Web 用户访问模式挖掘研究*)

陈新中　李　岩　杨炳儒
(北京科技大学信息工程学院　北京 100083)

Research on Web Usage Patterns Mining

CHEN Xin-Zhong　LI Yan　YANG Bing-Ru
(Information and Engineering School, University of Science and Technology Beijing 100083)
E-mail: chenxinzhong@sina.com

Abstract　With the rapid development of Internet, Web usage mining plays very important role in many fields including personalizing information service, improving designs and service of Web sites, developing the personal electric commerce, building adaptive Web sites, promoting the reputation and income of Web sites. The paper introduces the definition and classification of Web mining firstly, then the main technology and method of Web log preprocessing, the primary algorithm of Web usage mining, the evaluation method and important applications of Web usage mining are discussed in detail. At the end, the trend and research course concerning the Web usage mining are concluded.

Keywords　Data mining, AI, Web mining, Web usage mining

1　引言

目前 World Wide Web(WWW)已经发展成为拥有近亿个工作站、数十亿页面的分布式信息空间，在这个分布式信息空间中蕴涵着具有巨大潜在价值的知识，也带来了巨大的经济效益和社会效益。

对于不同层次、不同使用目的和爱好的浏览者需要个性化的信息服务，希望网站能够根据自己的浏览习惯，动态定制 Web 站点，实现个性化的浏览；对于网站的经营管理者来说，为提高网站的声誉和效益，需要了解其客户需要什么和想做什么，其中包括根据大多数客户的共同兴趣，开展有针对性的信息服务，以及对特定的用户开展个性化的信息服务和电子商务活动。

Web 服务器中的日志文件(Web Sever log)记录了每一位用户在访问本站点时的相关信息，包括：用户的 IP 地址、访问时间、访问的页面、访问的方式、HTTP 版本号、返回码、传输字节数、引用页的 URL 等。

然而对于一个热门的小型网站，其 Web 日志数据以每天数十兆的速度增长，人工分析和处理这些日志数据一般来说是不可能的。

解决这个问题的途径之一就是将传统的数据挖掘技术应用于从海量的 Web 日志数据中自动、快速地发现用户的访问模式，如频繁访问路径、频繁访问页组、用户聚类等。Web 用户访问信息挖掘所得到的模式既有助于提高网站的性能和安全性，也可以作为优化站点拓扑结构及页面之间的超链接关系的依据，也是在 Web 上进行市场开发和开展电子商务活动的依据，也可以作为网站为用户提供个性化服务和构建智能化 Web 站点的依据。

2　Web 挖掘综述

Web 挖掘就是从与 WWW 相关的资源和用户浏览行为中抽取感兴趣的、有用的模式和隐含的信息。

从上述定义中可知，Web 挖掘与传统的数据挖掘相比，有很多独特之处。首先，Web 挖掘的对象是大量、异质、分布的 Web 文档；其次，Web 在逻辑上是一个由页面节点和超链接构成的图；再者，Web 服务器日志记录了大量的用户访问站点时的信息。因此，Web 挖掘所得到的模式可以是关于 Web 内容的，也可以是关于 Web 结构的或是用户访问模式的。

按照挖掘对象的不同，可以将 Web 挖掘分为三大类[1]：Web 内容挖掘(Web Content Mining)、Web 结构挖掘(Web Structure Mining)和 Web 访问信息挖掘(Web Usage Mining)，如图 1 所示。

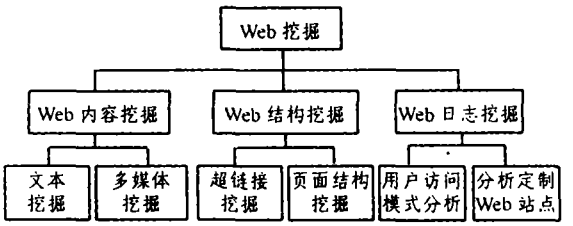

图 1　Web 挖掘的分类

Web 内容挖掘是指在人为组织的 Web 上，从文件内容及其描述中获取有用信息的过程。按照处理对象的不同，可以把 Web 内容的挖掘分为两类：①对于文本文档(包括 text, HTML 等格式)的挖掘即文本挖掘；②对于多媒体文档(包括 image, audio, video 等多媒体类型)的挖掘即多媒体挖掘；Web

