Web 日志挖掘中一种改进的会话识别方法

周爱武,程 博

(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:提出了一种改进的会话识别方法。该方法基于访问站点的首页和导航页,以首页或导航页作为新会话开始的标识。选取真实的 Web 日志,用 PL/SQL 编程实现改进的会话识别方法,并与现有方法进行比较。实验结果证明,改进的会话识别方法比现有方法识别会话更有效。

关键词:数据预处理;Web 日志;会话识别;站点首页;导航页

中图分类号:TP301 **文献标识码**:A **文章编号**:1674-7720(2010)15-0071-03

An improved method for session identification in Web log mining

ZHOU Ai Wu, CHENG Bo

(College of Computer Science & Technology, Anhui University, Hefei 230039, China)

Abstract: An improved session identification method was brought forward. The improved method was based on the site home page and navigation pages, using the home page or navigation pages as the identification of the new session beginning. Using a real Web log as experiment data, the improved session identification method was implemented by PL/SQL programming, and further compared with existed methods. Experimental results illustrate that the improved method can identify user sessions more effectively.

Key words: data preprocessing; Web log; session identification; site home page; navigation page

Web 日志挖掘现已成为 Web 挖掘研究的重点。其主要分为数据预处理、模式发现、模式分析 3 个阶段[1]。数据预处理阶段是要把从各种数据源得到的使用信息、内容信息和结构信息转换成模式发现阶段需要的数据抽象;模式发现阶段旨在使用各种数据挖掘技术发掘隐藏在数据背后的规律和模式;模式分析阶段旨在根据具体的实际应用,过滤掉在模式发现阶段没有用的规则或模式,并把有用的规则和模式转换为知识。

本文主要研究数据预处理阶段的会话识别。在分析现有的会话识别方法基础上,提出一种基于访问站点首页和导航页的改进会话识别方法,最后通过实验验证了改进的会话识别方法比现有方法更有效。

1 数据预处理

数据预处理是 Web 日志中最基础、最频繁的工作,是整个数据准备的核心工作。数据预处理的结果将直接影响到挖掘算法产生的规则和模式,因此预处理过程在整个 Web 日志挖掘过程中占据着非常重要的地位,是挖掘质量的保证。

数据预处理包括数据清理、用户识别、会话识别、路径补充和事务识别 5 个阶段[2]。(1)数据清理是指删除 Web 日志中与挖掘算法无关的数据;(2)用户识别是识别出访问网站的每个用户;(3)会话识别是在用户识别之后,把每个用户在一段时间内的访问序列进行分解,从而得到相应的会话。会话是指同一用户在一次浏览过程中连续请求的页面序列,它代表了用户对服务器的一次有效访问;(4) 路径补充是对识别出的用户会话进行优化的步骤,以使得其更加准确地描述用户的浏览请求;(5)事务识别是将用户会话进行语义分组,形成适合挖掘需要的事务。

2 会话识别分析

用户会话[3]是指用户从进入站点到离开站点期间所访问的一系列页面序列集合。可表示为:

$$Session \leqslant SessionID, \{(Pid_1, t_1)\cdots(Pid_k, t_k)\cdots(Pid_n, t_n)\}>, 1 \leqslant k \leqslant n \quad (1)$$

其中 $SessionID$ 是会话标识,$\{(Pid_1, t_1) \cdots (Pid_k, t_k)\cdots(Pid_n, t_n)\}$是此次用户会话的页面访问序列,$Pid$ 是访问页面的标识,t 是访问该页面的时间。(Pid_1, t_1)表示用户此次会话访问的第一个页面和时间,(Pid_n, t_n)表示用户

此次会话访问的最后一个页面和时间。

2.1 常用会话识别方法

目前常用会话识别方法主要有两大类:一类是基于时间阈值,另一类是基于用户访问页面时的参引页面。基于时间阈值的会话识别方法又可细分为以下3类:

(1)设定会话的持续时间阈值θ。即一个会话总的持续时间不超过θ。国外学者Catledge和Pitkow由实验得出θ设为25.5 min较好[4],许多商业产品都采用30 min作为缺省值。

(2)设定页面的访问时间阈值η[5]。假设(Pid_i, t_i)、(Pid_{i+1}, t_{i+1})为一个用户访问序列中的两条相邻访问记录。只有当$t_{i+1}-t_i \leqslant \eta$时,才认为这两条记录属于同一个会话。当$t_{i+1}-t_i > \eta$时,$(Pid_i, t_i)$是上一次会话的最后一条访问记录,而$(Pid_{i+1}, t_{i+1})$是新会话的第一条访问记录。一般$\eta$取10 min。

(3)上述方法(2)是对所有页面设定同一个页面访问时间阈值,并没有因页面的不同而不同。参考文献[6]中,根据统计的页面的访问时间,在正态分布的假设下为每个页面设定一个访问时间作为切分会话阈值,并结合页面内容及站点结构来确定页面重要程度,对该阈值进行调整。这是一种个性化的时间阈值设置方法。

