# Web日志挖掘分析的方法

日志文件的格式及其包含的信息  
①2006-10-17 00:00:00②202.200.44.43 ③218.77.130.24 80 ④GET ⑤/favicon.ico   
⑥Mozilla/5.0+(Windows；+U；+Windows+NT+5.1；+zh-CN；+rv：1.8.0.3)+Gecko/20060426  
+Firefox/1.5.0.3。  
①访问时间；②用户IP地址；③访问的URL，端口；④请求方法(“GET”、“POST”等)；  
⑤访问模式；⑥agent，即用户使用的操作系统类型和浏览器软件。

一、日志的简单分析  
1、注意那些被频繁访问的资源  
2、注意那些你网站上不存在资源的请求。常见的扫描式攻击还包括传递恶意参数等：  
3、观察[**搜索引擎**](http://lib.csdn.net/base/searchengine)蜘蛛的来访情况  
4、观察访客行为  
应敌之策：  
1、封杀某个IP  
2、封杀某个浏览器类型（Agent）  
3、封杀某个来源（Referer）  
4、防盗链  
5、文件重命名  
作用：  
1.对访问时间进行统计，可以得到服务器在某些时间段的访问情况。  
2.对IP进行统计，可以得到用户的分布情况。  
3.对请求URL的统计，可以得到网站页面关注情况。  
4.对错误请求的统计，可以更正有问题的页面。

二、Web挖掘  
根据所挖掘的Web 数据的类型，可以将Web 数据挖掘分为以下三类：Web 内容挖掘(Web Content Mining)、Web 结构挖掘(Web Structure Mining)、Web 使用挖掘(Web Usage Mining)（也称为Web日志挖掘）。  
①Web内容挖掘。Web内容挖掘是指从文档的内容中提取知识。Web内容挖掘又分为文本挖掘和多媒体挖掘。目前多媒体数据的挖掘研究还处于探索阶段,Web文本挖掘已经有了比较实用的功能。Web文本挖掘可以对Web上大量文档集合的内容进行总结、分类、聚类、关联分析,以及利用Web文档进行趋势预测等。Web文档中的标记,例如<Title>和<Heading>等蕴含了额外的信息,可以利用这些信息来加强Web文本挖掘的作用。   
②Web结构挖掘。Web结构挖掘是从Web的组织结构和链接关系中推导知识。它不仅仅局限于文档之间的超链接结构,还包括文档内部的结构。文档中的URL目录路径的结构等。Web结构挖掘能够利用网页间的超链接信息对搜索引擎的检索结果进行相关度排序,寻找个人主页和相似网页,提高Web搜索蜘蛛在网上的爬行效率,沿着超链接优先爬行。Web结构挖掘还可以用于对Web页进行分类、预测用户的Web链接使用及Web链接属性的可视化。对各个商业搜索引擎索引用的页数量进行统计分析等。   
③Web使用记录挖掘。Web使用记录挖掘是指从Web的使用记录中提取感兴趣的模式，目前Web使用记录挖掘方面的研究较多,WWW中的每个服务器都保留了访问日志,记录了关于用户访问和交互的信息,可以通过分析和研究Web日志记录中的规律,来识别网站的潜在用户;可以用基于扩展有向树模型来识别用户浏览序列模式,从而进行Web日志挖掘;可以根据用户访问的Web记录挖掘用户的兴趣关联规则,存放在兴趣关联知识库中,作为对用户行为进行预测的依据,从而为用户预取一些Web页面,加快用户获取页面的速度，分析这些数据还可以帮助理解用户的行为,从而改进站点的结构,或为用户提供个性化的服务。  
通过对Web服务器日志中大量的用户访问记录深入分析，发现用户的访问模式和兴趣爱好等有趣、新颖、潜在有用的以及可理解的未知信息和知识，用于分析站点的使用情况，从而辅助管理和支持决策。当前，web日志挖掘主要被用于个性化服务与定制、改进系统性能和结构、站点修改、商业智能以及web特征描述等诸多领域。

