基于校园网的用户行为数据分析系统的设计

严 楠,刘 涛
(安徽工程科技学院 计算机科学与工程系,安徽 芜湖 241000)

摘 要:数据分析系统是Web日志挖掘系统的一个重要组成部分,是模式分析的前序步骤,主要包括数据预处理和模式挖掘两个过程。数据预处理包括数据净化、用户会话识别和路径补充;模式挖掘包括事务识别、关联规则分析、序列模式分析、分类分析和聚类分析。在研究传统的分析方法的基础上,结合了一种改进的路径补充算法,经验证表明分析效率显著提高。

关键词:数据预处理;用户会话;事务;关联规则;模式挖掘
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2007)01-0239-03

Design of Data Analyzing System of Visitors' Behavior Patterns Based on Web of Campus

YAN Nan, LIU Tao
(Dept. of Computer Sci. and Eng., Anhui University of Eng. Technology and Science, Wuhu 241000, China)

Abstract: The system of data analyzing is a vital part of the system of Web usage mining, and data analyzing is the pre-step of the process of patterns analyzing. The data analyzing system includes two processes: data preprocess and pattern mining. The former includes following processes: data cleaning, user recognition, session recognition and path supplementation; and the latter includes: transaction recognition, association rules analyzing, sequential pattern recognition, classification analyzing and clustering analyzing. In this system, except of citing traditional methods of analysis, chose the method of finding the common ancestor which is the nearest in the process of path supplementation. Result of the application make known this algorithm is quite good.

Key words: data preprocess; user session; transaction; association rules; pattern mining

0 引 言

Web日志挖掘是将数据挖掘技术应用于Web服务器日志,通过分析日志文件发现用户访问站点的浏览模式。Web日志的挖掘在国外发展比较成熟,在国内的研究大约在20世纪90年代末开始起步,研究成果也较丰富,如比较优秀的算法有MFP算法、关联规则分析算法等,但这些算法都存在挖掘过程不可回溯的问题。

由于Web页面的多样性,数据分析处理技术就成为Web日志挖掘中的关键问题。文中在Web日志挖掘的过程中提出一种新的路径补充算法,消除Web站点中的页面对挖掘频繁访问页组的影响。最后对算法进行了验证和说明。Web日志挖掘主要包括:数据预处理、模式挖掘和模式分析可视化三个过程,如图1[1]所示。

Web日志文件是整个挖掘过程的主要源文件。需要指出的是Web日志文件本身并不是格式化的数据,在服务器端的Web日志文件一般是纯文本文件[2]。这里的Web日志是经过格式化的记录文件。Web日志文件的格式可以用表1表示。