另外一类识别用户会话的方法是基于用户访问页面时的参引页面[7],即引用页。描述如下:假设(Pid_i, t_i),(Pid_{i+1}, t_{i+1})为一个用户访问序列中两条相邻访问记录。其中(Pid_i, t_i)属于会话S。如果请求页面Pid_{i+1}的引用页面曾经在会话S中出现过,那么Pid_{i+1}就属于会话S;或者Pid_{i+1}的引用页面为空,且$t_{i+1}-t_i \leqslant \Delta$($\Delta$为时间延迟,一般取10 s),那么$Pid_{i+1}$属于会话S。

2.2 常用会话识别方法评估

第(1)、(2)两种方法使用单一时间阈值来识别用户会话显然是不合理的。方法(1)不能识别出访问时间大于30 min的会话,且识别不出两个连续较短的会话;方法(2)的不足在于,若一个用户在访问站点期间暂时离开电脑,但并没有退出站点,过10 min后回来继续浏览该站点,这实际上属于同一个会话,而方法(2)则会错误地认为用户开始了一个新的会话;方法(3)使用的统计学方法虽然大大减小了上限阈值,但仍然无法准确描述对页面感兴趣的用户阅读网页的平均时间,无法区分超短时间用户访问记录。

基于参引页面的会话识别方法引入了时间限制Δ,主要是考虑到下面这种情况:访问页面的引用页面为空,用户可能是通过点击浏览器上的"BACK"按钮,回溯到之前某个曾经浏览过的页面,进而访问到该页。这显然也是不合理的,用户从p页面回退到上级页面后,用户要在此页面搜寻到感兴趣的p页面,并点击链接进入该页面,所需时间一般不止10 s,且用户可能是回退多次后再点击链接进入p页面。因此,此处设置这个时间阈值并不合理。

3 改进的会话识别方法

3.1 会话划分思考

要准确地识别出用户会话,关键在于识别出两次相邻会话的分割点。即上一次会话结束时访问的页面及下一次会话开始时访问的页面。而找出新会话开始时访问的页面,也就意味着上一会话的结束。因此,研究重点放在寻找标记新会话开始的访问页面。

用户开始访问某一站点,一般是通过在浏览器的地址栏中输入站点的URL或是通过点击收藏栏中的收藏,通过站点的首页进入此站点的,此时用户也就开始了自己的一次会话。在Web服务器日志中,可以查看用户访问的URL是否是首页来判断用户的这种行为。当用户浏览完毕退出该站点,此时会话结束,而在Web服务器端日志中,无法判断这种用户行为。但当该用户下一次通过首页来访问站点时,在Web日志中发现用户又键入了首页URL,则很显然上一次会话在本条记录之前结束,本条记录标志用户开始了一个新的会话。

3.2 改进的会话识别方法

上述思想以访问站点的首页作为新会话开始的标记,基于这一前提用户开始访问站点时总是由站点首页进入站点。但真实的访问情况并不是所有的用户每次开始访问站点时都由首页进入。一般的站点分若干版块,而每一版块都有自己的导航页。如一门户网站有新闻、体育、娱乐各版块,有的用户只对体育感兴趣,那么他可能就会将体育版块的导航页做为收藏,每次访问站点时,点击收藏,直接进入体育导航页开始访问,而非先通过站点首页,再进入体育版块导航页。因此,识别用户会话,不能只以站点首页作为开始标记,还应考虑各导航页,因为很多用户是直接通过导航页访问自己感兴趣的页面而非站点首页。

改进的会话识别方法如图1所示,以站点首页或导航页作为新会话开始的标识。

改进的会话识别方法具体描述如下:

(1)首先用户访问序列中的第一条访问记录是第一个会话的开始序列,置入第一个会话中;

(2)读取用户访问序列中的下一条访问记录,直至序列中所有记录都处理完毕;

(3)判断本次访问的页面是否是站点的首页,若是首页,则当前会话结束,新会话开始,将该次访问置入新会话的访问序列中,然后转步骤(2)处理下一条访问记录。否则,转步骤(4);

(4)判断本次访问的页面是否是站点的导航页之一,若不是(即该页面为内容页),则将本次访问置入当前会话的访问序列中,然后转步骤(2)继续处理下一条访问记录。否则(即该页面是导航页之一),转步骤(5)判断它的上一条访问记录;

