文章编号：1000－5471(2008)06－0079－06

Web 使用挖掘中数据预处理的研究①

田倩飞， 左源瑞， 廖 鹏

西南大学 计算机与信息科学学院，重庆 400715

摘要：对 Web 使用挖掘中数据预处理阶段所采用的技术做了全面的研究，主要包括数据的来源及清理、用户识别、会话识别、路径补充等步骤及其所采用的技术. 在现有 Web 使用挖掘数据预处理的步骤上提出改进，即在路径补充后，利用最大向前序列法(MFR，Max Forward Reference)进行二次会话识别，克服了原有会话识别中最大向前序列法的缺点. 最后给出示例及其结果.

关 键 词：Web 使用挖掘；Web 服务器日志；数据预处理

中图分类号：TP393 **文献标识码**：A

Web 挖掘是运用数据挖掘技术从大量的 Web 数据中自动的发现和抽取信息的过程. Web 挖掘分为 3 大类，即 Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘[1]. Web 使用挖掘主要是通过挖掘网站的日志记录来发现用户访问 Web 页面的模式. 对用户访问模式进行分析，可得到用户的访问习惯、用户的兴趣所在，进而对网站的链接进行相应的修正，为用户提供个性化服务，向用户推荐 Web 页面，对不同的用户实行不同的促销策略等. 在 Web 使用挖掘中，数据预处理是保证其挖掘质量的关键. 数据预处理就是把数据源转化为适合进行用户模式挖掘的、准确可靠的规范数据. 数据预处理结果的质量直接影响着 Web 使用挖掘的成败. 因此，对数据预处理的研究目前已成为 Web 使用挖掘研究的焦点. 本文将探讨 Web 使用挖掘中预处理的过程及其所采用的技术.

1 Web 使用挖掘

Web 使用挖掘通常可以分为数据预处理、模式分析及知识发现这 3 个阶段. Web 使用挖掘的过程如图 1 所示.

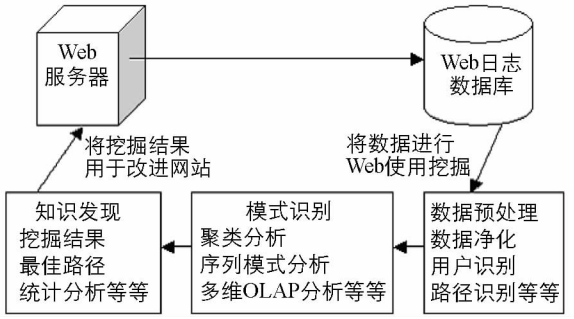

图 1 Web 使用挖掘的过程

2 Web 使用挖掘的数据预处理

2.1 Web 服务器日志

本文将 Web 服务器日志(Web Server Log)作为 Web 使用挖掘的重要数据来源. 最常见的日志格式有 2 种：通用日志格式(Common Log Format)和扩展日志格式(Extended Log Format). 其常用的字段和含义如表 1 所示.

① 收稿日期：2007－10－26

作者简介：田倩飞(1985－)，女，四川成都人，硕士研究生，主要从事数据挖掘、电子商务等研究.

表1 Web服务器日志文件格式[1]

域	描 述
日期	用户请求页面的时间、日期和时区
客户端IP地址	客户端主机的IP地址或DNS入口
用户代理	服务的提供者
请求	URL查询
参照	用户浏览的上页

如下所示是一个实际的日志：

192.168.216.186－－[26/Sep/2007：15：10：10－800]“Get about.html HTTP/ 1.1”200 4219/Index.html IE6.0 WinXP.

该日志表明一个IP地址为192.168.216.186的客户端在2007年9月26日15：10：10这一时刻使用IE6.0的浏览器在index.html页面上发出一个HTTP的Get请求，请求的目的是about.html.

2.2 数据预处理的步骤

Web使用挖掘中的预处理通常包括数据清理、用户识别、会话识别和路径补充等. 其步骤如图2所示.

