基于引用启发式和 URL 语义相结合的会话识别方法*

张 帅, 陈兴蜀, 童 浩, 崔晓靖
(四川大学 计算机学院, 成都 610065)

摘 要: 会话识别是 Web 日志的用户行为分析的关键步骤,精准的会话识别是有效进行用户行为分析的基础。已有的会话识别方法不能有效地动态适应不同的用户(如多 IP 单用户、单 IP 多用户)行为,在 Web 日志分析的基础上,提出了一种基于引用启发式和 URL 语义相结合的会话识别方法。实验结果表明,改进后的会话识别方法能更有效地识别出用户的真实会话。

关键词: Web 日志挖掘; 会话识别; 数据预处理; 引用启发式; URL 语义
中图分类号: TP311 **文献标志码:** A **文章编号:** 1001-3695(2014)01-0102-04
doi:10.3969/j.issn.1001-3695.2014.01.024

Methodology for session identification based on combination of referenced heuristic and URL-semantic

ZHANG Shuai, CHEN Xing-shu, TONG Hao, CUI Xiao-jing
(*College of Computer, Sichuan University, Chengdu* 610065, *China*)

Abstract: Session identification is an importent step in data preprocessing of Web log mining, the improvement of the accuracy of the session provides accurate and reliable data for user behavior analysis. Existing session identification methods can't adapt to the different users (eg, multi-IP-single-user, single-IP-multi-user) behavior dynamically. Based on Web log analysis, this paper presented a method of based on the combination of referenced heuristic and URL-semantic for session identification. Experimental results show that the presented approach can more effectively identify the user's session than the traditional method.

Key words: Web mining; session identification; data pre-processing; referenced heuristic; URL-semantic

0 引言

随着网络的日新月异,Web 日志挖掘已成为数据挖掘技术中越来越受重视的领域之一,挖掘中的预处理技术也变得非常重要。据统计,三分之二的数据挖掘分析家们认为在一个完整的数据挖掘过程中,预处理在时间上占据整个日志挖掘过程的 60% 以上[1],而且预处理的结果直接影响日志挖掘的质量,因此对于 Web 日志挖掘来说对数据预处理技术的研究很有意义。数据预处理一般包括数据清理、用户识别、会话识别和路径补全等。数据预处理的结果将直接影响 Web 日志挖掘的有效性和精准度,而会话识别的目的是将用户的所有访问序列分成多个单独的用户一次访问序列,会话的真实性和精准度是衡量预处理质量的重要指标,因此,会话识别是 Web 日志挖掘的重要处理步骤之一[2]。

用户会话直观地表述为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访问活动。目前国内外学者提出了多种用户会话的识别方法。最简单的会话识别是基于时间间隔的会话切分方法,这种方法给每次会话持续时间设置一个阈值,超出该阈值时间段的请求,即视为一个新的会话。会话时间的选择往往根据经验数值,如 30 min[3,4]、25.5 min[5]。文献[6]在固定时间阈值的基础上,通过分析日志样本得到不同时间段的动态时间阈值计算方法。文献[7]根据网站的拓扑结构及 cache(缓存)机制提出了 STT(session to transaction)算法识别会话。文献[8,9]提出了三种启发式的会话识别方法。文献[10]提出了一种利用最大前向引用将会话切分成事务的方法。文献[11]则用立方结构来描述网络日志数据,进而进行数据挖掘和在线数据处理。文献[12]提出了一种基于 URL 语义分析的用户会话识别方法,是基于良好组织的 Web 目录。文献[13]采用伪代码的形式概述了会话识别处理的过程,但没有涉及具体的会话识别方法。

这些方法虽然在一定程度上对会话进行了识别,但仍存在一些问题:a)不能很好地解决由于缓存的存在、网站拓扑信息错误量化(特别是网站架构在分布式系统上)等造成识别的准确率和精准度的下降;b)不能因正确有效地识别用户(如多 IP 单用户、单 IP 多用户)而造成会话质量的下降。然而正确有效地对 Web 日志文件进行会话识别,不仅可以为后续挖掘提供准确的数据源,有利于后续的挖掘算法分析,而且对于最终形

收稿日期: 2013-04-02; **修回日期:** 2013-05-12 **基金项目:** 国家自然科学基金面上资助项目(61272447)

作者简介: 张帅(1986-),男,四川巴中人,硕士研究生,主要研究方向为数据挖掘、网络安全(shuaikobe@163.com);陈兴蜀(1968-),女,教授,博士,主要研究方向为信息安全、计算机网络;童浩(1971-),男,硕士研究生,主要研究方向为信息安全、计算机网络;崔晓靖(1989-),女,硕士研究生,主要研究方向为数据挖掘、信息安全.