三、Web日志挖掘的方法  
（一）首先，进行数据的预处理。  
从学习者的访问日志中得到的原始日志记录并不适于挖掘，必须进行适当的处理才能进行挖掘。因此，需要通过日志清理，去除无用的记录；对于某些记录，我们还需要通过站点结构信息，把URL路径补充成完整的访问序列；然后划分学习者，并把学习者的会话划分成多个事务。  
（二）其次，进行模式发现  
一旦学习者会话和事务识别完成，就可以采用下面的技术进行模式发现。模式发现, 是对预处理后的数据用数据挖掘算法来分析数据。分有统计、分类、聚类、关等多种方法。  
① 路径分析。它可以被用于判定在一个站点中最频繁访问的路径，还有一些其它的有关路径的信息通过路径分析可以得出。路径分析可以用来确定网站上的频繁访问路径, 从而调整和优化网站结构, 使得用户访问所需网页更加简单快捷, 还可以根据用户典型的浏览模式用于智能推荐和有针对性的电子商务活动。例如：70% 的学习者在访问/ E-Business /M2时，是从/EB开始，经过/ E-Business /SimpleDescription，/ E-Business /M1；65%的学习者在浏览4个或更少的页面内容后就离开了。利用这些信息就可以改进站点的设计结构。  
② 关联规则。 使用关联规则发现方法，可以从Web的访问事务中找到的相关性。关联规则是寻找在同一个事件中出现的不同项的相关性，用数学模型来描述关联规则发现的问题：x=>y的蕴含式，其中x,y为属性——值对集(或称为项目集)，且X∩Y空集。在数据库中若S%的包含属性——值对集X的事务也包含属性——值集Y，则关联规则X=>Y的置信度为C%。  
③ 序列模式。在时间戳有序的事务集中，序列模式的发现就是指那些如“一些项跟随另一个项”这样的内部事务模式。它能发现数据库中如“在某一段时间内，客户购买商品A，接着会购买商品B，尔后又购买商品C，即序列A→B→C出现的频率高”之类的信息。序列模式描述的问题是：在给定的交易序列数据库中，每个序列按照交易的时间排列的一组交易集，挖掘序列函数作用是返回该数据库中高频率出现有序列。  
④ 分类分析。发现分类规则可以给出识别一个特殊群体的公共属性的描述，这种描述可以用于分类学习者。分类包括的挖掘技术将找出定义了一个项或事件是否属于数据中某特定子集或类的规则。该类技术是最广泛应用于各类业务问题的一类挖掘技术。分类算法最知名的是决策树方法，此外还有神经元网络、Bayesian分类等。例如：在/ E-Business /M4学习过的学习者中有40％是20左右的女大学生。  
⑤聚类分析。可以从Web访问信息数据中聚类出具有相似特性的学习者。在Web事务日志中，聚类学习者信息或数据项能够便于开发和设计未来的教学模式和学习群体。聚类是将数据集划分为多个类，使得在同一类中的数据之间有较高的相似度，而在不同类中的数据差别尽可能大。在聚类技术中，没有预先定义好的类别和训练样本存在，所有记录都根据彼此相似程度来加以归类。主要算法有k—means、DBSCAN等。聚类分析是把具有相似特征的用户或数据项归类,在网站管理中通过聚类具有相似浏览行为的用户。基于模糊理论的Web页面聚类算法与客户群体聚类算法的模糊聚类定义相同，客户访问情况可用URL(Uj)表示。有Suj={(Ci，fSuj(Ci))|Ci∈C}，其中fSuj(Ci)→[0，1]是客户Ci和URL(Uj)间的关联度：式中m为客户的数量，hits(Ci)表示客户Ci访问URL(Uj)的次数。利用Suj和模糊理论中的相似度度量Sfij定义建立模糊相似矩阵，再根据相似类[Xi]R的定义构造相似类，合并相似类中的公共元素得到的等价类即为相关Web页面。  
⑥统计。统计方法是从Web 站点中抽取知识的最常用方法, 它通过分析会话文件, 对浏览时间、浏览路径等进行频度、平均值等统计分析。虽然缺乏深度, 但仍可用于改进网站结构, 增强系统安全性, 提高网站访问的效率等。  
⑦协同过滤。协同过滤技术采用最近邻技术，利用客户的历史、喜好信息计算用户之间的距离，目标客户对特点商品的喜好程度由最近邻居对商品的评价的加权平均值来计算。  
（三）最后，进行模式分析。  
模式分析。基于以上的所有过程，对原始数据进行进一步分析，找出用户的浏览模式规律，即用户的兴趣爱好及习惯，并使其可视化，为网页的规划及网站建设的决策提供具体理论依据。其主要方法有：采用SQL查询语句进行分析；将数据导入多维数据立方体中，用OLAP工具进行分析并给出可视化的结果输出。（分类模式挖掘、聚类模式挖掘、时间序列模式挖掘、序列模式挖掘、关联规则等）

