Web日志挖掘中的会话识别方法研究

顾兆军,李晓红,王 伟,黄杰培
(中国民航大学 计算机学院,天津 300300)

摘 要:数据预处理是Web日志挖掘的首要环节,而会话识别是数据预处理中的关键步骤之一。为了更好地实现会话识别、提高会话识别的真实度,从而为后续的模式挖掘工作提供精确的挖掘数据,文中在分析了现有常用的会话识别方法后,提出了优化初始会话集的方法。在该方法中,首先初始会话集的产生采用传统的基于访问时间的方法,然后对初始会话集进行合并和断开操作,生成优化的会话集。最后,采用实验实现了该方法。实验结果表明会话质量得到了提高。

关键词:Web日志挖掘;数据预处理;会话识别;会话重组;访问时间
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2012)04-0045-05

Research on Method of Session Identification in Web Log Mining

GU Zhao-jun, LI Xiao-hong, WANG Wei, HUANG Jie-pei
(Department of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Data preprocessing is the first important in the process of Web log mining. At the same time, session identification plays a key role in data preprocessing. To better realize session identification and prepare for sequential work, propose a new method making use of access time and session reconstruction. In this method, the initial session sets are generated based on the access time. Then, the quality of session sets are optimized using a method of session reconstruction, that is, a method of union and rupture. At last, the method studied is carried out, and experimental results illustrate that the quality of session identification is more efficient.

Key words: Web log mining; data preprocessing; session identification; session reconstruction; visit time

0 引 言

Web信息具有复杂、多样性的特点,对Web信息的挖掘属于数据挖掘的范畴。Web数据挖掘就是从Web相关资源和用户对站点的访问信息中,发掘用户可能感兴趣的潜在信息和有意义的浏览模式。按照挖掘对象的不同,Web数据挖掘可分为三类:Web结构挖掘、Web内容挖掘和Web使用挖掘(Web日志挖掘)[1]。Web日志挖掘通常是从用户的Web浏览日志中发现用户群体的相似浏览行为和相似兴趣,以及某个特定用户的浏览习惯和兴趣爱好等,从而为用户提供个性化的信息服务,改进服务器的性能和结构。目前,多数的Web日志挖掘的主要数据来源于Web服务器日志,它顺序的记录了用户进行访问时对服务器的页面请求信息。日志文件在Web服务器上的数据存储格式多种多样,常见的日志数据格式主要有以下三种:CLF(通用日志格式)、ECLF(扩展的通用日志格式)和ExLF(W3C的扩展日志文件格式)[2]。

由于Web日志挖掘的数据是半结构化,甚至是无结构的数据,不同于数据挖掘的对象是结构化的数据库数据,但依然遵循数据挖掘的研究思路。Web日志挖掘过程一般需要4个阶段:源数据收集、数据预处理、模式发现和模式分析及应用[3]。源数据收集阶段:在Web日志挖掘中,原始数据直接来源于Web服务器。当客户请求某站点的Web服务器时,就会在站点服务器上产生相应的用户请求信息,同时,文件的有关信息(如服务器的版本号、文件的创建者、修改时间等)也会被同时记录。数据预处理阶段:通过上一步收集到的数据是杂乱的,必须对其进行处理,将分散的、凌乱的数据整理成包含有用信息的格式化数据。数据预处理阶段得到的格式化数据质量,将直接影响选取的挖掘算法执行的效果和质量,并对最终挖掘效果产生深远影响。模式发现:模式发现阶段就是利用一定的挖掘算法对序列集进行挖掘,产生推荐规则,以发现用户的访问模式。模式分析:一般来说在模式发现阶段使用数据挖掘技术得到的模式并不能让用户直接的使用,模式分析阶段完成的工作就是利用模式发现得来

收稿日期:2011-08-27;**修回日期**:2011-12-02
基金项目:中国民航局科研基金项目(MHRD200808)
作者简介:顾兆军(1966-),男,山东蓬莱人,教授,研究方向为计算机网络与信息安全、搜索引擎、民航信息系统;李晓红(1987-),女,安徽阜阳人,硕士研究生,研究方向为民航信息系统分析、网络与信息安全。