成准确可靠的用户行为模式也是极为重要的。会话识别的主要任务是从 Web 日志中识别出真实的会话,因此有效的会话识别决定了预处理的最终效果。因此,本文提出了一种基于引用启发式-URL 语义相结合的会话识别方法,在完成对 Web 日志数据清理和用户识别的基础上,根据引用启发算法生成 map session 列表,最后利用 URL 语义实现同一时间段邻域相关性会话的聚集。

1　本文的会话识别方法

1.1　会话识别整体框架

由于用户的个体差异,会话识别应根据个体不同在不同时间段内采用不同阈值划分会话。常用的 Web 用户会话识别方法,如时间阈值(timeout)、STT、单纯的引用启发式等,但是这些方法没有考虑用户的个体差异、个体行为,不能动态适应用户的随机访问;另外,用户在一定的时间内访问的网页具有一定的相关性。基于此,本文提出了一种基于引用启发式-URL 语义的会话识别方法。会话识别的主要步骤如图 1 所示,主要包括 map session 和 reduce session。Map session 是指根据时间阈值划分后的日志,在阈值内生成的会话;reduce session 是指对相邻时间内的会话合并生成的会话。Map session 根据固定的阈值划分日志,在划分后的日志中根据引用页面划分会话,对于阈值间的相邻会话进行连接、去重等操作生成若干个 session,而 reduce session 则利用 URL 语义将相关的 session 合并成一个 session。

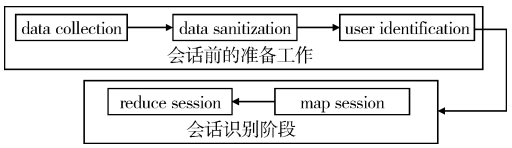

图 1　会话识别的主要步骤

1.2　会话识别前的准备工作

Web 日志挖掘的数据预处理的目的就是剔除网络日志中对挖掘过程无用的属性及数据,并将网络日志数据转换为挖掘算法可识别的形式保存。一般的网络日志数据挖掘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几个阶段。在会话识别前,Web 日志挖掘的数据预处理主要包括数据清理和用户识别两个步骤。

1.2.1　数据清理

数据清理的目的主要是为了删除 Web 日志中与挖掘算法无关的数据。由于用户访问时一般不会显式地请求页面上的图形文件、css 等,这些文件是根据 HTML 的超文本引用标记浏览器自动下载的。Web 日志挖掘的目的是获得用户的行为模式,这些文件与挖掘的用户行为无关,因此将 Web 日志中请求的 URI 后缀如 jpg、css、ico、gif、png 等对应记录删除。

1.2.2　用户识别

用户识别是将用户和请求页面相关联的过程,它是会话识别的基础。本文根据客户端 IP 地址、引用站点 URL、用户代理来识别,具体的规则如下:

a)首先判断用户的 IP 地址,不同的 IP 地址代表不同的用户;

b)当 IP 地址相同时,可以认为不同的操作系统或浏览器代表不同的用户;

c)在 IP 地址相同,用户使用的操作系统和浏览器也相同的情况下,则可以根据引用站点 URL 对用户进行识别:如果引用站点 URL 不能从已访问的任何页面到达,则认为这是一个新的用户。

上述规则并不能准确地标志用户,它主要用于 map session 阶段的用户标注。在规则 b)和 c)中,它能将单 IP 的用户暂时划分成多个不同的用户,虽然在 map session 阶段生成不同的会话列表,但在 reduce session 阶段会进行会话的合并操作,它会将同一时间段内的相似会话合并为同一会话,因此会话质量会进一步精确地提高。

