Web使用挖掘的应用研究*)

刘丽珍[1] 宋瀚涛[1] 陆玉昌[2]
(北京理工大学 北京100081)[1] (清华大学 北京100084)[2]

The Applied Research of Web Usage Mining

LIU Li-Zhen[1] SONG Han-Tao[1] LU Yu-Chang[2]
(Beijing Institute of Technology, Beijing 100081)[1] (Tsinghua University, Beijing 100084)[2]

Abstract Some effective and efficient knowledge patterns will be gained through searching, integrating, mining and analyzing on the Web. These useful knowledge patterns can help us to build so efficient Web site that WWW can service people well. In this paper we point out Web Usage Mining process influenced by Web site structure and content, and introduce the application of Web Usage mining in E-commerce. In the end a example of Web Usage Mining is given.

Keywords Web mining, Web usage mining, E-commerce

1 引言

Web数据的爆炸性增长导致了顾客信息的过量,合理地应用它们不仅可以使网站和商业公司在激烈的竞争中获益,还可以找到合作长久同时又有利可图的顾客。如何策略地解决这个问题取决于我们对Web使用挖掘技术的有效应用,在Web上运用数据挖掘技术发现和分析有用信息逐渐成为知识发现研究的重要方向。

Web是一个快速变化的信息源,不单单是网上内容的急剧膨胀,页面内容的改变也是极度频繁,新闻、股票市场、广告公司和网络服务中心都在一定的时间间隔内修改着他们的网上信息;另外,网页的链接和存取路径也常常被改变,还要面对各种不同的用户,而且用户的数量也在不断地增长。其使用兴趣和目的各不相同,如何才能找到用户感兴趣的信息?如何才能找到高质量的页面?

以上这些问题推动了在Internet上进行使用挖掘的研究,为了更好地管理Web站点的数据,使其高效地为网上用户提供有效的信息服务;挖掘用户感兴趣的内容;跟踪、分析用户的使用模式;提高用户使用网络的效率,我们要积极开展网络挖掘中使用挖掘的应用研究工作。Web使用挖掘的应用日益广泛,尤其是在电子商务的大力支持下,越发显示出蓬勃的生命力。它通过数据挖掘技术对Web上的数据进行挖掘,从而发现Web上的用户使用模式。

2 Web使用挖掘[9]

Web使用挖掘是从Web服务器中自动发现用户的访问模式。在Web服务器日志中自动搜集并记录着用户的访问操作,还有通过CGI记录的用户注册信息。通过对这些用户信息的分析,可以找出用户的访问模式,确定产品的市场战略,提高商业活动的效率,而且为站点的有效组织也提供了信息,还可以为特定的用户提供个性化的网络服务。

目前常用的工具有模式发现工具和模式分析工具,它们提供了用户行为的分析和数据的过滤,使用人工智能、数据挖掘、心理学和信息理论从数据集中挖掘知识。在访问模式发现以后,用相应的分析技术来理解、解释和显示这些模式。如使用OLAP联机分析处理技术,数据立方体简化用户使用模式的分析,还有用SQL查询发现知识等。

2.1 Web使用挖掘的框架[5,8]

Web使用挖掘框架主要包括三部分:

(1) 数据预处理,包括用户识别、操作识别、路径完善、事务标识、数据集成、数据转换,将Web日志转化成面向不同领域的适合数据挖掘的事务形式。

(2) 面向不同的领域采用数据挖掘算法,如关联规则挖掘、序列模式挖掘、路径分析挖掘、分类和聚类分析挖掘等。

(3) 模式分析的方法有:联机分析、可视化、知识查询和信息过滤。模式分析工具将抽象的使用模式以直观、容易理解的方式展现给分析者,分析者利用知识查询语言,根据需要对挖掘过程加以限制,得到感兴趣的使用模式。比如限定某一领域进行挖掘,然后就这一领域挖掘出来的使用模式进行分析,得出感兴趣的结果。

信息过滤分两部分:objective过滤和subjective过滤。objective过滤处理用不同模式发现关联的数值型度量的变化,比如:支持度和兴趣度;subjective过滤是用来处理使用挖掘通过分析网站内容和结构而形成访问网页的可信任度。对于Web使用挖掘,设想用网站结构和内容作为网站设计者的领域知识,在网页之间进行链接以提供这些页面的关联支持,那么在网页之间的拓扑链接越强,这些网页一起被访问的可信度也就越高。类似地,在同一个内容簇或同一类里的页面被认为在一起被访问的可信度远远大于不同簇或不同类中的页面。图1所示是Web使用挖掘系统框架。

