(5)

8 简琤峰, 谭建荣. 一种面向XML表达的Web数据模型. 计算机研究与发展, 2002, 39(2)

计算机科学

(1974年1月创刊)

第30卷第5期 (月刊)

2003年5月25日出版

ISSN 1002-137X

CN50—1075/TP

定价: 20.00元 国外定价: 5美元

邮发代号: 78—68

发行范围: 国内外公开

主管单位: 国家科学技术部

主办单位: 国家科技部西南信息中心

编辑出版: 《计算机科学》杂志社

重庆市渝中区胜利路132号 邮政编码: 400013

电话: (023) 63500828 E-mail: jsjkx@swic.ac.cn

社 长: 牟炳林

主 编: 朱宗元

印刷者: 重庆科情印务有限公司

总发行处: 重庆市邮政局

订购处: 全国各地邮政局

国外总发行: 中国国际图书贸易总公司 (北京399信箱)

国外代号: 6210—MO