基于 Web 的数据挖掘技术的应用研究①

Research and Application of Data Mining based on Web

何月顺 （南京航空航天大学 210016）
（江西抚州 东华理工学院 344000）
汤 彬 （江西抚州 东华理工学院 344000）
丁秋林 （南京航空航天大学 210016）

摘要：Web 是一个动态性极强的信息源，要访问、分析这些数据必须要研究异构数据的集成问题和选择合适的技术进行数据分析、集成和处理。文中介绍了多数据源数据仓库体系结构，多数据源数据的集成思想和实现的框架；分析了转换器在面向 Web 的数据挖掘中存在的不足和 XML 语言的技术特点；提出了应用 XML 技术对多数据源数据进行集成与转换以便构建数据仓库，同时给出了关键技术的实现方法。

关键词：XML 数据仓库 数据挖掘 Web 半结构化

1 引言

Web 上有海量的数据信息，是一个巨大的、分布广泛的和全球性的信息服务中心，它涉及新闻、广告、消费信息、金融管理、教育、政府、电子商务和许多其他信息服务。Web 还包括了丰富的动态超链接信息，以及 Web 页面的访问和使用信息，怎样对这些数据进行复杂的应用成了现今数据库技术的研究热点。例如，销售人员在向客户介绍产品时经常会遇到这样的问题：大部分客户在地理上非常分散，销售人员与客户在产品信息上的交流不是很直接，大部分客户只能依据产品样本选取产品，用电话、传真或电子邮件与销售人员进行联系、咨询，使得用户对产品了解非常模糊、不够形象和全面。尽管企业可以利用 Web 这种形式发布一些企业信息，提供产品资料下载，还可以在 Web 页面上提供一些用户接口，让用户可以对后台数据库进行操作（如查询），但效果并不是很明显，用户常常还是不能得到自己想要的数据，有时甚至得到的是垃圾数据。

企业如何利用 Internet 这个良好的交互工具提供更多的服务，使客户可以非常方便、更深入地了解企业产品，享受到更加人性化、专家级的咨询服务等，已经成为迫切需要解决的问题。

数据挖掘就是从大量的数据中发现隐含的规律性的内容，解决数据的应用质量问题。而 W eb 上的信息为数据挖掘提供了丰富的资源。相对于 Web 的数据而言，传统的数据库中的数据结构性很强，其中的数据为完全结构化的数据，而 Web 上的数据最大特点就是无序性和半结构化。显然，面向 Web 的数据挖掘比面向单个数据仓库的数据挖掘要复杂得多[1]。

2 基于 Web 的数据挖掘需要解决的几个问题

2.1 异构数据库环境

从数据库研究的角度出发，Web 网站上的信息也可以看作一个数据库，一个更大、更复杂的数据库。Web 上的每一个站点就是一个数据源，每个数据源都是异构的，因而每一站点之间的信息和组织都不一样，这就构成了一个巨大的异构数据库环境。如果想要利用这些数据进行数据挖掘，首先，必须要研究站点之间异构数据的集成问题，只有将这些站点的数据都集成起来，提供给用户一个统一的视图，才有可能从巨大的数据资源中获取所需的信息。其次，还要解决 Web 上

① 基金项目：国家"863"高技术项目支持，项目编号：863－511－810－041－03

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

的数据查询问题,因为如果所需的数据不能很有效地得到,对这些数据进行分析、集成、处理就无从谈起[6]。

2.2 半结构化的数据结构

半结构化数据有两层含义,一种是指在物理层上缺少结构的数据,另一种是指在逻辑层上缺少结构的数据。有一些结构化数据,如元组,为用于 Web 页面的显示而与 HTML 语言的标记符号嵌在一起,构成了物理上的半结构化数据。Web 中有大量丰富的数据:文本、图片、声音、图像等,这些数据多存在于 HTML 文件中,没有严格的结构及类型定义,这些都是逻辑层半结构化的数据[7]。Web 上的数据与传统数据库中的数据不同,传统的数据库都有一定的数据模型,可以根据模型来具体描述特定的数据。而 Web 上的数据非常复杂,没有特定的模型描述,每一站点的数据都各自独立设计,并且数据本身具有自述性和动态可变性。因而,Web 上的数据具有一定的结构性,但因自述层次的存在,从而是一种非完全结构化的数据,这也被称之为半结构化数据。半结构化是 Web 上数据的最大特点。

2.3 解决半结构化的数据源问题

Web 数据挖掘技术首先要解决半结构化数据源模型和半结构化数据模型的查询与集成问题。解决 Web 上的异构数据的集成与查询问题,就必须要有一个模型来清晰地描述 Web 上的数据。针对 Web 上的数据半结构化的特点,寻找一个半结构化的数据模型是解决问题的关键所在。除了要定义一个半结构化数据模型外,还需要一种半结构化模型抽取技术,即自动地从现有数据中抽取半结构化模型的技术[6]。面向 Web 的数据挖掘必须以半结构化模型和半结构化数据模型抽取技术为前提建立数据仓库。