2.2 Web使用挖掘的实验

采用分类和聚类的挖掘方法,通过访问网站的客户流量

*)基金项目:973国家重点基础研究项目(G1998030414)。刘丽珍 副教授,博士研究生,主要研究方向为网络挖掘。宋瀚涛 教授,博士生导师,主要从事多媒体与信息管理技术、网络通信技术的研究。陆玉昌 教授,从事WWW上的数据集成、数据仓储及知识发现的有效算法与软件系统等的研究。

的分析，得出群体客户的访问规律，使网站设计者可以根据客户的访问规律，在不同时间段内推荐不同质量的服务，有效提升网站的访问人数。步骤如下：

（1）了解上网客户的身份，进行特定客户群体分析，从访客的造访次数、停留时间及常访问的页面，找到有实用价值的客户。

（2）对网站特定主题内容和特定网页进行深度分析，如：国庆节活动、旅游介绍等，进一步了解网站内容与访问客户之间的互动关系，发现最吸引客户的商品和服务。

（3）通过参照访问客户参与网站活动效果及网页浏览状态，辅助网站内容规划，评估网站内容。

基于一些实验数据模拟，将一天24小时内某网站客户流量进行分析，用表1、图2表示出来，并分析出一天中的推荐服务量（图3）。

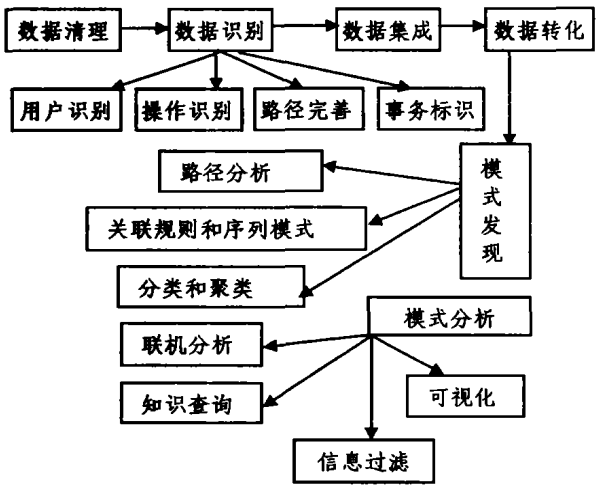

图1　Web 使用挖掘系统框架

表1　不同时间段的访问客户统计表

时间段	访客人数	时间段	访客人数
00:00--00:59	936	12:00--12:59	2466
01:00--01:59	725	13:00--13:59	1432
02:00--02:59	433	14:00--14:59	1649
03:00--03:59	389	15:00--15:59	1537
04:00--04:59	149	16:00--16:59	2361
05:00--05:59	118	17:00--17:59	2053
06:00--06:59	126	18:00--18:59	2159
07:00--07:59	235	19:00--19:59	1694
08:00--08:59	399	20:00--20:59	2078
09:00--09:59	1414	21:00--21:59	2120
10:00--10:59	2424	22:00--22:59	1400
11:00--10:59	2846	23:00--23:59	1463

