Web数据挖掘中数据异构问题解决方法的研究

李春梅[1] 李艾丹[2] 薛中玉[1] 韩 爽[1]
（1. 北京中机科海科技发展有限公司，北京 100048；2. 北京理工大学，北京 100081）

摘 要：Web是动态性极强的信息源，访问、分析信息必须研究异构数据的集成问题，并选择合适的技术进行数据分析、集成和处理。怎样对Web海量的数据信息进行深层次的应用已成为数据挖掘技术的研究热点。本文介绍了XML（可扩展标记语言）在Web数据挖掘中的应用，探讨了Web数据挖掘中的数据异构问题。通过XML技术建立数据抽取模型，解决互联网上绝大多数因异构、非结构化所导致的Web数据挖掘问题。

关键字：数据挖掘；半结构化；XML技术；数据抽取；模型
中图分类号：TP391 **文献标志码：**A DOI：10.3772/j.issn.1674-1544.2012.04.018

Research on Heterogeneous Data Problem Solving Method in the Process of Web Data Mining

Li Chunmei[1] , Li Aidan[2], Xue Zhongyu[1], Han Shuang[1]
(1.Beijing ZhongJiKeHai Technology Development Ltd., Beijing 100048;
2. Beijing Institute of Technology, Beijing 100081)

Abstract: The web was an information resource with dynamic state, to access and analyze the data we must study how to integrate heterogeneous architecture data and choose fit techniques to analyze, manage and integrate the data. How to apply plentiful web data to the field of web data mining has been brought into focus. The article discusses the data heterogeneity problem in Web by introducing the application of XML in the field of web data mining. By using XML technology a data extraction model is established for solving most of the difficulties in Web data mining caused by heterogeneous, unstructured problems on Internet.

Key words: data mining, semi-structured, XML technology, data extraction, mode

1 引言

在信息多元化的今天，Web为我们提供了涵盖电子商务、教育、经济、政务等方面的海量信息。随着Web技术的不断发展，跨地域、跨部门、跨平台的业务合作已成为当前企业或政府运营的主要模式。但是由于数据大多存储在关系数据库中，不同部门之间数据库结构存在异构性，导致数据表示格式不一致，给交换双方在识别对方所发送数据时带来一定的困难。数据库的数据模型能够描述特定的数据对象，但面向Web数据比面向单个数据仓库的数据挖掘要更加复杂[1]。XML（可扩展标记语言）具有与平台无关、易扩展和自描述性等优点[2]，为基于XML的异构数据处理的可行性提供了保证。

国内外关于数据挖掘中数据异构解决技术的研究很多。贝尔实验室的WenfeiFan在建立XML约束方面作了大量研究[3]。美国加利福尼亚大学洛杉矶分校(UCLA)DongwonLee在美国国防部高级研究项目(DARPA)和国家科学基金(NSF)的双重支持项目XPRESS XML中，提出了一系列XML与关系数据

第一作者简介：李春梅（1980- ），女，硕士，北京中机科海科技发展有限公司副总经理，高级工程师，主要研究方向：信息管理。

基金项目：国家国际科技合作计划项目“异构信息知识挖掘与可视化关键技术研究”（2010DFA14390）。
收稿日期：2012年3月26日。

库相互转化的算法及理论，如基于嵌套的面向XML的关系数据发布技术，基于约束的XML DTD（文档类型定义）模式向关系模式的转化等[4-5]。

国内主要以曾宇昆、方翔、谷长勇、孙宏伟等人的算法为代表[6]，他们大都使用DTD作为XML文档类型说明语言。然而，DTD不能表达元素中字符数据的数据类型，不支持命名空间，缺乏对XML文档的内容及语义的约束机制，只提供非常有限的几种数据类型，无法将两个有着完全相同内容的元素联系起来。许多中间件产品都提供了在关系数据库与XML文档之间转换数据的方法，如ASP2XML，DB2XM等等。各种主流的数据库产品也集成了这些中间件或提供了关系数据格式与XML数据格式的转换工具，如SQLServer2000、Oracle 8i/9i、DB2、Sybase等都增加了对XML的支持，但它们大都功能有限。

