# 面向电子商务的Web使用挖掘技术应用研究

朱志国,孔立平
(东北财经大学,辽宁 大连 116023)

**摘 要**:随着电子商务的深入发展,电子商务站点每天需要处理大量的数据,但数据资源中蕴涵的重要信息却至今未能得到充分的挖掘和利用。在日益激烈的电子商务市场竞争中,任何与消费者行为有关的信息对经营者来说都是非常宝贵的。企业了解用户的访问模式显得非常重要。给出Web使用挖掘的定义和完整模型框架,然后对Web使用挖掘中主要步骤的最新研究进展状况做详细的阐述和分析,其中包括:数据采集、数据预处理、模式发现、模式分析。最后对传统的和基于Web使用挖掘技术的电子商务结构模型做了对比,并深入分析了Web使用挖掘在电子商务的应用。

**关键词**:电子商务;Web数据挖掘;Web使用挖掘;模式发现
**中图分类号**:TP311;F724.6 **文献标识码**:A **文章编号**:1673-629X(2008)06-0228-05

# Research and Application of Web Usage Mining Technology Oriented E-Commerce

ZHU Zhi-guo, KONG Li-ping
(Dongbei University of Finance and Economics, Dalian 116023, China)

**Abstract**: With the deep development of E-commerce, lots of interactive data between the website and users are produced, however the implicit knowledge in the data have not fully mined and utilized. In the ever more intense competitiveness in E-commerce markets, it is important for E-commerce enterprise to understand users' navigational patterns in website. Firstly, presents the definition and full architecture of WUM. Secondly, many new advances in major steps of WUM are elaborated and analyzed, including data collection, data preprocessing, patterns discovering, patterns analysis. Finally, gives the comparison of traditional E-commerce structure model and WUM-based one and then the applications of the web usage mining in the field of E-commerce are deeply analyzed.

**Key words**: E-commerce; Web mining; Web usage mining; patterns discovering

## 0 引 言

随着Internet的迅猛发展,电子商务以其成本低廉、方便、快捷、安全、可靠、不受时间和空间的限制等突出优点成为互联网上发展最快的应用之一。2004年,美国B2B贸易的交易量达到3万亿美元,中国电子商务的交易总额达到4400亿元人民币。据预测,到2005年,世界各国公司通过因特网购买的商品和服务的贸易额可望达到4.3万亿美元,中国电子商务的交易总额将激增至6200亿元人民币[1]。

电子商务网站每天需要处理大量的数据,但数据资源中蕴涵的重要信息却至今未能得到充分的挖掘和利用。在日益激烈的电子商务市场竞争中,任何与消费者行为有关的信息对经营者来说都是非常宝贵的。为了解决数据爆炸但信息贫乏的现象,通过Web挖掘,经营者利用客户信息,分析和预测顾客行为,降低运营成本,提高竞争力。为了解决数据爆炸但信息贫乏的现象,通过Web使用挖掘技术,经营者充分利用电子商务网站服务器上收集到的客户信息,并且进一步分析和预测顾客行为,从而电子商务企业可以更加高效地利用顾客信息,同时及时地发掘出潜在的市场,提高自身的竞争力。

**收稿日期**:2007-09-11
**基金项目**:国家自然科学基金资助项目(70272050)
**作者简介**:朱志国(1977-),男,博士研究生,CCF会员,研究方向为信息系统工程、web数据挖掘。

## 1 Web使用挖掘技术

一般地,Web挖掘分为三类[2]:Web内容挖掘、Web结构挖掘和Web使用挖掘。Web内容挖掘是从文档内容或其描述中抽取知识的过程;Web结构挖掘是从WWW的组织结构和链接关系中推导知识。目前国际上对Web使用挖掘的研究比较多。

Web使用挖掘[3~5](WUM)是指能够从服务器、浏览器端的日志记录和用户的个人信息中自动发现隐

藏在数据中的模式信息,了解系统的访问模式以及用户的行为模式,从而做出预测性分析的技术。WUM的结果通常是用户群体的共同行为和共性兴趣,以及个人用户检索偏好、习惯和模式等。

