第35卷 第10期 计 算 机 工 程 2009年5月
Vol.35 No.10 Computer Engineering May 2009
·软件技术与数据库· 文章编号：1000—3428(2009)10—0073—02 文献标识码：A 中图分类号：TP393

# 改进的 Web 日志数据预处理技术

方元康[1]，胡学钢[2]，夏启寿[1]，朱 勇[1]
(1. 池州学院计算机中心，池州 247000；2. 合肥工业大学计算机与信息学院，合肥 230009)

**摘 要：**介绍 Web 日志挖掘预处理过程中一些特殊情况的处理方法，在会话识别阶段给出一种基于过滤框架网页与会话重组相结合的会话识别方法，在会话识别之前进行框架网页的过滤，根据传统的会话识别方法构造初始会话集，使用会话重组算法对初始会话集进行优化。实验结果显示，该方法提高了会话识别的质量。

**关键词：**Web 日志挖掘；数据预处理；frame 页面；会话识别

# Advanced Data Preprocessing Technology for Web Log

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

**【Abstract】** This paper introduces some special processing methods of preprocessing in the field of Web log mining. The algorithm based on frame page filtering and session identification is proposed at the stage of transaction recognition. Before identification, frame pages are filtered, and initial session sets are constructed by traditional method of session identification. The quality of session set is improved using a method of union and rupture. Experimental results show the algorithm enhances the quality of session identification.

**【Key words】** Web log mining; data preprocessing; frame page; session identification

一般来说，Web 日志挖掘[1]由 3 个阶段构成：数据预处理，模式发现和模式分析。作为模式发现的数据源，数据预处理的质量直接影响模式发现的最终结果。一个好的数据源不仅能够发现高质量的模式，而且还能提高 Web 日志挖掘的性能。因此，数据预处理是整个 Web 日志挖掘的基础，是 Web 日志挖掘质量保证的关键。本文讨论 Web 日志挖掘预处理的研究现状及方法，并在此基础上对用户会话构造算法进行了相应改进。

## 1 数据预处理流程

结合数据挖掘中遇到的问题，可以把预处理过程分为数据清洗、用户识别、会话识别、路径补充等步骤。

### 1.1 数据清洗

数据清洗是整个数据挖掘工作的基础，在任何形式的 Web 日志分析过程中，清除服务器日志中不相关的数据都是非常重要的。数据清洗包括删除一些对于分析没有意义的数据，去掉访问中出错的记录、用户请求方法中不是 GET 的记录。然后将处理后的数据导入关系数据库中，再进行进一步的知识发现[2]。

### 1.2 用户识别

由于本地缓存、代理服务器和防火墙的存在，使得有效识别用户的任务变得十分复杂。一般采用的方法是基于日志站点的方法，还可以使用一些启发性规则。例如：如果 IP 地址相同，但是代理信息变了，表明用户可能是在某个防火墙后内网的不同用户，则可以标记为不同的用户；还可以将访问信息、引用信息字段和站点拓扑结合，构造出用户的浏览路径，如果当前请求的页面同用户已浏览的页面没有链接关系，则认为存在 IP 地址相同的多个用户。使用这些规则并不能保证准确识别用户，因此，用户识别问题是个难点。

### 1.3 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同一次访问请求识别出来。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。

传统会话识别方法的不足之处在于：(1)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，因此会导致会话识别的效率大大降低；(2)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。按上述方法生成的会话集中的不真实的成分太多，将会直接影响 Web 日志的挖掘效果。

### 1.4 路径补充

在识别用户会话过程中的另外一个问题是确定访问日志中是否有重要的请求没有被记录，这就需要路径补充来完成。如果当前请求的页面与用户上一次请求的页面之间没有超文本链接，那么用户很可能使用了浏览器上“BACK”的功能调用缓存在本机中的页面。检查引用信息确定当前请求来自哪一页，如果在用户的历史访问记录上有多个页面都包含与当前请求页面的链接，则将请求时间最接近的作为当前请求