四、关联规则  
（一）关联规则  
顾名思义，关联规则（association rule）挖掘技术用于于发现数据库中属性之间的有趣联系。一般使用支持度（support）和置信度（confidence）两个参数来描述关联规则的属性。   
（二）Apriori方法简介  
Apriori算法最先是由Agrawal等人于1993年提出的，它的基本思想是：首先找出所有具有超出最小支持度的支持度项集，用频繁的(k—1)-项集生成候选的频繁k-项集；其次利用大项集产生所需的规则；任何频繁项集的所有子集一定是频繁项集是其核心。  
Apriori算法需要两个步骤：第一个是生成条目集；第二个是使用生成的条目集创建一组关联规则。当我们把最小置信度设为85%，通过关联规则的形成以及对应置信度的计算，我们可以从中得到以下有用的信息：  
1.置信度大于最小置信度时：我们可以这样认为，用户群体在浏览相关网页时，所呈列的链接之间是有很大关联的，他们是用户群的共同爱好，通过网页布局的调整，从某种意义上，可以带来更高的点击率及潜在客户；  
2.置信度小于最小置信度时：我们可以这样认为，用户群体对所呈列链接之间没太多的关联，亦或关联规则中的链接在争夺用户。

五、网站中Web日志挖掘内容  
　　（1）网站的概要统计。网站的概要统计包括分析覆盖的时间、总的页面数、访问数、会话数、惟一访问者、以及平均访问、最高访问、上周访问、昨日访问等结果集。  
　　（2）内容访问分析。内容访问分析包括最多及最少被访问的页面、最多访问路径、最多访问的新闻、最高访问的时间等。  
　　（3）客户信息分析。客户信息分析包括访问者的来源省份统计、访问者使用的浏览器及操作系统分析、访问来自的页面或者网站、来自的IP地址以及访问者使用的搜索引擎。  
　　（4）访问者活动周期行为分析。访问者活动周期行为分析包括一周7天的访问行为、一天24小时的访问行为、每周的最多的访问日、每天的最多访问时段等。  
　　（5）主要访问错误分析。主要访问错误分析包括服务端错误、页面找不到错误等。  
　　（6）网站栏目分析。网站栏目分析包括定制的频道和栏目设定，统计出各个栏目的访问情况，并进行分析。  
（7）商务网站扩展分析。商务网站扩展分析是专门针对专题或多媒体文件或下载等内容的访问分析。  
（8）有4个方向可以选择:①对用户点击行为的追踪，click stream研究；②对网页之间的关联规则的研究；③对网站中各个频道的浏览模式的研究；④根据用户浏览行为，对用户进行聚类，细分研究；（如果你能够结合现有的互联网产品和应用提出一些自己的建议和意见，那就更有价值了。）  
（9）发现用户访问模式。通过分析和探究Web日志记录中的规律，可以识别电子商务的潜在客户，提高对最终用户的服务质量，并改进Web服务器系统的性能。   
(10)反竞争情报活动。反竞争情报是企业竞争情报活动的重要组成部分。