*)基金项目：国家自然科学基金重点项目(698350010)、教育部科技重点项目(教技司[2000]175)。陈新中　博士研究生，主要研究方向：Web 挖掘、智能化 Web 站点。李　岩　博士研究生，主要研究方向：Web 挖掘、智能搜索引擎。杨炳儒　教授，博士生导师，研究方向：推理机制与知识发现、柔性建模与集成技术。

结构挖掘则是从 Web 的链接结构中获取有用的知识的过程。主要是通过对 Web 站点的结构进行分析、变形和归纳，将 Web 页面进行分类，以利于信息的搜索；Web 用户访问模式挖掘（也称为 Web 日志挖掘），是从 Web 的存取模式中获取有价值的信息或模式的过程。就是对用户访问 Web 时在服务器留下的访问记录进行挖掘，挖掘的对象是 Server Logs、Error Logs 、Cookie Logs 等日志信息以及用户的注册数据、文档、用户调查数据等。

3 Web 用户访问信息挖掘

3.1 Web 日志挖掘的数据源

WWW 的基本结构是：客户机——Web 服务器，或客户机——代理服务器——Web 服务器。从客户机、代理服务器和 Web 服务器上分别可以采集到单用户——多站点、多用户——多站点、多用户——单站点的用户访问信息。

Web 用户访问信息挖掘的数据源主要包括：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。

3.2 Web 日志文件的主要内容和格式

Web 服务器日志主要包括以下内容：用户的 IP 地址、时间戳、方法（如 GET、POST）、被请求文件的 URL、超文本传输协议（HTTP）的版本号、返回码（请求的状态，成功或错误码）、传输字节数、代理（用户使用的浏览器和操作系统的类型），有些扩展日志还包括参考页的 URL（用户从该页发出当前文件的请求）。

常用的日志格式包括：W3C 扩展日志文件格式、Microsoft IIS 日志文件格式、NCSA 通用的日志文件格式、ODBC 日志格式等。可以通过文件名来区分不同的日志格式，W3C 日志文件以 en 开始，最后是日期，例如，en000404；Microsoft IIS 格式文件以 in 开始，然后是日期，例如 in 000404；NCSA 文件格式以 nc 开始，例如 nc000404。

清单 1 显示的是 Microsoft IIS 日志文件格式的示例：

```
192.168.1.20,-,04/04/2001,9:26:32,W3SVC1,PENTHOUSE,192.168.1.23,230,
278,1109,200,0,GET,/Default.htm,-,
```

该日志文件记录了以下信息：Client IP address、Username、Request time、Request time、HTTP status code、Bytes receive。

从清单 1 中可以看出：发出请求的 IP 地址是 192.168.1.20。然后是日期和时间，以及记录日志的服务（W3SVC1），最后是 PENTHOUSE 的服务器 NetBIOS 名称。接下来显示的是服务器的 IP 地址 192.168.1.23。后面的一组数字是：230 说明请求所完成的时间长度，以秒为单位；278 是收到的字节数；1109 是发送的字节数。数字 200 指的是 HTTP 状态码，它后面跟的是 Windows 2000 状态码 0。然后我们看到的是 GET 声明，它仅仅是发送请求的类型，最后是请求的文件。

3.3 Web 日志挖掘系统的体系结构

Web 日志挖掘过程大体分为三个阶段：数据预处理、模式挖掘、模式分析和可视化，Web 日志挖掘的体系结构如图 2 所示。

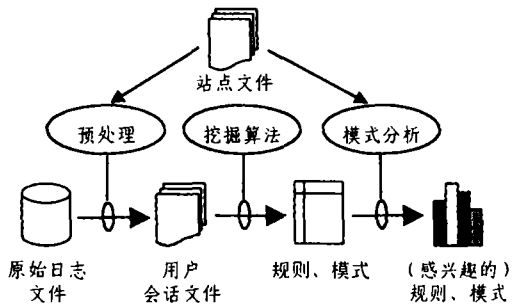

图 2 Web 日志挖掘系统的体系结构

3.4 Web 日志挖掘中的数据预处理[6,7]

数据预处理是指通过对原始的日志文件结合站点的结构和 Web 页面的内容，经过数据净化、用户识别、用户会话识别、用户访问路径补充和格式化等一系列处理过程，最终产生一个用户会话文件，作为用户访问模式挖掘阶段的输入。用户会话文件的主要内容为：访问 Web 站点的用户、请求的页面及顺序、每一页阅读的时间等。