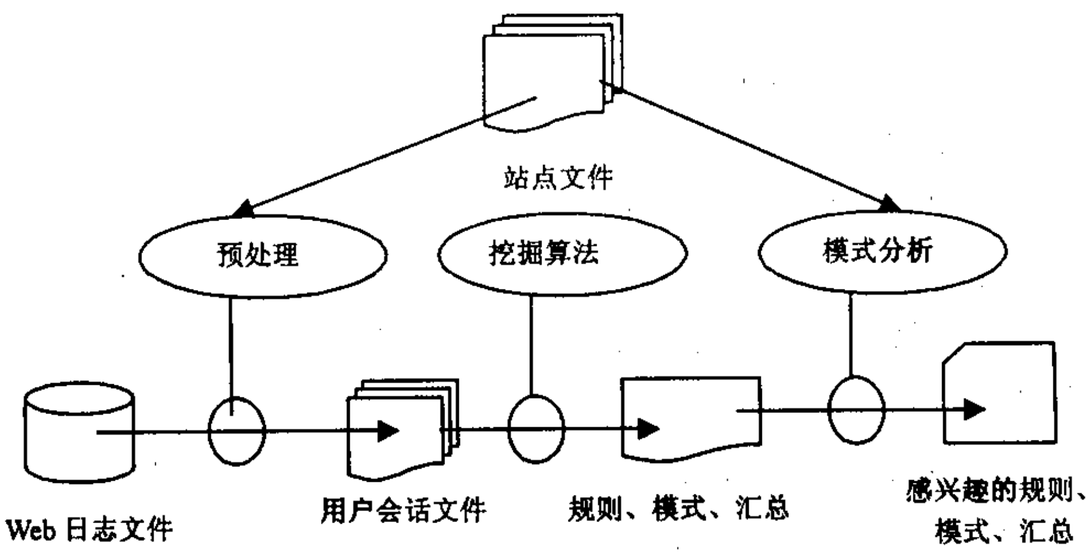

图1 Web日志挖掘的过程

表1 Web日志文件格式

字段名	含义	数据格式
User_IP	用户的IP地址	String
DateStamp	用户访问日期	Date
TimeStamp	用户访问时间	Time
Method	方法	Const int(常量定义)
Url	被请求文件URL	URL
HTTPVersion	用户HTTP版本	String
ReturnCode	服务器状态	int
Bytestransferred	传输字节数	int
User_Rerfer	参考页面URL	URL
Browserused	用户浏览器	String
ClientOperatingSystem	客户操作系统	String

这里的站点文件记录的是站点页面之间的链接关系。

收稿日期:2006-03-23
基金项目:安徽省高等学校省级自然科学研究项目(2005KJ065)
作者简介:严 楠(1979-),男,江苏南通人,助教,研究方向为数据挖掘及信息集成。

站点文件(WebSiteTable)的文件格式可以表示为DAG图:

$G = (V, E)$

其中,V 是站点所有页面的集合,E 代表超链接的边集。

1 改进的数据分析与处理算法

Web日志挖掘中的数据预处理工作关系到挖掘的质量。它包括以下几个过程:数据净化、用户识别、会话识别和路径补充[1]。

整个数据预处理的过程如图2所示。

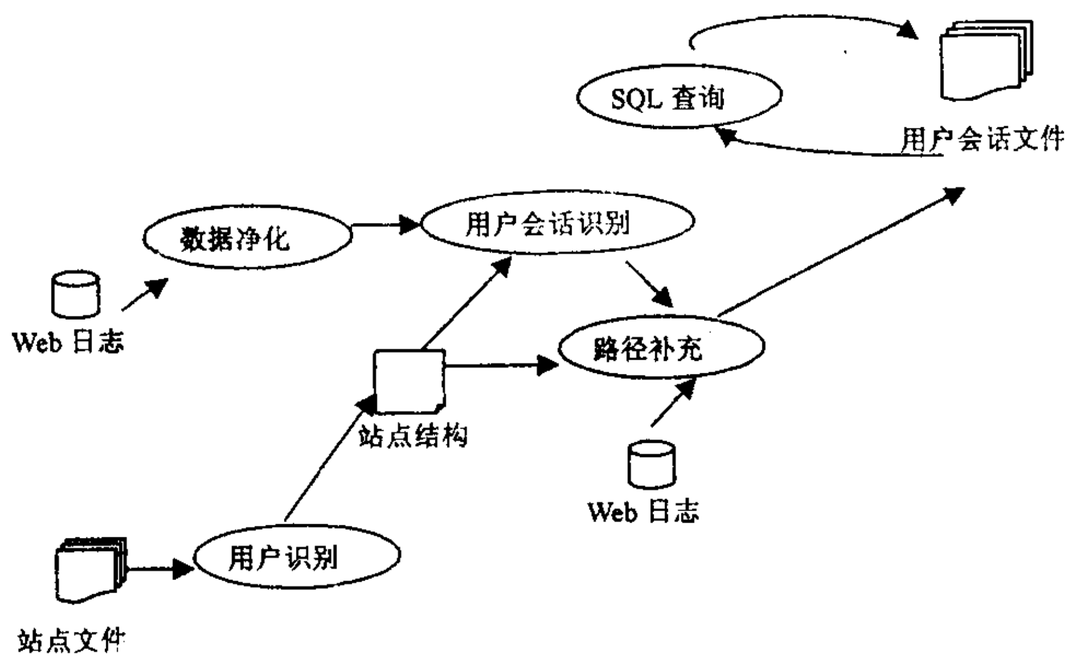

图2 数据预处理过程

1.1 数据净化

数据净化就是删除Web日志中与挖掘无关的记录,很多情况下,只有日志中的后缀为 .htm 的文件与用户的会话有关。数据净化的工作也就成为通过检查URL删除认为不相关的数据。例如:将文件后缀名为 gif,jpeg,map,cgi 的记录项删除。