六、相关软件及算法  
（一）相关软件：  
1.数据挖掘的专用软件wake。  
2.用OLAP工具  
3.已经有部分公司开发出了商用的网站用户访问分析系统，如WebTrends公司的CommerceTrends 3.0,它能够让电子商务网站更好地理解其网站访问者的行为，帮助网站采取一些行动来将这些访问者变为顾客。CommerceTrends主要由3部分组成：Report Generation Server、Campain Analyzer和Webhouse Builder。  
4.Accrue公司的Accrue Insight，它是一个综合性的Web分析工具,它能够对网站的运行状况有个深入、细致和准确的分析，通过分析顾客的行为模式，帮助网站采取措施来提高顾客对于网站的忠诚度，从而建立长期的顾客关系。  
（二）相关算法：  
1.运用各种算法进行数据挖掘：GSP算法, Prefixspana算法，  
2.关联规则分析：Apriori、FP-growth算法等。  
3.Apriori算法及其变种算法  
4.基于数据库投影的序列模式生长技术（database project based sequential pattern growth）  
5. Wake算法、MLC++等  
6. PageRank算法和HITS算法利用Web页面间的超链接信息计算“权威型”（Authorities）网页和“目录型”（Hubs）网页的权值。Web结构挖掘通常需要整个Web的全局数据，因此在个性化搜索引擎或主题搜索引擎研究领域得到了广泛的应用。  
7.参考检索引擎的挖掘算法，比如Apache的lucene等。

七、日志分析的价值或应用  
①在自己的网站上安装了网站统计的代码，如Google analytics、量子统计、百度统计、cnzz、51.la等，这些工具可以统计网站的流量，也就是网站上访客可看到的所有页面的访问量，但是这些统计工具都不能统计你主机上资源的原始访问信息，例如某个图片被谁下载了。  
②如果你的网站遭到了攻击、非法盗链和不良请求等，通过分析原始访问日志能大概分析出端倪来，例如：往主机上传了一个mp3，不幸被百度mp3收录，引来大量的盗链，导致我的主机流量猛增！通过分析日志，可以找出问题根源，删除了那个mp3，主机流量也降下来了。  
③分析访客来源（Referer）。这一段是告诉我们访客是从哪里来到这一个网页。有可能是网站其他页，有可能是来自搜索引擎的搜索页等。通过这条来源信息，你可以揪出盗链者的网页。  
④网站日志分析软件都能提供关于服务器的浏览量、统计网站所有页面和相关文件被显示的次数、访问最多的网页、客户端访问最频繁的文件、访问者的IP分布、每日访问统计、每周每月等的统计结果。1.访问者访问时段分析。结合IP地址和时段之间的关系可以将来访者大致的身份作一个基本的判断。如按上班前、工作期间、下班后、节假日等，可以针对访客的初步性质安排合适的内容，如产品信息和广告；2.访问者地区分布。分析通过将访问者的IP地址转换为地理区间可以分析出来访者的大致地理分布范围。  
⑤相关产品推荐。通过以上的关联分析，有了用户频繁访问路径和链接之间的兴趣度，可以构建个性化推荐系统模型。对于实证例子，我们可以在置信度高于最低置信度的相关链接之间，建立某种信息快速互联的桥梁，亦或是在网页规划中，充分考虑链接之间的关联关系，从而为更人性化、合理化的网页设计提供决策依据。如：当客户浏览/newimg/num1.gif时，有0.91的概率会浏览/newimg/num4.gif，那么，在两者之间就存在很高的关联性，从而我们有必要对这两个链接建立某种跟紧密的联系。  
⑥个性挖掘：针对单个用户的使用记录对该用户进行建模，结合该用户基本信息分析他的使用习惯、个人喜好，目的是在电子商务环境下为该用户提供与众不同的个性化服务。  
⑦系统改进：Web服务（数据库、网络等）的性能和其他服务质量是衡量用户满意度的关键指标，Web 用法挖掘可以通过用户的拥塞记录发现站点的性能瓶颈，以提示站点管理者改进Web缓存策略、网络传输策略、流量负载平衡机制和数据的分布策略。此外，可以通过分析网络的非法入侵数据找到系统弱点，提高站点安全性，这在电子商务环境下尤为重要。  
⑧站点修改：站点的结构和内容是吸引用户的关键。Web 用法挖掘通过挖掘用户的行为记录和反馈情况为站点设计者提供改进的依，比如页面连接情况应如何组织、那些页面应能够直接访问等。  
⑨智能商务：用户怎样使用Web站点的信息无疑是电子商务销售商关心的重点，用户一次访问的周期可分为被吸引、驻留、购买和离开四个步骤，Web用法挖掘可以通过分析用户点击流等Web日志信息挖掘用户行为的动机，以帮助销售商合理安排销售策略。  
⑩Web特征描述：这类研究跟关注这样通过用户对站点的访问情况统计各个用户在页面上的交互情况，对用户访问情况进行特征描述。