Web 用户访问模式挖掘系统框架模型研究

朱志国 [1,2]
（1. 大连理工大学管理学院，辽宁大连 116024；
2. 东北财经大学管理学与工程学院，辽宁大连 116023）

摘要：Web 用户访问模式挖掘技术可以从服务器、浏览器端的日志记录中自动发现用户的访问偏好、兴趣和趋势等信息，目前已经成为 web 挖掘领域的一个研究热点。文章首先给出 Web 访问模式挖掘系统的一般框架模型，然后介绍了框架模型中主要组成部分的工作原理，在此基础上，对 Web 访问模式挖掘系统中的一些关键技术的最新研究进展状况作了阐述和分析，其中包括数据采集、数据预处理、模式发现、用户可视化界面等，最后分析了未来该领域的研究重点作了展望。

关键词： Web 挖掘；Web 访问模式挖掘；数据预处理；模式发现；可视化
中图分类号：TP391 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2011.03.010

Research on Architecture Model of Web User Access Patterns Mining System

Zhu Zhiguo[1,2]
(1.Management School, Dalian University of Technology, Dalian 116024;
2.School of Management Science and Engineering, Dongbei University of Finance and Economics, Dalian 116023)

Abstract: Web user access patterns mining technology, currently a research hotspot in the field of web mining, is able to discover automatically such knowledge as the user access preference, interests and trends from the log records of the web server and browser sides. This paper starts with an introduction to the general architecture model of web user access patterns mining system, which is followed by a description on the working principle of the main components of this model. After that the paper elaborates and analyzes recent progress in some key technologies. Among them are data acquisition, data preprocessing, pattern discovery and visual user interface. The paper concludes with some prospects for the future research priorities in this field.

Keywords: Web mining, Web access patterns mining, data preprocessing, patterns discovery, visualization

1 引言

Web 上的数据以每天新增 100 万个页面的速度增长，页面数目已超过 10 亿。如何从这些位于分布式环境中的海量数据中挖掘和抽取潜在的、用户感兴趣的有用模式和隐藏的知识成为一个重要且非常有意义的课题。Web 访问模式挖掘技术 [1]（Web Access Pattern Mining，简称 WAPM）是解决这个问题的一个有效方法。这项技术能够从服务器、浏览器端的日志记录和用户的个人信息中自动发现和预测隐藏在数据中的模式信息，即用户群体的共同行为以及个人用户的检索偏好、习惯等。这种模式广泛应用于 Web 个性化服务、站点系统改进以及商业智能等领域。

作者简介：朱志国（1977- ），男，博士，东北财经大学副教授，研究方向：知识管理、信息系统工程、Web 数据挖掘。
基金项目：国家自然科学基金项目“人机协同思维中隐性知识共享管理方法研究”（70671016）。
收稿日期：2010 年 9 月 8 日。

目前，对于 Web 用户访问模式挖掘系统通用模型的研究与开发成果有 WebMiner、WEB-SIFT、WUM、LumberJack 等。WebMiner[2] 支持多种数据挖掘功能，并提供挖掘语言；WEBSIFT[3] 的主要特点是支持基于 Support Logic 的模式兴趣度自动识别和过滤出令人感兴趣的模式；最有影响力的是 WUM 系统 [4]，它把清理后的日志数据合成聚合树，挖掘和分析工作 的主要内容是使用 MINT 语言查询聚合树；LumberJack 系统 [5] 只支持对会话的聚类，提供各种分析报表。

本文在上述成果基础上，提出 WAPM 系统的一般框架模型，并对其中一些关键技术进展进行分析。

2 WAPM 一般框架模型

WAPM 系统框架包含数据预处理模块、访问模式挖掘模块以及用户界面模块 3 个主要组成部分（图 1）。这些模块主要功能如下。

（1）数据预处理模块。通过多种渠道采集用户的访问信息以及交易数据，经过一系列的清洗、识别、集成等步骤，得到集成数据，并对其进一步格式化，以适应不同需求偏好模式发现算法的需要。