对原始的 Web 日志进行预处理所得到的结果直接影响到挖掘算法产生的规则与模式，预处理是保证挖掘到正确、有用的用户访问模式的最为关键环节。

Web 日志数据预处理的具体流程如图 3 所示。

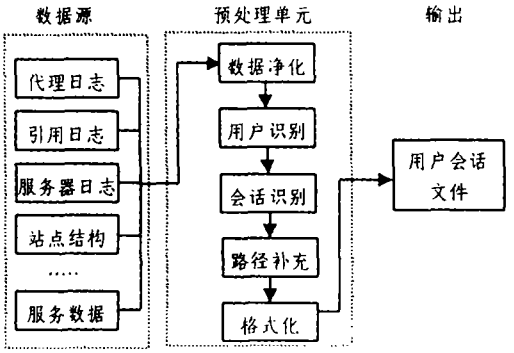

图 3 Web 日志预处理的具体流程

1）数据净化 数据净化是预处理的第一项任务，从原始的服务器日志文件中删除与挖掘任务不相关的数据项，对于任何 Web 日志挖掘方法都是极其重要的。只有当日志文件能够准确表示用户在访问 Web 站点的行为时，才能得到有意义的模式和统计结果。

Web 服务器上的每一个页面都有一个单独的链接，当用户请求一个 HTML 文档时，其所包含的图形、图像、商业广告和脚本文件也被自动下载。通常情况下，只有用户所请求的 HTML 才是有用的，应该被保留在预处理完成后所形成的用户会话文件中。用户访问模式挖掘的目的是要得到用户浏览 Web 站点时的行为视图，所以应该删除非用户直接请求的内容。通常的做法是通过检查 HTML 文件的后缀名来删除无关项，例如可以删除日志文件中后缀名为：gif、GIF、jpeg、JPEG、jpg、JPG 的图形文件，以及名为“count.cgi”的脚本文件。但也不能一概而论，例如：对于图形、图像为主要内容的 Web 站点来说，日志文件所记录用户访问的图形文件恰恰反映了用户的浏览行为，所以就不能删除 GIF 和 JPEG 文件。因此，Web 数据净化的实际删除操作中，应根据具体的挖掘任务结合网

站的性质来具体确定。

2)用户识别　完成了Web日志的数据净化处理后，接下来的任务就是唯一地识别用户。但是由于本地缓存、防火墙和代理服务器的存在，使得Web日志并没有精确记录用户的浏览行为，因此从日志中识别出每一个用户就比较困难。解决这个问题的最简单的方法就是依靠用户的合作。即使是基于日志/站点的方法，也有一些启发式规则帮助唯一识别用户：(1)对于具有相同IP地址的用户，如果代理日志表明用户所使用的浏览器或操作系统改变了，则认为不同的代理就表示不同的用户[2]；(2)将访问日志、引用日志和站点的拓扑结构结合，构造每一位用户的浏览路径。如果用户当前请求的页面同已浏览过的页面之间没有任何超链接关系，则认为另外一个用户在使用同一台计算机[3]；(3)文[4]提出了一种通过导航模式自动确定用户会话长度来识别用户的方法。

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划分为单个的会话(Session)。最简单的方法是采用超时识别用户会话，如果用户请求的页面之间的时间超过一定的间隔，则认为用户开始了一个新的会话。一些商业化的Web日志挖掘产品，通常以30分钟为界限。文[5]通过实验数据得出的时间界限为25.5分钟。通过分析Web日志得出用户访问模式的统计数据，则可以准确定位特定站点的时间界限，作为会话识别算法的输入。

