文章编号：1000－5471(2007)04－0103－04

电子商务网站的 Web 数据挖掘系统设计①

李 献 礼

长江师范学院 教育技术与信息中心，重庆 408003

摘要：在分析了电子商务网站 Web 数据源及其挖掘任务基础上，设计了一种电子商务网站的 Web 数据挖掘系统模型，详细阐述了模型的数据挖掘过程与关键技术.

关 键 词：数据挖掘；Web 挖掘；电子商务；划分

中图分类号：TP393 **文献标识码**：A

随着 Internet 的普及，电子商务的兴起，正在改变着人们的商务理念，经销商和客户之间通过互联网进行交易，节省了大量的费用和时间. 如何更快、更好地利用这一现代交易手段，缩短经销商和客户之间的距离是目前电子商务要解决的问题. 由于电子商务网站在运转的过程中积累下来的大量的有关用户商务行为数据给企业提供了丰富的信息，这些数据中隐藏着客户的商务行为模式. 对这些数据进行挖掘，可以找出这些隐藏的模式，企业用户可以根据挖掘的结果提出针对性的商务计划或者对网站进行改进，以推出个性化的页面，这又将促进电子商务更好地以客户为中心进行运转.

1 电子商务网站的挖掘任务与数据源

1.1 挖掘任务

本方案是针对建材的电子商务平台进行挖掘. 该平台为会员用户提供的功能主要有集中采购、分散采购、招标投标等. 用户访问的历史数据包括两部分，一是用户登录日志，二是用户进行商务活动的日志. 因此确定下述的挖掘的任务.

1) 通过对这些数据的挖掘，得到用户的商务行为模式的分类，并根据分类来预测用户的行为，根据这些预测给每个用户相应的个性化网页.

2) 通过挖掘，我们可以分析到网站的各功能模块的访问频率与顺序，根据这些结果对网站结构进行调整，提高客户访问的效率.

1.2 电子商务中 WEB 数据挖掘的数据源

电子商务中，客户的浏览信息通过 Web 服务器、代理服务器或者客户登记信息自动搜集，并保存在日志文件中，数据源主要有以下几种形式.

1.2.1 服务器端的网页数据以及日志文件

Web 结构挖掘是从 WWW 上的组织结构和链接关系中推导知识. 由于超文本文档间的关联关系使得 WWW 不仅仅可以揭示文档中所包含的信息，同时也可以揭示文档间的关联关系所代表的信息. 利用这些信息可以对页面进行排序，发现重要的页面. 挖掘 Web 结构的目的是发现页面的结构，在此基础上对页面进行分类和聚类从而找到权威页面.

① 收稿日期：2007－06－15
基金项目：重庆市教委科学研究项目(项目合同号：KJ071306).
作者简介：李献礼(1960－)，男，四川人，副教授，主要从事计算机网络应用，数据挖掘的研究.

WWW 上每一个提供信息资源的服务器上都有一个结构比较好的记录集，即 Web 访问日志. 每当有获取资源的请求到来时，Web 服务器都将记录和积累这些关于用户交互作用的数据. 分析不同的 Web 站点和 Web 访问日志可帮助人们理解用户行为和 Web 的结构，从而改进站点结构，为用户提供个性化的服务.

1.2.2 代理服务器端数据

代理服务器相当于在客户浏览器和 Web 服务器之间提供了缓存功能的中介服务器，它的缓存功能减少了 Web 服务器的网络流量，加快了网页的运行速度，同时将大量的用户访问信息通过代理日志的形式保存起来. 对此类信息的分析也可以助于对客户浏览习惯和目标的归纳和推测.

1.2.3 客户登记信息

客户登记信息是指客户通过 Web 网页在屏幕上输人并提交给服务器的相关信息. 若新客户到商城购物，首先要注册登录，然后才能购买商品，进行支付. 要想更好地了解客户，必须将客户的登记信息和访问日志结合起来分析，才能得出更准确的判断.

1.2.4 HTTP 请求信息[3]

辅之于监视所有到达服务器的数据，提取其中的 HTTP 请求信息. 此部分数据主要来自浏览者的点击流，用于考察用户的行为表现. 网络底层信息监听过滤指监听整个网络的所有信息流量，并根据信息源主机、目标主机、服务协议端口等信息过滤掉垃圾数据，然后进行进一步的处理，如关键字的搜索等，最终将用户感兴趣的数据发送到给定的数据接受程序存储到数据库中进行分析统计.

以上数据源可以分成内容数据、结构数据和用户使用数据.