（2）访问模式挖掘模块。将预处理过的数据输送进入集成访问信息数据库。接下来再利用一些访问模式挖掘算法，挖掘得到访问模式库。最终的模式结果通过可视化界面由用户进行评估。这个阶段的核心是挖掘内核。挖掘内核中的访问模式库是一个规则的集合，能够根据不同的挖掘要求选择最有效的挖掘算法或几种算法的序列组合，包括关联规则、聚类分类、序列模式等。

（3）用户界面模块。这个阶段使用图形用户界面，实现用户与系统之间的交互，帮助用户理解大量的复杂数据，为用户提供便利。在这个界面上，用户可以提出挖掘请求，设置挖掘参数，并可对返回的结果进行评估，对不满意的结果进行再次挖掘，直至得到满意的结果。

3 数据预处理模块

3.1 数据采集

在 WAPM 中，由于 HTTP 协议的无状态连接特性而很难得到准确的用户浏览信息，Robert Cooly[6] 和 Cyrus Shahabi[7] 提出从 Web 的结构出发多层次地进行 Web 站点信息采集。在 WAPM 中，用户访问数据的采集主要可以分为以下几种形式：

（1）服务器端数据采集：主要是从 Web 服务器日志中收集用户访问信息。这是执行 WAPM 的重要数据来源。表 1 是一条典型的扩展日志格式[8] (ECLF, Extended Common Log File) 的记录和提取出的相关信息。这些记录数据反映了多个用户（可能同时）对 Web 站点（单站点）的（存取）访问行为。由于 Web 环境中多级别缓存（如用户的本地缓存和代理服务器缓存）和防火墙的存在，所以日志中的网站使用数据并不完全可靠。

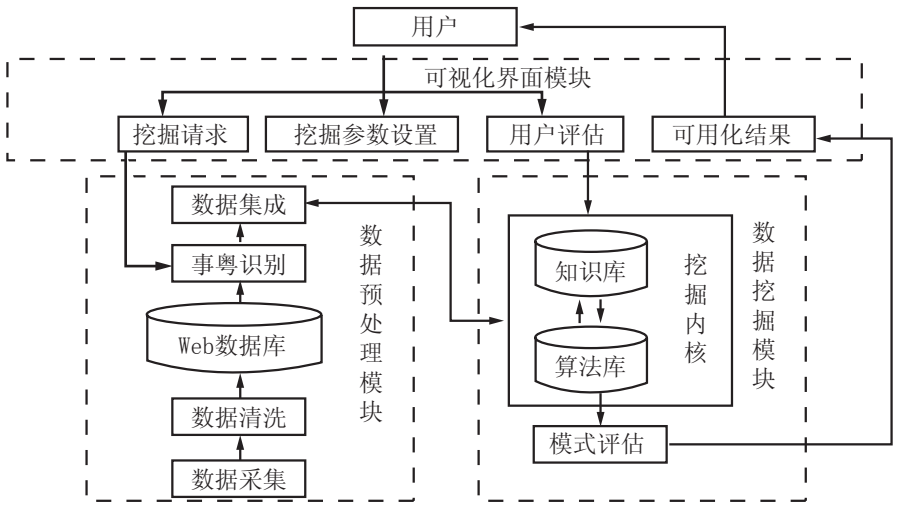

图 1 WAPM 系统框架模型图

表 1　Web 服务器 ECLF 记录格式及提取的信息

IP地址	221.202.41.83
访问时间	25/Dec/2003:05:00:00
访问页面	http://news.tom.com/piclib/491_11.html
Web服务器对于该请求返回的状态信息	200
返回给客户端的内容的大小	15763
该请求的引用地址	http://news.tom.com/pic
用户客户端类型	Mozilla/4.0 (compatible; MSIE5.0;Windows98;DigExt)

（2）代理服务器端的数据采集：代理服务器端可以得到从多个用户到 Web 服务器的访问记录 (不需用户许可)。若代理访问站点网页是通过 Web 应用程序动态生成的 , 对于用户的每次请求 , 代理需从 Web 服务器取得数据。该收集方法不能准确地确定浏览用户 , 对访问页面的采集不够全面 , 采集时间不准确。