列表可以根据正在分析的站点类型进行修改,例如:对于一个主要包含图形文档的站点、日志中的 gif 和 jepg 文件可能代表了用户的显示请求,此时就不能将图形文件删除。但很多时候也会忽略这点。

1.2 用户识别

用户识别的任务是识别出有哪些用户访问了网站,以及每个用户访问的页面序列。通常情况下,由于代理服务器的和防火墙的存在使用户识别变得复杂[3]。对于IP地址相同的访问者,可以用下面的方法区分用户。

1)如果IP地址相同,而Web日志文件记录的用户的浏览器或操作系统改变了(也就是用户代理改变了),就认为是不同的用户访问的。

2)将Web日志和站点的拓扑结构结合,构造用户的浏览路径。如果当前请求的页面同用户已经浏览的页面之间没有超链接关系,就认为存在另外具有相同IP地址的用户。这和后面的路径补充不同,路径补充时当前请求的页面同浏览过的页面存在超链关系。

1.3 会话识别

在跨越时间区较大的Web日志中,用户可能多次访问了一个站点。一个会话就是用户在一次访问过程中所访问的Web页面序列。

会话识别就是将用户访问记录分为若干个独立的会话序列[4]。这种以会话为基本单元有利于模式的挖掘和分析。

会话可以定义成一个三元组:

$S = <\text{User\_ID}, \text{User\_IP}, \text{time}_s, \text{time}_e, \{(\text{url}_1, \text{time}_1), \cdots, (\text{url}_m, \text{time}_m)\}>$

其中 User_ID 是用户标识,User_IP 是用户 IP 地址,time_s 是会话开始时间,time_e 是会话结束时间,$\text{url}_1 \cdots \text{url}_m$ 是用户的访问序列,其中 time_i 是访问 url_i 的时间。

怎样认为用户是多次而不是一次访问一个站点呢?可以认为用户访问一个页面后过了一个时间阈值如30min后又访问了一个页面,从用户访问后一个页面起,用户开始了一个新的会话。

1.4 路径补充

并不是所有用户真正访问页面的信息都会在Web日志文件中被记录。如果在Web日志中记录的当前请求页与用户上一次请求页之间没有超文本链接,那么用户很可能使用了浏览器上的"BACK"按扭调用缓存在本机中的页面,这时在Web日志文件中并不记录这些在缓存中被调用的页面;当然还有一种可能是用户直接输入了当前请求页的URL访问,一般可忽略这种情况。

