一种改进的Web日志会话识别方法

方元康[1,2],胡学钢[1],夏启寿[2]
(1.合肥工业大学 计算机信息学院,安徽 合肥 230009;
2.池州学院 计算机中心,安徽 池州 247000)

摘 要:会话识别是Web日志挖掘中的数据预处理中的一个重要步骤。文中提出了一种改进的会话识别方法。首先,在用户识别后,进行框架页面的过滤,从而大大地减少了实验产生的有效页面,然后为页面设置访问时间阈值,并根据页面内容及站点结构确定的页面重要程度对该阈值进行调整。通过实验证明,相对于传统的对所有页面使用单一的先验阈值进行会话识别的方法,该方法所得到的会话集更具有真实性。

关键词:Web挖掘; 数据预处理; 阈值;Frame页面;会话识别
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2008)11-0214-03

An Improved Method for Transaction Session Identification in Web Usage Mining

FANG Yuan-kang[1,2], HU Xue-gang[1], XIA Qi-shou[2]
(1. Computer & Information College, Hefei University of Technology, Hefei 230009, China;
2. Center of Computer Technology, Chizhou College, Chizhou 247000, China)

Abstract: Session identification is an important step in data preprocessing of web log mining, an access intervals-based improvement was carried out of transaction session identification in web usage mining. After identifying users, effective web pages in experiment are reduced greatly by filtering frame pages, and the access time threshold was adjusted by the web contents and site's structure on this condition. Compared to the traditional method that defines a uniform a threshold for all web pages experimentally, the approach presented can decide the access time threshold more accurately. Algorithm enhancing the quality of transaction session is proved by experiments.

Key words: Web mining; data preprocessing; threshold; frame page; session ident ification

0 引 言

Web使用挖掘[1]就是通过挖掘Web日志记录来发现用户访问Web页面的模式,从中可以提炼出设计者的领域知识、用户感兴趣程度、用户的访问习惯等,进而得到优化站点结构、开展个性化服务以及用户访问控制等对站点设计者、经营者有用的决策性信息。挖掘Web日志记录中最费时也是最关键的环节就是Web日志记录的数据预处理,而影响挖掘Web日志质量的关键因素就是在数据预处理中对会话识别的真实程度。

目前,会话的构造主要是基于启发式的方法:如基于时间的,依据站点结构的,给予引用的。

(1) H_{visit}: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 θ 则认为新的会话开始[2,3]。设 t_0 为会话初始页的时间戳,同一用户的一个URL请求的时间 t 如果满足 $t-t_0\leqslant\theta$,则被加入当前会话,第一个满足 $t_0+\theta<t$ 的页面成为下一个会话的初始页。一般 θ 取30min。

(2) H_{page}:给用户一个页面停留时间域值 $\triangle t$[4],如果2个连续请求的时间间隔没有超过这个值 $\triangle t$,这属于同一会话,否则分属于两个会话。$\triangle t$ 一般取10min。

(3) H_{Ref}:利用用户访问历史和参引页来划分[3],如果一个用户的请求不能通过参引页上的链接进入,则很可能属于另一个会话。即当前请求的参引页没有在前面访问过的页面中出现,则是一个新的会话开始。

(4) MF (maximal forward references):最大向前参引模型[5],即在一个用户会话里不会出现用户先前

收稿日期:2008-02-28
基金项目:安徽省自然科学基金项目(KJ2008B116);池州学院自然科学基金项目(XK0829)
作者简介:方元康(1968-),男,安徽池州人,硕士研究生,讲师,研究方向为数据挖掘;胡学钢,教授,博士,硕士生导师,研究方向为知识工程、数据挖掘、数据结构。

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,则表示当前会话结束,一个新的会话开始。

另外,文献[6]将(1)、(2)、(3)结合起来,生成基于时间和引用的启发式识别方法。

以上算法中,它们的不足之处在于两个方面:一方面可能使原本在同一个会话里的记录被划分到不同的会话中,也可能使原本不在同一个会话的记录被划分在同一个会话中;另一方面,由于用户会话产生的有效页面数比实际的有效页面数明显增多,因此,导致了会话识别的效率大大降低。如果按上述方法生成的会话集中的不真实的成分太多,那么将会使得挖掘出来的结果具有很小的理论价值,甚至失去理论价值。因此,文中提出了一种基于过滤框架网页与页面访问时间阈值相结合的会话识别方法。