3 XML 在 Web 数据挖掘中的应用

3.1 基于转换器的半结构化数据处理

在多数据源的数据仓库中对于半结构化数据的提取、表示和查询,一般通过包装器来处理。图 1 是多数据源信息集成系统的典型应用数据仓库体系结构。图中的信息源不仅指那些常见的数据库,也包括文件、HTML 文件、知识库、Legacy 系统等信息源。包装器负责把信息源的信息格式转换仓库系统使用的数据格式和数据类型。监视器部分负责自动检测信息中数据的变化并把这些变化上报给集成器[4]。每当有新信息源挂上仓库系统或信息源的相关信息发生变化时,新的或改变的数据就传给集成器,集成器对这些信息进行过滤、总结或与其他信息源进行合并处理,并安置在数据仓库中。为把新信息准确地集成到数据仓库中,集成器可能还要从原来或相关的其他信息源中获取进一步的信息。

转换器在进行信息转换时面临的主要问题是如何将半结构化的数据转换成结构化的信息。由于转换器是对应于数据源的,而如此庞大的 Web 数据源使得人工构造转换器的代价太大,因此需要有快速建造并自支维护转换器的工具,而目前这类工具不多且很不完善,因为在多数据源信息集成系统上的应用层中,用户提交查询后才能进行训练,在响应速度和处理效率上令人难以忍受,不是很实用。近年来 XML 及其相关技术的迅速发展,为多数据源信息集成系统中对半结构化数据的处理提供了解决的方法。

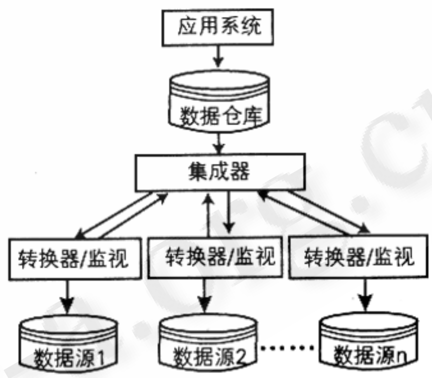


图 1 数据仓库体系结构

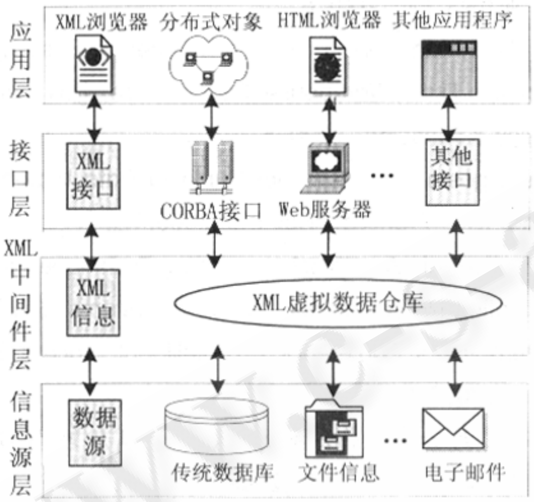
3.2 XML 在 Web 数据挖掘中的应用

促进 XML 应用的是那些用标准的 HTML 无法完成的 Web 应用。这些应用从大的方面讲可以被分成以下四类:需要 Web 客户端在两个或更多异质数据库之间进行通信的应用;试图将大部分处理负载从 Web 服务器转到 Web 客户端的应用;需要 Web 客户端将同样的数据以不同的浏览形式提供给不同的用户的应用;需要智能 Web 代理根据个人用户的需要裁减信息内容的应用。显而易见,这些应用和 Web 的数据挖掘技术有着重要联系,基于 Web 的数据挖掘必须依靠它们来实现。

XML 已经成为正式的规范,开发人员能够用 XML 的格式标记和交换数据。XML 在三层架构上为数据处

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

理提供了很好的方法。使用可升级的三层模型,XML 可以从存在的数据中产生出来,使用 XML 结构化的数据可以从商业规范和表现形式中分离出来。XML 在 Web 数据挖掘中可分为基于信息源层的数据集成(数据仓库构建);将集成处理结果发往 XML 虚拟数据仓库;基于数据仓库的数据挖掘与处理;将挖掘处理的结果送往应用层等四个过程,如图 2 所示。



2　XML 在 Web 数据挖掘中的应用层次与数据集成框架

下面主要介绍用 Java 程序通过 XML 对"XML 虚拟数据仓库"的数据的存取。

3.3 关键技术的实现

在 XML 中间件层主要涉及两个问题:针对每个数据的的转换器/监视器(如图 1 所示),它完成某种类型的数据源与虚拟数据库之间的双向映射;另一个 XML 的集成数据(虚拟数据仓库)公共模块的建立及管理。

