Web 挖掘在 Web 交易中的应用*

林 苗[1,2], 张广泉[1]
(1. 重庆师范大学 数学与计算机科学学院, 重庆 400047; 2. 闽江学院 数学系, 福建 福州 350108)

摘 要:基于 Web 的数据挖掘是一种结合了数据挖掘和互联网系统的热门研究课题。随着互联网的高速发展,Web 挖掘由于其独特的优点,在 Web 交易中扮演了越来越重要的角色。运用 Web 挖掘对 Web 交易服务器的日志文件和客户交易信息进行挖掘,有助于企业了解客户的访问行为,挖掘潜在客户群和开展有针对性的服务。对 Web 挖掘技术进行综述,并介绍了该技术在 Web 交易中的几个应用。

关键词:数据挖掘;Web 挖掘;Web 交易
中图分类号:TP391 **文献标识码**:A **文章编号**:1672-6693(2007)03-0038-04

Application of Web Data Mining on Web Transaction

LIN Miao[1,2], *ZHANG Guang-quan*[1]
(1. College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047;
2. Mathematics Department of Minjiang College, Fuzhou Fujian 350108, China)

Abstract: Data miming is the most effective means to solve the problem of "data blast", and it becomes one of the hottest research fileds. With the rapid speed development of Internet, Web mining which has unique advantages is playing more and more important role in the Web transaction application, and large volumes of data such as user address or URL requested are gathered automatically by Web servers and collected in access log files. In the Web transaction, the user's browsing behavior can be discovered by applying Web data mining technology to Web data, such as server logs. And it's helpful for the modern enterprises to provide personal information service and make their electronic commerce strategies. We first analyze the new characteristics of Web data mining on Web transaction and present its concrete usage in this domain. Then we discuss the method to be used in the Web transaction.

Key words: data mining; Web data mining; Web transaction

随着 Internet 的日益普及和电子商务的迅猛发展,人们的购物方式已经由传统的市场购买逐渐转到了网络电子市场的购买方式,这种电子化市场上的交易可以称为 Web 交易。在 Web 交易中,客户只要连接到在线市场的服务器上,就会在这个服务器上留下"足迹",这些足迹描述了客户与网站的联系,是企业需要挖掘的金矿。通过 Web 数据挖掘,企业可以根据客户的网上行为来探知客户的需求和客户的愿望,有针对地开展 Web 交易活动,以更好地满足客户需求,进而提升自己产品的市场竞争力。

1 Web 数据挖掘的概念及流程

1.1 Web 数据挖掘的概念与特点

Web 数据挖掘(以下简称 Web 挖掘)一般指的是 3 种完全不同的行为:结构挖掘、应用挖掘和内容挖掘。结构挖掘是用来提取网络的拓扑信息——网页之间的链接信息。应用挖掘是用来提取关于客户如何运用浏览器浏览和使用这些链接的信息。内容挖掘是用来提取文字、图片或其他组成网页内容成分的信息[1]。

* 收稿日期:2006-10-09
资助项目:重庆市自然科学基金(No. CSTC2006BB2259);重庆市教委科学技术研究项目(No. 040803);中国科学院计算机科学国家重点实验室开放课题(No. SYSKF0303)
作者简介:林苗(1981-),女,福建连江人,助教,硕士研究生,研究方向为软件理论与形式化方法。

Web挖掘的主要特点是对客户信息数据进行提取、变换、分析和处理,并从中提取出有利于商业决策的重要数据,其基本假定是"消费者过去的行为是其今后消费倾向的最好证明"[2]。

1.2 Web挖掘的流程

Web挖掘主要包括数据收集、数据预处理、模式识别和模式分析4个阶段,流程图如图1所示。

(1)数据收集。数据挖掘的基础是数据,客户在访问服务器时,会在服务器上产生相应的数据,这些数据都存放在Web服务器的日志记录中。

(2)数据预处理。Web挖掘中一个很重要的步骤就是为挖掘算法找到合适的数据。数据预处理就是将客户访问网站时留下的原始日志整理成事务数据库,以供挖掘使用。数据预处理主要有下面几个方面。

①过滤(Filtering)。数据预处理的首要步骤就是通过删除无关数据,合并某些记录等方法来过滤掉不需要的记录。

②反蜘蛛化(Despidering)。蜘蛛是一种半自动程序。只需给定一个起始链接(出发点),它就会自行决定此后的运行情况。蜘蛛程序会扫描起始页面包含的所有链接,然后访问这些链接指向的页面,再分析和追踪那些页面包含的链接。蜘蛛的行为与人的行为是完全不同的,在数据预处理中应该把蜘蛛的行为和客户的行为区分开来,并过滤掉蜘蛛行为在服务器上留下的记录。

