基于 Web 日志挖掘的网上学习行为研究

李晓昕[1,2],谢维奇[1,2]
(1. 驻马店职业技术学院 信息工程系,河南 驻马店 463000;
2. 山东大学 数学与系统科学学院, 山东 济南 250100)

摘 要:网络教育要想为学习者提供个性化的指导和服务,必须注重教学过程跟踪,注意对学生学习行为的分析。Web 服务器日志中记录了访问者的所有信息,通过数据挖掘的方法可以获得需要的有用知识,并由此得到用户的访问模式。文中使用 Web 日志挖掘的方法分析学生的网上学习行为,通过数据过滤、用户辨别和会话辨别,采用模糊集和粗糙集的方法获得访问用户的聚类和分类等有用信息。实验证明,通过 Web 日志挖掘的方法,可以更好地了解学生的学习偏好,提高教学服务质量。

关键词:Web 日志挖掘;网上学习行为;模糊聚类;粗糙集
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2011)12-0073-04

Research on E-Learning Behavior Based on Weblog Mining

LI Xiao-xin[1,2], XIE Wei-qi[1,2]
(1. Department of Information Engineering, Zhumadian Vocational and Technical College, Zhumadian 463000, China;
2. School of Mathematics and System Sciences, Shandong University, Jinan 250100, China)

Abstract: In order to provide personalized guidance and service for learners, online education must focus on tracking the process of teaching, pay attention to analyze the student's learning behavior. Web server logs keep visitor's all information, so useful knowledge needed can be gotten by data mining , and thus the user's access patterns. It uses analysis of Web log mining methods of online learning behavior of students, through data filtering, user identification and session identification, use fuzzy set and rough set way to get access to the user useful information such as clustering and classification. Experiments show that, can better understand the learning preferences, can improve the teaching quality of service through Web log mining.

Key words: Weblog mining; e-learning behavior; fuzzy clustering; rough set

0 引 言

计算机技术的发展为网络远程教育提供了良好的技术支持。网络、多媒体及虚拟现实等技术相结合,可以模拟教学课堂的现场,学生可以根据自己的需要主动选择学习的内容及课程的进度。教师与学生之间可以通过网络环境的支持,在任何时间、任何地点进行信息交流,不必拘泥于课堂上有限的时间。这种模式打破了传统教育的局限,具有传统教育所不可比拟的优点。中央电大的远程开放教育和其他普通高校开设的网络学院便是网络远程教育的一种尝试。现在的网络教学系统所提供的教学资源虽然可以满足不同的学习者在不同的学习时间的个别化学习的需求,但大多数只是将教学资源放到了网络上,没有考虑到学习者的个性化特征。教师难以对学生学习的过程进行控制,而学习者自己由于能力所限,也难以对学习过程进行自我控制,所以大多数网络教学的效果是难以保证的。

因此,建立双向互动网络教学模式,注重教学过程跟踪,注意对学生学习行为的分析,根据学生的实际情况进行个性化的指导,对建立个性化的网络学习环境,提高网络学习质量大有益处。

现有的网络教学平台大都对学生的访问信息进行记录,可以对这些数据进行分析。另一方面,注意到用户每次访问网站的信息都保存在网站服务器日志中,这些信息能较为全面地反映用户的需求,所以可以通过它对用户访问的内容、行为和次数等进行分析,从而获取有用的用户信息,并能总结出用户的访问模式。更重要的是,通过对这些用户特征的理解和分析,对学生的学习状况、学习需求、能力差异、学习进度和兴趣爱好进行跟踪,对教学资源的访问模式进行分析,可以掌握远程学习者的学习情况,改进教学网站的设计,合

收稿日期:2011-04-22;**修回日期**:2011-07-27
基金项目:河南省基础与前沿技术研究计划项目(112300410056);驻马店市基础与前沿技术研究计划项目(11705)
作者简介:李晓昕(1976-),女,河南驻马店人,讲师,CCF 会员,研究方向为数据挖掘、粗糙集理论及应用。