图1给出了一个面向电子商务的Web使用挖掘系统的比较完善的框架模型图。从图中可以看出它包含了数据采集、数据预处理、模式发现和模式分析4个主要的阶段。

### 1.1 数据采集

源数据收集在Web使用挖掘中是第一步,数据源主要包括:Web服务器日志(包括服务器日志、引用日志和代理日志)、Web站点的拓扑结构和站点文件、用户的注册信息、用户调查信息、cookies,以及与网站服务相关的数据库数据等。但主要是服务器日志。

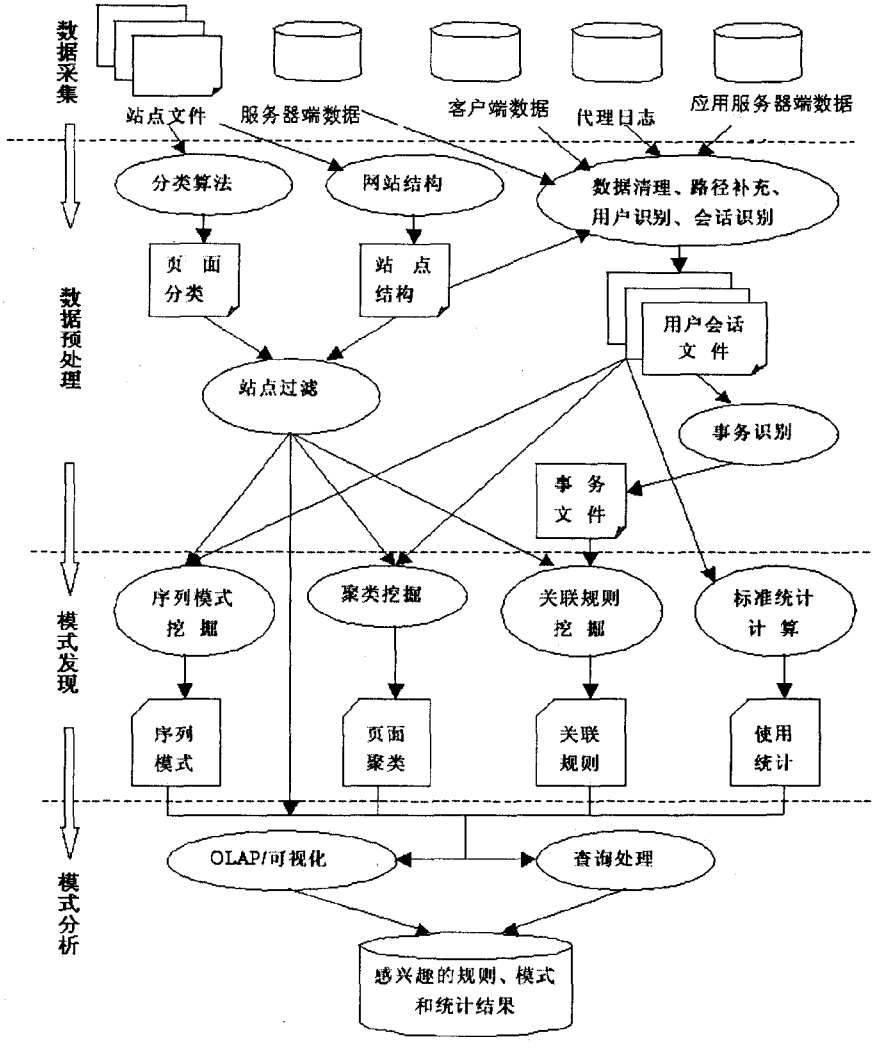


图1　Web使用挖掘框架模型图

### 1.2 数据预处理

WUM中对数据进行预处理,其目标是将包含在多种数据源中的信息转化为适合数据挖掘和模式发现所必需的数据抽象概念,然后在事务数据库上实施挖掘算法,以期最终获得有价值的规律。因为预处理的结果直接影响到挖掘算法产生的规则和模式,可以说预处理过程是WUM质量保证的关键。

通常WUM的预处理过程[6]包括数据清理、用户识别、会话识别、路径完善等几个步骤:

1)数据清理(data cleaning)。数据清理解决"脏数据(dirty data)"的问题,消解数据中的不一致性,并将多个数据源中的数据统一成一个数据存储。比如,将不同服务器上格式和描述都不同的原始数据规范化,去除日志文件中包含gif,jpeg,gif,map的文件名的项目。可以预先定义一个缺省的规则库(例如下面的算法1)来帮助删除记录。另外,还可以预先将网站分为一般网站、图片网站、音频网站等,分别建立对应的规则库,然后按照该类网站的规则库进行数据清理。

