评测Web使用分析中会话识别的准确度

石 晶* 龚震宇 裘杭萍 张毓森
(解放军理工大学指挥自动化与计算机学院 南京 210016)

【摘要】目前用于用户会话识别的方法主要有两类：基于时限的会话识别和与拓扑结构(超链接)结合的会话识别，这两类方法都是在用户识别的基础上对用户活动作出猜测而得到的。该文提出了一套用于对这些启发式方法所获得的数据的准确程度进行量化的评测系统，不同的估测方法反映不同的数据挖掘应用的需要。最后通过一个实际站点的数据说明了评测系统的识别结果是准确的。

关 键 词 Web使用挖掘; 会话识别; 日志; 启发式规则
中图分类号 TP392

Measuing Accuracy of Sessionizers for Web Usage Analysis

Shi Jing Gong Zhengyu Qiu Hangping Zhang Yusen
(PLA University of Science and Technology Nanjing 210016)

Abstract This paper describes timeout-based sessionizing mechanisms and topology-aware heuristics which now used to identify user sessions. The Sessionizing tools are based on heuristic rules and on assumptions about the site's usage ,and therefore prone to error.The paper proposes a formal framework composed of a set of measures for the evaluation the accuracy of sessionizing tools.The different measures reflect the requitements of different web usage analysis applications.Experiment using the log data of a real web site shows the use of the measures.

Key words Web usage mining; identifying session; log; heuristic rule

Web使用记录挖掘，是将数据挖掘的技术应用于Web服务器上的日志文件，从而发现用户访问Web页面的模式。通过分析和探究Web日志记录中的规律，可以识别电子商务的潜在客户，增强对最终用户的因特网信息服务的质量和交付，并协助站点管理员优化站点。

Web使用挖掘的过程一般分为预处理阶段、挖掘算法实施阶段、模式分析阶段。其中，预处理阶段中的用户会话识别是信息分析的基础。用户会话是一个用户在一次访问请求的所有Web页面。会话识别是否准确直接决定了后续的数据挖掘处理结果是否有意义。由于本地缓存、代理服务器和防火墙的存在，使Web日志中记录的数据并不精确，增加了识别用户会话的难度。

目前，已经提出了一些技术来解决上述问题，如基于时限的会话识别机制和与拓扑结构(超链接)结合的启发式规则[1]。但是，上述方法都是启发式的，是在用户识别的基础上对用户的活动作出猜测。在利用数据挖掘技术作进一步的分析之前，首先必须对由这些启发式规则预处理获得的数据所存在的错误进行量化。针对这个问题，本文提出了由多种评测方法组成的一套评测系统。

1 用户识别与用户会话的重建

Web使用挖掘不需要了解用户的身份，但需要区分出不同的用户。最常用的解决方法是使用

2002年1月18日收稿
* 女 26岁 在职博士生

cookies。但是cookie仅能为用户识别服务，并不能识别一次会话的终点和下一次会话的起点(即用户会话的“边界”)，也不能推测出用户在一次会话中的所有活动。

本文着重讨论有关用户会话的重建问题，即推测用户从进入到其离开站点期间所进行的一组活动，讨论时假定已经通过可靠方法完成用户识别，因而不再考虑由于用户识别错误给用户会话识别带来的影响。下面是对本文中几个常用语的解释：

用户会话(访问)：用户从进入到其离开站点期间所进行的一组活动[2]。由于一个用户可能多次访问一个站点，Web日志文件中记录了关于一个用户的多次会话。

用户活动日志(user activity log)：同一个用户的一组顺序记录的活动日志。

会话化(sessioning)：将每个用户活动日志分割成会话的过程。

启发式会话化(sessionizing heuristic)：在对用户行为的假想基础上对会话的推测。

1.1 启发式会话识别

启发式会话识别将用户活动日志分割成一组“构造会话”集。“真实会话”是用户一次实际访问活动的集合，试验数据由测试站点的Web服务器提供。评测方法通过对构造会话与真实会话之间的差距进行量化来评估会话识别方法的质量。

启发式会话识别可以分为两类：面向时间和面向引用。面向时间的启发式方法通过对一次会话中站点持续时间或页面停留时间加以界限来进行会话识别。用户对一个站点的访问持续时间平均为25.5 min[3]。在大多数应用中，采用30 min作为一次会话的最长持续时间[3,4]。在一次访问中，用户对于某个页面的阅读与处理时间在一个时间范围内变化。如果在两次请求之间的时间间隔过长，则认为后一次请求是一个新的会话的开始。页面停留时间会根据页面内容和应用目的的不同而不同，这类启发式方法已使用[1,5]。

