Web 日志挖掘中路径补充的影响评估①

蔡卫欣[1]，冯振宇[1]，杨 剑[2]

[1](中国民航大学 航空工程学院，天津 300300)
[2](石家庄铁道大学 四方学院，石家庄 050000)

摘 要：Web 用户访问多是匿名访问，Web 日志挖掘的主要目标是从 Web 访问记录中抽取用户行为模式，通过分析挖掘结果理解用户的行为，从而改进站点的结构。Web 日志挖掘第一步是进行数据预处理。数据预处理是 Web 页面分析中最耗时的阶段，首先研究了数据预处理的过程，包括数据清洗、用户识别、会话识别、路径补充。提出了一种路径补充的算法，并通过实验验证了路径补充对规则抽取的数量和规则抽取的质量均有显著影响的假设，评估了 Web 日志挖掘中路径补充的作用，这也为 Web 日志挖掘中数据预处理应进行到何种程度提供了实验基础。

关键词：Web 日志挖掘；数据预处理；路径补充；评估

Evaluation of Impact of Path Completion in Weblog Mining

CAI Wei-Xin[1], FENG Zhen-Yu[1], YANG Jian[2]

[1](College of Aeronautical Engineering, CAUC, Tianjin 300300, China)
[2](Shijiazhuang Railway Institute, Shijiazhuang 050000, China)

Abstract: Web user access is almost anonymous access. The main goal of weblog mining is to extract users' behavior patterns from the Weblogs, and then understand users' behavior by analyzing the mining results to improve the structure of the site. The first step of weblog mining is data preprocessing. Data preprocessing is the most time consuming stage in web page analysis. This paper first studies the process of data preprocessing, including data cleaning, user identification, session identification, path completion. A path completion algorithm is proposed. The paper poses the hypothesis that the path completion has a significant impact on rule extraction quantity and quality, and then experimental verification is conducted to assess the effect of path completion in weblog mining. The experiment result also provides an experimental basis to what extent data preparation should be carried out.

Keywords: Weblog mining; data preprocessing; path completion; evaluation

1 引言

进行数据分析的前提是有充足的高质量数据。如果输入有误，不管应用多么完善的算法，输出也是错误的。Web 日志挖掘更是这样，所以在对日志文件进行挖掘前，需要彻底进行数据预处理。

数据预处理是 Web 页面分析中最耗时的阶段，本文的目的是通过实验找出在日志挖掘中数据预处理应进行到什么程度，更精确地讲，本文旨在评估路径补充即重现 Web 访问者活动的影响，这些活动代表 Web 用户的行为模式。

2 数据预处理中的相关问题

数据预处理包含数据清洗、用户识别、会话识别、路径补充等过程。

数据清洗是指删除 Web 服务器日志中与挖掘算法无关的数据项。由于 Web 使用挖掘主要是对用户浏览行为的研究，所以只有使用准确描述用户浏览行为的数据进行挖掘，才能发现正确的规则和模式[1]。

① 收稿时间:2010-07-12;收到修改稿时间:2010-09-05

Web 服务器日志文件是匿名用户的访问记录，匿名也给用户身份的唯一识别带来了困难，所以需要重现每个用户的活动。客户端数据收集机制有助于解决用户识别问题，但是涉及隐私无法展开，加上缓存、代理服务器(包括网吧、局域网等环境)和网络防火墙的存在，会使识别用户这一步变得比较复杂[2]。

会话是指用户对服务器的一次有效访问，通过其连续请求的页面，可以获得用户在网站中的访问行为和浏览兴趣。日志文件中不同用户访问的页面当然属于不同的会话。当某个用户的页面请求在时间上跨度比较大时，就有可能是该用户多次访问同一个网站，我们可以将用户的访问记录分成多个会话来处理[3]。

由于缓存等原因使得访问日志中并没有完全记录用户的访问行为，路径补充就是要将用户会话中的访问路径补全，从而更好地反映用户的访问过程。

3 实验过程

3.1 研究方法

本实验通过以下步骤完成：

(1) 数据获取-为了获得需要的数据(IP 地址，访问数据和时间，URL 地址)，定义日志文件中待观察的变量

(2) 建立数据矩阵-从日志文件(访问信息)和站点拓扑结构(Web 内容信息)