2)用户识别(User Identification)。用户识别是从日志中识别出每个访问网站的用户。最常被WUM工具使用的技术就是基于日志/站点的方法,并辅助一些启发式规则帮助识别用户。

启发式规则的核心思想:

(1)不同的IP地址代表着不同的用户;

(2)用户的IP地址相同,但相应的代理日志表明用户的浏览器类型或操作系统发生了改变,则认为代表着不同的用户;

(3)用户的IP地址相同,用户使用的操作系统和浏览器也相同的情况下,则根据网站的页面链接结构对用户进行识别(如果用户请求的某个页面不能从已访问的任何页面到达,则认为这是一个新的用户)。

要说明的一点是,这些仅是帮助识别用户的启发规则,并非使用了这些规则就能准确地识别出用户。在用户识别的过程中,还会产生一些问题,典型的有:

①单IP地址/多服务器会话。Internet业务供应商(ISP)为用户提供了许多用于上网的代理服务器。因此在同一时间段内可能有许多不同用户通过同一代理服务器(单IP地址)存取同一网站。

②多IP地址/单服务器会话。一些ISP和私用工

具会为来自不同用户的每次请求随机分配 IP 地址池中的某一个,在这种情况下,一次单独的服务器会话可能会有多个 IP 地址。

③多 IP 地址/单用户。一个用户从不同机器上网会在不同会话中使用不同地址,这就使得追踪同一用户的重复访问变得很困难。

④多代理/单用户。某用户在同一机器上打开多个浏览器窗口,访问 Web 站点的不同部分,或打开不同的浏览器进行访问,将产生单个用户的多个服务器会话。

3)会话识别(Session Identification)。在跨越时间区段较大的 Web 服务器日志中,用户有可能多次访问了该站点。会话识别的目的就是将用户的访问记录分为单个的会话。

用户会话 S 可以定义为:

$$S=<\text{UserId},\{(\text{Pid}_1,\text{time}_1),\cdots,(\text{Pid}_k,\text{time}_k)\}> \quad (1)$$

令 $RS=\{(\text{Pid}_1,\text{time}_1),\cdots,(\text{Pid}_k,\text{time}_k)\}$

$S=<\text{UserId},RS>$

其中:UserId 是用户标识。RS 是用户在一段时间内请求的 Web 页面的集合。RS 包含用户请求页面的标识符 Pid 和请求时间 time。

通常可以采用超时方法识别用户会话,对于超时阈值的设定,有两种方法,一种是设定整个用户会话的超时时间,则(1)式中的用户会话必定满足下面的条件(其中 $T$ 为预先设定的超时阈值):$\text{time}_k-\text{time}_1\leqslant T$。

另一种方法是设定相邻请求之间的超时时间,如果两页间请求时间的差值超过一定的界限就认为用户开始了一个新的会话,则(1)式中的用户会话必定满足下面的条件(其中 $T$ 为预先设定的超时阈值):$\text{time}_i-\text{time}_{i-1}\leqslant T$,其中 $1<i\leqslant k$。

超时阈值的设定直接影响 Web 日志数据预处理的结果输出,设定不同的超时阈值,就会产生不同的用户会话文件,从而最终影响 Web 日志的挖掘结果。

### 1.3 模式挖掘

数据预处理完成之后,在此基础上,Web 使用数据的模式发现采用的算法有:统计分析、关联规则挖掘、路径分析、时序模式发现、聚类和分类算法等。具体如下:

1)统计分析。统计方法是从 Web 中提取有用信息最常用的一种技术。通过对 Session 文件的分析,可以对感兴趣的信息进行统计,一般包括各种统计数据,如最频繁访问的 $N$ 个页面、每页平均浏览时间、网址路径平均访问长度等,也可能涉及一些关于限制的错误分析,如统计非法 IP、无效 URL 和未授权访问等。