本文以Web数据挖掘中出现的数据异构作为研究对象，针对目前异构数据处理的需求，利用XML技术解决Web数据挖掘中数据异构的问题，使数据提供者能方便地将数据库资源发布使用，同时使数据使用者能够方便透明地访问分布式异构数据库资源。采用XML作为中间数据表示格式，数据交换双方通过本地关系数据库与XML的相互映射，使用XSLT解决数据异构问题，实现数据挖掘中异构数据问题的处理。

2　Web数据挖掘技术

2.1　Web数据挖掘

Web数据挖掘是由Oren Etzioni在1996年首先提出的[7]。根据对象不同，Web数据挖掘分为Web使用记录挖掘、结构挖掘和内容挖掘。Web使用记录挖掘是从“访问印记”中获取有价值的信息，预测用户网上行为，根据用户的需求调整网站结构；Web结构挖掘是通过网页超链接发现相互联系的结构，通过隐藏的结构模型，对Web页面重新分类，找到相似网址；Web内容挖掘是从描述中抽取有价值信息的过程，是一种基于网页内容元素对象的Web挖掘。Web数据挖掘分类如图1所示。

2.2　Web数据挖掘主要技术

2.2.1 聚类分类技术

聚类是对记录分组，把相似的记录聚在一个集里。分类则是将数据集中的一些能描述数据类的属性提取出来，利用基于归纳的学习方法得到分类模型，使所有新加入的数据都能够分到某个已知的数据类别中，实现对新数据的预测。聚类和分类的不同是聚类不依赖于预先定义好的类，不需要训练集，在分类基础上挖掘出共同特征，根据对象的特征控制同一类数据。该技术可以根据访问用户对个人信息或访问模式进行分析，挖掘出可理解的知识模式。

2.2.2　关联规则技术

数据关联是数据库中存在的一类重要的可被发现的知识，关联规则就是要找出隐藏在数据间的相互关系，发现大量数据中项目集之间满足如最小支持度和最小置信度的所有关联或相关联，找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数，因此关联分析生成的规则带有可信度。在个性化信息检索中，用户对信息的需求强调兴趣信息的相关性，而非相异性。从用户访问序列数据库的序列项中挖掘出相关规则。

2.2.3　可视化技术

可视化数据挖掘技术是建立在可视化和分析

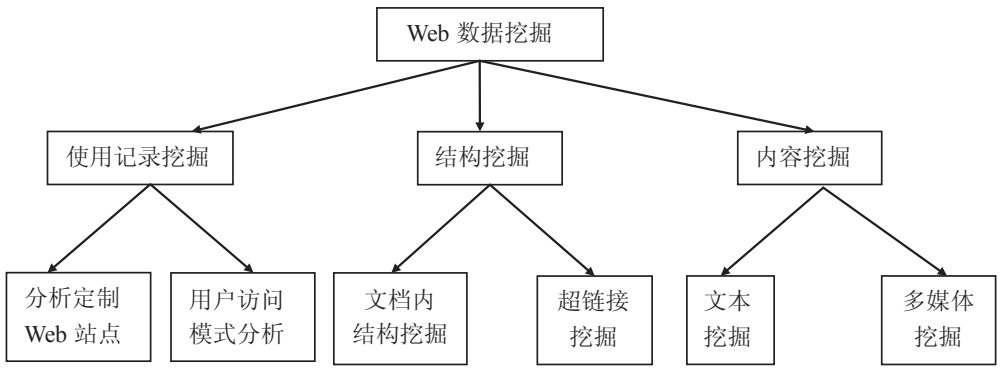

图1　Web数据挖掘分类

过程基础上，是对数据挖掘结果的表示方式，显示数据的功能性，加强数据挖掘处理。信息的可视化侧重于多维的标量数据，研究的重点在于设计和选择合适的显示方式表示庞大的多维数据及相互之间的关系，解决如何实现模型中的映射、变换和交互控制，使数据和挖掘结果更容易理解，允许对结果进行比较和检验。可视化技术突破了仅由文字、图表、数据、视频表达的空间概念，直接将真实空间与虚拟网络空间相结合。揭示了信息在网络空间中的多维特征，使网络信息分布可视化，更清楚地分析数据。