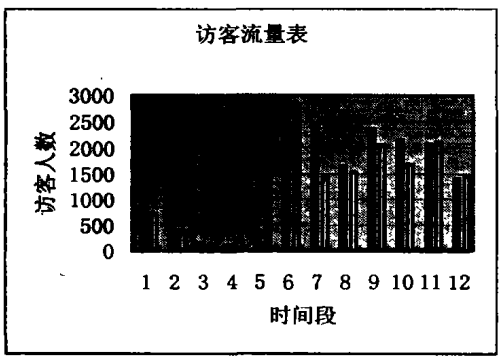

图2　访客流量表

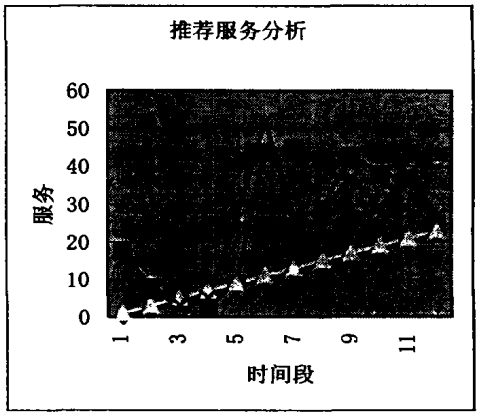

图3　推荐服务分析表

3　网站结构和内容对使用挖掘的影响

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。图4所示的是 Web 使用挖掘的过程，从中不难看出，网站的结构和内容对整个 Web 使用挖掘过程的每个重要阶段都是关键性的数据源。在 Web 上有三种数据：内容数据、结构数据和用户使用数据。内容数据是指网页上实际存在的数据，是供网上用户使用的，通常是由文本和图像组成的；结构数据是用来组织内容的一种描述性的数据，主要是指页与页之间的超链接；而用户使用数据是指 Web 页面的使用模式，比如：IP 地址、页面引用和访问时间等数据。使用数据通常源于普通和扩展的服务器日志。以上三种数据组建了数据提取、页面浏览、点击流和会话。页面浏览是指客户端用户一次点击网页的行为，一系列的页面浏览构成点击流。

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。

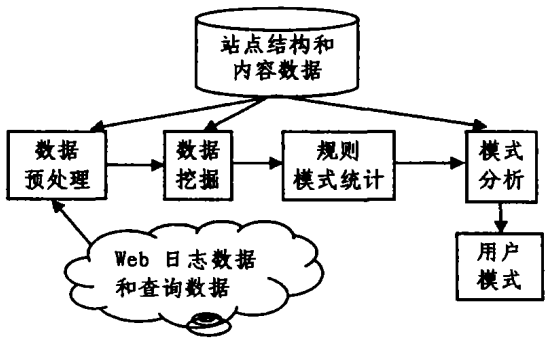

图4　Web 使用挖掘过程

4　使用挖掘在电子商务中的应用

由于迅速发展的电子商务竞争异常激烈，所以网站销售

商一定要做好快速迎合在线顾客需求的准备。在线销售的伸缩性使得人们能够监控销售，并且及时了解价格调整和产品服务的可适应性。另外，通过对销售信息的挖掘能展现影响产品所有方面的重要趋势和模式，包括：货运、销售和库存等。

日益增长的网站访问信息和飞速发展的数据挖掘技术使得网站能够真正地为它的在线顾客提供个性化服务，市场应该使网站施效于它真正的客户和利益。在一个动态的强竞争的网络环境中，电子商务必须通过较好地理解访问频繁的客户和最有利可图的顾客的行为，才能取得它们的竞争优势。要想了解客户的访问行为就必须通过使用挖掘去挖掘你的网站数据，使网站的努力集中在有利可图的顾客和前景上。

目前大部分的公司都有自己网站产生出来的巨大数量的用户信息，因而大型的电子商务网站需要有适合大量数据的挖掘工具，希望能通过数据挖掘得到益处。另外 Web 是个理想的市场环境，其中每笔交易都能被获取和存储，通过 Web 上的使用挖掘可以使网站达到以下目的：

·识别 Web 客户的关键特性。

·测试和决定哪个市场活动影响力最大。

·识别出对新产品特别有兴趣的客户。

·降低商品的价格，改善和客户的关系。

·改善网站广告和销售过程。

使用挖掘在电子商务中的具体挖掘步骤如图5所示。

(1)*纵览数据* 通过对顾客的可视纵览调查，能揭示一定顾客的特性统计，从而为网站设计者和市场经营者提供一些直接的战略战策。

(2)*分析数据* 将数据按照不同类型的分析方法分成不同的簇，进行分类聚类的挖掘。

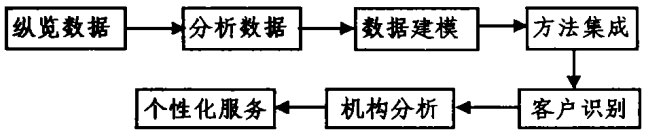

图5 使用挖掘在电子商务中的应用

(3)*数据建模* 建立网站访问者的行为模型，发现和学习访问者的特性和在线行为。例如，通过检验访问者的特性、点击广告和进行在线购物的习性，结合数据挖掘工具建模和预测网站访问者的行为，将个性化的服务推荐给有极大兴趣并且有购买欲望的顾客。一旦网站和顾客建立了相互了解的关系，那么这些顾客将可能永远是网站的老顾客。