2)关联规则。在 Web 使用挖掘中,关联规则主要用于发现用户之间、页面之间以及用户浏览页面和网上行为之间存在的潜在关系。最为著名的关联规则挖掘方法是 R. Agrawal 提出的 Apriori 算法。无论哪种算法,关联规则的发现都遵循两个步骤:第 1 步是迭代识别所有的频繁项目集,要求频繁项目集的支持率不低于用户设定的最小支持度(具体见下面定义 1);第 2 步是从频繁项目集中构造可信度不低于用户设定的最小置信度(具体见下面定义 2)。

定义 1:支持度(Support)是指交易集 $T$ 中包含 $X$ 和 $Y$ 的交易数与交易数据库 $D$ 中所有的交易数之比,记为 $\text{Support}(X\Rightarrow Y)$,即:

$$\text{Support}(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|D|$$

定义 2:置信度(Confidence)是指在交易集 $T$ 中包含 $X$ 和 $Y$ 的交易数与包含 $X$ 的交易数之比,记为 $\text{Confidence}(X\Rightarrow Y)$,即:

$$\text{Confidence}(X\Rightarrow Y)=|\{T:X\cup Y\subseteq T,T\in D\}|/|\{T:X\subseteq T,T\in D\}|$$

3)序列模式。序列模式与关联模式相仿,差别在于序列模式把数据间的关联性与时间先后顺序联系起来。即不仅需知道事件是否发生,而且需要确定事件发生的时间先后。可以把它看成是一种增加了时间属性的特定关联模型。

下列 4 个规则中,①是序列模式,②是关联规则,比较①和②便不难发现,①考虑了访问的先后顺序,而②则没有考虑时间因素。

①访问页面 P1 和 P3 之后有 35% 的用户又访问了页面 P5;

②访问页面 P1 和 P3 的用户中有 35% 的用户也访问了页面 P5;

4)聚类。聚类是将数据点集合分成若干类或簇[7](cluster),使得每个簇中的数据点之间最大程度地相似,而不同簇中的数据点最大程度地不同,从而发现数据集中有效的、新颖的、可以理解的数据模式分布。在 WUM 中,聚类技术是对符合某一访问规律特征的用户(页面)进行用户(页面)特征挖掘。

5)路径分析。使用路径分析技术进行 WUM 时,从 Web 站点拓扑结构图抽象而来的有向图中挖掘出最频繁的路径访问模式或大参引访问序列。Zaiane 等人[8]从原始日志数据中导出最大向前引用序列 MFR 的过程,实际上就是在构造用户的访问子图。路径分析可以用来确定网站上最频繁的访问路径,从而调整站点的结构。例如,可以得到类似如下的信息:

(1)访问网站的用户中有 25% 是从页面 B 开始的;

(2)有15%的用户访问路径C→M→D→E;

针对(1),可以在页面B上直接添加一些想给用户直接传达的信息,或者通过链接指向相应的页面,从而提高该信息的点击率;针对(2),可知路径C→M→D→E为频繁访问路径,可以在这几个页面上添加其它超链接或者促销信息,从而增加其它信息的访问频度。

### 1.4 模式分析

挖掘出来的用户行为模式(集合),需要合适的工具和技术对其进行分析、解释和可视化,从中筛选出有趣(有用)的模式,使之成为人们可以理解的知识,否则挖掘出来的模式将得不到很好的应用。具体包括:

1)可视化技术。与其他数据挖掘应用领域内一样,Web使用挖掘技术与可视化技术的结合还刚起步。Web使用挖掘领域内的可视化技术主要分为基于点和基于序列两类。基于点的可视化技术适合显示数据对象的各种统计值,例如产品或页面的访问次数,页面间转移的频率或者次数等。而基于序列的可视化方法着重表现用户行为序列特征,用各种方法描绘用户的访问序列。

2)知识查询技术。自动搜索相关的规则、模式以及其它的知识,可以帮助分析用户的目标,用智能的方式回答查询。目前研究人员已经在SQL语言的基础上提出几种适合在数据挖掘过程中使用的查询语言,如DMQL;也有专门为Web挖掘而定义的WebSSQL,WebLQM和Squeal等。

## 2 基于Web挖掘的电子商务模型

### 2.1 传统电子商务的结构模型

在传统电子商务的结构模型(如图2所示)中没有引进与数据挖掘相关的组件,移动设备和PC客户端通过互联网与Web应用服务器通信,应用服务器与后台的数据库服务器进行信息交流。该应用系统通常工作在分布环境中。

