基于XML的WEB数据挖掘技术

徐振航 刘莉芹 (武汉科技学院经济管理学院 430073)

摘要：数据挖掘技术的核心部分已发展了近十年，研究领域涉及数理统计、人工智能、机器学习等。当今，随着人们对数据需求的不断加强，以及WEB技术的飞速发展，使得数据挖掘技术又向前迈进了一步。面向WEB的数据控制是目前数据挖掘技术的一大热点，但由于WEB数据存在方式的特殊性，使WEB数据控制变得十分复杂。而XML的出现，为WEB数据挖掘技术带来了巨大的发展契机。

关键词：数据挖掘 数据库 XML WEB

1 引言

面向WEB的数据挖掘要比面向单个数据仓库中的数据挖掘要复杂的多，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，同时可以很好的定义和解释相关的查询语言。而WEB上的数据以多种形式存在，没有特定的模型来描述，每一个站点上的数据都是由站点开发人员自行设计与组织，并且数据本身还存在着自我描述性和动态可变性，因而WEB上的数据是一种介于结构化与半结构化之间的数据，这可以称之为半结构化数据。如何用一个模型来清晰的描述WEB上的半结构化数据，是进行WEB数据挖掘的关键。现在的许多WEB站点上的信息，多用HTML来描述，因而只能在浏览器中提供数据的显示方式，要想真正做到准确、高效的挖掘数据非常困难。XML是由W3C定义的一个新的的标记语言，其TAG具有语义，由用户定义，能够反映一定的数据的含义，XML的文件描述的语义非常清晰，很容易与关系数据库中的属性一一对应，并且能够支持十分精确的查询，由此可见，XML能为WEB数据挖掘带来新的解决方法。

2 面向WEB的数据挖掘

WEB上有海量的数据信息，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。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。充分利用有用的数据，废弃虚伪无用的数据是数据挖掘技术的最重要的应用。相对于WEB的数据而言，传统的数据库中的数据结构性很强，即其中的数据为完全结构化的数据，而WEB上的数据最大的特点就是半结构化，所谓半结构化是相对于完全结构化的传统数据库的数据而言。显然，面向WEB的数据控制比面向单个数据仓库的数据挖掘要复杂的多。

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东西。其次，还要解决WEB上的数据查询问题，因为如果所需的数据不能很有效的得到，对这些数据进行分析、集成、处理就无从谈起。

2.2 半结构化的数据结构

WEB上的数据与传统的数据库中的数据不同，传统的数据库都有一定的数据模型，可以根据此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

的描述WEB上的数据，正对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，即自动的从现有数据中抽取半结构化模型的技术。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。

3 XML与WEB数据挖掘技术

以XML为基础的新一代的WWW环境是直接面对WEB数据的，不仅可以很好的兼容原有的WEB应用，而且可以更好的实现WEB中的信息共享与交换。XML可看作一种半结构化的数据模型，可以很容易的将XML的文档描述与关系数据库中的属性一一对应起来，实施精确的查询与模型抽取。

3.1 XML的产生与发展

XML(eXtensible Markup Language)是由万维网协会(W3C)设计，特别为WEB应用服务的SGML(Standard General Markup Language)的一个重要分支。总的来说，XML是一种中介标示语言(Meta-markup Language)，可提供描述结构化资料的格式，详细来说，XML是一种类似于HTML，被设计用来描述数据的语言。XML提供了一种独立的运行程序的方法来共享数据，它是用来自动描述信息的一种新的标准语言，它能使计算机通信把INTERNET的功能由信息传递扩大到人类其他多种多样的活动中去。XML由若干规则组成，这些规则可用于创建标记语言，并能用一种被称作为分析程序的简明程序处理所有新创建的标记语言，正如HTML为第一个计算机用户阅读INTERNET文档提供一种显示方式一样，XML也创建了一种任何人都能读出和写入的世界语。XML解决了HTML不能解决的两个WEB问题，即INTERNET发展速度快而接入速度慢的问题，以及可利用的信息多，但难以找到自己需要的那部分信息的问题。XML能增加结构和语义信息，可使计算机和服务器即时处理多种形式的信息。因此，运用XML的扩展功能不仅能从WEB服务器下载大量的信息，还能大大减少网络业务量。