1　一种优化的会话识别方法

采用基于过滤框架网页与页面访问时间阈值相结合的会话识别方法。首先,在用户识别后,进行框架页面的过滤,从而大大地减少了实验产生的有效页面,然后为页面设置访问时间阈值,并根据页面内容及站点结构确定的页面重要程度对该阈值进行调整。通过实验证明,相对于传统的对所有页面使用单一的先验阈值进行会话识别的方法,该方法所得到的会话集更具有真实性。

1.1　会话识别前的数据预处理

在会话识别前,Web 日志挖掘的数据预处理主要包括以下两个步骤:数据清理、用户识别。

1.1.1　数据清理

Web 日志记录中包括用户 IP 地址、用户 ID、用户请求访问的 URL 页面、请求方法、访问时间、传输协议、传输的字节数、错误代码、用户代理等属性[7]。数据清理根据需求对原始日志文件进行处理,主要删除后缀为 gif,jpg,jpeg 的图片文件,以及 cgi,js,js 的脚本文件。

1.1.2　用户识别

用户是指通过一个浏览器访问一个或几个服务器的个体。在实际使用中惟一确定一个用户很难,一个用户可以通过几个代理或机器访问服务器。一般最常被 Web 日志挖掘工具使用的技术是基于日志/站点的方法,可以使用一些启发式规则帮助识别用户。

1) 如果用户的 IP 地址不同,则认为是不同的用户;

2) 如果 IP 地址相同,而代理(agent)日志中表明用户的浏览器或操作系统改变了,则可以假设为两个不同的用户;

3) 将访问日志、引用日志和站点拓扑结构结合,构造用户的浏览路径。如果当前请求的页面同用户已浏览的页面间没有链接关系,则认为存在 IP 地址相同的多个用户。

通过这 3 条规则,结合用户提交的查询信息便可以给不同的用户赋予不同的用户 ID 号。

1.2　过滤框架页面

HTML 规范通过"Frame"标记支持多窗口页面,每个窗口里装载的页面对应一个 URL,要说明的一点是: Subframe 页面同时又可以是包含子窗口的 Frame 页面[8]。

当用户访问的 URL 对应的是一个 Frame 页面时,浏览器通过解释执行页面源程序,会自动向 Web 服务器请求该 Frame 页面中包含的所有 Subframe 页面,这一个过程可以重复进行,直到所有的 Subframe 页面被请求。在图 1 的例子中,如果用户请求 A 页面,其它五个页面(B, C,D, E, F)也一起被请求。由此导致对 Frame 页面与其 Subframe 页面的请求记录总是一起出现在 Web 日志中。在这样的用户会话文件上进行数据挖掘, Frame 页面和其 Subframe 页面作为频繁访页组出现的概率很高,但 Frame 页面与它的 Subframe 页面之间的关系在创建 HTML 文件时就是已知的,因此, Frame 页面和其 Subframe 页面同时出现在挖掘结果中使得挖掘出来的频繁访问页组的兴趣性下降。如果在会话识别后,把 Frame 页面和其 Subframe 页面作为一个整体考虑,即用户对 Frame 页面的请求就是请求多窗口页面。从全局上看,这样处理可以有效地消除 Subframe 对日志挖掘的影响,从而提高挖掘结果的兴趣性。

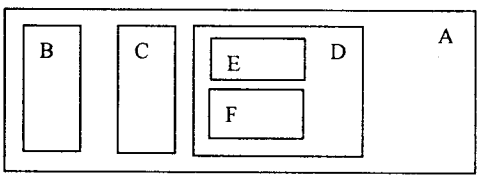

图 1　多窗口页面示意图

1.3　基于访问页面时间阈值的会话识别

1.3.1　为每个页面生成一个合理的访问时间阈值

该方法通过使用访问时间间隔超出某阈值 δ 来识别会话,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合理的访问时间作为切分阈值,并给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。

定义:链接内容比 R_{LCR} 是指页面链入(L_I)链出(L_O)数与页面内容之比,记页面大小为 S_{DS},β 为页面

R_{LCR} 对页面访问时间阈值 δ 的影响因子，α 是平滑系数，实验表明 α 选择为 1.25 比较合适。

计算公式如下：

$$R_{LCR} = 2(0.618L_I + 0.382L_O)/S_{DS} \quad (1)$$

$$\beta = 1 - \exp(-\mathrm{sqrt}(\mathrm{sqrt}(R_{LCR}))) \quad (2)$$

$$\delta = \alpha t(1 + \beta) \quad (3)$$

1.3.2 基于时间阈值 δ 生成用户会话集

要设置每个页面的访问时间阈值 δ，首先要获得统计后的页面的访问时间 t，并结合页面的 R_{LCR} 影响因子 β 调整 δ。统计后页面的访问时间 t 的集合记为 $S_t = \{t_1, t_2, \cdots, t_n\}$，页面的影响因子 β 集合记为 $S_\beta = \{\beta_1, \beta_2, \cdots, \beta_n\}$。