1.3　Map session 阶段

与已有的基于引用启发式算法不同,map session 首先根据时间阈值将日志分块,在每个块中根据引用站点 URL 以及客户端 IP 地址的拓扑邻近性划分会话,最后对相邻的会话进行连接、去重操作。这里的 IP 地址的拓扑邻近性是指同一物理位置中 IP 地址间的相邻性。

定义 1　用户访问记录(UVR)。UVR = 〈URI, ReferURI, Date〉,其中 URI 表示请求页面,ReferURI 表示用户浏览的上一页面,即引用站点 URL,Date 表示请求时间。它与 UID 合称为一条访问日志记录(raw log),这里 UID 指的是用户标志,用于用户集合的划分。

定义 2　用户集合(US)。设用户在一定的时间内在日志中留着连续 n 条记录,$US(i) = \langle UID, tS, tE, UVR_1, UVR_2, \cdots, UVR_n \rangle$,这里 i 对用户集合进行标注,并随时间流增加而增加,$tS = UVR_1.Date$,$tE = UVR_n.Date$,$0 < tE - tS \leqslant \delta$,$\delta$ 为程序设定的时间阈值。

定理 1　对于 $\forall i$、$\forall j$,如果 $US(i).UID = US(j).UID$ 且 $i \neq j$,当 $US(i).tS < US(j).tS$,则 $US(i).tS < US(i).tE < US(j).tS < US(j).tE$。

证明　根据定义 2 可知,用户集合是按照时间阈值划分的。当 $i \neq j$,则

$$US(i) \cap US(j) = \varnothing$$

又

$$US(i).tS < US(j).tS$$

$$US(i).tE < US(j).tS$$

所以 $US(i).tS < US(i).tE < US(j).tS < US(j).tE$

定义 3　用户会话集合(USS),$USS = \langle USID, UVR_1, \cdots, UVR_n \rangle$,其中 USID 为用户会话集合标志符,$n$ 为 USID 连续访问 Web 记录于 Web 日志的连续条记录数。USS 满足如下规则:

a)$UVR_1.Date < \cdots < UVR_n.Date$;

b)对于 $\forall \kappa$,如果 $UVR_\kappa \in USS_i$,则 $UVR_\kappa \notin USS_j (i \neq j)$;

c)由于垃圾访问记录的存在,即被清理的日志记录,UVR 不一定在用户会话中。

Map session 分为两个阶段,第一阶段是阈值内的会话划分,第二阶段是阈值间相邻会话的连接、去重。首先对 Web 日志每隔 δ 进行分割。然后根据 IP 地址、操作系统版本和浏览器信息,得到用户集合 US。得到 US 后,根据引用站点 URL 进行用户会话识别,生成阈值内的 map session。接着对阈值间相邻会话进行连接、去重,具体的规则如下:

a)如果用户访问的网页 UVR_i、URI 为网站主页,且满足下列条件之一,则属于当前会话:

(a)$UVR_i.ReferURI \neq \varnothing$;

(b) UVR_i. ReferURI = ∅, 且 $\exists UVR_{i-1}$ (UVR_i. Date − UVR_{i-1}. Date$\leqslant\delta$);

b) 如果用户访问的网页 UVR_i. URI 不为网站主页,且 UVR_i. ReferURI $\in$ { URI | UVR_{κ_1}. URI, …, UVR_{κ_n}. URI }, $0 \leqslant UVR_{\kappa_n}$. Date − UCR_{κ_1}. Date$\leqslant\delta$, $\kappa_n = i-1$,则属于当前会话。

c) 当用户访问网页的引用页面 UVR_i. ReferURI 不存在,且 $\nexists\, t$ 满足 UVR_i. Date $< \delta$,使得 t 对应于一条用户访问记录 UVR,则根据 IP 地址决定 UVR_i 的归属问题;如果存在拓扑结构相邻的 IP(通过 IP 所属的地域位置、默认网关等信息判断 IP 的邻近性,实验部分用到了四川大学网上调查(http://my-dc. scu. edu. cn/)的数据库,该数据库中保存了每个校园 IP 所处的地域位置、默认网关等信息),它所对应的用户访问记录 UVR_k,使得 UVR_i. Date − UVR_k. Date $< \delta$,其中 UVR_k. URI = UVR_i. URI,则将 UVR_i 归为 UVR_k 所在的会话中。

