

文章编号:1001-9081(2010)04-1056-03

## 利用图片类日志信息改进会话识别质量

范纯龙,姜宏飞,李 华

(沈阳航空工业学院 计算机学院,沈阳 110136)

(zgr\_mzd@163.com)

**摘 要:**数据预处理是 Web 日志挖掘的基础,而会话识别则是数据预处理的关键步骤,其质量严重影响 Web 日志挖掘的结果。在分析现有会话识别方法的基础上,提出了利用数据预处理中废弃的图片等日志数据,并结合扩展 Web 图结构,从页面分组规则和路径补全算法两个方面改进会话识别质量,并通过实验证实该方法对改善会话识别质量是有效的。

**关键词:**会话识别;数据预处理;Web 图结构;路径补全;数据清洗

**中图分类号:** TP393.09 **文献标志码:** A

## Use of picture log information in improving session identification quality

FAN Chun-long, JIANG Hong-fei, LI Hua

(School of Computer Science, Shenyang Institute of Aeronautical Engineering, Shenyang Liaoning 110136, China)

**Abstract:** Data pre-processing is the basis for Web log mining, and session identification is a key step in data preprocessing, so session identification quality seriously influences Web log mining results. The paper analyzed the current session identification methods and proposed to improve session identification quality by pictures log data abandoned in data pre-processing. With reference to the expansion of Web graph structure, the improvement was made from such two aspects as page grouping rules and path completion algorithm. The method is experimentally proved to be effective to improve the session identification quality.

**Key words:** session identification; data pre-processing; Web graph structure; path completion; data cleaning

### 0 引言

Web 日志挖掘过程由数据预处理、模式发现和模式分析 3 部分组成。其中数据预处理的质量将直接影响到后面两阶段的分析和挖掘效果,而会话识别则是数据预处理质量的核心因素。

会话是用户对服务器的一次有效访问,它包括会话期间内用户访问的页面集和遍历的页面间路径关系两个方面。识别出一组准确、可靠的会话集是下一步数据分析的基础,会话识别的数据来源是经过数据清洗和用户识别后的日志数据。现有的数据清洗方法主要是删除 Web 日志中与挖掘算法无关的数据<sup>[1]</sup>,这些数据一般包括图片日志、cgi 脚本日志、网络爬虫日志、状态码值大于 299 的日志和 GET 以外的服务请求日志。用户是通过一个浏览器访问一个或几个服务器的个体,用户识别一般依据由 IP 地址、日志的代理属性和参考域属性 3 方面确立的启发式规则<sup>[2-3]</sup>进行,用户识别的实质是对清洗后的日志进行分组。

会话识别是在用户日志分组的基础上进一步分析出同一个用户的多次会话过程。本文结合常规浏览器访问请求的特点,利用数据清洗时废弃的部分日志数据,并结合扩展的 Web 拓扑结构<sup>[4]</sup>,改进会话的页面集确定方法和路径补充算法,最终实现提高会话识别质量的目标。

### 1 会话识别现状分析

在较长时间段内,用户可能多次访问某站点,Web 日志记

录了这些访问过程的大量信息,但 Web 的缓存机制和协议特点,导致 Web 日志记录与会话间不是简单的对应关系,这造成会话页面集确定及遍历路径关系识别的困难。

用于挖掘的 Web 日志一般包含请求 IP、返回字节数、请求时间、请求方法、服务响应状态码、请求域和参考域等属性。请求域是用户请求的当前页,参考域指明当前请求是用户在访问参考域页面过程中发出的。

目前,会话页面集的确定方法主要是启发式的,常用会话生成方法有以下 4 种:Timeout<sup>[1]</sup>,如果用户在整个站点的停留时间或相邻页面请求的间隔时间大于域值;参考域<sup>[2]</sup>,Web 日志的参考域不在该用户已访问页面集内;最大向前引用模型<sup>[5-6]</sup>,用户在浏览一个网页后按下“返回”按钮;Web 拓扑结构法<sup>[7-10]</sup>,以登录等功能页面来标记会话分界点,进而确定日志记录的会话边界。这 4 种会话识别方法不足之处在 3 个方面:1) Timeout 法中的域值受访问内容和用户习惯等多种因素影响,难于统一;2) 参考域伪造和缓存严重干扰参考域会话边界识别;3) 标记分界法过分依赖选择的页面且缺少通用性。

会话的访问路径主要依靠 Web 日志记录的参考域和请求域关系来确定。但因各级缓存的影响,导致 Web 日志记录的用户访问路径信息存在路径缺失的情况,为此需要进行路径补充。路径补充时依次扫描会话页面集中的每个页面,如果相邻页面间没有直接的引用关系,则认为使用过 Back 功能,然后将距当前请求最近的来源页面作为当前请求的前驱