4)路径补充　能够可靠识别不同用户的访问操作的另外一个重要、但很艰巨的任务就是确定Web日志中是否未完整记录用户的访问行为，路径补充的任务就是将这些遗漏的请求补充到用户会话文件中。

由于本地缓存和代理服务器缓存的存在，用户可以使用浏览器上的"Back"按键，调用存储在缓存中的页面，从而使服务器日志遗漏用户请求的一些页面。

解决由于缓存机制所带来的用户访问路径遗失的问题，可以通过结合站点的拓扑结构、引用日志和时间信息进行补充。如果用户请求的Web页面与当前已请求的最后一个Web页之间没有超链接关系，需要检查引用日志记录，以确定该请求来自何处。如果该请求缓存的Web页是用户以前访问过的，则可以认为用户点击了"Back"按纽，执行了返回操作；如果利用引用记录也不能确定，则需要利用站点的拓扑结构进行分析。如果缓存中包含了一个以上的用户请求Web页，则提取时间最近的Web页，添加到用户会话文件中。

5)格式化　当以上对服务器日志的预处理步骤完成后，就进入了数据预处理的最后一个处理过程，将用户会话针对挖掘活动的特定需要进行格式化。例如，在对关联规则的挖掘中，不需要关于网络协议和时间的信息，则可以在用户会话文件中将其删去。

3.5　用户访问模式挖掘算法

由于Web数据的特殊性，数据挖掘的一些成熟算法不能直接应用到Web数据的挖掘。Web数据挖掘的方法和算法涉及诸多领域的知识，如统计学、数据挖掘、机器学习和模式识别等。目前已经用于Web日志和用户会话文件的分析及用户行为模式的挖掘方法主要有以下几种[7]：

1)统计分析　是分析用户访问站点的行为数据的最常用的方法。通过分析用户浏览页面的时间、用户的浏览路径和路径长度等信息，可以获得用户访问站点的基本信息，如页面访问次数，日平均访问人数，最受用户欢迎的页面等；也可以进行有限的错误分析，如非法用户登录等。统计分析的结果可以用于提高网站的性能、安全性以及优化站点结构和市场决策。

2)关联规则　指发现用户会话中经常被用户一起访问的页面集合，这些页面之间并没有顺序关系。如果关联规则中的页面之间没有超链接，则这是一个我们感兴趣的关联规则。挖掘关联规则通常使用Apriori算法或其变形算法。

关联规则既可以作为站点设计人员优化站点的参照，也是在Web上进行市场开发和商务决策的依据。同时关联规则还可以作为启发式规则为远程客户预取可能请求的页面，减少服务器的响应时间，以减少用户的等待时间。

3)聚类　聚类分析是把具有相似特征的用户或数据项归类。在Web日志挖掘中，聚类分析主要有两类：用户聚类和页面聚类。用户聚类将具有相似浏览行为的用户归类。利用这类知识可以在电子商务中进行市场分割或者为用户提供个性化Web页面内容；页面聚类则是将内容相关的页面归类。页面聚类的结果可以供搜索引擎使用，用以根据用户查询的信息或历史记录，建立与相关HTML页面间的超链接。

4)分类　是将数据项划分成预先定义的类别。在Web日志挖掘领域中，分类主要是按照用户特征数据将用户归属到既定的用户类。分类技术要求选择和抽取特征属性来描述指定的用户类别。分类的方法主要包括决策树分类法、贝叶斯分类法、最近邻分类法和Support Vector Machine等。

5)序列模式　指在时序数据集中发现在时间上具有先后顺序的数据项。在Web日志挖掘领域中，序列模式识别指寻找用户会话中在时间上有先后关系的页面请求。利用发现的序列模式可以预测用户即将可能请求的页面，这样就可以针对特定的用户组在页面中放置不同的广告条来增加广告的点击率(click through)。其它方面的序列模式有：趋势分析，转折点监测，相似性分析等。