d) 如果不满足上述三种情况,则新建一个 map session。

上述规则的优点在于它能根据用户访问的时间快速生成阈值内的会话。对于相邻阈值间的同一用户会话,如果后一会话中访问记录的引用页面在前一会话的用户访问记录中,则连接为同一会话。对于生成完全相同的用户会话,只保留其中一个。

1.4 Reduce session 阶段

由于 map session 是按照 δ 和引用页面划分的,这样就存在同一个会话被划分成若干个具有相同用户 ID 的会话;另外,由于多个用户使用代理服务器,在 map session 中,简单地将相同用户 ID 划分成一个会话,也是不现实的。下面则使用基于 URL 语义的方法将相同 ID 的不同 map session 会话重新划分。

在大多数网站中,考虑到信息的组织、查阅和检索的高效性,大都是基于主题层级结构方式的 Web 目录组织网络[14,15]。因此,Web 目录可以用来描述一个页面的内容[16],表征 URL 的语义。

定义 4 Web 路径用于表示能够表征 URI 的 Web 目录,每个目录的长度为上一目录长度 + 1。URL 间的距离 Dis 定义如下:

$$\mathrm{Dis}(\mathrm{URI}_i,\mathrm{URI}_j)=\begin{cases}\alpha\times\tau\times\dfrac{\min((L_i-L_c),(L_j-L_c))}{\exp(L_c)} & \tau\leqslant 1\\ 1 & \tau>1\end{cases}\quad(0\leqslant\alpha\leqslant 1)\tag{1}$$

其中:$\tau=\dfrac{|\mathrm{URI}_i.\mathrm{Date}-\mathrm{URI}_j.\mathrm{Date}|}{\vartheta\times\delta}$ ($\vartheta>1$), L_i、L_j 和 L_c 分别代表 URI_i 的 Web 路径长度、URI_j 的 Web 路径长度以及两者共同路径的长度,α 为调节因子。式(1)中之所以考虑时间,是因为对于网站的访客来说,同一会话中 URL 的访问时间间隔保证在 $\vartheta\times\delta(\vartheta>1)$ 内,ϑ 的主要作用是为了识别更长的会话。显然,Dis 的取值范围为[0,1],值越小表示越匹配。两个会话(USS_1, USS_2)的距离矩阵定义如下(这里假设 USS_1 的长度大于 USS_2 的长度):

$$D^{(m,n)}_{\Delta(\mathrm{USS}_1,\mathrm{USS}_2)}=\begin{pmatrix}\mathrm{dis}_{1,1} & \cdots & \mathrm{dis}_{1,n} & \cdots & \mathrm{dis}_{1,m}\\ & \ddots & \vdots & \vdots & \vdots\\ & & \mathrm{dis}_{n,n} & \cdots & \mathrm{dis}_{n,m}\end{pmatrix}\tag{2}$$

其中:$\mathrm{dis}_{i,j}=\mathrm{Dis}(\mathrm{USS}_1.\mathrm{UVR}_i.\mathrm{URI},\mathrm{USS}_2.\mathrm{UVR}_j.\mathrm{URI})$, m、n($m\geqslant n$)分别表示两个会话的记录数,$i\in[1,m]$, $j\in[1,n]$。

为了计算两个会话的差异程度,定义两个会话的语义离散度为

$$\sigma_\Delta=\mathrm{argmin}\sum_{i=1,j\in[1,m]}^{n}\mathrm{dis}_{i,j},0\leqslant j-i\leqslant m-n,\frac{d_j}{d_i}>0\tag{3}$$
$$\text{s. t.}\qquad \sigma_\Delta<n\times\omega$$