在识别用户会话中的另一个重要问题是确定访问日志中是否有重要的请求没有被记录。这就是路径补充要做的工作。

假如Web站点页面链接拓扑结构如图3所示。

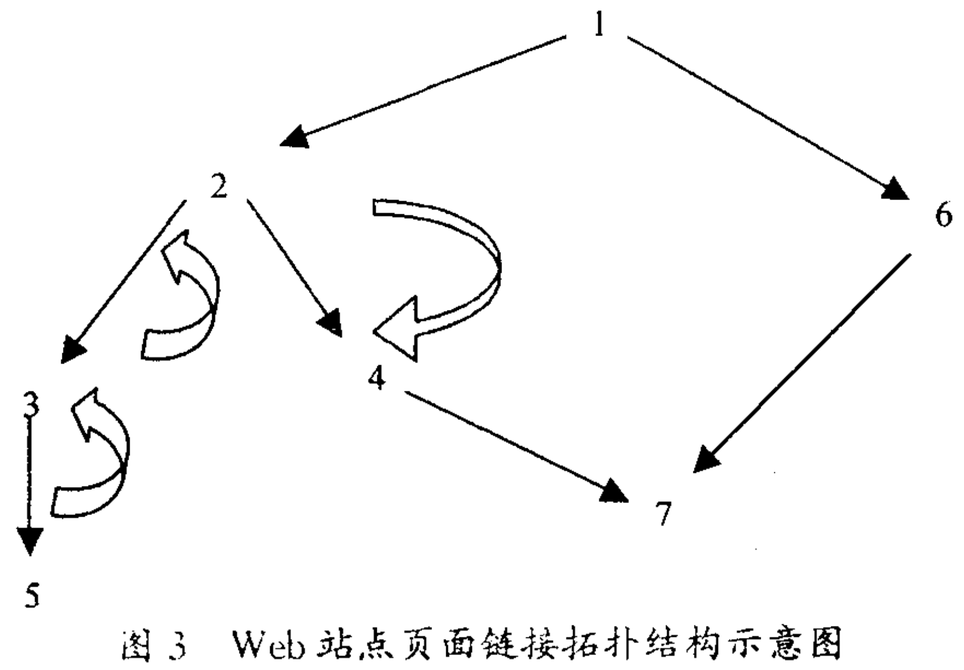

图3 Web站点页面链接拓扑结构示意图

假如用户识别后的结果为一个用户的访问序列为:1-2-3-5-4-7。会发现用户访问是从5直接到4,而5和4之间并没有链接关系,这时可能是用户直接输入4的URL进行访问的,这里忽略这种情况。可认为是用户访问5时,按了浏览器上的BACK按钮,退回到2再访问4的。这时就认为用户实际的访问路径为1-2-3-5-3-2-4-7,也就是路径补充要做的工作。

若想把从页面5到页面4的路径补充完整,可以先找到5和4的最近共同祖先,然后补充从5到祖先和祖先到4的路径。如图3中5和4的最近共同祖先是2,路径被补充为1-2-3-5-3-2-4-7。

路径补充后的文件可以和用户会话文件格式一样,但

是注意,这时对被补充的页面的访问时间等信息并不会被记载,而且这个数据从 Web 日志文件上是无法得知的。

2　算法实现

用户会话对数据挖掘来讲显得粗糙,不够精确。需要把会话进一步分解成具有一定语义的事务。事务是用户真正访问一个页面的过程,用户要访问一个内容页,他最有可能是通过一系列的导航页到达内容页,也可能是他直接就访问内容页。内容页一般是用户关心的信息;导航页可以看作为使用户找到内容页的链接页面。

最大向前路径(MFP,maximal forward path)是在用户会话中的第一页到回退的前一页组成的路径[5]。对于每个用户会话,从开始页面为起点,每个最大前向引用路径即为一个事务。例如,用户会话为 1－2－1－3－4－3(见图 3),则以 1 为起点,最大前向引用路径为 1－2;1－3－4。因为在路径补充后,这是用户真正要访问内容页(2 和 4)的访问路径。

做出如下分析:

1)在用户刚开始访问时,用户的访问路径总是从导航页走向内容页,设立一个标志 flag,并用 flag 为 true 标识访问路径的这种走向。并用 flag 为 false 表示从内容页到导航页的反走向。

2)在用户会话的下一个页面中突然出现与前面访问的序列有相同页面的时候,如上面用户会话序列中第二个 1 页面出现与第一个 1 页面相同,则会有下面的情况存在:

a. 当 flag 为 true 即用户在此之前一直在走从导航页到内容页的路,这时从导航页到内容页的访问路径结束,前面走过的路径被划分成一个事务。用户可能会退回到导航页去访问另一个内容页,以开始新的访问事务。

b. 当 flag 为 false 即用户在此之前一直在走从内容页到导航页的路,这时从内容页到导航页访问路径结束,用户即将开始新的访问事务。用户在此之前一直在走从内容页到导航页的路径没有意义,应当被忽略。在算法中:

输入:会话识别文件、Web 日志文件

加工:路径补充

输出:用户会话文件

基于此,给出基于 Java 语言实现的算法:

```
public int pathAdd() {
  UserSessionRecordClass prior, current; //声明同一会话中前后访问的两条记录
  for (int i=1;i< UserSessionTable.RecordNumber;i++){
  UserSessionRecordClass first = ust. selectFromUserSessionTable(1);
  //first 为第一条记录
  UserSessionRecordClass s= ust.selectFromUserSessionTable(i);
  //s 为当前记录
  prior= first;
  while(s.Session_ ID= = first.Session_ ID){//当记录序列为同一个会话时
    current= s;
    if(Array[prior.URL_ ID][current.URL_ ID]= =1){
    //若前后两条记录存在超链关系
    prior= current;//继续下去,不需补充路径
    } else { //若前后无超链关系
    ust.insert(prior.Session_ ID,
      wst.getRecords(prior.URL_ ID,
      wst.getNearParent(prior.URL_ ID,
      current.URL_ ID)));
    }
  }
  }
  return 1;
}
```