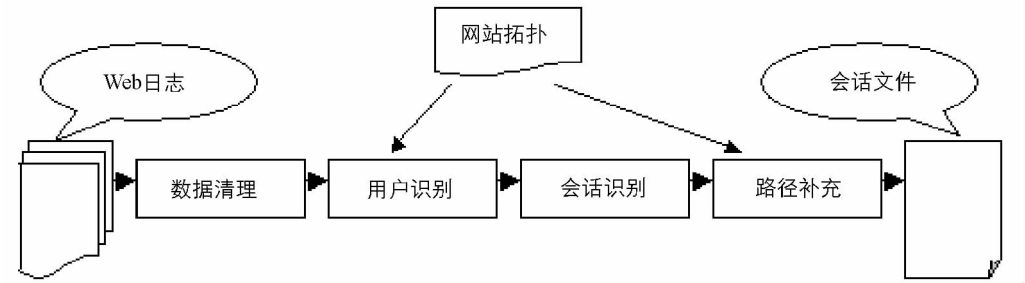

图2 数据预处理的一般步骤

2.3 数据预处理技术分析和步骤改进

将对Web使用挖掘数据预处理步骤中的相关技术进行具体分析，并对原有的预处理步骤提出改进，即在路径补充后，利用最大向前序列法进行二次会话识别. 改进后的步骤如图3所示.

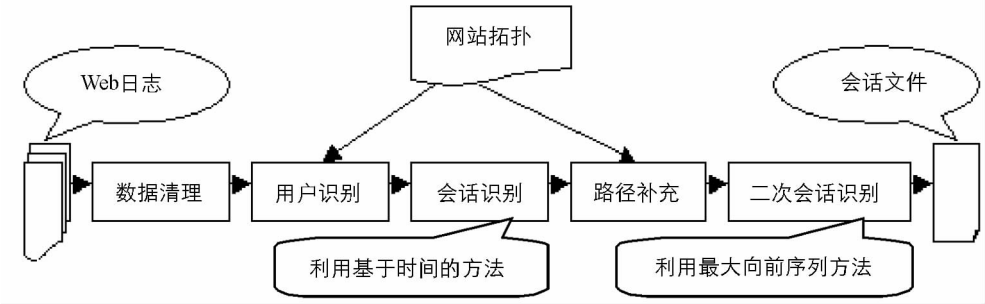

图3 改进后的数据预处理

2.3.1 数据清理

数据清理的任务是去除与挖掘过程不相关的和冗余的日志项. 有3种类型的不相关或冗余的数据需要清理，即错误请求、robots请求和嵌在HTML文件中的附属资源[2].

2.3.1.1 错误请求 错误请求是带有“error”或“failure”状态的项，这些请求对挖掘过程是无用的. 可通过检查请求状态码去除这种请求. 状态码是200时，表示服务器已经成功地响应浏览器的请求，一切正常. 一般地说，以2开头的状态代码表示成功，以3开头的状态代码表示由于各种不同的原因用户请求被重定向到了其他位置，以4开头的状态代码表示客户端存在某种错误，以5开头的状态代码表示服务器遇到了某个错误.

2.3.1.2 Robots请求 Web robots（也称为Spider）是一种扫描Web站点以抽取其内容的软件工具. Spider自动地扫描一个Web页面的所有超链接. 搜索引擎比如Google周期性地使用Spider获取一个Web

站点的所有页面以更新其搜索索引. 可以通过查找请求页面“robots. txt”的所有主机去除 robots 请求.

2.3.1.3　附属资源　HTTP 协议是一种面向非连接的协议，每次客户连接请求完所要的网页后，服务器会自动与客户断开连接，同时被申请的网页文件连同文件上的图片、声音和脚本代码一起被下载到了客户端. 由于 Web 使用挖掘的主要目的是得到用户行为的信息，因而可以通过检查 URL 后缀名的方法来删除非用户请求的文件记录. 比如，所有文件后缀名为 gif,jpeg,GIF,JPEG,jpg,JPG,css 和 map 的日志项都可以清除掉. 常用的脚本比如后缀为“. cgi”的请求文件也可清除掉.