表 2　Web 使用挖掘数据采集技术汇总

采集特点 \ 采集内型		Web服务器端	代理服务器端	客户端
用户	单			●
	多	●	●	
站点	单	●		
	多		●	●
使用强制性?		否	是	是
面向整个Web?		否	是	是
影响Web服务?		是	否	否
采集时间准确?		否	否	是
准确识别用户?		否	否	是

（3）客户端的数据采集：客户端的浏览路径采集比服务器端的采集更具优越性，因为它建立在用户的行为源上，可以准确地捕捉用户的行为、浏览路径和浏览时间。但是这种方式有可能会侵犯用户的隐私，因此需要用户的许可。客户端的用户浏览路径采集包括 Java Applet、Plug-in、网页跟踪帧和修改浏览器 [9] 等技术。

综上所述，服务器端数据采集最为方便，但是访问页面的采集不够全面 , 采集时间不准确。代理服务器端的优势是能在不需用户许可的情况下采集到多个 Web 服务器的访问记录。客户端采集建立在用户的访问行为源上，因此可以准确地捕捉用户的行为信息，但是用户隐私保护存在障碍。表 2 对这些采集技术进行对比总结。

3.2　数据预处理

WAPM 系统中对数据进行预处理的目的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须的数据抽象概念 , 然后在事务数据库中实施挖掘算法。由于预处理的结果直接影响挖掘算法产生的规则和模式，这个过程是整个系统质量保证的关键。通常 WAPM 的预处理过程包括数据清理、用户识别、会话识别和事务识别等几个步骤。

（1）数据清理：数据清理解决“无用数据”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，文献 [10] 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。

算法 1：

选择记录属性。在 Web 日志中，选用如下属性：

A={IP,r.date,r.time,Request,Size,Referer,Agent}。

删除无用记录。建立一个删除列表：

DT = {.GIF .JPEG .JPG .gif .jpeg .jpg}。

凡是对后缀名在删除列表中的文件的申请记录均应删除。

具体过程如下 :

输入：T1.log 中的所有记录集 L;

输出 : 数据清理后的关系表 T2.log // T2.log 表包含 A 中的字段；

For all $l_i \in L$ // 依次处理 L 中每条记录

{if T1 请求的文件后缀名不在 DT 中

Then 选取 T1 中所有属于 A 的字段值并存入 T2.log 中 ;}

（2）用户识别：用户识别是从日志中识别出每个访问网站的用户。最常被 WAPM 系统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 [11]。

需要说明一点，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识别用户。在用户识别的过程中，还会产生一些问题，典型问题如下 [12]：

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。

（3）会话识别：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会话。

用户会话 S 可以定义为：

$S = <UserId,\{(Pid_1,time_1,),\cdots\cdots,(Pid_k, time_k)\}>$ （1）

令 $RS = \{(Pid_1,time_1,),\cdots\cdots,(Pid_k, time_k)\}$

$S = <UserId, RS>$

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。由于 RS 包含用户请求页面的标识符 Pid 和请求时间的标识符 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 2 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):$time_k - time_1 \leqslant T$。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，则公式 (1) 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): $time_i - time_{i-1} \leqslant T$，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。

（4）事务识别：事务识别指将页面访问序列划分为代表 Web 事务或用户会话的逻辑单元。和用户 Session 识别不同的是，它以事务为单位，只包含与事务相关的页面。事务识别方法中最简单的莫过于时间窗口法，即定义一个时间长度，该时间片段内用户浏览的所有页面都归为一个事务。而比较常用的则是最大向前参考法 [13]。具体做法是，从用户访问的首页开始，到第一个回退动作为止定义为一个事务，接下来的第一个向前动作引发下一事务，直到下个回退动作产生，周而复始，将用户访问页面序列划分为一个个事务。比如，一个用户在一次浏览过程中请求了 ABCBCDE 页面，根据最大向前参考法，用户访问过的访问服务器会话期间应该是 ABC 和 BCDE。

4 模式挖掘模块

数据预处理阶段完成之后，常用的 WAPM 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。

4.1 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。最近也有独立 Agrawal 的频集方法的工作 [14]，以避免需要大量空间存储中间结果和反复扫描数据库而带来的算法上的缺陷。无论哪种算法，关联规则的发现都遵循 2 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；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。