2.2.4 决策树技术

在给定已知分类类别的数据集之后，决策树算法通过自上而下的方法生成一个树状图，通过对大量数据有目的地分类，从中找到一些有价值的、潜在的信息。根据信息论原理，寻找数据库中具有最大信息量的字段，建立决策树的一个结点，再根据字段的不同取值建立树的分枝。建好模型在预测新数据集时，只需要从根结点开始依次向下判断，直至达到叶结点，得到一个该数据样本的类别。

3 数据挖掘中的数据异构问题

数据异构问题主要表现为数据库的异构，异构数据库可以实现数据的共享和透明访问，每个数据库在加入异构数据库系统之前本身就已经存在，拥有自己的DBMS（数据库管理系统）。异构数据库的各个组成部分具有自身的自治性，实现数据共享的同时，每个数据库系统仍保有自己的应用特性、完整性控制和安全性控制。异构数据库的异构性主要体现在3个方面。一是基础操作系统的异构：各个数据库系统运行的操作系统可以是Unix、Windows、Linux等。二是计算机体系结构的异构：各个参与的数据库可以分别运行在大型机、小型机、工作站、PC机或嵌入式系统中。三是存储模式的异构：一般的存储模式包括关系模式、对象模式、对象关系模式和XML文档树型模式等几种，其中关系模式为主流存储模式。即便是同一类存储模式，他们的模式结构可能也存在差异，例如Oracle所采用的数据类型与SQLServer所采用的数据类型并不是完全一致的。

解决Web数据挖掘中数据异构的方法有以下4种。

（1）对象－关系映射方法。关系数据库数据是结构化数据，XML文档属于半结构化数据，为了在数据库和XML文档之间传递数据，必须在文档结构和数据库结构之间建立映射。在基于XML的数据处理系统中，用户通过客户端向数据交换子系统提交数据请求，系统从关系数据库中提取用户请求中所涉及的关系模式，对所提取的关系模式进行切割，将一个复杂的关系模式切割为若干相互独立的子关系模式，并结合用户数据请求信息，针对各子关系模式构造其所对应的SQL语句，实现用户请求到SQL语句序列的转换。所有支持XML的关系型数据库和某些中间件都可以使用对象－关系的映射方式，通过执行各子关系模式所对应的SQL语句从数据库中抽取数据，根据所提取的关系模式构造对应的XMLSchema。将XML文件中的数据视为特定的对象树模型，通过传统的对象－关系映射技术或SQL命令的对象视图将该模型映射到关系型数据库。

（2）将XML数据导入到Oracle数据库。数据库厂商Oracle开发了转换XML到数据库表中的辅助工具。Oracle XML SQL Utility把XML文档元素建模为一组嵌套的表，通过使用Oracle对象数据类型建模套入元素。“XML−to−SQL”可能要求数据模型改进或重新构造最初的XML文档，使用Oracle XML Save 来存储XML文档到对象关系模型中。用SQL命令创建数据库和数据库表，再逐字段解析各字段的值，并用SQL命令将这些值插入到新建的数据表，保存到相应的数据库中。