**基金项目：**国家自然科学基金资助项目(050504F)；安徽省教育厅自然科学基金资助项目(XK0829, KJ2008B45ZC) 池州学院自然科学基金资助项目(2007XJ015)
**作者简介：**方元康(1968－)，男，讲师、硕士研究生，主研方向：数据挖掘；胡学钢，教授、博士；夏启寿、朱 勇，讲师、硕士
**收稿日期：**2008-11-11 **E-mail：**fyk80@163.com

的来源，如果引用信息不完整，则可利用站点的拓扑结构来代替。

## 2 基于过滤框架网页与会话重组的会话识别算法

针对传统的会话识别的缺陷，本文提出一种基于过滤框架网页与会话重组相结合的会话识别方法。该算法分为 2 步：(1)在完成具体的用户识别之后，过滤掉大量的框架网页，再使用页面固定时间阈值的方法产生初始会话集；(2)使用会话重组算法对产生的初始会话集进一步地进行优化。实验数据显示，与传统的会话识别算法相比，该方法所得到的会话集更具有真实性，同时也提高了会话识别的效率。

### 2.1 过滤框架页面及初始会话集的产生

HTML 规范通过“Frame”标记支持多窗口页面[3]，每个窗口里装载的页面对应一个 URL，Subframe 页面同时又可以是包含子窗口的 Frame 页面。例如，若 A, B, C 这 3 个页面的关系是 A 包含 B、B 包含 C，则 A 是 Frame 页面，B, C 是 Subframe 页面，但 B 既是 Frame 页面，也是 Subframe 页面。Frame 页面与其 Subframe 页面总是一起出现在用户会话中。在对 Web 服务器日志进行挖掘试验时，大量的 Frame 和 Subframe 页面使得会话识别的结果偏离了真实性。Web 日志挖掘的目的是发现未知的用户行为模式，而 Frame 页面和 Subframe 页面的关系是已知的事实，因此，应当消除 Subframe 页面对会话识别的影响。Frame 页面和其中的 Subframe 页面作为一个多窗口页面展现在用户面前，因此，在数据预处理阶段将 Frame 页面和其中的 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。从全局而言，这样处理可以有效地消除 Subframe 页面对会话识别的影响，最终提高挖掘结果的兴趣性。

在完成用户识别及过滤框架页面以后，使用页面固定时间阈值的方法构造初始会话集 $H$。具体做法如下：给用户 1 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$\triangle t$，则属于同一会话；否则分属于 2 个会话。$\triangle t$ 一般取 10 min。

### 2.2 初始事务会话集合 $H$ 的进一步优化

在初始会话集中，可能存在这样的情形：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。因此，要对这些不真实的会话进行连接与拆分，使它们尽可能地接近真实会话。

#### 2.2.1 会话的连接与拆分

在会话识别中，可能存在实际会话$<L_1,\cdots, L_i, L_j,\cdots, L_n>$(其中 $L_1, L_2,\cdots, L_n$ 都是记录)被划分成$<L_1,\cdots,L_{i-1}, L_i>$和$<L_j,\cdots, L_{n-1}, L_n>$2 个事务会话。由于 $L_i, L_j$ 在一个实际会话里，表明用户还没有转向另一个话题，或用户还没有离开站点。简单地说，就是在网站的拓扑结构中，从 $L_i$ 到 $L_j$ 有直接或间接的链接。基于上述事实，在对会话进行优化时，如果遇到会话边界 $L_i$ 和 $L_j$(形式为$<\cdots, L_i>$和$<L_j,\cdots>$)，则分 2 种情况考虑将 $L_i$ 和 $L_j$ 连接成一个会话。

在会话识别中，也可能有 2 个实际会话$<L_1,\cdots L_{i-1}, L_i>$和$<L_j,\cdots L_{n-1}, L_n>$ (其中 $L_1, L_2,\cdots, L_n$ 都是记录)被划分成$<L_1,\cdots, L_i, L_j,\cdots, L_n>$一个会话。$L_i, L_j$ 虽然是一个用户在服务器上记录的 2 条连续记录，但是由于不在同一个实际会话里，表明用户已经转向另一话题。该用户可能是通过一定量的后退，也可能是直接输入网址等方法到达记录 $L_j$ 的页面。基于上述事实，在优化时，对于会话内部记录 $L_i$ 和 $L_j$(形式为$<\cdots, L_i, L_j, \cdots>$)，如果通过 $L_i$ 或 $L_i$ 前面的 $L$ 条记录不可以寻迹到 $L_j$，则将 $L_i, L_j$ 拆分。