③客户的认证。这包括两个层次:一个是识别出同一客户在一次浏览中为了建立会话而发出的页面请求,另一个是识别在多次站点浏览中的同一客户。这些主要可以通过注册,Cookie和内嵌ID,分析客户IP的方式来认证。

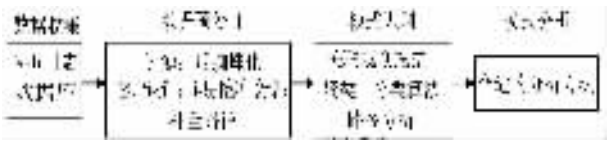

图1 Web挖掘流程图

④识别客户会话(Sessionization)。客户会话是指客户在一次访问中访问的所有Web页面,它反映了一个访问者对网站浏览的理解。不同客户访问的页面属于不同的会话。如果同一客户访问的页面跨时间较长,则认为开始一个新的会话。

⑤补全路径(Path Completion)。客户端缓存的存在,使得客户浏览页面时可能使用浏览器的后退功能,造成路径的缺失。因此需要根据客户访问路径的前后页面进行推理,补全访问路径。

(3)模式识别。该阶段是运用各种挖掘算法与技术对预处理后的数据进行挖掘,生成模式。常见的模式识别方法有关联规则挖掘、聚类、分类算法和路径分析技术等。

①关联规则挖掘。关联规则关注的是事务间的关系。在Web挖掘中,关联规则挖掘就是挖掘出客户在一个访问期间在服务器上访问的页面之间的关系。挖掘发现的关联规则往往是指支持度和置信度超过预设阈值的一组访问网页。客户与Web企业交易的很大一部分是由客户关联购买行为引发的,关联购买行为在客户与Web企业的交易行为中占着极大的比例。因此,对客户关联购买行为进行挖掘分析,将有助于客户资源价值的挖掘,增加客户的关联消费。

②聚类和分类算法。聚类技术是对符合某一访问规律特征的客户进行客户特征挖掘。分类技术主要是根据客户群的特征来挖掘客户群的访问特征(某些共同的特性),这些特征可用于把数据项映射到预先定义好的类中去[3-4]。在对Web用户访问信息的挖掘中,利用分类技术可帮助企业找到潜在客户。先对已经存在的客户群进行分类,然后通过衡量某客户和已有客户群共性的吻合程度来衡量该新客户的适宜度,再根据适宜度是否超过预设阈值可以判断该客户是否是一个潜在客户,并采取相应的营销策略。

③路径分析。路径分析技术是Web应用挖掘所特有的。Web站点的拓扑结构就是一幅定义在站点各页面之间联系的有向图,客户在一段时间内的访问模式是它的子图。客户访问频繁的有向边是频繁路径,通过路径分析技术,可以挖掘出Web交易网站中客户频繁访问的重要页面。

(4)模式分析。该阶段是实现客户访问模式的分析,它的基本作用是排除模式识别中没有价值的规则或模式,从而将有价值的模式提取出来[5]。

2 Web挖掘在Web交易中的应用

Web挖掘技术在Web交易中应用十分广泛,本文主要讨论以下3方面的应用。

(1)分析客户关联购买行为。客户与企业交易的一大部分是由客户关联购买行为引发的,即客户购买某种产品后,他会相应地同时或随后再购买某

些产品,如买了电脑的客户,通常随后会再购买软件。因此,根据客户关联购买行为,企业可以在客户购买某产品的同时,或在该产品被购买后的一长段 $t1$(对某种产品客户的再购买时间) $< t2$(产品已使用时间)时间里向客户推销该产品的关联产品。这里就需要利用 Web 挖掘技术中的关联分析来寻找产品间的有效关联规则。

该问题可以具体表述为:设 $I = \{i_1, i_2, \cdots, i_n\}$ 是 n 个不同商品的集合,D 是针对 I 上商品的交易集合,D 中每一项交易事件均包含若干个商品项目 I_a,$I_a \subset I$,要寻找的产品关联规则表示为:$A \Rightarrow B$,其中 A、$B \subset I$,并且 $A \cap B = \phi$,规则 $A \Rightarrow B$ 在交易集 D 中的支持度 $Sup(A \Rightarrow B) = P(A \cup B)$,置信度 $Conf(A \Rightarrow B) = P(B|A)$。关联分析的任务是在给定 I 和 D 后,找出所有的产品 A 和 B,它们具备关系 $A \Rightarrow B$,且 $Sup(A \Rightarrow B) \geq S_0$,$Conf(A \Rightarrow B) \geq C_0$,这里 S_0、C_0 为事先设定的阈值。