理配置教学资源,为学习者提供更好的服务。

1 Web日志挖掘

Web日志挖掘[1,2]通过分析Web访问日志中信息的规律发现用户的访问模式,找出隐藏的用户,从而提高对最终用户的服务质量,并可优化Web服务器的性能。

1.1 Web日志挖掘的主要内容

可以从分析系统性能、改进系统设计和理解用户目的三个方面对Web日志进行研究。但基于研究内容不同,采取的主要技术也不一样。

基于分析系统性能的研究,主要是对Web日志数据进行多方面的统计,如:页面的访问次数、访问量和访问路径等,通过分析统计的结果,可以对服务器的镜像站点和缓冲等设置进行相应的调整。

基于改进系统设计的研究,主要是如何以Web日志数据为依据在组织和表现形式上对Web服务器进行自动或半自动调整。

基于理解用户目的的研究,主要是采用算法从Web日志中找出频繁的用户访问模式或访问路径。这个方向也正是所要研究的。

1.2 Web日志挖掘的过程

Web日志挖掘的过程一般分为三个阶段,即数据收集和预处理、模式发现和模式分析。

Web日志挖掘的重要阶段是日志数据预处理,它的工作量占整个数据挖掘过程的50%,直接影响数据挖掘的质量。数据预处理主要是从原始Web日志中找出能为用户浏览模式发现算法所使用的规范化数据,算法处理结果的准确度与可信度将直接受这些数据的影响[3,4]。这个阶段主要包括数据过滤、用户辨别、会话辩别和路径完善等过程。数据过滤就是去掉挖掘过程中不需要的数据。用户辩别主要是分析多个用户通过防火墙或代理服务器访问站点的情况,从而将用户和请求页面相关联。在此过程中,不仅需要Web日志,还需要知道网站的拓扑结构。会话辨别是在一段时间内对一个用户的全部请求页面进行分解得到用户会话。路径完善是将本地或代理服务器缓存中的遗留请求补充完整的过程。

模式发现和模式分析过程主要是利用数据挖掘算法去分析过滤后的数据,从而发现知识。目前常用的Web日志数据挖掘方法主要包括统计、关联规则、聚类、分类和访问序列等。

(1)分类:分类是基于一定的标准将数据项进行划分。常用的分类方法主要包括粗糙集、决策树,贝叶斯法等方法。

(2)聚类:聚类分析是在相似的基础上对数据进行分类,没有具体的分类标准,目的是分析不同数据项的相似性,它是Web日志挖掘的主要任务之一。在Web日志挖掘中,聚类分析主要用于用户聚类和页面聚类。用户聚类是根据用户的访问动作,分析用户的会话,寻找具有相似访问模式的用户,并将其分类;页面聚类是通过分析页面被访问的情况,寻找被同一个用户访问的页面,并将其分类,被分类后的页面具有相同或相似的特征。聚类分析包括粗糙集理论、统计法和模糊理论等。

(3)访问序列:在一段时间内,具有一定先后时序关系的数据项就是一个访问序列。在Web日志挖掘领域中,主要是寻找在一段时间内用户先后访问的一系列请求页面,通过分析这些请求页面来预测用户将来可能的请求。

2 Web日志数据预处理

2.1 Web日志的组成

Web日志完整、详细地记录了访问用户的浏览行为,是Web日志挖掘的主要数据来源,通常以纯文本的形式存储在默认的存储位置:% systemroot% \system32\logfiles\,扩展名为.log。日志文档常见的日志记录中通常包含data(日期)、time(时间)、c-ip(客户端IP)、cs-username(客户端用户名)、s-ip(服务器IP)、s-port(服务器端口号)、cs-method(客户端请求方法)、cs-uri-stem(URL资源)、cs-uri-query(URL查询)、sc-status(协议状态)和cs(User-Agent)(客户端代理)。