的数据,并结合网站结构,以得到有一定"语义"的浏览模式,呈现给用户。

1 数据预处理

Web日志记录了用户对网站的每一次点击访问,但由于各种原因,Web日志中存在很多不完整或者错误的记录数据,这些数据不但对后续的挖掘无用,还会增加处理的复杂性,产生严重的后果。因此在进行数据挖掘之前必须对数据进行预处理。数据预处理一般包含以下几个步骤:数据清洗、用户识别、用户会话识别和路径补充等[4]。

1)数据清洗:数据清洗的目的就是删除Web服务器日志中对挖掘算法无用的字段和记录信息,也就是根据具体的分析要求将Web日志文件数据中冗余的、不一致的、不相关的数据去除掉,同时检查是不是有错误的数据以及一些没有被日志记录下来的数据。可按照如下方式进行处理:

(1)删除客户请求方法(cs-method)中不是GET的记录;

(2)删除sc- status中显示出错的记录,也就是协议状态为400-599日志记录要清除掉;

(3)当用户请求网站的某一页面时,并且所需要的不是该网页上的图片或者音频文件时,删除日志中有关页面样式、音频、图片、脚本文件等记录信息,即删除日志中文件中后缀是css、jpg、bmp、gif、cgi、js等的记录,以节省存储空间,有效使用挖掘数据;

(4)删除自动程序的请求记录,如网络机器人、蜘蛛和爬虫程序进行网页抓取时所产生的访问日志。

2)用户识别:用户识别是指从日志中的每条记录中识别出不同用户。常采用的启发式规则是:如果用户的IP地址不同,则为不同用户;如果用户的IP地址相同但所使用的浏览器版本或者代理服务器不同,则为不同的用户;如果两者均相同,但用户当前请求的网页和所有已经访问的页面之间不存在直接的超链接关系,也标识为不同的用户[5]。此启发规则只是帮助识别用户,并不是能准确地识别用户。

3)会话识别:对用户进行识别后,可以把该用户在网站上的所有访问序列按时间戳的顺序联系起来,这样得到的用户会话序列,极为粗糙不准确。用户很可能在访问了网站后,相隔很长时间再次访问,很明显,简单地按时间戳排列得到的会话序列并不是用户一次访问站点的浏览行为,而是多次的拼接。用户会话指用户在一次访问站点过程中,从进入该网站到离开网站期间对网站的一系列浏览行为。所以,会话识别的目的就是把按时间戳排列得到的会话序列划分为多个独立的、一次性访问的序列。如何区分用户的每个独立会话,Web Server日志中并没有显示记录。会话识别步骤是数据预处理过程中的重点和难点。

4)路径补充:由于各种原因,Web Server中记录的用户请求有可能存在遗漏,路径补充就是补全遗漏的客户请求,确保用户访问路径的完整性。

通过上述4个步骤,Web日志预处理后,就可得到相对完整的用户会话集,把它们按照一定的规则存在数据库中,以备模式挖掘阶段使用。

2 会话识别分析

用户会话(Session)是用户在一次访问站点过程中,从进入该网站到离开网站期间一系列浏览行为。一次用户会话就是单个用户在浏览Web站点时完整的访问路径[6]。

定义1 Web服务器日志可以看成是按照时间戳排序的集合,表示为:

$$L=\{l_1,l_2,\cdots,l_i,\cdots,l_n\}(1\leqslant i\leqslant n) \tag{1}$$

$|L|=n$,即日志集合L所包含的日志数目为n。集合L的每个元素l具有以下性质[6]:

1) $l_i=\{\text{userid},\text{url. refer},\text{time}\}(1\leqslant i\leqslant n)$,这里的userid是经过用户识别后赋予每个用户唯一的标识符。

2) $l_i.\text{time}<l_j.\text{time}(1\leqslant i<j\leqslant n)$。

定义2 会话s表示为:

$$s=\{\text{userid},(<l'_1.\text{url},l'_1.\text{time},l'_1.\text{status}>,\cdots,<l'_k.\text{url},l'_k.\text{time},l'_k.\text{status}>,\cdots,<l'_m.\text{url},l'_m.\text{time},l'_m.\text{status}>) \tag{2}$$