收稿日期:2009-10-12;修回日期:2009-12-17。 基金项目:辽宁省教育厅基金资助项目(2009B140)。

作者简介:范纯龙(1973-),男,辽宁营口人,副教授,硕士,主要研究方向:信息安全、入侵检测;姜宏飞(1982-),女,辽宁葫芦岛人,硕士研究生,主要研究方向:信息安全、访问控制;李华(1987-),男,辽宁大连人,主要研究方向:信息安全、入侵检测。

节点补充到会话路径中;如果日志不完整,找不到请求的来源页面,则采用当前站点的拓扑结构来补充。这种方法存在两点不足:一是无法补充路径缺失长度大于2的情况;二是用户回退操作与目标请求之间发生的日志信息没有有效利用。

会话识别涉及问题的主要根源有两个方面,首先是缓冲区技术导致 Web 日志数据不完整,而预处理过程中进行的数据清洗又加剧了数据缺损;再有是分析技术中没有充分利用浏览器特性、日志记录和 Web 拓扑结构信息。如请求数据没有改变时服务器会返回 304 状态代码,这是主流浏览器设置 If-Modified-Since 头信息的结果,这些信息在数据清洗时都被废弃了。

因此,多角度的利用数据清洗中废弃的日志数据,减少信息损失;将图片、js 等文件数据引入拓扑结构中,丰富信息的表达形式,将为改进会话识别质量提供数据和结构基础。

## 2 会话识别过程的优化

### 2.1 确定会话页面集

目前实践中确定页面集的方法主要是 Timeout 方法,本文以相邻访问页面 Timeout 方法为基础,着重解决时间阈值受用户习惯和网站内容影响而难于统一的问题。

首先在清洗数据时保留了服务器返回状态码为 304 的页面请求,因为浏览器访问网站时,它会询问 Web 服务器该请求文件是否修改,若没有修改服务器返回的状态码就是 304;然后利用 Web 拓扑结构防范参考域伪造问题;最后利用统计理论中的 3 倍标准差原理,利用启发式规则动态设置访问相邻页面时的时间阈值,以适应客户访问行为和网站内容的不确定性。

页面集划分的具体过程分为以下 3 个步骤。

1) 页面请求的参考域如果不是网站内页面,说明可能是首次启动浏览器或由搜索引擎等其他途径进入 Web 站点,开始一个新会话,这里分析的页面请求包括服务器响应状态码为 304 的页面请求。

2) 页面请求的参考域引用关系与 Web 拓扑关系不相符,说明存在参考域伪造,开始一个新会话。

3) 设请求页面长度为标准页面长度的  $\beta$  倍,访问标准页面的平均期望时间为  $t_0$ ,用户已访问页面的标准差为  $\delta$ ,则依据概率中 3 倍标准差原理定义相邻页面间隔  $\Delta t = \min\{\beta * t_0 + 3\delta, 30\}$ 。实验中,访问 10 KB 页面的  $t_0$  值一般设置为 2 min,  $\Delta t$  的变动范围通常为 5 min ~ 15 min。

### 2.2 改进路径补充算法

现有的路径补充算法没有充分利用 Web 拓扑结构和图片日志等辅助信息,为改进路径补充算法,首先在数据准备时,保留状态码为 304 的日志作为辅助数据,包括图片、样式表、小程序等日志记录数据,仅舍弃错误数据,这样做的本质是根据处理目的地变化进行多层次的数据清洗,而不是一次完成数据清洗工作;其次,扩展 Web 拓扑结构,仍然以页面拓扑结构为基础,但将图片、小程序等辅助数据也引入拓扑结构中,因为辅助数据不包含链接,所以它们彼此间没有关系,而仅能通过若干特定页面节点内的链接才能到达,这个包含了辅助数据的 Web 拓扑结构就是网站的扩展 Web 拓扑结构。

辅助数据的特点使得辅助数据的每个节点将同某些页面节点关联,对其访问也只能通过访问页面拓扑结构中与该辅

助节点相关的关联节点才能实现,这里将访问特定辅助数据节点的页面节点集合称为该辅助节点的关联集。如图 1 所示,页面拓扑图由节点  $\{A, B, C, D, E, F, G, H, K, T1, T2, T3, T4\}$  构成,辅助数据由节点  $\{P1, P2, P3, P4\}$  构成,对节点  $P1, P2, P3, P4$  的访问只能通过其对应的节点关联集  $\{A, D, T1\}$ 、 $\{D, F, T2\}$ 、 $\{E, H, T3\}$ 、 $\{C, D, E, T4\}$  实现。另外,某些样式表文件、广告配置文件等辅助数据都会设置成每次访问都刷新或固定时间间隔刷新的工作方式,从而避开缓存,而在 Web 服务器日志上留下访问痕迹,这些痕迹日志数据虽然不是页面文件日志,但可以从侧面反映用户访问中经过的可能页面。