Data和time表示页面被访问的时间;c-ip是访问用户的Ip地址;cs-username是用户名,对于匿名访问来说,它常为空,只有在用户访问的页面需要验证时才有用;s-ip和s-port是用户访问的服务器的Ip地址和端口;cs-method是用户访问的请求类型,主要有Post和Get两种类型;cs-uri-stem(URL资源)是用户访问的页面路径,根据URL可以了解用户获得了哪些信息;sc-status表示服务器的响应状态,如:2**表示用户请求成功;4**表示用户端错误;5**表示服务器错误等,cs(User-Agent)表示用户浏览器的基本属性,如:Ip、Os和Browser等。

2.2 Web日志数据预处理

(1)数据过滤。

数据过滤就是检查Web日志中URL的后缀,删除对Web日志挖掘认为意义不大的数据。根据前面对Web日志文件的分析,发现Web日志中URL的后缀名为gif、jpeg、js、cgi、css、txt等格式的文件,还有一些模板框架文件,这些文件对Web日志挖掘意义不大,所以首先要对Web日志数据进行数据过滤才能使用。

在文中收集的日志中删除了 GET 以外的其他动作。sc-status 中只保留请求成功的数据。另外,由于搜索引擎的 Web Robot 对网站的浏览是不带任何感性色彩的,所以日志中的 Web Robot 的请求也要被过滤掉,只保留那些能说明访问过的页面的文件,例如后缀为 htm,asp,jsp 等的文件。

(2)数据属性的约简。

文中收集到的是不同用户访问同一台服务器的 Web 日志,所以对于不同用户的访问,服务器端 Ip 地址和端口号都是相同的,这对所研究的日志挖掘意义不大。对于匿名用户来说,username 常为空,另外,只研究请求成功的访问用户。

经分析,所关心的主要内容是:什么用户(c-ip)在什么时间(time)利用什么样的浏览器(cs(User-Agent))访问了哪些页面(cs-uri-stem),经属性约简后,得到的 weblog 日志记录中主要包含: time(时间)、c-ip(客户端 IP)、cs-uri-stem(URL 资源)、cs-uri-query(URL 查询) 和 cs(User-Agent)(客户端代理)。

(3)用户辨别[5]。

由于本地缓存、公司防火墙和代理服务器的存在使从日志中识别每个访问网站的用户变得复杂,而且 Username 中的内容常是空白的,所以在这里采用基于 IP 和 Agent 的方法进行用户辨别,认为拥有相同的 IP 和 Agent 的用户就是同一个用户[6]。

(4)会话辨别。

会话识别的目的就是将用户的访问记录分为单个会话,一般采用超时辨别法,当用户的访问时间间隔大于阈值时(时间一般设置为 30 分钟),可以判断用户进行了两次会话。

文中的数据采用的是驻马店职业技术学院网站(www.zmdvtc.cn)2010 年 10 月 27 日的日志,文件 1.28M,共 7418 条数据。经过数据过滤、属性简化、用户辨别和会话辨别,在日志中保留记录 2743 条,识别用户 467 个,识别会话 760 个,访问页面 287 个。

3　模式发现及结果

3.1　聚类分析

由于数据量大,这里选取 8 个用户和 10 个页面(如表 1 所示)采用模糊聚类[7~9]的方法进行分析(U 表示用户,P 表示用户访问的页面)。

把表 1 中的数据作为原始的 Web 数据矩阵 M,经过数据标准化及用直接海明距离法 $r_{ij} = 1 - C \times \sum_{k=1}^{m} |x_{ik} - x_{jk}|$ 处理后得到 M 的相似矩阵 M'(见图 1),直接海明距离法中的参数 $c = 0.01$。

表 1　用户访问页面信息表

	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9
U0	1	0	0	0	0	0	0	0	0	0
U1	17	1	1	0	0	1	1	1	2	12
U2	0	0	0	1	1	0	0	0	0	0
U3	2	0	0	0	0	0	0	0	0	1
U4	0	0	0	1	1	0	0	0	0	0
U5	2	0	0	0	0	0	0	0	0	0
U6	0	0	0	1	1	0	0	0	0	1
U7	1	0	0	0	0	0	0	0	0	0