数据清理算法总结如下：

输入：原始日志文件

输出：清理后的日志文件

算法：逐条读取原始日志文件

```
if (status==200                    //请求得到正常响应
&& request! = robots. txt          //非 robots 请求
&& URL 后缀名! =. gif||. jpeg||. GIF||. JPEG||. JPG||. jpg||. css||. map||. cgi)
```

从该日志记录中选取与挖掘任务相关的属性并加入清理后的日志文件

else 过滤该日志记录并转向下一条日志记录

直至最后一条日志记录

2.3.2　用户识别

用户识别是从日志文件中的每一条记录中识别出响应的用户. 由于缓存、代理服务器和防火墙的使用，使得识别用户变得很复杂，可能出现以下 4 种混淆情况[3]：

① 不同的用户可能在同一时间通过代理服务器访问 Web 服务器；

② 同一用户可能在不同的机器上访问 Web 服务器；

③ 同一用户可能在同台机器上使用不同的浏览器和不同的操作系统访问 Web 服务器；

④ 不同的用户使用同一台机器浏览同一站点.

针对上面 4 种情况，采用响应的启发式规则来识别用户，具体规则如下所示：

① 不同的 IP 地址代表着不同的用户；

② 如果 IP 地址相同，但 Agent 信息中如浏览器软件或操作系统不同，则可以假设为不同的两个用户.

③ 如果 IP 地址和 Agent 信息都相同则判断每一个请求访问的页面与访问过的页面之间是否有链接. 如果一个请求访问的页面与已经访问过的所有的页面之间并没有直接的链接，则假设在访问 Web 站点的机器上同时存在着多个用户. 在本规则中，需用网站的拓扑结构图对用户进行识别.

需要说明的是：使用上述规则并不能完全识别出用户，如用户通过直接在地址栏里输入 Url 来访问一个页面，若此页面与他之前已访问的页面之间没有任何链接，就会被认为不同的用户. 为了更准确的识别出用户，可以结合其他方法，如采用 Cookies，客户端 Agent 以及注册使用等方法.

2.3.3　会话识别

会话(Session)是指用户在一次访问网站期间从进入网站到离开网站所进行的一系列活动. 在跨度时间段较大的 Web 服务器日志中，用户可能多次访问了该站点，会话识别的任务就是把属于同一用户的每次访问请求识别出来.

现常用的会话识别方法有 2 种，分别为基于时间的方法和最大向前序列法[4-5]. 基于时间的方法又可以分为基于整个会话时间的方法和基于相邻页面访问时间的方法. 具体如下：

① 基于整个会话时间的方法：给用户在整个站点的停留时间一个上界，如果超过这个域值，则认为新的会话开始. 此域值一般取 30 min.

② 基于相邻页面访问时间的方法：给用户一个页面停留时间域值，如果 2 个连续请求的时间间隔没有

超过这个值，则属于同一会话，否则分属于两个会话. 此域值一般取 10 min.

③ 最大向前序列法：在一个用户会话里不会出现用户先前已经访问过的页面. 如果用户在向前浏览到一个网页时，按下了“返回”按钮或某个链接而到达已经访问过的页面，则表示当前会话结束，一个新的会话开始.

2.3.4　路径补充

路径补充的目的是为了补全访问日志中没有记录的用户请求，获得用户的完整的访问路径，这样才能更准确地发现用户的访问模式. 在前一步骤会话识别中曾提到过，用户在浏览网页时，由于本地缓存和代理服务器缓存的存在，使得用户通过按下浏览器上的“后退”按钮可得到相应页面，而在服务器日志文件中却没有记录. 比如：访问 ABAC，由于缓存的存在，只能记录到 ABC，需要使用路径补充算法来进行补充. 目前大多数路径补充算法都使用网站的拓扑结构来进行补充，方法的优点是准确率比较高，但是随着网站的信息越来越多，网站的拓扑结构会越来越大，这样就会很费时.