XML中的标志(TAG)是没有预先定义的，使用者必须要自定义需要的标志，XML是能够进行自解释(Self Describing)的语言。XML使用DTD(Document Type Definition 文档类型定义)来显示这些数据，XSL(eXtensible Style Sheet Language)是一种来描述这些文档如何显示的机制，它是XML的样式表描述语言。XSL的历史比HTML用的CSS(层叠式样式表Cascading Style Sheets)还要悠久，XSL包括两部分：一个用来转换XML文档的方法；一个用来格式化XML文档的方法。XLL(eXtensible Link Language)是XML连结语言，它提供XML中的连结，与HTML中的类似，但功能更强大。使用XLL可以多方向连结，且连结可以存在于对象层级，而不仅仅是页面层级。由于XML能够标记更多的信息，所以它就能使用户很轻松地找到他们需要的信息。利用XML，WEB设计人员不仅能创建文字和图形，而且还能构建文档类型定义的多层次、相互依存的系统、数据树、元数据、超链接结构和样式表。

3.2 XML的主要特点

正是XML的特点决定了其卓越的性能表现。XML作为一种标记语言，有许多特点：

(1) 简单，XML经过精心设计，整个规范简单明了，它由若干规则组成，这些规则可用于创建标记语言，并能用一种常常称作分析程序的简明程序处理所有新创建的标记语言。XML能创建一种任何人都能读出和写入的世界语，这种创建世界语的功能叫做统一性功能。如XML创建的标记总是成对出现，以及依靠称作统一代码的新的编码标准。

(2) 开放，XML是SGML，在市场上有许多成熟的软件可用来帮助编写、管理等，开放式标准XML的基础是经过验证的标准技术，并针对网络做最佳化。众多业界顶尖公司，与W3C的工作群组并肩合作，协助确保交互作业性，支持各式系统和浏览器上的开发人员、作者和使用者，以及改进XML标准。XML解释器可以使用编程的方法来载入一个XML的文档，当这个文档被载入以后，用户就可以通过XML文件对象模型来获取和操纵整个文档的信息，加快了网络运行速度。

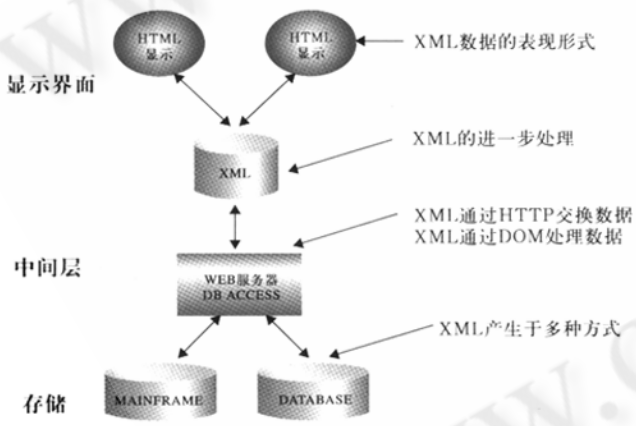
(3) 高效且可扩充，支持复用文档片段，使用者可以发明和使用自己的标签，也可与他人共享，可延伸性大，在XML中可以定义无限量的一组标注。XML提供了一个标示结构化资料的架构。一个XML组件可以宣告与其相关的资料为零售价、营业税、书名、数量或其他任何数据元素。随着世界范围内的许多机构逐渐采用XML标准，将会有更多的相关功能出现：一旦锁定资料，便可以使用任何方式透过电缆线传递，并在浏览器中呈现，或者转交到其他应用程序做进一步的处理。XML提供了一个独立的运用程序的方法来共享数据，使用DTD，不同组中的人就能够使用共同的DTD来交换数据，你的运用程序可

以使用这个标准的DTD来验证你接受到的数据是有效的，也可以使用一个 DTD 来验证你自己的数据。

(4) 国际化，标准国际化，且支持世界上大多数文字。这源于依靠它的统一代码的新的编码标准，这种编码标准支持世界上所有以主要语言编写的混合文本。在 HTML 中，就大多数字处理而言，一个文档一般是用一种特殊语言写成的，不管是英语还是日语或阿拉伯语，如果用户的软件不能阅读特殊语言的字符，那么他就不能使用该文档。但是能阅读XML语言的软件就能顺利处理这些不同语言字符的任意组合。因此，XML 不仅能在不同的计算机系统之间交换信息，而且能跨国界和超越不同文化疆界交换信息。

3.3 XML 在 WEB 数据挖掘中的应用

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。数据的集成、发送、处理和显示是下面过程中的每一个步骤，看下图:



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的联系，基于WEB的数据挖掘必须依靠它们来实现。

XML 给基于 WEB 的应用软件赋予了强大的功能和灵活性，因此它给开发者和用户带来了许多好处。比如进行更有意义的搜索，并且 WEB 数据可被 XML 唯一的标识。没有XML搜索软件必须了解每个数据库是如何构建的。由于不同来源数据的集成问题的存在，现在搜索多样的不兼容的数据库实际上是不可能的。XML能够使不同来源的结构化的数据很容易的结合在一起。软件代理商可以在中间层的服务器上对从后端数据库和其他应用处来的数据进行集成。然后，数据就能被发送到客户或其他服务器做进一步的集合、处理和分发。XML的扩展性和灵活性允许它描述不同种类应用软件中的数据，从描述搜集的WEB页到数据记录，从而通过多种应用得到数据。同时，由于基于XML的数据是自我描述的，数据不需要有内部描述就能被交换和处理。利用XML用户可以方便的进行本地计算和处理，XML格式的数据发送给客户后，客户可以用应用软件解析数据并对数据进行编辑和处理。使用者可以用不同的方法处理数据，而不仅仅是显示它。XML文档对象模式(DOM)允许用脚本或其他编程语言处理数据，数据计算不需要回到服务器就能进行。XML可以被利用来分离使用者观看数据的界面，使用简单灵活开放的格式，可以给WEB创建功能强大的应用软件，而原来这些软件只能建立在高端数据库上。另外，数据发到桌面后，能够用多种方式显示。

XML还可以通过以简单开放扩展的方式描述结构化的数据，XML补充了HTML，被广泛的用来描述使用者界面。HTML描述数据的外观，而XML描述数据本身。由于数据显示与内容分开，XML定义的数据允许指定不同的显示方式，使数据更合理地表现出来。本地的数据能够以客户配置、使用者选择或其他标准决定的方式动态地表现出来。CSS和XSL为数据的显示提供了公布的机制。通过XML，数据可以粒状的更新。每当一部分数据变化后，不需要重发整个结构化的数据。变化的元素必须从服务器发送给客户，变化的数据不需要刷新整个使用者的界面就能够显示出来。但在目前，只要一条数据变化了，整一页都必须重建。这严重限制了服务器的升级性能。XML也允许加进其他数据，比如预测的温度。加入的信息能够进入存在的页面，不需要浏览器重新发一个新的页面。XML应用于客户需要与不同的数据源进行交互时，数据可能来自不同的数据库，它们都有各自不同的复杂格式。但客户与这些数据库间只通过一种标准语言进行交互，那就是XML。由于XML的自定义性及可扩展性，它足以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。总之，在这类应用中，XML解决了数据的统一接口问题。但与其他的数据传递标准不

同的是：XML并没有定义数据文件中数据出现的具体规范，而是在数据中附加TAG来表达数据的逻辑结构和含义。这使XML成为一种程序能自动理解的规范。

XML应用于将大量运算负荷分布在客户端，即客户可根据自己的需求选择和制作不同的应用程序以处理数据，而服务器只须发出同一个XML文件。如按传统的“CLIENT/SERVER”工作方式，客户向服务器发出不同的请求，服务器分别予以响应，这不仅加重服务器本身的负荷，而且网络管理者还须事先调查各种不同的用户需求以做出相应不同的程序，但假如用户的需求繁杂而多变，则仍然将所有业务逻辑集中在服务器端是不合适的，因为服务器端的编程人员可能来不及满足众多的应用需求，也来不及跟上需求的变化，双方都很被动。应用XML则将处理数据的主动权交给了客户，服务器所作的只是尽可能完善、准确地将数据封装进XML文件中，正是各取所需、各司其职。XML的自解释性使客户端在收到数据的同时也理解数据的逻辑结构与含义，从而使广泛、通用的分布式计算成为可能。

XML还被应用于网络代理，以便对所取得的信息进行编辑、增减以适应个人用户的需要。有些客户取得数据并不是为了直接使用，而是为了根据需要组织自己的数据库。比方说，教育部门要建立一个庞大的题库，考试时将题库中的题目取出若干组成试卷，再将试卷封装进XML文件，接下来在各个学校让其通过一个过滤器，滤掉所有的答案，再发送到各个考生面前，未经过滤的内容则可直接送到老师手中，当然考试过后还可以再传送一份答案汇编。此外，XML文件中还可以包含进诸如难度系数、往年错误率等其他相关信息，这样只需几个小程序，同一个XML文件便可变成多个文件传送到不同的用户手中。

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的结合在一起，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。XML的扩展性和灵活性允许XML描述不同种类应用软件中的数据，从而能描述搜集的WEB页中的数据记录。同时，由于基于XML的数据是自我描述的，数据不需要有内部描述就能被交换和处理。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。相信在以后，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。■