以下是SQLServer数据库数据转化成XML文件及架构，XML数据保存到Oracle数据库的部分程序代码[8−9]。

```
//读取表名，填充到Dataset ds中
name=(Session["name"]).ToString();
DataSet myDataSet=new DataSet();
DataSet ds=new DataSet();
SqlConnection my Connection=new SqlConnection(Co
nnectionString+name);
String SQL="select name from sysobjects where
type='U'"
SqlDataAdapter myComm=new SqlDataAdapter(SQL,
myConnection);
myComm.Fill(ds,"mydata");
//解析多表
for(int i =0；i<ds.Table[0].Rows.Count；i++)
{
```

```
        myConnection.Open();
        string strSQL=
        "select * from"+ds.Tables[0].Rows[i][0].
        ToString();
        SqlDataAdapter myCommand=new SqlDataAdap
        ter(strSQL,myConnection);
        myCommand.Fill(myDataSet,ds.Tables[0].
        Rows[i][0].ToString());
}
//将DataSet中的数据保存为XML数据，并保存到指定的文件夹
myDataSet.WriteXmlSchema(Server.MapPath("dataDB//")+name+".xsd");
myDataSet.WriteXml
(Server.MapPath("dataDB//")+name+".xml",XmlWriteMode.IgnoreSchema);
XmlDataDocument datadoc=new XmlDataDoeument(myDataSet);
Datadoc.Save(Server.MapPath("dataDB//")+nanle+"1.xml");
myConnection.Close();
//读取XML数据填充到DataSet ds中
DataSet ds=new DataSet();
ds.ReadXml(Server.MapPath("dataDB//")+(Session["name"]).ToString()+".xml");
dbName=(Session["name"]).ToString();
//用SQL新建数据库
String strCommand="create database"+"_"+dbName;
OleDbCommand myCommand=new OleDbCommand(strCommand,myConnection);
myConnection.Open();
myCommand.ExecuteNonQuery();
myConnection.Close();
//在新建数据库中根据从DataSet中解析得到的字段名及其类型创建数据表
//将记录写入新建的数据表中
DataTable myTable=ds.Tables[n];
dtName=myTable.TableName;
string colName,colNamel,colT;
……
myOleDbDataAdapter.InsertCommand=new SqlCommand
("Insert into"+"_"+dtName+"VALUES"+"("+colNamel+")",myConnection);
……
string col=colName.Substring
(0,colName.Length-colName.Substring(colName.IndexOf(",")).Length);
colName=coLName.Substring(colName.IndexOf(",")+1);
string coll="@"+col;
string ct=colT.Substring(0,colt.Length-colT.Substring(colt.IndexOf(",")).Length);
colT=colt.Substring(colt.IndexOf(",")+1);
ct="SqlDbType. "+ct:
workParam=myOleDbDataAdapter.InsertCommand.Parameters.AddWithValue(coll,colT);
workParam.SourceColumn=col;
workParam.SourceVersion=DataRowVersion.Current;
……
myOleDbDataAdapter.Update(ds,dtName);
```

采用类似的方法，可以实现XML文件与其他关系数据库之间的相互转换。

（3）使用XSLT（扩展样式表转换语言）解决数据异构问题。XSLT定义了用于将XML源文档转换成结果文档的指令元素。对基于Web的数据应用而言，无论采用何种数据格式，只要能将其转换成HTML（超文本标记语言）数据格式，就可以利用一般的数据库技术进行处理，所以利用XSLT技术作为中转可以实现XML数据类型到HTML数据类型的转化，从而完成异构数据的转换，最终将数据显示在浏览器上。

（4）XML与XSLT生成跨平台SQL实现异构数据转换。由于XML的数据描述性和扩展性，本方法采用XML文件存储数据库信息，以XSLT样式作为模板来解析XML文件。采用模板映射法将源数据库数据信息存储到一个良构的XML文件中，从XML文件中提取需要的数据，按照XSLT样式表模板文件所定义的格式组合生成所需的SQL脚本文件，将SQL脚本导入到目标数据库中运行，即可生成目标数据库系统中的数据库文件，完成数据的转换。

由于数据是统一导出导入，XML数据存储文件和XSLT模板文件的结构相对比较稳定；XSLT文件使用XSLT解析器解析，所有逻辑问题都集中在XSLT模板中，并且XSLT文件开发起来比较简单，

从而降低了系统的开发难度；XSLT模板文件的可扩展性强，可按需求编写XSLT模板文件，无需更改应用程序，增加了系统的可维护性。

4 XML技术

4.1 处理异构数据的基本过程

传统的关系数据库之间的数据交换大多采用文本文件作为中间媒介，但是，文本文件只能实现单表间的简单信息的交互。Excel电子表格以二进制格式存储，其优于传统纯文本文件的一个好处是可以对数据分类（数值型、文本型等），所以电子表格也为关系数据库之间进行信息交换提供了方便，但随着计算机网络的不断发展，电子表格在数据交换中的应用远不如XML。

利用XML文档作为中间形式，对异构数据格式进行转换，从而被其他的系统接收，实现异构数据源之间的数据交换。这种基于复制技术的异构数据转换方法既保持了各数据库相对独立性和自治性，又使各异构数据源实现了信息集成。