因此可利用日志文件中的引用记录和用户的历史访问记录对用户访问路径进行补充[6]. 该方法的思想是：假如 P 和 Q 是同一用户的两个连续的请求，Q. refer≠ P. url (Q. refer≠null)，则页面 Q 不是从页面 P 直接到达，这时在当前用户会话中查找等于 P. refer 和 Q. refer 的记录，假如找到的记录分别为 S. url 和 T. url，如果两者相等，则直接插入；否则，把 S 到 T 的记录插入到 P 和 Q 之间. 如果在当前会话里面找不到，则在当前用户的其他会话里面查找；如果存在多个会话包含记录 S. url 和 T. url，则取时间和 Q 最近的会话进行补充.

2.3.5　二次会话识别

在第一次会话识别中，利用最大向前序列法有一个缺点，即由于客户端缓存的存在，服务器端可能没有用户回退访问的记录，最大向前序列法不能将会话进行完全地识别. 因此，作者决定首先采用基于时间的方法进行会话识别，接着进行下一步路径补充后，用户回退访问等已经补充完整，再利用最大向前序列进行二次会话识别，就可以识别出新的更为精确的会话. 路径补充后的最大向前序列算法如下所示：

输入：路径补充后的访问序列 S1

输出：二次会话识别后的访问序列 S2

算法：逐条读取 S1 中的记录

if 有重复的访问页面

将此序列从重复处截断，得到更细小精确的序列记录入 S2

else 将原序列记录入 S2 并转向下一条序列记录

直至 S1 中最后一条序列记录

3　示　　例

如图 4 为某网站的部分拓扑结构图，表 2 为数据清理后的部分日志记录.