这里假设,同一个 IP 地址,如果用户代理中浏览器版本或者操作系统是不同的,但它们的访问时间相邻并且 $\sigma_\Delta<n\times\omega$ (ω 是预先设定的阈值,该阈值可以根据实际情况进行设置),则认为是同一个用户会话,否则就不相同。σ_Δ 的计算算法如下。

```
输入:两个会话 s1、s2 的 URL 列表,且 n≤m。
输出:s1、s2 的离散度。
i,k,minscore = 0;
while(i < n)
    j = i;
    while(j < m)
        score s[i][j] = URIi 与 URLj 的离散值
        if j == i; then
            minscore += score s[i][j];
        end if
        j = j + 1;
    end while
    i = i + 1;
end while
if m > n;then
    while(k < n)
        minrow = 1.0 * (k + 1);
        j = k + 1;
        while(j < = m - n + k + 1)
            if minrow > socres[k][j - 1];then
                minrow = scores[k][j - 1];
            end if
            if k < n - 1 then
                scores[k + 1][j] += minrow;
            end if
            j = j + 1;
        end while
        k = k + 1;
    end while
    minscore = minrow
end if
RETURN minscore
```

2 实验与结果分析

对于会话识别方法的评价,本文采用如下的方法:设 α 表示真实的会话个数,β、γ 分别表示通过会话识别方法 π 识别出的会话个数及真实会话个数,其中方法 π 的真实会话个数的识别方法根据文献[17]的方法判定。定义:

$$A_\pi=\frac{\gamma}{\beta}$$
$$R_\pi=\frac{\gamma}{\alpha}$$
$$H_\pi=\frac{2}{\frac{1}{A_\pi}+\frac{1}{R_\pi}}=\frac{2A_\pi R_\pi}{A_\pi+R_\pi}\tag{4}$$

当 $H_\pi > H_\tau$,则表示方法 π 优于方法 τ。

实验数据来源于四川大学官方网站的 Web 日志(2012. 6.

17)。通过分析 Web 日志,会话时间都小于 20 min,大部分会话时间都小于 10 min。对于 6 月 17 日的 Web 日志,经过人工识别,产生的真实会话 $\alpha = 4\ 659$ 个,其中会话时间小于 10 min 的达到 4 285 个,占总会话个数的 91.97%。因此,将 map session 阶段时间阈值 δ 设为 10 min,式(1)中的 ϑ 是为了识别更长的会话,这里 $\vartheta = 2$,使得 map session 阶段能够识别更完整的会话。图 2 展示了 ω 分别取值 0.05、0.1、0.15、02、0.25、0.3 时识别的会话个数 β 及真实会话个数 γ 的变化趋势。从图 2 中可以发现,$\omega = 0.15$ 后,β、γ 的个数趋于平衡,不再发生变化。在 ω 较小时,由于 reduce session 阶段不会将 URL 语义相近的会话合并为一个会话,真实会话的个数会相对较低,因此总体趋势 β 的值随着 ω 的增大而减少,γ 的值随着 ω 的增大而增大。当 $\omega = 0.15$ 时,会话识别效果趋于稳定。已有的会话识别方法和改进后的方法的实验结果如表 1 所示,它们的 H_π 如图 3 所示。

表 1　实验结果

方法	会话个数				
	β	γ	A_π	R_π	H_π
时间阈值	6 616	4 193	63.37%	90.00%	74.38%
引用启发	4 525	3 907	86.34%	83.86%	85.08%
本文方法	4 290	3 996	93.15%	85.77%	89.31%

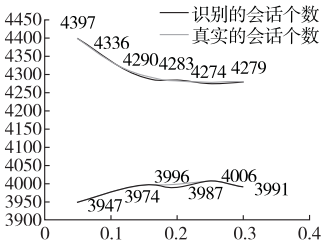

图 2　ω 不同取值 β 和 γ 的变化情况

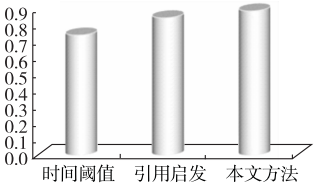

图 3　各种方法的 H_π 比较

实验表明:

a)本文的方法在精准度和识别率都有提高;

b)相对于固定时间阈值的会话识别,本文的方法可以识别出更长的会话,更长会话的生成是在 reduce session 阶段生成的,因此 β 大幅减少;

c)相对于引用启发的识别方法,本文的方法在 map session 阶段根据引用页面划分会话,然后通过 reduce session 将相关的会话合并,这样可以识别出多 IP 单用户、单 IP 多用户的会话,在一定程序上解决了因代理等造成的会话难划分的问题,进一步提高了会话的识别率和精准度。