M' =

$$\begin{bmatrix} 1.00 & & & & & & & \\ 0.84 & 1.00 & & & & & & \\ 0.96 & 0.80 & 1.00 & & & & & \\ 1.00 & 0.84 & 0.96 & 1.00 & & & & \\ 0.96 & 0.80 & 1.00 & 0.96 & 1.00 & & & \\ 1.00 & 0.80 & 0.96 & 1.00 & 0.96 & 1.00 & & \\ 0.96 & 0.80 & 1.00 & 0.96 & 1.00 & 0.96 & 1.00 & \\ 1.00 & 0.84 & 0.96 & 1.00 & 0.96 & 1.00 & 0.96 & 1.00 \end{bmatrix}$$

图 1　Web 模糊相似矩阵 M'

采用直接聚类法对模糊相似矩阵 M' 进行聚类分析。

$\lambda = 1$ 时,用户分为 $\{U0, U3, U5, U7\}$, $\{U2, U4, U6\}$, $\{U1\}$。

$\lambda = 0.96$ 时, 用户分为 $\{U0, U3, U5, U7, U2, U4, U6\}$, $\{U1\}$。

可以得到结论,在研究的 8 个学员中, $U0, U3, U5$ 和 $U7$ 这四位学员学习习惯比较接近,而 $U2, U4$ 和 $U6$ 比较相近。

3.2　基于粗糙集的分类

依然选用表 1 中的 8 个用户和 10 个页面,建立用户访问页面信息系统 IS = (U, C, V, f),其中 $U = \{U0, U1, \cdots, U7\}$, $C = \{P0, P1, \cdots, P9\}$, 对应的差别矩阵[10~13] 见图 2。

∅							
C-{p3, p4}	∅						
{p0, p3, p4}	C	∅					
{p0, p9}	C-{p3, p4}	{p0, p3, p4, p9}	∅				
{p0, p3, p4}	C	∅	{p0, p3, p4, p9}	∅			
{p0}	C-{p3, p4}	{p0, p3, p4}	{p9}	{p0, p3, p4}	∅		
{p0, p3, p4, p9}	C	{p9}	{p0, p3, p4}	{p9}	{p0, p3, p4, p9}	∅	
∅	C-{p3, p4}	{p0, p3, p4}	{p0, p9}	{p0, p3, p4}	{p0}	{p0, p3 p4, p9}	∅

图 2　用户访问页面信息系统的差别矩阵

从差别矩阵可以看出，该信息系统的核为 CORE (C) = {P0,P9}。计算可得属性约简为{P0,P3,P4,P9}。这样知道，在研究的 10 个页面中，P0 和 P9 非常重要，P3 和 P4 也比较重要，这四个页面经常被访问，可以通过修改网站的结构，把他们放在比较容易访问的位置，便于学习者浏览。

3.3 学习者访问路径

通过对每个会话的分析，得到远程学习者在本网站的访问序列(见表 2)，也就是他在这个网站的访问路径。通过对路径的分析可以知道学习者在网站的访问习惯、访问偏好等，可以帮助我们调整网站的布局，提高学习者的学习效率。

表 2 用户访问序列

SessionID	UserID	Web 访问序列
S0	U1	P1,P2,P5,P6,P7,P8,P9,P0,P012,P0,P013,P014,P015,P016,P017,P018,P019,P020,P021,P022,P023,P024,P025,P9,P0,P012,P014,P0
S010	U1	P057,P058,P014,P073,P0,P010,P014,P076,P077,P078,P079,P080,P081,P082,P083
S457	U112	P9,P011,P3,P4,P9,P011,P3,P4
.....	.....	.....