其中$1\leqslant k\leqslant m$,且$l'_k.\text{userid}=\text{userid}$,$l'_k.\text{time}<l'_j.\text{time}$。会话$s$的长度用$\text{length}(s)=m$表示,即会话$s$所包含的页面请求数目。$l'_k.\text{status}$表示该记录在该会话中的位置信息:会话第一条记录值为1($l'_1.\text{status}=1$);会话的最后一条记录值为2($l'_m.\text{status}=2$);否则值为0($l'_k.\text{status}=0$)。

显然,不同用户产生的页面访问序列属于不同的会话。在Web日志中,当一个用户访问时间跨越较大时,很可能并不是用户的一次访问,用户也有可能多次访问了该站点。倘若不考虑时间因素的影响,可能得到的结果是:识别出的访问序列包含了某个用户多次访问的情况,未把每次访问情况区别开。因此,需要通过某种识别方法一一辨别用户每次访问序列。

2.1 基本会话识别方法介绍

会话的识别有多种算法,有的基于访问时间,有的基于站点拓扑结构。目前,常用的会话识别算法有三种,如表1所示,前两种基于访问时间,后一种基于站点拓扑结构[7,8,9]。

目前大部分研究都只基于一种会话算法,这样得出的会话集会与真实会话之间存在较大差异。例如,某一用户中途有事情,离开网站一段时间,回来后继续访问,这样,θ和δ算法就会把同一会话分割成两个会话;而另一用户在访问过程中由于对某个网页的偏爱,保留了某一网页,下一次访问时直接打开该网页继续访问,如果只采用算法三,又把同一用户的时差相隔很久的两次会话放在了同一会话中。由此可见,使用单一的传统会话算法,将直接影响数据预处理的结果,从而影响Web日志挖掘的质量。因此,文中提出了基于访问时间的一次会话算法和利用合并和断开操作的二次会话算法,以产生更为真实的会话集。

表1　常用的会话识别方法

属性	基于访问时间的方法		基于站点结构的方法
	1	2	
特点	设置用户在整个站点的停留时间上限值为θ	设置用户在一个页面的停留时间上限值为δ	一个用户的请求不能通过其参考页上的链接进入,识别为一个新会话
会话识别标准	设t_0为会话初始页面的时间戳,若一个页面请求的时间戳t满足$t-t_0\leqslant\theta$,则被加入这个会话;第1个满足$t_0+\theta>t$的页面成为下一会话的起始页面	设t'为当前会话最后一个请求页面的时间戳,若下一请求的时间戳t''满足$t''-t'\leqslant\delta$,则加入当前会话,否则开始一个新会话	设p,q为两个连续请求的页面,t_p、t_q为p、q的时间戳,时间限制$\Delta=t_p-t_q$,属于会话s,如果q的参考页在s中,或$t_p-t_q\leqslant\Delta$,则加入当前会话s,否则q为一个新会话起始页
应用	θ取30min	δ取10min	Δ取10s

2.2　新会话识别算法描述

首先,初始会话集的产生采用传统的基于访问时间的方法,综合考虑用户在整个站点的访问时间和两个相邻页面的访问时间,然后对初始会话集进行合并和断开操作,即会话重组[10],生成优化的会话集。

2.2.1　使用基于访问时间的方法产生初始会话集

文中产生初始会话集的方法是采用基于访问时间的方法,一次会话总时间不超过30min,连续两个页面请求的时间不超过10min。得到的初始会话集很可能存在这样的问题:本来属于两个不同会话的被划分到同一个会话中,而原本属于同一个会话的两条记录又被划分到两个不同的会话中。所以,文中采用合并和断开的方法对得到的初始会话集进行进一步优化,以产生更为真实的会话集。

2.2.2　使用会话重组的方法优化初始会话集

1)合并。

采用综合考虑的基于访问时间的会话识别方法,产生的初始会话集,很可能使原本属于同一个会话的两条记录被划分到两个不同的会话中。例如,实际会话是$<L_1,\cdots,L_i,L_j,\cdots,L_m>$,但由于用户的会话总时间超过了30min,或者对页面L_i比较感兴趣,浏览时间超过了10min,实际会话就会被划分成$<L_1,\cdots,L_i>$和$<L_j,\cdots,L_m>$两个会话[11]。对于上述情况,当遇到会话边界L_i和L_j(形式为$<\cdots,L_i>$和$<L_j,\cdots>$),并且L_j的参引页refer与L_i所在的会话集中任一url相同,则将L_i和L_j所属的会话归并到同一会话中。