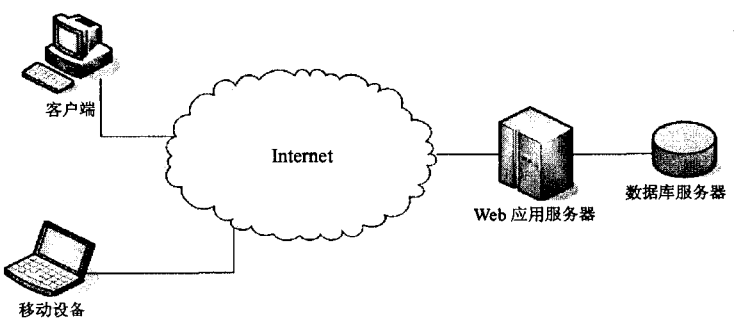


图2 传统电子商务结构模型

应用服务器为应用业务提供了一个运行环境。在这个环境中,各层应用和商业服务分离为各种组件,这些组件通过网络相互通信。应用软件在Web应用服务器及其内嵌的JVM(Java Virtual Machine,Java虚拟机)中运行。这些服务器的组件利用网络基础架构提供的目录和安全服务,通过HTTP或IIOP(Internet Inter-ORB Protocol)与客户和其他组件通信。

### 2.2 基于Web使用挖掘的电子商务模型

基于Web挖掘技术的电子商务模型(如图3所示)除了包括传统电子商务模型中的基本构件外,添加了知识库服务器。应用服务器除了和数据库服务器交流外,也可以和知识库服务器进行信息交流,例如应用服务器向知识库服务器提出数据挖掘请求后,知识库服务器通过数据挖掘引擎,对数据库进行数据挖掘处理,结果返回给应用服务器。

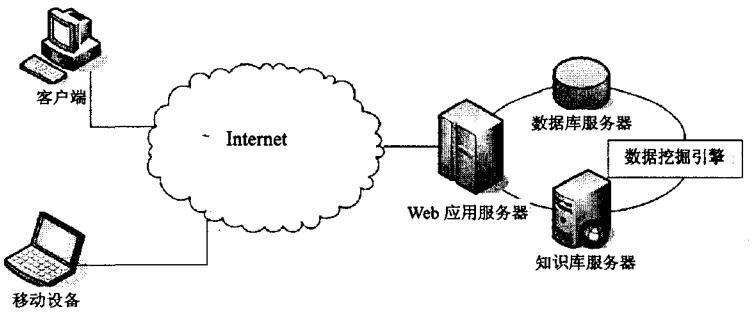


图3 基于Web使用挖掘的电子商务结构模型

## 3 Web使用挖掘在电子商务中的应用

通过收集、加工和处理涉及消费者消费行为的大量信息,确定特定消费群体或个体的兴趣、消费习惯、消费倾向和消费需求,进而推断出相应消费群体或个体未来的消费行为。然后对所识别出来的消费群体进行特定内容的定向营销,节省成本,提高效率,从而为企业带来更多的利润。

a.优化Web站点。Web设计者不再完全依靠专家的定性指导来设计网站,而是根据访问者的信息来设计和修改网站结构和外观。站点上页面内容的安排和链接就如超级市场中物品的摆放一样,把相关联的物品摆放在一起有助于销售。网站管理员也可以按照大多数访问者的浏览模式对网站进行组织,按其所访问内容来裁剪用户与Web信息空间的交互,尽量为大多数访问者的浏览提供方便。

b.设计个性化网站。强调信息个性化,识别客户的喜好,使客户能以自己的方式来访问网站。对某些用户经常访问的地方,有针对性地提供个性化的广告条,以实现个性化的市场服务。

c.留住老顾客。通过Web挖掘,电子商务的经营

者可以获知访问者的个人爱好,更加充分地了解客户的需要,根据每一类(甚至是每一个)顾客的独特需求提供定制化的产品,有利于提高客户的满意度,最终达到留住客户的目的。

d.挖掘潜在客户。通过分析和探究 Web 日志记录中的规律,可以先对已经存在的访问者进行分类,确定分类的关键属性及相互间关系。然后根据其分类的共同属性来识别电子商务潜在的客户,提高对用户服务的质量。