2 系统结构设计与挖掘过程

2.1 系统结构设计

本系统[1,2]由数据挖掘系统、用户和电子商务系统三部分组成. 其中数据挖掘系统是一个独立的三层结构. 最底层是数据层，即数据库和其他异构数据源. 这些数据源来自电子商务系统的关系数据库系统以及用户提供的知识等. 它的中间层是挖掘层，有多维数据库和挖掘引擎. 顶上层是用户界面，用来与用户交互. 整个挖掘系统的如下图 1 所示.

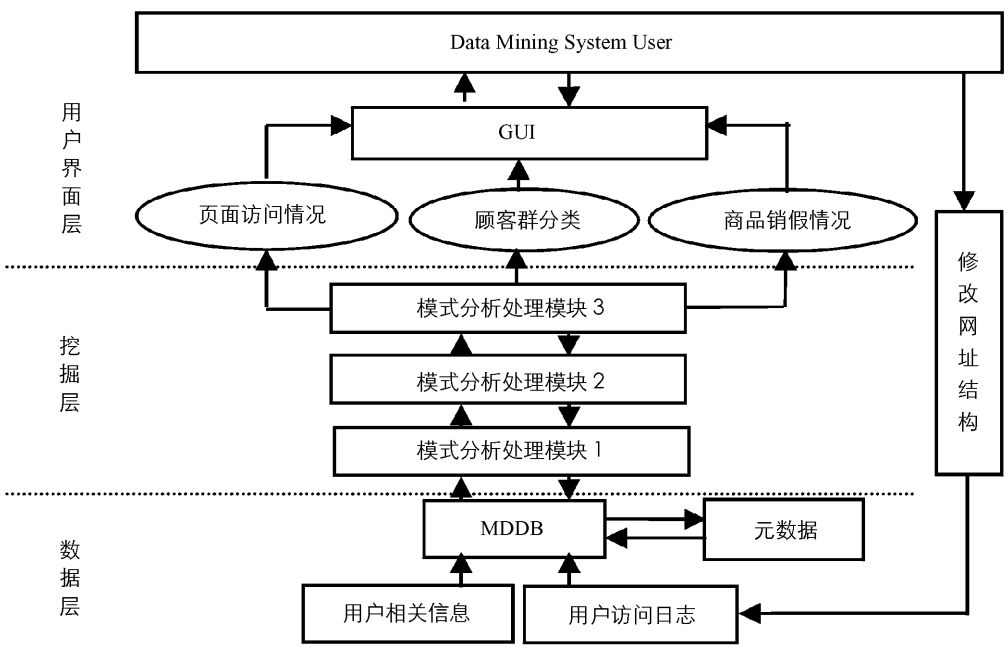

图 1 电子商务网站数据挖掘系统模型

从上图可知：电子商务系统是向客户提供服务的系统，它会产生许多信息，给挖掘系统提供了数据的源泉. 而用户的作用就是与挖掘系统交互，对挖掘提供指导，并根据挖掘出来的知识，将结果人为地反馈

给电子商务系统.

2.2　挖掘方法的选择

根据挖掘的任务，确定挖掘的算法. 由于我们要对用户分类，分类的标准是未知的所以采用聚类分析. 聚类分析就是将数据对象分组成多个类或者簇，在同一个簇中的对象之间具有较高的相似度，而不同的簇中的差别较大. 相异度是来衡量对象之间相异程度的属性，一般通过对象的属性值来计算. 通常用欧几里得距离来作为相异度. 本系统采用基于划分的 k -中心点的方法.

2.3　挖掘的过程

对在线访问客户数据的挖掘有两种办法，一是客户访问信息的挖掘，另一部分是客户登记信息的挖掘. 其挖掘的过程如下：

2.3.1　数据预处理

我们所得到的数据一般都具有不完全性、冗余性和模糊性，要使挖掘内核更有效地挖掘出知识，就必须为它提供干净、准确、简洁的数据 . 预处理主要对用户访问日志进行数据清洗、用户惟一性识别、用户会话识别、完善访问路径和事务识别等处理.

2.3.2　模式发现

模式发现阶段就是利用挖掘算法挖掘出有效的、新颖的、潜在的、有用的及最终可以理解的信息和知识. 在经过数据预处理阶段后，可根据不同的需求选择模式发现技术[4]. 如统计分析、关联规则、时序模式、路径分析及聚类、分类技术.

2.3.3　用户交互

数据挖掘过程需要用户交互. 这种交互主要是两方面：一是用户向挖掘系统提供参数或者约束条件，指导挖掘；二是因为挖掘的目的是不确定的，所以挖掘有时会返回很多的结果，其中大部分不是用户所想要的，用户只希望看到他想得到的那一部分结果集. 同时，对于挖掘系统来说，指定约束条件，也会有助于对结果的剪枝，淘汰不必要的结果，提高效率. 因此，用户在挖掘开始前和进行的过程中都要根据需要给挖掘系统提出要求或者输入参数.