4.2 聚类模式

在 WAPM 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。通常可以将用户浏览页面的总和视为数据空间 [15]，构造一个 URL_UserID 关联矩阵 $M_{m\times n}$, 如公式 (2) 所示 。其中 ,$h_{i,j}$ 是 j 客户在一段时间内访问第 i 个 URL 的次数 ; 每一行向量 M[1,j] 表示所有客户对 URL “1” 的访问情况 ; 每一列向量 M[i,1] 表示客户 “1” 对该商务站点中所有的 URL 的访问情况。因此 , 可以这样认为 : 行向量既代表了站点的结构 , 又蕴涵有客户共同的访问模式 ; 而列向量既反映了客户类型 , 也勾勒出了客户的个性化访问子图。那么 , 再使用一些度量方法 (例如 Hamming 距离) 分别度量行向量和列向量的相似性 , 就可以得到两种类型的聚类 : 使用聚类 (用户聚类) 和网页聚类。使用聚类主要是把具有相似特性 (或浏览模式) 的用户聚集在一组。这类知识对电子商务和为用户提供个性化的服务特别有用。网页聚类可以找出具有相关内容的网页组，有助于提高网上搜索引擎的准确度。

4.3　分类模式

在 WAPM 中 , 分类技术可以预先把页面分到不同的类中，这样在分析以往的访问记录得知某用户以前经常访问某一类的网页之后，便可以将该类网页中还没有被该用户访问过的页面推荐给用户，节省了用户搜寻所需信息的时间，同时增强了网站的个性化服务意识 [16]。最为典型的决策树学习系统是 ID3，它采用自顶向下不回溯策略，能保证找到一个简单的树。算法 C4.5 和 C5.0[17] 都是 ID3 的扩展，它们将分类领域从类别属性扩展到数值型属性。

$$M_{mxn}=\begin{bmatrix} h_{1,1}\,h_{1,2}\ldots h_{1,j}\ldots h_{1,n} \\ h_{2,1}\,h_{2,2}\ldots h_{2,j}\ldots h_{2,n} \\ \vdots\quad\vdots\quad\vdots\quad\vdots \\ h_{i,1}\,h_{i,2}\ldots h_{i,j}\ldots h_{i,n} \\ \vdots\quad\vdots\quad\vdots\quad\vdots \\ h_{m,1}\,h_{m,2}\ldots h_{m,j}\ldots h_{m,n} \end{bmatrix}\text{URL ID} \qquad (2)$$

5　用户界面模块

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 [18] 和基于序列 [19] 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的序列特征，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、模式以及其他的知识，可以帮助分析用户的目标，用智能的方式回答查询。建立一个多层数据库（MLDB），用数据库技术来管理 Web 的元数据（Meta Web）是其中的一个方法 [20]，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[21] 等。

6　小结与展望

WAPM 系统旨在通过挖掘用户访问信息，得到一些有用模式和规律 , 以使网站有针对性地完善自身 , 更好地服务于用户并取得较好的经济效益。本文给出了 WAPM 模型框架，并对其中主要业务过程的一些关键技术进展进行了评述。WAPM 技术作为一个新兴研究领域，采用了来自多个领域的技术和先验知识，虽然取得很多突破性的进展，但是在未来的研究当中，还有一些热点和难点方向。

(1) 数据收集与预处理。在 WAPM 过程中，如何既不侵犯用户的个人隐私 , 又能尽量收集到完整的网站访问日志 , 同时又能保证服务器的工作效率和服务质量。有关数据收集的专门工具和技术正在研究中。

(2) 用户界面。开发一个智能化模式分析工具，将不同数据源挖掘出来的模式进行集成，并提供集统计分析、可视化分析技术以及过滤、解释功能等为一体的模式分析功能。

参考文献

[1] Sergey Brin, Rajeev Motwani. What Can You Do with a Web in Your Pocket[C]// Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. USA: Institute of Electrical & Electronics Engineer, 2002: 42−54.

[2] Robert C, Bamshad M, Jaideep S. Web Mining: Information and Pattern Discovery on the World Wide Web[C]. IEEE International Conference on Tools with Artificial Intelligence. USA: IEEE Press, 2006:558−567.