XML技术已渐成规范化，开发人员能够利用XML进行格式标记和数据交换。XML在三层架构上为Web数据处理提供了很好的方法。使用可升级的三层模型[10]，XML可以把数据从商业规范和表现形式中分离出来，为数据挖掘任务提供统一有效的数据集。在Web上使用XML交换数据的步骤如图2所示。

根据转换规则构造新类型结构目标树，转换过程就是从源树生成结构树的过程。在结构树的构造中，源树可以被过滤和重新排序，还可以增加任意的结构。XML源文档被解析成DOM树存放在内存中，接着对文档进行分析，每个DOM树中的节点都会与一个模式相比较，目标树根据格式化处理，在显示器上输出。XML解析器能够验证XML文档结构，并为访问提供一个编程界面。通过使用文档对象模型处理和编辑方法，开发人员能够处理XML结构树的元素。XML数据只包含事实不显示信息，可以使用XSL（可扩展样式表语言）来描述XML文档如何显示。

4.2 数据抽取模型

如何用一个模型清晰地描述Web上的半结构化数据，实现异构数据的集成，是进行数据挖掘的关键，也是当前网络信息管理和应用方面的研究热点。Web站点上的数据信息一般采用HTML描述，信息只能在浏览器中提供数据的显示方式，要想在这种方式下获得数据描述，真正做到准确、高效的挖掘是不可能的，必须寻求新的解决途径。XML是一种层次结构数据模型，实现了数据与形式的分离，通过转换可以将XML文档中的标记与关系数据库中的属性对应起来，可以支持精确的查询。XML具有强大的数据描述和数据抽取功能，主要通过4个环节来实现：标示数据源并映射成XHTML（可扩展超文本置标语言）；查找数据内的引用点；映射成XML；合并结果并处理数据。基于XML的数据抽取过程如图3所示。

4.3 与传统方法的比较优势

传统的异构数据转换是借助专有工具包来实现的，需要数据库管理人员掌握较丰富的程序设计知识，使用过程比较繁琐，效率低，二次维护不方便。在进行Web页面信息挖掘时，站点的数据是由开发人员自行设计的，表面上看，XML文件与HTML文件比较相似，都以一对相互匹配的起始和结束标记符来标记信息，但二者功能不同。

HTML是一种描述网页文档的标记语言，提供了如何在浏览器中显示信息的方式，而没有反映数据本身所包含的语义。XML是一种完全面向数据语义的标记语言，它取消了显示样式与布局描述能