3　结束语

本文从多环节、多方面对 Web 日志挖掘的数据预处理部分中的会话识别进行了研究。针对现有处理方法存在的缺陷,提出了引用启发式与 URL 语义相结合的方法。实验结果显示,此方法有效地提高了会话识别的精度,使得会话识别出来的结果更接近于用户的真实会话。Web 日志数据的可靠性是对 Web 日志挖掘及用户行为分析的重要前提和基础,而会话识别的质量是决定后续挖掘效率和质量的基础及关键,因此,提高会话识别的质量是至关重要的。本文针对多 IP 单用户、单 IP 多用户会话识别,较以前的识别方法有一定的提升。

由于数据的爆炸式增长,Web 日志挖掘更会成为研究的热点。本文只针对会话识别作了研究,因此,如何更有效地改进 Web 日志预处理中的其他技术环节,确保数据的正确性仍然是今后 Web 日志挖掘研究的重要方向。

参考文献:

[1] TANASA D, TROUSSE B. Advanced data preprocessing for intersites Web usage mining[J]. IEEE Intelligent Systems, 2004, 19(2): 59-65.

[2] HOFMANN T. Latent semantic models for collaborative filtering[J]. ACM Trans on Information Systems, 2004, 22(1): 89-115.

[3] ISHIKAWA H, OHTA M, YOKOYAMA S, *et al*. On the effectiveness of Web usage mining for page recommendation and restructuring [C]//Lecture Notes in Computer Science, vol 2593. 2003: 253-267.

[4] GOBINATH R, HEMALATHA M. Improved preprocessing techniques for analyzing patterns in Web personalization process[J]. International Journal of Computer Applications, 2012, 58(3): 13-20.

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

[6] 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.

[7] 马瑞民,李向云. Web 日志挖掘中数据预处理技术的研究[J]. 计算机工程与设计,2007, 28(10): 2358-2360.

[8] BERENDT B, MOBASHER B, NAKAGAWA M. The impact of site structure and user environment on session reconstruction in Web usage analysis[C]//Proc of WEBKDD. 2002: 159-179.

[9] SPILIOPOULOU M, MOBASHER B, BERENDT B. A framework for the evaluation of session reconstruction heuristics in Web usage analysis[J]. INFORMS Journal on Computing, 2003, 15(2): 171-190.

[10] CHEN M, PARK J, YU P S. Data mining for path traversal patterns in a Web environment[C]//Proc of the 16th IEEE ICDCS. 1996: 385-392.

[11] ZAIANE O R, XIN Man, HAN Jia-wei. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs[C]//Advances in Digital Libraries Conference. 1998: 19-29.

[12] 朱志国. 基于 URL 语义分析的 Web 用户会话识别方法[J]. 大连理工大学学报,2011, 51(3): 440-446.

[13] RAMYA C, SHREEDHARA K S, KAVITHA G. Preprocessing: a prerequisite for discovering patterns in Web usage mining process[C]// Proc of the 3rd IEEE ICIME. 2011: 730-734.

[14] McCALLUM A, NIGAM K, RENNIE J, *et al*. Building domain-specific search engines with machine learning techniques[C]//Proc of AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace. 1999: 28-39.

[15] JUNG J, YOON J. Collaborative information filtering by using categorized bookmarks on the Web[C]//Proc of the 14th International Conference on Applications of Prolog. Tokyo: The Prolog Association, 2001: 343-357.

[16] LABROU Y, FININ T. Yahoo! as an ontology: using Yahoo! categories to describe documents[C]//Proc of the 8th International Conference on Information Knowledge Management. 1999: 180-187.

[17] 石晶,龚震宇,裘杭萍. 评测 Web 使用分析中会话识别的准确度[J]. 电子科技大学学报,2002, 31(3): 281-285.