算法步骤如下：

1) 根据日志文件 log 统计得到 t 的集合 S_t，按用户将日志文件划分，即用户识别，根据 1.3.1 节的方法统计页面的访问时间得到 S_t；

2) 根据影响因子集合 S_β 和 α 值调整 S_t 得到页面访问时间阈值集合 S_δ，按式(3) 结合 α 和 S_β 计算得到 S_δ；

3) 根据 S_δ 重新划分日志文件得到用户的会话集合。

2 实验数据及分析

本实验数据来源于池州学院网站：211.86.192.12，服务器日志数据为 2007 年 10 月 24 日～11 月 3 日。实验中将基于固定时间阈值的会话识别算法与基于页面访问时间阈值的会话识别算法这两种算法结果进行比较分析(见表 1)。

表 1 训练数据与测试数据实验统计结果

	训练数据	测试数据
实验数据中共有访问记录数	708912	69852
数据清洗后共有有效页面数	55442	4265
页面过滤后共有有效页面数	31369	2536
基于 ip、agent 和 referer 方法划分的用户数	21492	1739
基于固定页面时间阈值(10min)方法划分的会话数	24627	1990
基于访问页面阈值方法划分的会话数	25667	2075
页面过滤后的会话数	24605	2038

从表 1 统计结果中得到以下结论：

1) 通过框架页面的过滤，会话个数基本不变，但是大大减少了有效页面数，提高了会话识别的效率，更重要的是提高了挖掘结果的兴趣性。

2) 相对于使用固定会话长度的识别方法，改进的方法可以识别长会话。因为会话长度固定的识别方法会话最长为 θ，一般设置为 25.5 min，而实际应用中会有较多的长会话存在，改进的方法就可以识别出这些长会话。

3 结束语

采用基于过滤框架网页与页面访问时间阈值相结合的会话识别方法，发现 Frame 页面极大地影响了挖掘结果的兴趣性。文中在现有数据预处理技术的基础上提出先消除 Frame 页面对挖掘结果的影响，再使用页面访问时间阈值来进行会话识别，经过实验数据的检验，改进后的数据预处理技术使得挖掘结果的兴趣性显著提高。

参考文献：

[1] Han Jia - Wei, Meng Xiao - Feng, Ang Jing. Research on Web Mining[J]. Journal of Computer Research & Development, 2001, 38(4): 405 - 414.

[2] Fu Y, Sandhu K, Shih M. A generalization - Based Approach to Clustering of Web Usage Session[C]// Proc 1999 KDD Workshop Web Mining, LNCS 1863. [s.l.]: Springer - Verlag, 2000: 21 - 28.

[3] Cooley R, Mobasher B, Srivastava J. Data Preparation for Mining World Wide Web Browsing Patterns[J]. Knowledge and Information system, 1999, 1(1): 5 - 32.

[4] 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.

[5] Chen M S, Park J S, Yu P S. Data Mining for Path Traversal Patterns in a Web Environment[C]// Proc 16th Int'l Conf Distributed Computing System(ICDCS96). [s.l.]: IEEE CS Press, 1996: 385 - 392.

[6] 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)：11 - 14.

[7] Srivastava J, Cooley R, Deshpande M, et al. Web Usage Mining: Deiscovery and Applications of Usage Patterns from Web Data[J]. Proc ACM SIGKDD, 2000, 1(2): 12 - 23.

[8] 金松河，钱慎一，张素智. Frame 页面过滤算法在 Web 日志挖掘预处理中的应用[J]. 云南民族大学学报：自然科学版，2006，15(1)：63 - 65.

(上接第 213 页)

Commondity Grid Kit[J]. Concurrency and Computation: Practice and Experience, 2001, 13(15): 1045 - 1055.

[4] The globus 联盟. Package org.globus.ftp[EB/OL]. 2004 - 10. http://www.globus.org/cog/distribution/1.2/api/org/globus/ftp/package - summary.html.

[5] Foster I, Kesselman C. 网格计算[M]. 第 2 版. 金 海，袁平鹏等译. 北京：电子工业出版社，2004：233 - 237.