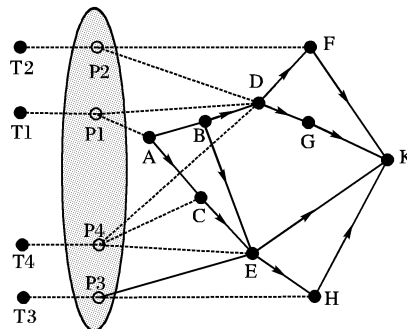


图1 路径补充拓扑结构

如图 1 所示,路径补充问题的一般示例描述如下,对于一个给定的请求页面,其参考域页面为 K,且日志中无 K 的日志记录,但存在 A 的记录,A 可以沿 Web 拓扑关系到达 K,其中,由 A 到 K 的路径长度小于给定阈值。一般情况下,从节点 A 到达节点 K 有多条路径,因此需要在这些路径中确定一个访问 K 的路径,并将确定的路径作为到达 K 的路径补充到会话路径中。目前实际中,阈值设置一般为 1,选择方法是选取时间最晚的日志记录对应的页面节点作为路径补充,即相当于从 E、F、G、H 中选择一个页面节点,但路径缺失步长一般很难确定,所以补充效果一般。

当缺失路径长度大于 1 时,如图 1,从 A 到 K 有 6 条可选路径,一般最短路径 ABEK 是最佳选择。如果在该用户的 Web 日志中,A 的日志记录后面出现了 P3 的日志记录——该记录可能是因为用户访问页面 T3 导致的对 P3 的访问,且服务器状态码为代表正确传送文件的 200,则说明用户以前没有访问过 P3 的关联集  $\{E, H, T3\}$  中的任何一个节点,而路径 ABEK 因包含节点 E,故不是一个合理的补充路径;所以可能的路径就只能是 ABDFK 和 ABDGK 两条了,若在该用户的 Web 日志中还有 P2 的日志记录,且服务器状态码为代表文件未修改的 304,则说明用户以前通过访问 P2 关联集  $\{D, F, T2\}$  中的某个页面请求过 P2 文件,那么在这两条路径中,应该优先选择包含节点 F 的路径,即路径 ABDFK。

如上例,利用数据清洗中废弃的辅助数据,可以更好地选取补充路径。这种作用主要有两个方面,一是删除部分不可能路径,二是可对多个候选路径进行评价,这两方面都会改善路径补充的质量,是对时间因素和最短路径因素的有益补充。

为描述算法,对需要补充路径的请求作如下假设,SL 为候选路径集合,path 为 SL 中的一条路径,PS(path) 为路径所包含的页面节点集合;SP 为 SL 中所有路径相关的页面节点集合的并,W 为与处理用户相关的日志记录集合,log 为一条日志记录,P(log) 为该日志对应的节点,S(log) 代表其状态码;x 为辅助页面节点,则  $G(x)$  代表由扩展 Web 拓扑得到的

该节点对应的关联集。则补充路径的选择算法为:

1. begin
2. for each  $\log \in W$
3. if  $P(\log)$  为页面节点 then continue;
4. if  $S(\log) = 200$  then  $SP = SP - G(P(\log))$ ;
5. if  $S(\log) = 304$  then
6. for each  $\text{page} \in SP \wedge \text{page} \in G(P(\log))$
7. page 的引用次数加 1;
8. next page;
9. next log;
10. for each  $\text{path} \in SL$
11. if  $PS(\text{path}) - SP \neq \emptyset$  then  $SL = SL - \{\text{path}\}$
12. else  $SUM(\text{path 上节点的引用次数})$ ;
13. next path;
14. 保留  $SL$  中的最短路径集;
15. 保留  $SL$  中  $\text{path}$  引用次数最高的路径集;
16. if  $SL = \Phi$  then 原请求为参考域伪造;
17. else 任选  $\text{path} \in SL$  为结果路径;
18. End;

上面的算法描述中,2~9步是利用辅助数据删除不可能经过节点和计算页面节点的访问频度;9~13步删除不可能路径,并计算每个路径中包含的所有节点的引用次数总和,用该和值代表该路径可能的访问频度;14~17步完成结果路径的选择。

### 3 实验分析

实验数据来自实验室网站的数据,网站地址为 [www.redccp.com](http://www.redccp.com),服务器日志采用扩展日志格式。网站有 621 个页面和 350 个图片等辅助文件资源,Web 日志的日数据量约 2000 多条,因为访问人员类型简单,访问量较低,网络爬虫访问数据约占总访问量的 65% 左右,即日均人为访问约 700 行记录。