面向引用的启发式会话识别考虑页面间的链接关系，它基于这样一种推理：用户经常是通过超链接来访问新页面，而不是手工输入请求页面的URL地址。下面将讨论基于引用的启发式会话识别方法的性能评测。一个URL请求页面的“引用页(referrer)”是指包含这个URL请求页面链接的页面。如果会话中的所有页面都没有包含某个页面链接，则基于引用的启发式方法将这个页面请求划分至新的会话中去。

1.2 几种被选取进行评估分析的会话识别方法

下面将对面向时间和面向引用的启发式会话识别方法的性能进行评测。每种启发式方法对由Web日志文件经过用户识别后所得到的用户活动日志进行分析，得出构造会话集。

H1：面向时间的启发式方法：一个会话的持续时间不超过阈值q。t_0：一个构造会话c的第一个URL请求的时间戳。t：一个待划分的URL页面请求P请求的时间戳。如果$(t-t_0\leq q)$则P属于构造会话c。第一个时间戳大于(t_0+q)的页面请求Q是下一个新构造会话的第一个成员。

H2：面向时间的启发式方法：一个页面的停留时间不超过阈值d。t'：最新加入一个构造会话c的URL请求的时间戳。t''：一个待划分的URL页面请求P请求的时间戳。如果$(t''-t'\geq d)$则P属于构造会话c。否则请求P是下一个新构造会话的第一个成员。

H3：基于引用的启发式方法：两个连续的页面请求P、Q，其时间戳分别为t_p、t_q。P属于会话S。如果Q的引用页面已经包含在会话S中，或者引用页面无法确定，并且$(t_q-t_p\leq\Delta)$，Δ是指定的时延，那么Q也被划分入会话S中。否则Q被划分入一个新的构造会话。

日志文件将“无法确定”的引用页面表示为“–”。下列情况可能出现“无法确定”页面。

1) 起始页面的引用页面，或刚刚跳转至子站点或外部服务器时所进入的页面。
2) 在地址栏中直接敲击或通过书签记录的URL地址访问的页面。
3) 在一次会话中装载的一个含有帧页面集的页面。
4) 通过“BACK”按钮所访问的所有页面。

“帧页面集”是指包含一组帧页面的页面。当对页面发出访问请求后，帧页面集的全部帧页面被依次载入，所有帧页面都具有同样的引用页。但日志文件并不能反应这一点，所以在此基础上进行基于引用页面的启发式会话识别时可能产生错误。

2　评测启发式会话识别方法的准确性

2.1　基本实体

指定U作为网站的URL集，它包括所有静态Web页面和所有利用脚本语言动态生成的URL。

服务器日志文件L是U中的URL根据访问时间戳排序后所组成的文件。将L中的第i个成员l看作$L[i]$。l包含请求页面的URL，请求的时间戳，主机的标记、引用页面的URL和用户所使用的代理等信息。L被划分成由真实会话组成的集合R。由启发式方法划分L得到的构造会话所组成的集合被定义为C。

2.2　日志文件与会话的关系

一种启发式方法h对L运用划分机制，生成一种构造会话集C_i。理想的启发式方法ih将生成真实会话集，即$C_{ih}=R$。这里的评测方法用于对构造会话与真实会话的匹配程度进行量化估测，即对“真实会话的重建”程度进行评测。

定义 1　如果一个真实会话的所有成员都包含在某个构造会话中，则这个真实会话被称为“完全重建”。

一个n个成员的真实会话被称为“完全重建”，当且仅当存在一个有m个成员的构造会话c，满足：$\forall i=1,2,\cdots,n\exists j\in\{1,2,\cdots,m\}:r[i]=c[j]$

构造会话保持了真实会话的成员序列。即，一组连续的请求序列A，B，C，若A，C是属于同一个构造会话c，则B也一定属于c。如果真实会话r被构造会话c“完全重建”的话，则也可以称为“连续重建”。

2.3　吻合度的评测

估测方法M通过比较C_h与R之间的差别来对方法h评分。将h的得分表示为$M(h)$，$M(h)=[0,1]$，则$M(ih)=1$，其中R集是已知的。估测方法分为两类：一类是绝对型估测，它根据被完全推测正确的真实会话的数量来评分；另一类是渐进型估测，它根据真实会话被正确推测的程度来评分。

2.3.1　绝对型估测

根据定义1所指的完全重构的真实会话，基本型的绝对估测方法M_{cr}的定义如下：

$M(h)$＝在C_h中包含的完全重构的真实会话的数量/真实会话集R的成员个数$|R|$。

根据对真实会话和构造会话关系统的不同限制条件，派生出以下三种绝对型估测方法。

1) 估测方法$M_{cr,\text{start}}$的真实会话r满足：$(\forall i=1,2,\cdots,n\exists j\in\{1,2,\cdots,m\}:r[i]=c[j])\wedge(r[1]=c[1])$

2) 估测方法$M_{cr,\text{end}}$的真实会话r满足：$(\forall i=1,2,\cdots,n\exists j\in\{1,2,\cdots,m\}:r[i]=c[j])\wedge(r[n]=c[m])$