(4)*方法集成* Web 使用挖掘并不是一个孤立的单一过程，而是对网站各个有影响的方面都要进行分析的综合过程。在这个分析的基础上，发现和控制货运周期和具体产品的趋势，发现的模式能提示网站目前需要哪些货源，以确保产品和服务的迅速递送。

(5)*客户识别* 要通过使用网站所产生的事务和顾客数据了解购买者是谁，最喜欢买什么?这就要求我们结合多方因素为每个访问者建立一个唯一的记录，获取并分析每个购物者的行为信息。

(6)*机构分析* 分析作为一种反馈系统服务于电子商务网站，能影响网站的设计、销售、库存和市场经营。对客户信息的整体机构的分析挖掘，可以使网站了解购买者的个人信息特点和购买物品的价格特点；了解所处热卖的产品，及时调整库存、货运和计划定单；并制定出浮动价格、增强广告设计、奖励促销等一系列营销策略。

(7)*个性化服务* Web 允许人们按照自己的喜好定制新闻、天气、市场和股票报告，但是人们必须要提供个人信息，以便于网络了解我们的偏好，适时地推荐具有个性化特点的产品和服务。因而网站要积累大量的用户信息，为用户提供感兴趣的产品和服务，建立牢固可靠的贸易关系，个性化网站的设计应用。最大的利益取决于网站和顾客信息的集成，这就要求在个性化基础上挖掘相互之间的作用与影响，并建立用户的个人偏好记录。通过讨论、聊天和邮件的方式学习顾客，并且可以进行相互影响的交流，从而形成网站的个性化服务。

Web 通过加大顾客的自由选择权而促进了商务的发展，网站要递送个性化的服务就需要挖掘 Web 使用数据发现顾客的特性，依靠这些数据的积累，进一步拓展网站的个性化功能。

应答式的个性化服务将会变得很规范。顾客很愿意以最少的代价去寻找所需的产品、服务和信息，因而商业网站必须结合公司的库存数据库，以适应不同个性用户的产品需求，同时也可以通过交叉推荐的方式来进行推荐。总之，个性化是通过网站和顾客之间的联系，使用积累信息进行挖掘，又递送商品服务于顾客的过程。

结束语 Web 挖掘是一个新兴的有巨大发展前景的研究领域，其技术在国内外有着广泛的应用。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。另外，通过 Web 使用的个性了解，比较已存在顾客的综合个性，能在已有顾客的知识帮助下发掘出潜在的新顾客的个性、生活方式和特点。Web 使用挖掘作为一个新兴的研究领域，其应用技术依然面临着很多挑战。

参 考 文 献

1 Han Jiawei, Kambr M. Data Mining. Beijing: Higher Education Press, 2000, 550

2 Mena J. Data Mining Your Website, America, 1999, 368Witten Frank, Data Mining, 1999

3 Wang Weiqiang. Text Ming on the Internet. Computer Science(计算机科学), 2000

4 Wang Jichen. Research on Web Text Mining. Journal of computer Research&Development, 2000, 37(5)

5 Chen En hong. Web Usage Mining: Discovering User Behavior Patterns From Web Data. Computer Science(计算机科学), 2001, 28(5)

6 Yang Xiaohua. Hyperlink Structure Mining of Web Sites. Computer Project&Application, 2001, 8

7 Wang shi. Web Mining. Computer Science(计算机科学), 2000, 27(4)

8 Wang Xiaoyan. Web Usage Mining, 2000, 3

9 Linoff G S, Berry M J A. Mining the Web, America, 2001, 348

10 Cooley R, Tan P-N, Srivastava J. Websift: The Web site information filter system. In WEBKDD, San Diego, CA, 1999

11 Liu Lizhen. The Research of Web Mining. In: The 4th World Congress on Intelligent Control and Automation, 2002

12 Cooley R, et al. Data preparation for mining world wide Web browseing patterns. Knowledge and Information Systems, 1999

13 Fayyad U, et al. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996, 39(11)

14 Jahn U, Schnattinger K. Deep knowledge discovery from natural language texts. In: Proc of the 3rd Int'l Conference Knowledge Discovery and Data Mining; Newport Beach, 1997