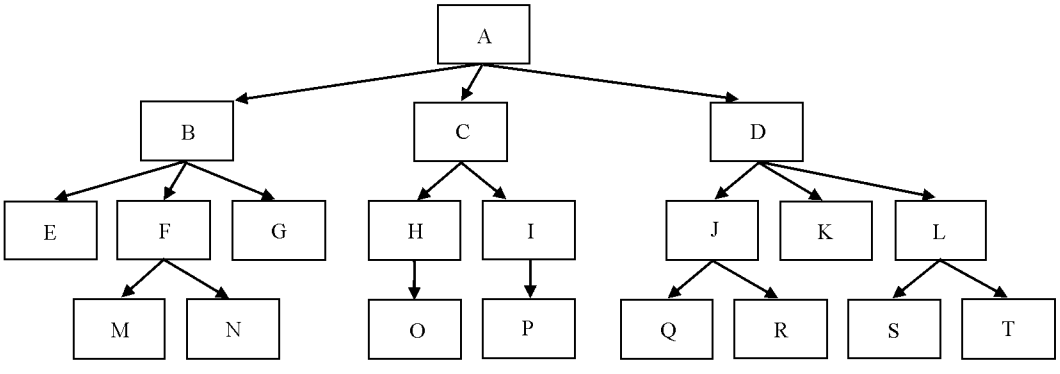

图 4　网站部分拓扑结构图

表 2　部分日志记录

编号	IP 地址	日期	请求	参照	代理
1	192.168.215.166	26/JAN/2008：21：02：12＋0800	A	－	IE6.0/WINXP
2	192.168.215.166	26/JAN/2008：21：02：53＋0800	B	A	IE6.0/WINXP
3	192.168.215.166	26/JAN/2008：21：03：30＋0800	L	－	IE6.0/WINXP
4	192.168.215.166	26/JAN/2008：21：33：37＋0800	F	B	IE6.0/WINXP
5	192.168.215.166	26/JAN/2008：21：34：05＋0800	T	L	IE6.0/WINXP
6	192.168.215.166	26/JAN/2008：21：34：16＋0800	D	－	IE6.0/WIN2000
7	192.168.215.166	26/JAN/2008：21：34：58＋0800	J	D	IE6.0/WIN2000
8	192.168.215.166	26/JAN/2008：21：35：29＋0800	R	J	IE6.0/WIN2000
9	192.168.215.166	26/JAN/2008：21：35：36＋0800	N	F	IE6.0/WINXP
10	192.168.215.166	26/JAN/2008：21：36：21＋0800	C	A	IE6.0/WINXP
11	192.168.215.166	26/JAN/2008：21：37：02＋0800	H	C	IE6.0/WINXP
12	192.168.215.166	26/JAN/2008：21：37：46＋0800	I	C	IE6.0/WINXP

利用数据预处理技术对上述日志进行预处理，表 3 列出了每一步处理后的结果. 由此结果可知，利用二次会话识别能够识别出更精确的会话.

表 3　日志的预处理步骤及结果

步　骤	结　果
数据清理	A－B－L－F－T－D－J－R－N－C－H－I
	A－B－F－N－C－H－I
用户识别	L－T
	D－J－R
	A－B
会话识别	F－N－C－H－I
	L－T
	D－J－R
	A－B
路径补充	F－N－F－B－A－C－H－C－I
	L－T
	D－J－R
	A－B
	F－N
二次会话识别	F－B－A－C－H
	C－I
	L－T
	D－J－R

4　小　　结

研究了 Web 使用挖掘中的预处理步骤及方法，针对会话识别中最大向前序列的缺点，提出在路径补充后对会话进行二次识别. 作者的后续工作是将本文的预处理方法运用到更大的数据集上进行处理，并与其他方法进行比较. 在具体的实施过程中，还存在一些其他问题，如对框架式的网页处理等，已有学者提出 Frame 页面过滤算法[3]. 作者将针对数据预处理中的实际问题继续研究，期待能提出更准确、快速的数据预处理方法，以使 Web 使用挖掘能取得更好的效果.

参考文献:

[1] 夏火松. 数据仓库与数据挖掘技术 [M]. 北京: 科学出版社, 2004: 207－219.

[2] Zidrina Pabarskaite. Implementing Advanced Cleaning and End-user Interpretability Technologies in Web log mining [J]. Journal of Intelligent Information systems, 2007, 2: 79－104.

[3] 张 沫, 陈基漓, 阮百尧. Web日志挖掘中数据预处理技术的研究 [J]. 计算机与数字工程, 2007, 9: 90－92.

[4] Robert W C. Web Usage Mining: Discovery and Application of Interesting patterns [J]. Ph. d. thesis, Graduate School of University of Minnesota, 2000: 12－23.

[5] Mobasher B. A RoadmaP to More Effeetive Web PersonaliZation [J]. Integrating Domain Knowledge with Web Usage Mining. Proceedings of the International Conference on Internet Computing, 2003: 237－249.

[6] 李烈彪, 张海鹏, 周亚峰. Web日志挖掘中数据预处理方法的研究 [J]. 计算机技术与发展, 2007, 7: 45－48.

Research on Data Preprocessing in Web Usage Mining

TIAN Qian-fei, ZUO Yuan-rui, LIAO Peng

School of Computer and Information Science, Southwest University, Chongqing 400715

Abstract: This paper researches the techniques in data preprocessing of Web usage mining, mainly including the steps and their techniques: sources of data, user identification, session identification and path completion. To give new steps of preprocessing in Web usage mining, that is, using max forward reference method (MFR) to do a session reconstruction. This method can overcome the backwardness of MFR in the first session identification. At last the paper gives an example as well as its results.

Key words: web usage mining; web server log; data preprocessing

责任编辑 陈绍兰