表 1 是一个假设的某家电企业在线交易中其客户购买数据的一部分,本文以该数据为例来说明数据挖掘技术中的关联分析在客户关联购买行为分析中的应用。

表 1　某在线家电企业的部分客户购买数据

客户交易项目	成交次数
电视机、音响	100
DVD、音响	200
电视机、DVD、音响	100
音响	100
功放	100
电视机、功放	100
DVD、功放	200
电视机	40
DVD	60
合计	1000

根据表 1 数据,构造商品间的关联表,如表 2 所示,以音响列为例,表中数据表示音响交易中与电视机一起交易的有 200 次,与 DVD 一起交易的有 200 次,单独购买音响的有 150 次,共 550 次交易涉及到音响。其他各行列的含义与此相同。

表 2　客户关联购买表

	音响	功放	单独购买	合计
电视机	200	100	40	340
DVD	300	200	60	560
单独购买	100	100		
合计	600	400		

这里,I = {电视机、DVD、音响、功放},D 为客户的交易集,共 1000 项,可以挖掘出以下规则:电视机⇒音响,$support = 0.2$,$confidence = 0.588$;电视机⇒功放,$support = 0.1$,$confidence = 0.294$;DVD⇒音响,$support = 0.3$,$confidence = 0.536$;DVD⇒功放,$support = 0.2$,$confidence = 0.357$;音响⇒电视机,$support = 0.2$,$confidence = 0.333$;音响⇒DVD,$support = 0.3$,$confidence = 0.5$;功放⇒电视机,$support = 0.1$,$confidence = 0.25$;功放⇒DVD,$support = 0.2$,$confidence = 0.5$;

若设最小阈值 S_0 为 0.2,C_0 为 0.5,则有效的关联规则有电视机⇒音响,DVD⇒音响,音响⇒DVD,功放⇒DVD。以 DVD⇒音响为例,关联规则说明有 30% 的客户在购买 DVD 后购买音响,这个结论的可信度为 53.6%。据此,企业就能相应地安排营销策略,例如将 DVD 和音响放在一起销售,当客户购买一种产品后,适当地向其推荐另一种产品,以增加客户的关联消费。

(2) 寻找潜在客户群。客户是企业的重要资产,因此大多数企业都希望在稳住老客户的基础上不断增加新客户的数量。利用 Web 挖掘技术对客户信息进行挖掘可以帮助企业寻找潜在的客户。对这类客户实施一定的策略,使他们尽快成为在册客户群体,对一个 Web 交易网站来说,这也许就意味着订单数的增多和效益的增加。文献[6]给出了寻找潜在客户的一种较常见方法。除此之外,还可以通过测量客户的适宜度方法来寻找潜在客户。该方法是通过衡量某客户和已有客户群共性的吻合程度来衡量该客户的适宜度,并根据适宜度是否超过预设阈值来确定该客户是否是一个潜在客户。

假设某 Web 交易网站经过数据挖掘统计发现,其客户群中 56% 的客户拥有本科文凭,41% 的客户是女性,年收入高于 3 万的客户占客户总人数的 20%,而年收入高于 8 万的则占客户总人数的 7%。现在调查两个参与者:(a) 李红,本科毕业,女,年收入 5 万;(b) 王军,高中毕业,男,年收入 2 万。表 3 给出了度量他们符合客户特征程度的一个方法。

表 3 根据符合各种属性的客户比例计算出了李红和王军两个参与者的分值。结果表明李红得分 2.1,王军得分 2.76,显然王军比李红更吻合当前的客户描述,假设事先设定的适宜度阈值 $t = 2.5$,则由于王军得分 $2.76 > t$,可以认为王军是一个潜在的客户。因此,可以对他实施一定的策略,使他尽快成为在册客户。该度量方法比文献[6]中的方法更简单,易于理解。

(3)发现重要页面。Web挖掘还可以帮助企业挖掘出Web交易网站中客户频繁访问的页面,这些页面可以被认为是重要页面。在用路径分析技术进行Web挖掘寻找重要页面的过程中,最常用到的工具是图。一个交易网站可以用一个有向图来表示,

表3 Web交易网站数据挖掘统计

	本科文凭	女	年收入>3万	年收入>8万	总分数
已有客户	56%	41%	20%	7%	
YES分数	0.56	0.41	0.20	0.07	
NO分数	0.44	0.59	0.80	0.93	
李红分数	0.56	0.41	0.20	0.93	2.1
王军分数	0.44	0.59	0.8	0.93	2.76