(3) 各个层次的数据预处理

(4) 数据分析-在各个日志文件中寻找用户行为模式

(5) 理解输出数据-定义假设，从分析结果建立数据模型

(6) 比较数据分析结果

3.2 研究过程

Taucher 和 Greenberg 证明超过 50%的页面访问是通过 Backward 按钮进行的[4]。这是浏览器缓存的问题，在 Backward 时，不会访问 Web 服务器，服务器上当然没有任何该页面的日志文件，路径补充重点是把用户通过 Back 按钮访问的路径记录补全。本文提出了一种路径补充的算法—补充父节点法。算法思想：通过观察用户浏览网页过程发现，无论用户如何访问页面，假设节点 x、y 之间需要补充，则在路径补充的时候都应先找末节点 y 的父节点，然后用前一节点 x 的父节点与其匹配，如果相等，则将父节点补充进去。否则再寻找 x 的父节点的父节点，以此类推到最终相等为止，然后将所有用过的父节点全部补充进去。

补充父节点法的具体算法如下所示:

```
输入: P[n] // P[n]长度为 n 的访问路径集合
输出:CP[m]  //CP[m]表示路径补充完后长度为 m 的访问路径集合
CP[m]=P[0];
For(i=1;i<n;i++) //对 P[n]中每一个访问节点，循环补充路径
 {
     Int j=0;
Node[j]=SearchFN(P[i]);    //调用寻找父节点函数，找出 P[i]的父节点赋给 Node
     If(CP[m]==Node[j])
CP[--m]= P[i]; //如果 Node 跟集合 CP[]中最后一个节点相同，将 P[i]添加到 CP[]中
     Else
     {
while(Node1[k]!=Node[j])//循环寻找父节点直到匹配为止
     {
Node1[k]=SearchFN(CP[m])   //如果不同，则从集合 CP[]中的最后一个节点开始循环寻找父节点与 Node[j]比较
            While (Node1[k]!=Node[j])
            {
              Temp=SearchFN(node[k++]);
              Node[k]= Temp;
             }
            Temp =SearchFN(Node[j++]);
            Node[j]=Temp;
          }
          For(s=0;s<=k;s++)
               CP[++m]=Node1[s];
          For (n=j ; n-- ;n>=0)
               CP[++m]=Node[n];
     }
}
```

本实验中，将数据清洗、用户识别、会话识别之后的日志文件称为 File1。用户识别的一种方法是基于使用的 Web 浏览器，为了突出路径补充的影响，将这

种方法用于会话识别，对会话识别的结果 File1 进行更改后，得到的文件称为 File2。对 File2 应用补充父节点法进行路径补充后，得到的文件称为 File3。

我们做了如下假设：路径补充对规则抽取的数量有显著影响；路径补充对规则抽取的质量有显著影响。

4 实验结果

本实验针对某高校网站为期十天的使用日志文件进行处理分析。

4.1 路径补充对规则抽取数量的影响

表 1 数据可见：路径补充后，访问记录数增加了约 70%，序列平均长度从 4 增加到 6。而在会话识别中，基于不同浏览器识别会话后，用户访问序列数只增长了 6%。

表 1　各文件的访问数量和序列根据

	Web 访问计数	用户访问序列计数	频繁序列计数	序列平均长度
File1	46256	11024	83	4
File2	46256	11705	76	4
File3	78018	11705	95	6

从该网站十天的日志文件中挖掘出频繁访问序列，在满足最小支持度(取 s=0.03)的频繁序列中，通过分析产生序列规则，见表 2。

表 2　在特定文件中发现序列规则

Body	⇒	Head	File 1	File 2	File 3
（A10）	⇒	(A50)	1	1	1
（A10）	⇒	(A68)	1	1	1
（A10）	⇒	(A71)	1	1	1
（A10）	⇒	(A71)，（A10）	0	0	1
…	⇒	…	…	…	…
（B4）	⇒	（A25）	1	0	0
（B4）	⇒	（A29）	1	1	0
（B4）	⇒	（A50）	1	1	1
（B4）	⇒	（A50），（A29）	1	1	0
（B4），（A50）	⇒	（A29）	1	1	0
…	⇒	…	…	…	…
（C5）	⇒	（A10）	1	1	0
（C5）	⇒	（A15）	1	1	1
（C5）	⇒	（B4）	0	0	1
（C5）	⇒	（C6）	1	0	1
衍生序列规则计数			45	40	84
衍生序列规则百分比			47.4	42.1	88.4
Cochran Q 检验			Q=55.26984，df=2，p<0.000000		
Kendall 和谐系数			0.29089		