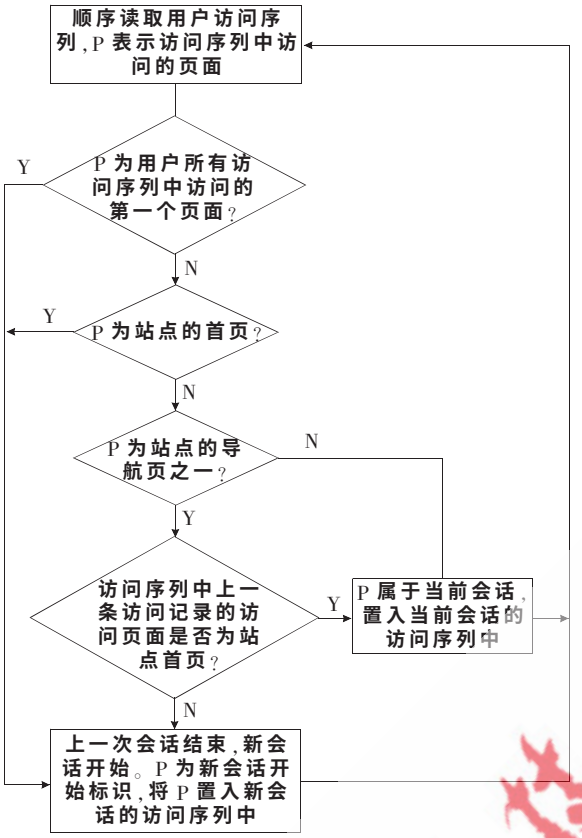

图 1 改进的会话识别方法

(5)判断上一条访问记录，若上一条访问记录访问的页面是首页，则本次访问记录和上次访问记录同属一个会话；若上一条访问记录访问的页面不是首页，则本次访问就标识了新会话的开始，将其置入新会话的访问序列中。转步骤(2)，处理下一条访问记录。

4 实验与结果分析

4.1 实验过程

4.1.1 数据准备

选用了安研星空站点 http://www.ahusky.cn/从 2009 年 2 月 17 日至 2009 年 3 月 5 日的 Web 服务器日志，共计 1 251 331 条记录，作为实验数据，如下图 2 所示。

图 2 原始 Web 日志

4.1.2 会话识别

将这些 Web 访问日志通过 SQL Loader 载入 Oracle 数据库中，经过数据清理，共有有效访问记录 35 273 条，存放在表 log 中，如图 3 所示。

此处，以 Web 访问日志中的 IP 地址作为用户标识，利用 Oracle PL/SQL 编程实现上述改进的会话识别算法。为了与其他的会话识别方法进行比较，分别用 2.1 节中的方法(1)和方法(2)对同样的 Web 日志进行会话识别，其中方法(1)取时间阈值 30 min，方法(2)取时间阈值 10 min。实验结果如下表 1 所示。

图 3 经过数据清理后的 Web 日志

表 1 实验结果

	识别出的会话总数	识别会话准确率/%(用户 220.178.4.195)
方法(1)(θ=30 min)	5 073	62.47
方法(2)(η=10 min)	5 226	64.85
改进的会话识别方法	11 325	82.19

4.2 实验分析

通过实验发现，改进的会话识别方法识别出的会话数(11 325 条)要远多于方法(1)(5 073 条)和方法(2)(5 226 条)。另外，为了比较这三种会话识别方法识别会话的准确率，将三种方法中识别出的关于用户 220.178.4.195 的会话分别与原始的 Web 日志记录比较，发现改进的会话识别方法识别会话的准确率 (82.19%) 也要高于方法(1)(62.47%)和方法(2)(64.85%)。由此可见，改进的会话识别方法能够识别出更多的会话，且识别会话的准确率也更高。

数据预处理阶段的会话识别为模式分析阶段提供了挖掘数据，即每一个有效的用户会话，因此它直接影响到模式分析阶段能否发现有效的模式。本文提出的基于站点首页和导航页的改进会话识别方法能识别出更多的会话，识别会话的准确率更高。

参考文献

[1] SRIVASTAVA J, COOLEY R. Web usage mining: Discovery and applications of usage patterns from Web data[C]. SIGKDD Explorations, 2000.

[2] COOLEY R, MOBASHER B, SRIVASTAVA J. Data preparation for mining world wide web browsing patterns[J]. Knowledge and Information Systems, 1999,1(1):5-32.

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

[4] CATLEDGE L, PITKOW J. Characterizing browsing strategies in the world wide Web[J]. Computer Networks and ISDN Systems, 1995,27(6):1065-1073.

[5] SPILIOPOULOU M, MOBASHER B, BERENDT B, et al. A framework for the evaluation of session reconstruction

heuristics in Web usage analysis [J]. Informs Journal of Computing, 2003,15(2):171-179.

[6] 严奉华,刘建平,杨凡丁.改进的 Web 访问日志会话识别算法[J].计算机工程与设计.2008,29(22):5685-5687.

[7] 熊忠阳,周亚峰.Web 访问挖掘的预处理技术的研究[J].计算机技术与发展 2007,17(8):14-18.

(收稿日期:2010-03-01)

作者简介:

周爱武,女,1965 年生,副教授,主要研究方向:数据库与 Web 技术、数据仓库与数据挖掘。

程博,男,1985 年生,硕士研究生,主要研究方向:数据库与 Web 技术。