2)断开。

在一次会话识别的过程中,也有可能将原是两个或者多个不同会话反而划分到同一个会话中。例如,有两个实际会话$<L_1,\cdots,L_i>$和$<L_j,\cdots,L_m>$被划分成一个会话$<L_1,\cdots,L_i,L_j,\cdots,L_m>$。$L_i$和$L_j$是Web Server日志中的两条相邻有用记录,但事实上不属于同一个会话,用户已经从一个话题转向了另一话题[11]。通过对用户浏览行为的分析,很多用户在浏览时经常返回首页页面,开始一个具有新意义的会话集,因此,对于会话内部记录L_i和L_j(形式为$<\cdots,L_i,L_j,\cdots>$),如果L_j为网站首页[12],则把该会话划分为形式为$<\cdots,L_i>$和$<L_j,\cdots>$的两个会话。

由以上分析可以看出,采用优化初始会话集得到的会话会更接近事实会话,更有利用Web日志挖掘的研究。

2.3　会话识别算法

算法1 使用基于时间的方法产生初始会话集SR

Function Session1(L:日志集合,SR:初始会话集)

{$i=1$; $k=1$;

While (|UserSet|) Do //|UserSet|为所有已识别出的用户数

{$j=1$;

SR_k:={useri,($<l_j$.url,$l_j$time,1>)};// 存放第i个用户的第一条访问路径

While ($j\leqslant$useri.$|L|$) Do//遍历用户i的访问记录

{ If

($(l_{j+1}.\text{time}-SR_k[l_1^k.\text{time}]\leqslant\theta)\wedge(l_{j+1}.\text{time}-SR_k[l_{\text{length}(SR_k)}^k.\text{time}]\leqslant\delta)$)

Then

$SR_k:=SR_k\cup<l_{j+1}.\text{url},l_{j+1}.\text{time},0>$;// 满足以上条件把第$j+1$条记录的URL和时间戳加入当前会话$SR_k$

Else

{$k:=k+1$;

SR_k:={useri,($<l_{j+1}$.url,l_{j+1}.time,1>)};

$SR_{k-1}[l_{\text{length}(SR_{k-1})}^{k-1}.\text{status}]=2$;// 否则用户$i$开始一个新的会话,并且上一个会话的最后一个记录的状态置为2

}

$j:=j+1$;

}

$i:=i+1$;

$k:=k+1$;

$SR_{k-1}[l^{k-1}_{length(SR_{k-1})}.status]=2$;

//新用户必然开始一个新的会话

}}

算法2使用合并和断开重组 SR 产生更真实的会话集 S

Function Session2(SR:初始会话集,S:优化的会话集)

{$i=1$; $k=1$;

While ($|SR|$) Do // $|SR|$ 为初始会话集总数

{$j:=i$;

While ($SR_{i+1}.userid=SR_i.userid$) Do

{If ($SR_{i+1}.[l^{i+1}_1.url]$ 所对应的 refer 与 SR_i 中任一 url 相同)

Then{ $SR_{i+1}.[l^{i+1}_1.status]=SR_i.[l^i_{length(SR_i)}.status]=0$;

Length(SR_i)+=Length(SR_{i+1});

$i:=i+1$;} //合并操作

}

$S_k=SR_j$;

For ($a=1;a\leqslant length(SR_j);a++$)

{If($SR_j[l^j_a.url]$ = 'index.asp')

Then

{

$k:=k+1$;

$S_k[l^k_1.url]=SR_j[l^j_a.url]$;

$S_k[l^k_1.status]=1$;

$S_{k-1}[l^{k-1}_{length(S_{k-1})}.status]=2$;

continue;}//断开操作

$k:=k+1$;

}

$i:=i+1$;

}}

3 实验与结果分析

3.1 实验数据准备

实验所选取的日志数据是中国民航大学易航网(http://www.ehang.com.cn/)的日志数据,由于数据量较大,实验只选取了2011年3月17日的数据,经过处理后存入SQL数据库中的有用日志数据共28235条,图1是存入log表中的数据,log表的定义为:id(日志记录标识),user_id(用户标识),page_id(请求页URL标识),refer_id(参考页URL标识),begin_time(开始访问时间)。