4 结束语

通过 Web 日志挖掘的方法可以跟踪学习者的网上学习行为，了解学生的学习偏好，便于我们更好地提高教学支持。另一方面，由于网络环境的复杂性，以及学习者上网习惯的关系，想从海量的 Web 日志中发现需要的知识有相当的难度，这需要不断探索新的方法。

参考文献：

[1] Han Jiawei. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2004.

[2] Liu Bing. Web 数据挖掘[M]. 北京：清华大学出版社，2009.

[3] 周 勇，刘 锋. 基于粗糙集的 Web 结构挖掘[J]. 计算机技术与发展，2008，18(3)：151-153.

[4] 朱鹤祥. Web 日志挖掘中数据预处理算法的研究[D]. 大连：大连交通大学，2009.

[5] 张春生，庄丽艳. 基于兴趣的 Web 挖掘中用户身份的识别新方法[J]. 计算机技术与发展，2009，19(5)：62-64.

[6] Kosala R, Blockeel H. Web Mining Research: A Survey (2000)[J]. SIGKDD Explorations - Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2000, 2(1): 1-15.

[7] 许海洋. 模糊聚类分析在数据挖掘中的应用研究[J]. 计算机工程与应用，2005，17：177-179.

[8] 王 伟，高 亮，吴 涛. 一种基于模糊聚类的离散化算法[J]. 计算机技术与发展，2008，18(3)：53-55.

[9] 高新波. 模糊聚类分析及其应用[M]. 西安：西安电子科技大学出版社，2004.

[10] 苗夺谦. 粗糙集理论、算法与应用[M]. 北京：清华大学出版社，2008.

[11] Pawlak Z. Rough Set[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[12] Ziarko W. Variable Predsion Rough Set Model[J]. Journal of Computer and System Sciences, 1993, 46(1): 39-59.

[13] Bean C. Autonomous Clustering Using Rough Set Theory[J]. International Journal of Automation and Computing, 2008, 5(1): 90-102.

(上接第 72 页)

基于分布式和位集合的频繁项集生成算法，解决在数据源分布、数据量巨大的情况下，如何高效地进行频繁项集挖掘的问题。

参考文献：

[1] Han J, Kamber M. DataMining: Concepts and Techniques [M]. Beijing: Higher Education Press, 2001.

[2] 陈 伟. Apriori 算法的优化方法[J]. 计算机技术与发展，2009，19(6)：80-83.

[3] 黄端琼，陈崇成，黄洪宇，等. 基于映射位集合的遥感图像关联规则挖掘[J]. 计算机应用，2005，25(7)：1592-1594.

[4] 刘永彬，秦亮曦，王永卿，等. 基于 Apriori 和位集合的关联规则应用[J]. 微计算机信息，2007(33)：141-143.

[5] Agrawal R, Srikant R. Fast algorithm for mining association rules[C]//Proc of the 21st VLDB Conference. Zurich, Switzerland: [s. n.], 1995: 487-499.

[6] Ashok S, Edward O, Shamkant N. An efficient algorithm for mining association rules in large databases[C]//Proc of the 21st VLDB Conference. Zurich, Switzerland: [s. n.], 1995: 432-443.

[7] 徐章艳，刘美玲，张师超，等. Apriori 算法的三种优化方法[J]. 计算机工程与应用，2004，36(2)：190-202.

[8] 钱光超，贾瑞玉，张 然，等. Apriori 算法的一种优化方法[J]. 计算机工程，2008，34(23)：196-198.

[9] 董 杰. 基于位表的关联规则挖掘及关联分类研究[D]. 大连：大连理工大学，2009.

[10] 钱少华，蔡 勇，钱雪忠. 基于数组的 Apriori 算法的改进[J]. 计算机应用与软件，2006，23(3)：111-113.

[11] 郭云峰，张集祥. 对关联规则挖掘中 Apriori 算法的一种改进[J]. 杭州电子科技大学学报，2009，29(2)：60-63.

[12] 袁万莲，郑 诚，翟明清. 一种改进的 Apriori 算法[J]. 计算机技术与发展，2008，18(5)：51-53.