网站页面定义成结点,页面之间的超链接定义成图中的边。Web挖掘[7-8]就是从图中确定最频繁的路径访问模式,确定重要页面。这样,企业就可以将重要的商品信息和最新促销信息放在这些重要页面上,从而能让尽可能多的客户浏览这些信息,并最终达到增加Web交易的目的。

一个交易网站可以用一个有向图$G=(V,E)$来表示,其中V是页面的集合,E是页面间超链接的集合,页面抽象为图中的顶点,超链接抽象为有向边。根据某时期访问该站点的所有用户ID和相应的访问次数可以建立以客户ID为行,各页面的地址为列,元素的值是客户访问次数的关联矩阵P。假设该段时间里,有m个客户访问该Web交易站点,且该站点共有n个页面。

$$P=\begin{pmatrix} v_{11} & v_{12} & \cdots & v_{ij} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2j} & \cdots & v_{2n} \\ \square & & & & & \\ v_{i1} & v_{i2} & \cdots & v_{ij} & \cdots & v_{in} \\ \square & & & & & \\ v_{m1} & v_{m2} & \cdots & v_{mj} & \cdots & v_{mn} \end{pmatrix}$$

则v_{ij}表示客户i在这段时间T内访问第j个页面的次数,其中$i\in[1,m]$,$j\in[1,n]$。行向量$P[i,\cdot]$表示客户i对所有页面的访问情况,是客户访问本站点的个性化子图,$P[\cdot,j]$表示客户群对·页面的访问情况。

发现重要页面的算法可以描述为以下3个步骤。

(1)计算客户群体对第j个页面的访问情况

$$c_j=\sum_{i=1}^{m} v_{ij}, j\in[1,n]$$

逐个计算客户群体对各个页面的访问情况,所得结果构成页面访问集合$C=\{c_1,c_2,\cdots,c_n\}$。

(2)求出第j个页面的权重$weight_j$的值

$$weight_j=\frac{c_j}{\sum_{j=1}^{n} c_j}, j\in[1,n]$$

逐个计算各个页面的权重,所得结果构成权重集合$WEIGHT=\{weight_1,weight_2,\cdots,weight_n\}$。

(3)对$WEIGHT$集合中n个页面的$weight$值按从大到小的顺序进行排序操作,权值越大的页面就是越重要的页面。

3 小结

Internet已经让世界紧密地联系在一起,网络上几乎与业务相关的所有活动都可以被记录在数据库中[9],网络数据库中包含了Web企业获得胜利的秘密。利用Web挖掘,可以帮助企业分析客户关联购买行为,有针对地推荐商品,挖掘潜在客户群,获得网站重要页面信息以调整网站信息分布等,具有较强的现实意义。随着Internet的进一步发展,Web挖掘在个性化信息服务、开展有针对性的电子商务、构建智能化的Web站点,提高网站的声誉和效益等方面将起到极其重要的作用[10]。

参考文献:

[1] GORDON S L, MICHAEL J A B. Mining the Web: Transforming Customer Data into Customer Value [M]. USA: John Wiley & Sons Inc, 2004.

[2] 王书舟,高中文. Web使用挖掘技术在电子商务中的应用[J]. 微机发展,2003,13(2):41-43.

[3] FAYYAD U. From Data Mining to Knowledge Discovery: an Overview [C]. In#Proc. ACM KDD, 1994.

[4] BUCHNER A, MULVENNA M D. Discovering Internet Marking Intelligence Through Online Analytical Web Usage Mining [J]. SIGMOD Record, 1998.

[5] 王书舟,高中文. Web日志挖掘在电子商务中的应用研究 [J]. 计算机系统应用,2006(1):52-55.

[6] 邹显春,谢中,周彦晖. 电子商务与Web数据挖掘 [J]. 计算机应用,2001,21(5):21-23.

[7] 吕佳. 基于免疫聚类的Web日志挖掘[J]. 重庆师范大学学报(自然科学版),2007,24(2):32-35.

[8] 张德然. 可靠性统计与数据挖掘 [J]. 西华师范大学学报(自然科学版),2005,26(3):334-337.

[9] 陈莉. 数据挖掘与虚拟数据 [J]. 四川师范大学学报(自然科学版),1998,21(6):657-661.

[10] 韩家炜,孟小峰,王静,等. Web挖掘研究[J]. 计算机研究与发展,2001,38(4):405-414.

(责任编辑 游中胜)