因为主流浏览器的缓存默认配置为三天数据,故实验中选取从 2009 年 7 月 6 日到 2009 年 7 月 11 日的 Web 日志数据进行分析,共收集日志记录 12436 条,剔除爬虫访问和错误数据后,有效的待分析数据 5012 条,共区分不同用户 139 个。

实验内容包括用户会话页面集的确定和访问路径关系重建两个方面。确定页面集将基本 Timeout 分割法和 3 倍标准差法对比,前者收集会话 264,最大页面数量 25,后者收集会话 231,最大页面数量 33,二者交集为 213,其中页面数量 < 3 的会话交集为 187,即交集部分主要是访问量较短的会话,另外基本法将同一会话分割为多个会话的情况相对严重;访问路径重建中,首先采用标准差法确认页面集,然后将 1 步路径补充法和文种提出的多步路径补充法相比较,这里多步补充限定路径最多补充 3 步。前者需要补充路径 16 条,后者需要补充路径 22 条,其中补充路径长度大于 1 的 5 条,另有 1 条参考域伪造路径,是伪装爬虫的访问数据,在数据清洗时并没

有正确清除。结果反映二者的差别主要是前者对于补充路径长度大于 1 的情况自动分割为两个不同的会话,所以补充路径数量反而较少。

### 4 结语

会话识别是数据预处理工作的关键所在,本文从确定页面集和路径补充两个方面论述改进会话识别的方法。首先,辅助数据作为 Web 日志记录内容的重要组成部分,包含了用户访问信息和系统设置信息,对其利用是有客观的数据基础的;其次,3 倍标准差法在确定页面集时,更多地考虑了用户访问习惯对 Timeout 取值的影响,对于个别长文特别明显,降低了会话被错误切分的情况;再有,改进的路径补充算法大大改善了长路径缺失的补充效果。后续工作主要是更好地利用超文本协议和浏览器特性改善会话识别和相关数据挖掘的效果;在实际网站应用中,网站规模对扩展 Web 拓扑结构的复杂性影响巨大,加之日志信息规模的快速膨胀,如何改善算法的时间空间效率将是关键问题。

#### 参考文献:

- [1] 方元康,胡学钢,夏启寿. Web 日志预处理中优化的会话识别方法[J]. 计算机工程,2009,35(7):49-51.
- [2] 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.
- [3] 方成效,袁可风. Web 日志挖掘的数据预处理研究[J]. 计算机与现代化,2006,23(4):79-81.
- [4] 丁国栋,王斌,白硕. Web 超链挖掘:中国境内 Web 图结构研究[J]. 计算机工程,2005,31(14):24-26.
- [5] 陈子军,王鑫昱,李伟. 一种 Web 日志会话识别的优化方法[J]. 计算机工程,2007,33(1):95-97.
- [6] 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.
- [7] KHASAWNEH N, CHAN CHIEN-CHUNG. Active user-based and ontology-based Web log data preprocessing for Web usage mining [C]// Proceedings of the 2006IEEE/WIC/ACM International Conference on Web Intelligence. New York: ACM,2006:325-328.
- [8] CIMIANO P, HOTH O A, STAAB S. Learning concept hierarchies from text corpora using formal concept analysis[J]. Journal of Artificial Intelligence Research,2005,24(1):305-339.
- [9] BANKO M, CAFARELLA M, SODERLAND S, et al. Open information extraction from the Web[C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc.,2007:2670-2676.
- [10] SUCHANEK F M, IFRIM G. Combining linguistic and statistical analysis to extract relations from Web documents[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM,2006:712-717.

(上接第 1055 页)

- [5] NOLL J, CALVET J, MYKSVOLL K. Admittance services through mobile phone short messages [C]// Proceedings of the International Multi-Conference on Computing in the Global Information Technology. Washington, DC: IEEE Computer Society, 2006:77.
- [6] RONGYU H, GUOLEI Z, CHAOWEN C. A PK-SIM card based end-to-end security framework for SMS[J]. Computer Standards and Interfaces, 2009,31(4):629-641.
- [7] 孙亮,张来顺,赵国磊. 移动警务安全短消息通信系统设计[J]. 计算机技术与发展,2006,16(3):173-175.
- [8] 施寒潇,吕强. 一个 SMS 增值应用系统[J]. 计算机工程,2003,29(15):128-130.
- [9] 贾宏宇,赵俊峰. 短消息平台的原理与设计[J]. 小型微型计算机系统,2003,24(5):819-824.