6)依赖性建模　是Web挖掘领域另外一种非常有用的模式识别方法。其目的在建立一种模型，该模型可以表示Web域中不同变量之间的重要依赖关系。例如，根据用户在线购物的行动过程(从偶然的访问者变成真正潜在的购买者)，人们希望构建一种用于描述用户购物不同阶段行为的模型。有几种概率论的方法可以用来模拟用户的浏览行为，如隐马尔可夫模型、贝叶斯信度网。这种模式不仅可以提供分析用户行为的理论框架，而且对于预测未来的Web消费具有潜在的作用。这些信息也可以用于提高站点的在线销售额或为用户浏览导航提供方便。

3.6　模式分析及可视化

模式分析的主要任务就是通过领域专家和机器的评价，从Web日志挖掘所得到的模式中过滤掉冗余、无关、和常识性的规则和模式，找出用户感兴趣的模式。通常采用的方法有两种：一种方法是采用SQL查询语句进行分析；另外一种方法是将数据导入多维数据立方体中，而后利用OLAP工具进行分析并提供可视化的结果输出。

经过模式分析所得到的有价值的模式，采用可视化的技术以图形界面的方式表示给使用者。

4　Web日志挖掘的应用

日志挖掘技术可以应用在网站的个性化服务、站点系统性能改进以及智能化等方面[1,6,7]。

(下转第43页)

计出来的兴趣度相同。实验结果见表1。

从实验结果可得出如下结论：首先，CRI方法能区别J-measure方法和GACE方法所不能区别的不同例外关联规则。例如，如果使用J-measure方法，对第1,2,5,9个规则估计的兴趣度都是一样的，但是CRI方法能够区分这些规则集相应的兴趣度大小。再看GACE方法，第11,12个规则集估计的兴趣度一样，但是CRI方法能够区分这些规则集相应的兴趣度大小；其次，两个等价的规则集，CRI方法估计出来的兴趣度相同。1和11,2和4,3和6,5和7,8和10,9和12都是等价的规则集，但是J-measure方法和GACE方法却对某些等价的规则集算出不同的兴趣度，而CRI方法对这些等价方法估计出来的兴趣度都是相同的；再次，CRI方法基本上能够找到J-measure方法和GACE方法找出的所有兴趣度高的规则。

结论 本文讨论了例外关联规则的挖掘方法，并引入兴趣度的概念来测量例外关联规则所包含的信息量。本文对估计兴趣度的方法做了改进，根据常识关联规则与参照关联规则来度量一条例外关联规则的相对兴趣度，从而能更好地反映例外关联规则所包含的信息量，挖掘出真正感兴趣的例外关联规则。本文给出了具体的实现算法，实验结果表明该方法达到了较好的效果。

参考文献

1 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. 20th Int'l Conf. Very Large Data Bases (VLDB), 1994. 487～499

2 Liu H, Lu H, Feng L, Hussain F. Efficient search of reliable exceptions. In Methodologies for Knowledge Discovery and Data Mining, LNAI 1574(PAKDD), Berlin: Springer, 1999. 194～203

3 Padmanabhan B, Tuzhilin A. A belief-driven method for discovering unexpected patterns. In: Proc. Fourth Int'l Conf. Knowledge Discovery and Data mining(KDD), 1998. 94～100

4 周欣，沙朝锋，朱扬勇，施伯乐. 兴趣度——关联规则的又一个阈值. 计算机研究与发展，2000，37(5)：627～633

5 Suzuki E, Kodratoff Y. Discovery of surprising exception rules based on intensity of implication. In Principles of Data Mining and Knowledge Discovery, LNAI 1510 (PKDD), Berlin: Springer, 1998. 10～18

6 Smyth P, Goodman R M. An information theoretic approach to rule induction from databases. IEEE Trans. Knowledge Discovery and Data Eng, 1992, 4(4): 301～316

7 Suzuki E. Discovering unexpected exceptions: a stochastic approach. In: Proc. Fourth Int'l Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery(RSFD), 1996. 225～232

(上接第29页)

1)个性化服务 通过用户聚类，根据同组中其他用户的页面请求的页面为用户动态生成页面；根据页面聚类，决定当前页面包含哪些页面的超链接，自动修改站点的拓扑结构；将当前用户会话与已有的页面聚类进行最佳匹配，为用户推荐其尚未浏览并且与当前页面之间没有直接链接的页面。