3　实验结果分析

数据分析与处理技术的改进算法已经在安徽工程科技学院的 Web 服务器日志上进行了验证。试验的环境是在 Xeon 2.0GHz 处理器和 512MB 内存上进行的,运行的平台为 Solaris X86。实验数据是长度为 15MB 的日志,记录了从 2/Dec/2005－08:12:24 到 6/Dec/2005－08:30:20 时间段的页面请求,其中包含 13 万条记录。日志数据中有 302 个不同的 HTML 页面,从中识别出 1 642 个用户会话。通过补充路径算法比较一般数据分析与处理技术技术与改进技术。图 4 列出了在一般算法中时间阈值为 60min,改进算法中时间阈值为 30min 的情况下实验结果数据。从实验结果不难看出在更短的时间阈值内改进的算法能够提升站点的结构,较有效地发掘用户的行为。

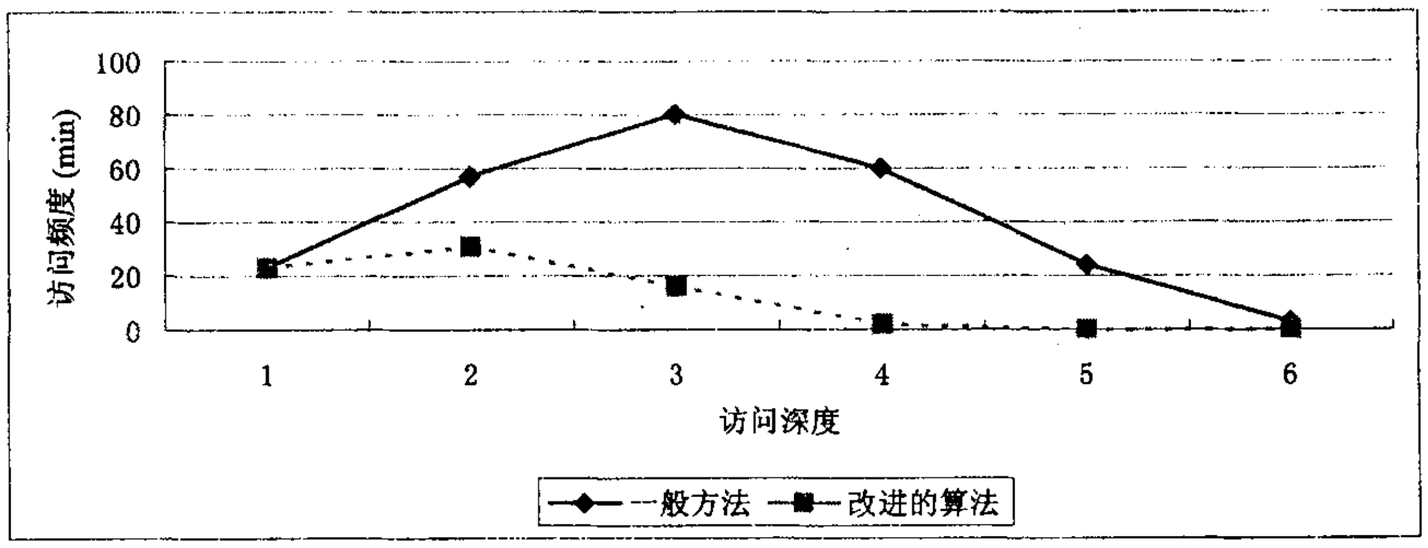

图 4　一般方法和改进算法的实验结果的比较

4　结束语

基于校园网的用户行为模式发现和分析是 Web 日志挖掘的一个应用,通过对用户的访问行为模式的分析,将有利于改善网站网页链接结构,开展有针对性的电子商务活动。文中在现有数据分析与处理技术的基础上提出改进了的路径补充算法,删除用户会话中出现的无意义的页

(下转第 244 页)

时,系统即会出现连接被拒绝的错误;