e.延长客户驻留时间。在电子商务中,为了使客户在网站上驻留更长的时间就应该了解客户的浏览行为、知道客户的兴趣及需求所在,并根据需求动态地向客户做页面推荐,调整 Web 页面,提供特有的一些商品信息和广告,以使客户满意。

f.降低运营成本。通过 Web 挖掘,公司可以分析顾客的将来行为,进行有针对性的电子商务营销活动;可以根据关心某产品的访问者的浏览模式来决定广告的位置,增加广告针对性,提高广告的投资回报率;可以得到可靠的市场反馈信息,降低公司的运营成本。

g.增强电子商务安全。Web 的内容挖掘还包括挖掘存有客户登记信息的后台交易数据库。客户登记信息在电子商务活动中起着非常重要的作用,特别是在安全方面,或者在对客户可访问信息的限制方面。

h.提高企业竞争力。分析潜在的目标市场,优化电子商务网站的经营模式。根据客户的历史资料不仅可以预测需求趋势,还可以评估需求倾向的改变,有助于提高企业的竞争力。

## 4 结束语

Web使用挖掘技术是 WWW 技术和数据挖掘技术的结合,是当今世界上的热门研究领域,其研究具有广阔的应用前景和巨大的现实意义。对 Web 网站上电子商务过程中产生的数据进行挖掘、发现知识,有利于信息的准确检索、个性化的信息服务、改进门户站点的设计、制定针对性的销售策略、构建智能化 Web 站点、提高网站的声誉和效益。总之,Web 挖掘有效地支持了电子商务中 CRM,ERP 和 SCM 等关键的商务流程,是电子商务营销创新的重要技术手段。

**参考文献:**


[1] 易观国际.互联网研究系列报告:电子商务(2006).[EB/OL].2006.http://www.analysys.com.cn.

[2] Pitkow J. In search of reliable usage data on the WWW[C]// In:Proc of 6th Int'l World Wide Web Conf. Santa Clara,California:[s.n.],1997.

[3] Cumming G,Hits J,Isses M. A year watching the web[C]// In:Proc of 6th Int'l World Wide Web Conf. Santa Clara,California:[s.n.],1997.

[4] Perkowitz M,Etzioni O. Adaptive web sites:Conceptual cluster mining [C]//In:Sixteenth International Joint Conference on Artificial Intelligence. Stockholm:[s.n.],1999.

[5] Madria S K,Bhowmick S S,Kngetal W. Research issue in web data mining[C]//Proc. of Data Warehousing and Know ledge Discovery,first Intel. Conf,DaWak'99. Birmingham:[s.n.],1999:303-312.

[6] Zaiane O,Xin M,Han J. Discovering applying olap and data mining technology web access patterns and trends by applying olap and data mining technology on web logs[C]//In proceedings Advances in Digital Libraries Conference (ADL). Paris:[s.n.],1998:19-29.

[7] 宋擒豹,沈钧毅. Web 页面和客户群体的模糊聚类算法[J].小型微型计算机系统,2001,22(2):229-231.

[8] Zaiane O R,Han J. Resource and knowledge discovery in global information systems:A preliminary design and experiment [C]//In:Proc of KDD95. Montreal,Canada:[s.n.],1995:331-336.


(上接第 227 页)

使逻辑网络拓扑发生变化,因此这种无关性使得网络编码在实际的 P2P 流媒体系统中应用具有重要意义。

**参考文献:**


[1] Ahlswede R,Cai N,Li S Y R. Network Information Flow[J]. IEEE Trans on Information Theory,2000,46(4):1204-1216.

[2] Ganesh A J,Kermarrec A M,Massoulie L. Peer-to-Peer Membership Management for Gossip-Based Protocols[J]. IEEE Transaction on Computers,2003,52(2):139-149.

[3] Gkantsidis C,Rodriguez P R. Network Coding for Large Scale Content Distibution[C]//In IEEE Incofom. Miami,FL:[s.n.],2005.

[4] Lin M J,Marzullo K. Directional Gossip:Gossip in a Wide-Area Network[R]. CS1999-0622. California:Univ. of California,1999.

[5] Zhang X,Liu J,Li B. CoolStreaming/DONet:A data-driven overlay network for live media streaming[C]//In:Znati T. Proc. of the IEEE INFOCOM. Miami:IEEE Press,2005.