3.2 会话识别结果分析

把文中算法应用于清洗后的日志数据,得到如图2所示的会话识别结果:

id	user_id	page_id	refer_id	begin_time
[redacted]	1	1	2	2011-3-17 8:35:05
8a8598452fb39d4	1	3	2	2011-3-17 8:35:06
8a8598452fb39d4	2	2	4	2011-3-17 8:35:11
8a8598452fb39d4	2	5	2	2011-3-17 8:35:13
8a8598452fb39d4	3	6	2	2011-3-17 8:35:14
8a8598452fb39d4	3	1	6	2011-3-17 8:35:15
8a8598452fb39d4	3	1	6	2011-3-17 8:35:15
8a8598452fb39d4	4	7	2	2011-3-17 8:35:15
8a8598452fb39d4	4	1	7	2011-3-17 8:35:16
8a8598452fb39d4	4	1	7	2011-3-17 8:35:16
8a8598452fb39d4	2	1	2	2011-3-17 8:35:17
8a8598452fb39d4	5	2	4	2011-3-17 8:35:19

图1 存入数据库中的清洗后的日志数据

数据预处理->用户会话

会话编号	用户编号	访问路径
6	5	6,1
7	6	1,11,13,14
10	8	7,15,17,29,10,9,8,53,7,15,17,29,10,9,8,53,7,15,17,29,26,10,9,8,53,150,10
11	9	9,8,2,26,10,11,13,9,8
12	10	1,11,32,25,13,14
20	11	6,1,11,1,6,1
21	17	16,21,11,25,13
33	24	18,8,52,56,52,63,52,74,52,63,52,63,117,52,63,151,152,52,6,52,133
34	24	233,68,8,133,63,8,68,8,63,20,252,133,260
43	29	22,24,11,25,13,1,32,14,16,149,16,21

图2 数据预处理后的用户会话序列

会话识别质量的评测标准不同,有的关注连续性,有的关注完整性。目前使用最为广泛的是精确度和查全度。精确度是用完全构建的会话数(即事实会话数目 R 与识别算法识别出的会话数 R_h 的交集)与识别算法识别出的会话数 R_h 的比值表示:$precision(h)=\frac{|R_h\cap R|}{|R_h|}$;查全度是用完全构建的会话数与事实会话数 R 的比值来表示:$recall(h)=\frac{|R_h\cap R|}{|R|}$[13]。

实验同时使用了3种会话识别方法,以比较不同会话识别算法的精确度和查全度。方法1采用文中算法,方法2采用站点总停留时间不超过 θ 的算法,方法3采用页面停留时间不超过 δ 的算法。由于真实会话数不易得到,这里需要人工计数,以userid=6的用户为样本(userid是用户识别阶段已识别出的用户标识)。

表2 会话识别结果

方法	识别出的会话数	精确度	查全度
1	113	78.26%	86.27%
2	91	74.73%	66.67%
3	95	69.49%	61.76%

由表2可以看出,文中算法无论在精确度和查全度上都优于方法2和方法3,从而有效地提高了会话

识别质量,为后续的模式挖掘工作打下了良好的基础。

4 结束语

Web 日志挖掘是当前研究的热点,数据预处理是 Web 日志挖掘首先要解决的问题,在数据预处理的各个步骤中,会话识别起到至关重要的作用,会话识别的质量直接决定着数据预处理质量的好坏,从而影响着 Web 日志挖掘的最终结果。文中重点研究了预处理阶段的会话识别算法,提出了基于访问时间的一次会话识别和利用断开和合并操作的二次会话识别的新会话识别方法,弥补了传统会话识别算法中真实性较低的不足,提高了数据预处理的质量。同时,文中算法也存在着不足,基于时间的一次会话所采用的访问时间是先验值,会与实际试验数据存在不匹配的地方,而且算法的效率还有待于提高,需要进一步深入的研究。

参考文献:

[1] 韩家炜,孟小峰. Web 挖掘研究[J]. 计算机研究与发展,2001,38(4):405-414.