(2)高级组合查询以 20 并发用户开始并发,当并发量增加至 300 时,系统响应不正常,时而能全部并发,时而不能全部并发,并会出现连接被拒绝的错误。

结论:如果访问量很大,系统并不能很好地持续工作,系统的功能还有待提高。

4 结束语

负载测试是为了测量 Web 系统在某一负载级别上的性能,以保证 Web 系统在需求范围内能正常工作,不过,基于 Web 系统的测试、确认和验收是一项重要而富有挑战性的工作,由于 Web 页数目多且变动频繁,对其测试如果单靠手工是无法进行的,必须有测试工具的参与,LoadRunner 可以说是一个很好的自动化测试工具,在实际的测试环境中,可以通过 LoadRunner 测试的结果,分析出系统的性能瓶颈,从而提高系统的性能,但是对于测试得到的数据如何分析,并找出系统中的瓶颈,还是需要一定的经验积累。

参考文献:

[1] Patton R. 软件测试[M]. 北京:机械工业出版社,2002.

[2] Ash L. Web 测试指南[M]. 李 昂译. 北京:机械工业出版社,2004.

[3] Mosley D J, Posey B A. 软件测试自动化[M]. 邓 波,黄丽娟,曹青春等译. 北京:机械工业出版社,2003.

[4] 二炮. LoadRunner 自动化测试工具的应用[EB/OL]. 2004-05. http://www.sztest.net/.

[5] Sunshinelius. 让 LoadRunner 走下神坛[EB/OL]. 2005-10. http://www.51testing.com.cn/.

(上接第 241 页)

面,消除了无意义的页面对挖掘结果冗余的影响。经过实验数据的检验,表明改进后的数据分析与处理技术使得挖掘结果更有利于发现用户的行为模式。

参考文献:

[1] 陆丽娜,杨怡玲,管旭东,等. Web 日志挖掘中的数据预处理的研究[J]. 计算机工程,2000,26:66-72.

[2] Yang Qiang, Wang Ke. Web-Log Cleanung for Constructing Sequential Classifiers[J]. Applied Artificial Intelligence, 2003,17(5):431-441.

[3] 陈云芳,王汝传,柯行斌. 基于用户行为分析的入侵检测应用模型的研究[J]. 微机发展,2004,14(2):124-127.

[4] 张智颖,梁 伟. Web 使用挖掘中的数据预处理算法研究[J]. 微型机与应用,2004,21(8):11-15.

[5] Spiliopoulou M. Web Usage Mining for Web Site Evaluation [J]. Commuications of ACM,2000(8):94-123.

2007 全国开放式分布与并行计算学术年会征文通知

由中国计算机学会开放系统专业委员会主办、广西大学计算机与电子信息学院承办的“2007 全国开放式分布与并行计算学术年会(DPCS2007)”将于 2007 年 10 月 12-15 日在广西南宁市广西大学召开。本次年会录用的论文将由《小型微型计算机系统》和《微电子学与计算机》以正刊方式发表,欢迎大家积极投稿。

征文范围:

(1) 开放式分布与并行计算模型、体系结构、算法及应用;

(2) 开放式网络、数据通信、网络与信息安全、业务管理技术;

(3) 开放式海量数据存储与 Internet 索引技术,分布与并行数据库及数据/Web 挖掘技术;

(4) 开放式机群计算、网格计算、Web 服务、P2P 网络及中间件技术;

(5) 开放式移动计算、移动代理、传感器网络与自组网技术;

(6) 分布式人工智能、多代理与决策支持技术;

(7) 分布、并行编程环境和工具;

(8) 分布与并行计算算法及其在科学与工程中的应用;

(9) 开放式虚拟现实技术与分布式仿真;

(10) 开放式多媒体技术与流媒体服务,包括媒体压缩、内容分送、缓存代理、服务发现与管理技术。

论文必须是未正式发表的、或者未正式等待刊发的研究成果。来稿一律不退,请自留底稿。会议将评选优秀论文,予以奖励并推荐到一级学报发表。征文投稿截止日期:2007 年 6 月 15 日;论文录用通知日期:2007 年 7 月 10 日。论文投稿需提交激光打印稿一式 2 份和电子版 WORD 文件。

论文投寄地址:广西南宁市大学东路 100 号 广西大学计算机与电子信息学院 钟诚 李陶深收 邮编:530004

Email:dpcs2007@sina.com 电话:钟诚 0771-3236396,13607819333 李陶深 0771-3236627,13768301390

专委会联系人:南京大学计算机系 陈贵海 电话:025-58916715 电子邮件:gchen@nju.edu.cn