综上所述，通过连接和拆分 2 种操作对会话进行优化，能使所得到的会话更接近实际会话。

#### 2.2.2 会话重组的算法描述

会话重组的算法描述如下：

(1)输入初始会话集 H, $H=\{h_1, h_2,\cdots, h_n\}$, $h_i=\{L_1, L_2,\cdots, L_m\}$ //$h_i(1\le i\le n)$是属于 H 中的一个会话，L 是属于某一个会话 $h_i$ 中的 //一条记录。

(2)依次读入会话集 H 中的会话，以及依次读入同一个会话的每一条记录 L

(3)DO CASE

(4) CASE $L_i, L_j$ 是 2 个连续会话中的尾记录 $L_i$ 和首记录 $L_j$

(5) IF 模式 $L_i$->$L_j$ 是该用户经常访问的模式(即频繁访问模式)

(6) 将 $L_i, L_j$ 连接 //合并了 2 个连续的会话

(7) ELSE

(8) IF 在网站的拓扑结构中，$L_i$ 或 $L_i$ 前面的 L 条记录能链接到 $L_j$

(9) 将 $L_i, L_j$ 连接 //合并了 2 个连续的会话

(10) ELSE

(11) $L_i, L_j$ 仍然是 2 个连续会话中的尾记录 $L_i$ 和首记录 $L_j$

(12) ENDIF

(13) ENDIF

(14) CASE $L_i, L_j$ 是同一个会话中的 2 条连续记录

(15) IF $L_i, L_j$ 的时间间隔>设定的阈值&&在网站的拓扑结构中， $L_i$ 或 $L_i$ 前面的 L 条记录不能链接到 $L_j$ //若 $L_i, L_j$ 分属 2 个会话，则 $L_i, L_j$ 间隔时间较长。

(16) 将 $L_i, L_j$ 拆分 //$L_i, L_j$ 被拆分到 2 个会话中

(17) ELSE

(18) $L_i, L_j$ 仍是同一个会话中的 2 个连续记录 $L_i$ 和 $L_j$

(19) ENDIF

(20) ENDDO

(21)输出优化后的会话集 H’

## 3 实验数据及分析

本文的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-2007 年 11 月 3 日。实验中将基于引用会话识别算法、基于固定的先验阈值(10 min)的会话识别算法、基于过滤框架网页与会话重组的会话识别算法 3 种会话识别算法进行比较。

本文使用目前常用的评价标准[5]：会话被算法 $h$ 完整重建的程度。一般使用精确度和查全度这 2 个指标来衡量重建程度。精确度是用完全构建的会话数目与构造生成的总会话 $Rh$ 数目的比值表示：$precision(h) = |Rh\cap R|/|Rh|$，查全度是完全构建会话数目与真正的会话 $R$ 数目的比值表示：$recall(h) = |Rh\cap R|/|R|$。在对各种算法进行比较时，以基于引用的方法为基准。实验数据如表 1 所示。

**表 1 各种会话识别方法的比较结果**

| 会话构造方法 | 有效页面数 | 会话数 | 会话交集数 | 精确度/(%) | 查全度/(%) |
|---|---|---|---|---|---|
| 基于引用 | 135 086 | 26 809 | 26 809 | 100.000 | 100.000 |
| 基于固定时间阈值 | 135 086 | 45 702 | 12 309 | 26.933 | 45.913 |
| 基于过滤框架网页与会话重组 | 95 904 | 57 406 | 20 208 | 35.202 | 75.378 |

由表 1 可知，基于过滤框架网页与会话重组的会话识别算法的精确度与查全度都比基于固定时间阈值的会话识别算法的高。

(下转第 77 页)