[3] Robert Cooley, Tan Pangning, Jaideep Srivastava. Discovery of Interesting Usage Patterns from Web Data [M]. Germany: Springer−Verlag, 2000.

[4] Spiliopoulou M, Faulstich L C. WUM: A Tool for Web Utilization Analysis[C]. EDBT Workshop WebDB'02, Verlagia, Spain. Germany: Springer, 2002:184−193.

[5] Ed H Chi, Rosien A, Jeffrey H. Intelligent Discovery and Analysis of Web User Traffc Composition[C]. WEBKDD 2002−Mining Web Data for Discovering Usage Patterns and Profiles, 4th International Workshop, Edmonton, Canada. Germany: Springer, 2002: l−16.

[6] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns[J].The Journal of Knowledge and Information Systems, 1999,1(1):45−52.

[7] Shahabi C, Zarkesh A , Adibi J , et al. Knowledge Discovery from Users Web−page Navigation [C]// Proceedings of the IEEE RIDE 97 Work Shop. USA: Institute of Electrical & Electronics Engineer,1997:204−210.

[8] Luotonen A. The Common Log File Format [EB/OL]. [2010−10−15]. http://www.w3.org/pub/www/.

[9] Hu M, Liu B. Mining and Summarizing Customer Reviews[C]// Proceedings of the 10th ACM SICKDD International Conference on Knowledge Discovery And Data Mining. New York: ACM Press, 2004:168−177.

[10] Jindal N, Liu B. Review Spam Detection[C]// Proceedings of the 16th International Conference on World Wide Web. New York: ACM Press, 2007:1189−1190.

[11] Berendt B, Mobasher B, Spiliopoulou M. Measuring the Accuracy of Sessionizers for Web Usage Analysis[C]// Proceedings of the Workshop on Web Mining at the First SIAM Lztemational Conference on Data Mining. Berlin: Springer,2001, April:7−14.

[12] Spiliopoulou M, Mobasher B, Berendt B. A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis[J]. INFORMS Journal of Computing, Special Issue on Mining Web−Based Data for E−Business Applications,2005,15(2):171−190.

[13] Chen M S, Park J S. Data Mining for Path Traversal Patterns in a Web Environment[C]// Proceedings of the 16th International Conference on Distributed Computing Systems. 1996: 385−392.

[14] Bamshad Mobasher. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns[C]// Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX297). USA: Institute of Electrical & Electronics Engineer,1997:108−132.

[15] Hogo M, Snoek M, Lingras P. Temporal web usage mining[C]//Proceedings of the IEEE/WIC International Conference on Web Intelligence. USA: IEEE Press, 2003:450−453.

[16] Matsumoto S, Takamura H, Okumura M. User classification using word sub−sequences and dependency sub−teed[C]//Proceedings of the 9th Pacific−Asia Conference on Knowledge Discovery and Data Mining. Berlin:Springer Verlag,2005,301−311.

[17] Wang Xiaoguo, Huang Shaokun, Zhu Wei,et al.Method of Building Decision Trees of Customer Classification by Using C4.5 Algorithm[J]. Computer Engineering, 2003, 29(14):211−216.
〔王晓国 , 黄韶坤 , 朱炜 , 等 . 应用 C4.5 算法构造客户分决策树的方法 [J].2003,29(14):211−216.〕

[18] Bettina Berendt. Detail and Context in Web Usage Mining: Coarsening and Visualizing Sequences[C]. Heidelberg: Springer−Verlag , 2002:1−24.

[19] Robertson George G, Mackinlay Jock D. Cone Trees: Animated 3D Visualizations of Hierarchical Information[C]// Proceedings of CHI’91 Conference on Human Factors in Computing Systems. Heidelberg: Springer−Verlag, 2001:189−194.

[20] Sadri R, Zaniolo C, Zarkesh A. Optimization of Sequence Queries in Database Systems[C]// Proceedings of the Twentieth ACM SIGMOD−SIGACT SIGART Symposium on Principles of Database Systems. New York: ACM Press, 2001:71−81.

[21] Maciej Zakrzewicz. Sequential Index Structure for Content−Based Retrieval[C]. PAKDD 2001.Berlin:Springer, 2001:306−311.