2)系统改进 用户对网站服务的满意程度与系统的性能密切相关，包括服务器缓存、网络传输、负载平衡、数据分布等。根据代理服务器的访问日志预测用户的请求在时间和空间的分布，这种预测可以帮助代理服务器选择页面预取和缓存的策略；从服务器日志中挖掘路径配置文件，作为创建动态HTML页面的依据，所以用户请求动态页面之前已经将动态页面生成，从而减少服务器的响应延迟时间。

3)网站拓扑结构改进和构建自适应站点 根据用户访问模式所描述的用户访问站点时的行为信息，可以修改站点页面内容、链接结构和构建自适应站点的信息。

4)网络安全 随着Internet的迅速发展，网站的安全性成为关注的焦点，同时也成为制约基于网络的应用的主要因素。由于Web日志挖掘技术可以用于监视非法登录、黑客入侵等，所以将Web日志挖掘技术和网络技术相结合是提高门户网站安全性的较佳的解决方案。

5 Web用户访问模式挖掘的发展方向

Web用户访问信息挖掘是一个较新的研究领域，具有广阔的发展和应用前景，应该指出的是，面对日益增加的商业需求，Web用户访问信息挖掘技术还有许多问题需要解决，有待这一领域的研究者深入研究。将来很有用的几个研究方向是：

(1)Web日志挖掘中内在机理及新的挖掘体系和结构的研究；

(2)用户访问模式库的动态维护和更新、模式(知识)的评价体系和评价方法；

(3)挖掘算法在海量数据挖掘时的适应性和时效性研究；

(4)智能站点服务个性化和性能最优化的研究；

(5)关联规则和序列模式在构造自组织站点的研究[8]；

(6)分类在电子商务市场智能提取中的研究[8]。

结束语 目前，国内外在Web用户访问信息挖掘领域的研究尚处于起步阶段，还没有形成比较成熟的理论和统一的体系。本文通过阐述Web用户访问信息挖掘中数据预处理的主要流程和技术方法、用户访问信息挖掘算法、模式分析方法以及Web用户访问信息的应用，旨在对Web日志挖掘的框架、流程、技术方法及用户访问信息挖掘的研究和发展方向做全面概括。

随着Internet的进一步发展，Web用户访问信息挖掘在个性化的信息服务、改进门户站点的设计和服务、开展有针对性的电子商务、电子政务、构建智能化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，Web用户访问信息挖掘技术将成为重要的研究课题和方向。

我们正在承担教育部科技司重点项目："远程教育网关键技术——信息挖掘和智能搜索工具的研究"(教技司[2000]175)。结合我们在数据挖掘内在机理研究方面所取得的成果，试图提出一种新型的Web用户访问信息挖掘的结构框架——基于双库协同机制的Web用户访问信息挖掘，我们将在以后的文章中详细论述。

参考文献

1 Kosala R, Blockeel H. Web Mining Research: A Survey. SIGKDD Exllorations, 2000, 2(1)

2 Pitkow J. In search of relable usage data on the www. In: Sixth Intl. World Wild Web Conf. Santa Clara, CA, 1997. 451～463

3 Pirolli P, Pitkow J, Rao R. Silk from a sow's ear: Extracting usable structures from the web. In: Proc. of 1996 Conf. on Human Factors in Computing Systems (CHI-96), Vancouver, British Columbia, Canada, 1996

4 Cooley R, Mobasher B, Srivastava J. Grouping web page references into transactions for mining world wide web browsing patterns: [Technical Report TR 97-021]. University of Minnesota, Dept. of Computer Science, Minneapolis, 1997

5 Catledge L, Pitkow J. Charactering browsing behaveious on the World Wild Web, Computer Networks and ISDN Systems, 1995, 27(6): 1065～1073

6 Cooley R, et al. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1999, 1

7 Srivastava J, et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Appear in SIGKDD Explorations, 2000, 1(2)

8 王实，高文，李锦涛. Web数据挖掘. 计算机科学，2000，27(4)