(1) 转换器/监视器中的双向映射。首先用 XML 描述集成数据,用 XSL 定义用户视图,用 XML 文档和格式文件 DTD 表示集成模式和数据源之间的映射。各个数据源的模式通过相应的转换器/监视器溶入全局模式。在这个过程中,转换器/监视器数据源中的数据结构转换为一个 DOM 对象。一个关系转换器/监视器能决定关系模型和 DOM 对象数据模型之间的映射。

例如:有 Corp(CorpID char(6), CorpName char(20), CorpAddress char(40), JuridicPerson char(10),…)和 Client(CorpID char(6),ClientName char(20), ClientType char(8), ClientAddress char(40), Price number(10,2),Number number(10),ToDate date,…)两个关系表,关系表与 XML DTD 之间的映射关系如下:

```
<! ELEMENT Corp(CorpID, CorpName, CorpAddress, JuridicPerson)>
<! ELEMENT Corp CorpID IDREF #REQUIRED>
<! ELEMENT Corp CorpName (#PCDATA)>
…
```

在 XML 文档与数据库进行双向转化的过程中,元素结点 Corp 和 Clien 对应数据库中的表,而 CorpID, CorpName 等对应表的列。由于 Clien 为 Corp 的子结点,因此用 CorpID 建立关联。

(2) XML 虚拟数据仓库。如何实现对各个数据源的集成存取,或者说将用户对集成视图的操纵转换成对数据源的操纵,包括两个方面:一个是将用户对集成模式的访问转换成数据源可以执行的请求;另一个是将各种数据源返回的数据转换成集成模式的表示形式。反之亦然。具体步骤如下:

- 从数据仓库中读取数据,生成 XML 文件,统一的格式表示。
- 将 XML 转换成一个 DOM 对象模型,为上层提供访问服务。

具体程序实现可以采用 ASP 或 Java 作为设计语言,利用通过 XML 的 DOM 来操作 XML 文档。

下面的 Java 程序利用 DOM 型的 Parser 读入 XML 文件,生成 DOM 树型结构,并根据树型结构进行相应的操作和处理。

```
…
public class XMLDataReader{
  public static void main(String args[]){
    String xmlFile = "mydata.xml";//用 XML 定义的数据表与 XML DTD 之间的映射文件
    DOMParser parser = new DOMParser();
      Try{ Parser.parser(xmlFile);} //把 XML 文件传给 DOMParser
          catch (SAXException SE){SE.printstackTrace();}
          catch(IOException ISE){ISE.printstackTrace();}
    Document doc = parser.getDocument();//从
```


©《计算机系统应用》编辑部 http://www.c-s-a.org.cn

parser 取得树型资料

```
NodeList nl = doc. getElementsByTagName ("文档"); // 从树型结构中取得名字为 //"文档"的所有节点的集合
    for(int num =0; num < nl. getLength ( ); num + +){
        Node textNode = nl. item ( num ) . getChild-nodes( ) . item(0);
        System. out. println ( textNode. getNodeValue ( ) ); }
    }
  }
```

4 结束语

面向 Web 的数据挖掘是一项复杂的技术,由于 Web 数据挖掘比单个数据仓库的挖掘要复杂的多,因而面向 Web 的数据挖掘成了一个难以解决的问题。而 XML 的出现为解决 Web 数据挖掘的难题带来了机会。由于 XML 能够使不同来源的结构化的数据很容易地结合在一起,因而使搜索多样的不兼容的数据库能够成为可能,从而为解决 Web 数据挖掘难题带来了希望。作为表示结构化数据的一个工业标准,XML 为组织、软件开发者、Web 站点和终端使用者提供了许多有利条件。随着 XML 作为在 Web 上交换数据的一种标准方式的出现,面向 Web 的数据挖掘将会变得非常轻松和有效。

参考文献

1 Jiawei Han and Micheline Kamber. Data Mining: Concept and Techniques [M]. Morgan Kaufmann Publishers, Inc. 2001. 8.

2 W. H. Inmon Ken Rudin. Christopher K. Buss Ryan Sousa. Data Warehouse Performance. [M] Jhon Wiley & Sons, Inc. 2000. 5.

3 Lou Agosta. The Essential Guide to Data Warehousing[M]. Prentice – Hall PTR, 2000. 11.

4 Jing S, et al. An Object – wrapping Technique for Integrating Nontraditional Database Systems [J]. IEEE International Conference on Intelligent Process System. , 2002. 07.

5 邓苏、刘青宝、陈卫东著,数据仓库原理与应用[M],电子工业出版社,2002.3。

6 李军怀等, XML 在异构数据集成中的应用[J],计算机应用,2002,22(9),11 – 12。

7 戴青云等,Web – based 多集成系统的研究[J],计算机科学,200[illegible],30(9),120

©《计算机系统应用》编辑部 http://www.c-s-a.org.cn