本系统挖掘后返回的是聚类分析的结果，由于数据是多维的，因此不能使用二维的图形来表示簇. 挖掘系统把簇用列表形式表示，使用中心点的值来代表簇中的所有对象，并配以对簇的统计信息. 用户根据结果，还将根据实际情况进行进一步的处理，针对客户的分类情况，并采取相应的措施.

2.3.4　用户与电子商务系统的交互

挖掘系统与电子商务系统不直接发生联系，这个过程由用户来完成. 得到挖掘结果以后，用户还需要对结果进行进一步的处理，分析每一个簇所代表的客户的性质，而且可能还需要对簇的结果进行修改.

在确定了客户类别以后，用户即根据客户类别的性质对网站进行修改，使得客户进来以后，网站可以分辨出它所属的类别，然后根据类别来给出提示，引导客户进入自己最需要的功能. 这样数据挖掘的一次生命周期就完成了.

3　挖掘过程中的关键问题及分析

3.1　模式分析工具

需要有集统计技术、可视化技术等为一体的工具辅助人们理解用户的访问模式. 可视化技术在其他领域的应用已取得了巨大的成功，因此理解 Web 用户的行为也是一个自然的选择. 该系统用服务器日志扩展一系列的网页访问模式称作网页路径，在分析的过程中过滤无关的 Web 页面，使人只分析有意义的部分，最终形成一个可视化的结果——一个有向环图.

在上图中的接点是页面，边是页面之间的超级链接. 此外，联机分析处理(OLAP)技术也可以应用到模式分析中来.

3.2　挖掘出来的知识的解释

无论在一般的数据挖掘领域还是 Web 挖掘中都需要有能解释挖掘出来的知识工具. 显然，这些工具要实现的功能不仅是过滤已经发现的访问模式，而且要具有有关特定问题的知识，能够指出问题的所在. 比

如，在Web挖掘中，在挖掘出的访问模式、网站拓扑结构或从用户行为模式的基础上开发智能工具，它可以建议把网站某一物理链接进行改变.

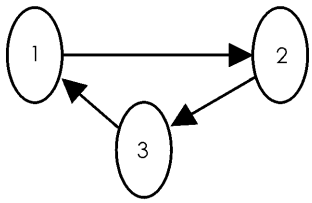

图2 有向环图

3.3 网站结构和内容对使用挖掘的影响[5]

目前有许多工具可以进行数据的清理和Web服务器日志中的会话识别，还有大量的数据挖掘算法从预处理后的数据集中发现用户使用模式和预测趋势，但最终Web使用挖掘的效果依然不能令人满意，其中一个重要的原因就是人们忽视了对使用挖掘效果起着重要影响作用的网站结构和内容.

网站结构和内容的处理是一个内部关联的任务. 网页如何链接取决于网页的浏览方式，网站内容的创建技术又决定着网站的内容和结构，而不同的用户则决定着网站主页内容的设计. 因此，网站的结构、内容和用户的使用有着密不可分的联系，网站的结构和内容影响着Web使用挖掘的不同阶段，页面文件在语义上依赖着网站内容，而网站内容的决定是一个手工过程，取决于创建网站的技术和分析目的.

4 结 束 语

电子商务的Web挖掘是一个有巨大发展前景的研究领域. 电子商务通过Web上的使用挖掘所提供的足够的知识，可以锁定相当数量的顾客进入商务关系中，以改善销售状况和保存客户关系，从而增加市场效益. 同时，通过Web使用的个性了解，比较已存在顾客的综合个性，挖掘出潜在的新顾客的个性、生活方式和特点.

参考文献：

[1] Barnsh adMobster, RCooley, jSrivastava. DataPreparation for Mining World Web Browsing Patterns [J]. Journal of Knowledgeand Information system, 1999, 1(1): 35－38.

[2] 幻张娥，冯秋红，宜葱玉，等. Web使用模式研究中的数据挖掘 [J]. 计算机应用研究，2001，18(4)：80－84.

[3] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques[M]. 北京：机械工业出版社，2000.

[4] Imon W H. Building the DataWarehouse[M]. 北京：机械工业出版社，2000.

[5] 刘丽珍，宋渤涛，陆玉昌. Web使用挖掘的应用研究[J]. 计算机科学，2003(1)：30.

Solution of E-commerce Model Based on WEB Data Mining

LI Xian-li

Center of Education Technology and Information, Yangtze Normal University, Chongqing 408003, China

Abstract: Based on an analysis of data sources of the electronic commerce websites and the task of their mining, this paper puts forward an application model, which uses data mining techniques in E-commerce and, finally, elaborates the process of data mining and its key techniques.

Key words: data mining; Web mining; E-commerce; partition

责任编辑 汤振金