[2] 钱小军,平玲娣,潘雪增. Web 文本挖掘技术研究及其实现[D]. 杭州:浙江大学,2002.

[3] Hseush W, Pu C. A Practical Technique for Asynchronous Transaction Processing[C]//Proceedings of ICDCS. [s. l.]: [s. n.],1995:110-117.

[4] 王听忠,王　辉,武新梅,等. 基于协同推荐的 Web 日志预处理过程[J]. 微计算机信息,2006(22):150-151.

[5] Lin Haibin, Keselj V. Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests[J]. Data & Knowledge Engineering,2007,61(2):304-330.

[6] 赵　伟,何丕廉,陈　峡,等. Web 日志挖掘中的数据预处理技术研究[J]. 计算机应用,2003,23(5):62-66.

[7] 朱　岩,杨永田,张玉清,等. 基于层次结构的信息安全评估模型研究[J]. 计算机工程与应用,2004,40(6):40-43.

[8] Wang Xidong, Ouyang Yiming, Hu Xuegang, et al. Discovery of User Frequent Access Patterns on Web Usage Mining[C]//The 8th International Conference on Computer Supported Cooperative Work in Design. [s. l.]:[s. n.],2004:765-769.

[9] 欧阳一鸣,汪曦东,郭　骏. Web 使用挖掘数据预处理中的会话构造[J]. 计算机工程与应用,2005,41(25):148-151.

[10] 张海强,胡学龙. 一种基于引用日志文件的启发式会话识别算法[J]. 扬州大学学报,2007(3):57-61.

[11] Facca F M, Lanzi P L. Mining interesting knowledge from Weblogs: a survey[J]. Data and Knowledge Engineering, 2005,53(3):225-241.

[12] 戴智丽,王鑫昱. 一种基于动态时间阈值的会话识别方法[J]. 计算机应用与软件,2010,27(2):244-246.

[13] Spiliopoulou M, Mobasher B, Berendt B, et al. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis[J]. INFORMS Journal on Computer,2003, 15(2):171-175.

(上接第 44 页)

3 结束语

上述实验结果表明,基于像素积分投影的印刷体维吾尔文切分方法,对于图 2 中的维文字体有比较好的切分效果,其他印刷体维吾尔文字体也可以采用这种方法。但是这种方法采用的是像素积分投影,由于不同字体的行列像素厚度有一定的差异,所以对于其他维吾尔文字体就需要根据字体像素厚度情况设定不同的阈值。

参考文献:

[1] 尹　芳,王卫兵,陈德运. 印刷体英文文档识别系统的设计与实现[J]. 哈尔滨理工大学学报,2008,13(6):9-12.

[2] 罗剑锋. 字符图像的分割与识别[D]. 北京:北京理工大学,2003.

[3] 吴晓峰. 基于单词全局特性的印刷体英文单词识别系统研究[D]. 广州:中山大学,2005.

[4] 欧　珠,普次仁. 印刷体藏文文字识别技术研究[D]. 西藏:西藏大学,2009.

[5] 哈力木拉提,阿孜古丽. 多字体印刷维吾尔文字符识别系统的研究与开发[J]. 计算机学报,2004,27(11):1480-1484.

[6] 王　华,丁晓青. 多字体多字号印刷体维吾尔文字符识别[J]. 清华大学学报,2004,44(7):946-949.

[7] 袁保社,吾守尔·斯拉木. 一种手写维吾尔文字母识别算法[J]. 计算机工程,2010,36(2):198-188.

[8] 董国君. 印刷体俄文文字识别研究[D]. 乌鲁木齐:新疆大学,2009.

[9] 求是科技,苏彦华. Visual C++数字图像识别技术典型案例[M]. 北京:人民邮电出版社,2004.

[10] Albadr B. A Segmentation-Free Approach to Text Recognition with Application to Arabic Text[D]. Washington: University of Washington,1995.

[11] An K H. Concurrent pattern recognition and optical character recognition[D]. USA: University of North Texas,1991.

[12] Majid M. Altuwaijri. A Parallel Recognition System for Arabic Cursive Words with Neural Learning Capabilities[D]. USA: The Graduate Faculty of The University of Southwestern Louisiana,1995.