表 2 中，没有进行路径补充的文件 File1 和 File2，序列规则的分析结果具有高度的一致性。大多数规则都是从路径补充的文件中抽取的，确切地 File3 共抽取了 84 个规则，占所有发现规则总数的 88%多。

根据表 2 中 Cochran Q 检验的结果，零假设在 1%置信度水平下被拒绝，零假设认为发现规则的概率不取决于 Web 日志挖掘中各个层次的数据预处理。

Kendall 和谐系数代表各个待检验文件中发现规则数目的一致性程度。表 2 中的和谐系数是 0.29，和谐系数为 1 代表完全一致，0 代表不一致。

通过多重比较，根据发现规则的平均概率，File1 和 File2 可以归为一类。统计证明在 0.05 的显著性水

平下，File3 和其他文件发现规则的平均概率有明显区别。

表 3 待检验文件分组

File	均值	1	2
File 2	0.421	****	
File 1	0.474	****	
File 3	0.884		****

从以上三个表的数据分析可以初步得出结论：路径补充对规则抽取数量的影响很大(File3 对比 File1，File2)，而会话识别时基于使用的浏览器不同对规则抽取数量没有显著影响(File1 对比 File2)。

4.2 路径补充对规则抽取质量的影响

序列规则的质量从两方面评估：

① 支持度

② 置信度

序列规则分析的结果表明路径补充后，不仅发现规则的数量不同，质量也有区别。

Kendall 和谐系数代表各个待检验文件之间在发现规则的支持度方面的一致程度。表 4 中和谐系数值是 0.36。

经过多重比较，根据发现规则的平均支持度，所有的文件(File1，File2，File3)都划分到同一个组内，统计没有发现各个文件在支持度方面存在重大差别。

表 4 衍生规则支持度的同类分组

支持度	均值	1
File 2	3.032	****
File 1	3.161	****
File 3	3.207	****
Kendall 和谐系数		0.3633

表 5 衍生规则置信度的同类分组

置信度	均值	1	2
File 2	37.545	****	
File 1	37.959	****	
File 3	41.831		****
Kendall 和谐系数		0.1650	

但是从置信度值方面，各个文件发现规则的质量不同。表 5 中和谐系数几乎是 0.17。通过多重比较，根据发现规则的置信度，File1 和 File2 可以归为一类。统计证明在 0.05 的显著性水平下，File3 和其他文件发现规则的平均置信度有明显区别。

从以上两个表的数据分析可以初步得出结论：路径补充对规则抽取质量的影响非常大(File3 对比 File1，File2)，而会话识别时基于使用的浏览器不同对规则抽取质量没有什么影响(File1 对比 File2)。

总之，本实验的结论验证了路径补充对规则抽取的数量和质量均有显著影响。本文提出的路径补充算法是切实有效的，各表中的数字充分说明了这一点。由于该高校网站具有的随机性和代表性，本文的结论具有一定的适用性。但是，由于实验条件所限，未能在更多的日志服务器中进行验证。

5 结论

本文首先研究了数据预处理的过程，提出了一种路径补充的算法，并通过实验验证了路径补充对规则抽取的数量和质量均有显著影响的假设。这也为 Web 日志挖掘中数据预处理应进行到何种程度提供了实验基础。但是，Web 日志挖掘领域浏览路径补充仍然是一个有待进一步探讨的问题，本文的路径补充算法也有待进一步完善，以期能在实际 Web 挖掘中得到应用和检验。由于实验环境和实验数据的单一和局限，本文的实验结果也请专家给出宝贵意见。

参考文献

1 闫永权.基于频繁访问模式树的 Web 使用挖掘研究[硕士学位论文].长沙:湖南大学,2006.

2 赵伟,何王廉,陈霞,谢振亮.Web 日志挖掘中的数据预处理技术研究.计算机应用,2003,23(5):62－67.

3 缪勇.匿名用户浏览路径挖掘研究与实现[硕士学位论文].南京:南京理工大学,2006.

4 Taucher L, Greenberg S.Revisitation patterns in world wide web navigation. Proc. of Int. Conf. CHI'97, 1997 Atlanta.