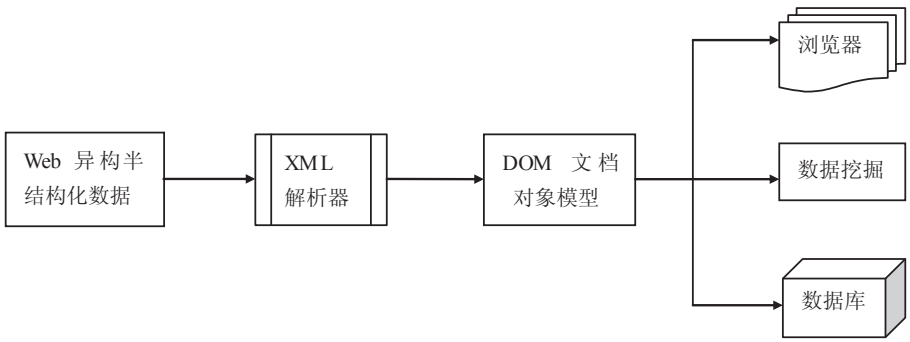

图2 XML交换数据的步骤

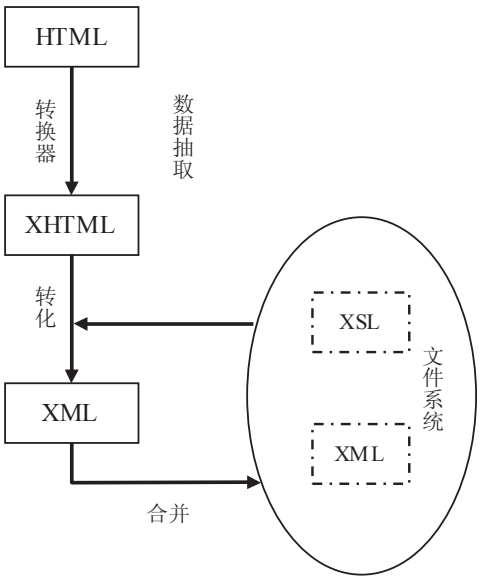

图3　基于XML数据抽取过程

力，突出了数据的语义与元素结构描述能力。XML通过自身就可以轻松地进行数据操作，可以多种方式显示，也可以由其他应用软件进行深入的处理，提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。XML本身就是一类描述性语言，在数据传送过程中，XML始终保留了诸如父/子关系的数据结构，多个应用程序可以共享和解析同一个XML文件，不必使用传统的字符串解析或拆解过程。

在超链接方面，HTML虽然可以链接本机或其他主机上的文件，但只能指定单向且固定的链接位置，提供服务的计算机所采用的OS平台及所用的数据库管理系统不同，直接数据交换困难，必须找到一种可以表达数据的统一标准。XML可以建立多重链接，除目标网页位置外，同时可提供如何从其他网址链接的信息，可以进一步指定目标网址找到后的动作，是否自动显示或搬运到原有的文件内。由于XML的自定义性及可扩展性，能够表达各种类型的数据，客户收到数据后可以进行处理，也可以在不同数据库间进行传递，XML解决了数据的统一接口问题。

5　结论

虽然基于Web数据挖掘技术的应用研究取得了一定进展，但Web数据的特殊性，致使数据控制和数据集成仍然非常困难。XML的系统独立性使得XML数据能够在不同的系统中使用各种编程语言解析和处理，屏蔽异构数据库间系统环境的差异。本文探讨关系模式与XML模式的特点以及实现关系模式到XML模式映射中存在的难点，建立基于XML的层次结构数据模型，实现数据与形式的分离，通过转换可以将XML文档中的标记与关系数据库中的属性对应起来，可以支持精确的查询，运用XML技术有效地获取Web上的数据。分析XML技术与传统方法在数据挖掘中的比较优势。利用XML技术定位和抽取Web数据，XML+XSLT生成跨平台SQL脚本，准确高效地从Web页面中提取有用信息，在一定程度上解决了Web数据挖掘面临的异构信息难以提取的问题。

参考文献

[1] Han Jiawei, Kamber Micheline. Data Mining: Concept and Tbchniques[M]. San Francisco: Morgan KaUfmann Publishers. Inc. 2001.

[2] Shanmugasundaram Jayavel, Tufte Kristin, He Gang, et al. Relational Databases for Querying XML Documents: Limitations and Opportunities[C]//Edinbergh, Scotland: Proceeding of the 25th International Conference on Very Large DataBases(VLDB). 1999: 302−314.

[3] Fan W, Simeon J. Integnty Constraints for XML[J]. Journal of Computer and System Science(JCSS), 2003, 66(1):254−291.

[4] Lee Dongwon, Chu Wesley W. Constraints−preserving Transformation from XML Document Type Definition to Relational Schema[C]//Salk Luke City, Utah: Proceedings of the 19th international conference on Conceptual Modeling(ER), 2000:323−338.

[5] Lee Dongwon, Mani Murali, Chu Wesley W. Conversions Methods between XML Schema and Relations Models[M]//Knowledge Transformation for the Semantic Web. Amsterdan: IOS Press, 2003.

[6] 方翔, 李伟生. 关系模式到XML模式的影射[J]. 计算机应用研究, 2002(1):130−132.

[7] Etzioni O, Mine G, Widener T. The World Wide Web: Quagmire or Gold Mine[J]. Communication of the ACM, 1996, 39(11):65−68.

[8] Dalvi Dinar, Gray Joe. NETXML高级编程[M]. 英宇, 林琪, 费广正, 译. 北京: 清华大学出版社, 2002.

[9] Rirdan Rebecca M. ADO. NET程序设计[M]. 李高健, 译. 北京: 清华大学出版社, 2002:23−25.

[10] Coyle Frank P. XML、Web服务和数据革命[M]. 袁勤勇, 莫青, 译. 北京: 清华大学出版社, 2003.