3) 估测方法$M_{cr,\text{start}-\text{end}}$只考虑那些第一个和最后一个成员都与某个构造会话的相应成员相同的真实会话。即r满足：$(\forall i=1,2,\cdots,n\exists j\in\{1,2,\cdots,m\}:r[i]=c[j])\wedge(r[1]=c[1])\wedge(r[n]=c[m])$

2.3.2　渐进型估测

在许多情况下，构造会话与真实会话部分相同。下面讨论估测真实会话与构造会话的重叠度。

定义 2　真实会话与构造会话的重叠度是两者相同成员的数量除以真实会话成员总数。即，对一个含有n个成员的真实会话r和一个含有m个成员的构造会话c，重叠度$\deg_o(r,c)$为

$$\deg_o(r,c)=\frac{|\{i\in\{1,2,\cdots,n\}\,|\,\exists j\in\{1,2,\cdots,m\}:r[i]=c[j]\}|}{n}$$

为了计算一个真实会话r的重叠度，需要一个函数f，用来在各个构造会话与这个真实会话的重叠度之间进行选择。f可被定义为所有构造会话与这个真实会话的重叠度的平均值或最大值，f为最

大值，即

$$f(r,h,\deg_o)=\max_{c\in C_h}\{\deg_o(r,c)\}$$

最后，利用函数 g来对所有真实会话的重叠度进行总计，从而对一个启发式会话识别方法进行评分。g 可以是真实会话的重叠度的平均值或最大值，g 为平均值。即

$$g(h,\deg_o)=\mathrm{avg}_{r\in R}\{f(r,h,\deg_o)\}$$

在上述定义的基础上，提出渐进型估测方法：$M_o(g)$通过函数 g 计算真实会话集R与C_h 中构造会话的重叠度，来获得启发式会话识别方法h的评分。

$$M_o(g)(h)=g(h,\deg_o)=\mathrm{avg}_{r\in R}\{f(r,h,\deg_o)\}$$

如果一个真实会话被一个构造会话完全推测，则其重叠度为1。

上述估测方法并没有考虑构造会话的长度。如果某个构造会话错误地包含一些本应分配到别的构造会话的成员时，重叠度并不改变，从而不会改变对启发式方法h的评分。这样有可能会推导出一些不合逻辑的页面关系。下面提出的估测方法将考虑这种情况。

定义 3　真实会话与构造会话的重叠度是两者相同成员的数量除以真实会话与构造会话的并集中的所有成员的数量，则重叠度

$$\deg_s(r,c)=\frac{w_a}{n+m-w_a}$$

估测方法$M_s(g)(h)$ 的定义与 $M_o(g)(h)$ 相似。不同的估测方法能够从根据数据挖掘分析的不同目的来对各种启发式会话识别方法的性能加以区别。例如，对于调查一个站点的主要入口的分析而言，则需要这样一种启发式方法，能够在用估测方法 $M_{cr,\mathrm{start}}$ 评测时获得较好的评价。对于另外一种用于改进页面设计的应用而言，需要查找哪些页面是用户离开站点前所访问的，其启发式会话识别方法必须以较好的评分通过 $M_{cr,\mathrm{end}}$。即使两种启发式方法在完全推测方面的性能都不理想，但是能够提供较大重叠度的方法将能生成更多的用于分析页面与对象关系的重要信息。

3　启发式会话识别方法的评估

3.1　评估实验

实验数据包含2001年10月18日和19日发生的174 663个页面请求。为了正确的反应用户的浏览行为，站点禁止使用缓存。这个网站是基于cookie的，并且服务器在内部生成会话id号。每当一个浏览器实例被激活时(即使来源于同一个用户)，就创建一个新的真实会话。cookie只用于用户识别，会话记录机制的工作与用户cookie无关。如果存在两个活动时间相重叠的浏览器实例，则会生成两个会话id号。由于这些请求可以被认为是一个活动的一部分，则将它们合并入一个真实会话。

实验共获得4 400个用户和14 279个真实会话，将会话的边界标记去除，再应用第1.2节中所讨论的启发式方法进行会话识别，得到构造会话。对于基于会话持续时间进行会话识别启发式方法，选择30 min作为一个会话的最长持续时间，记作h1-30。对于基于页面停留时间进行会话识别的启发式方法，选择10 min作为页面的最长停留时间，记作h2-10。对于基于引用信息的启发式方法h-ref，设置Δ为10。图1显示了测试结果。其中M_cr、M_crs、M_cre、M_crse分别为第2.3.1节中的四种绝对型估测方法：M_{cr}、$M_{cr,\mathrm{start}}$、$M_{cr,\mathrm{end}}$、$M_{cr,\mathrm{start-end}}$；$M_o$、$M_s$分别为第2.3.1节中的两种渐进型估测方法：$M_o(g)(h)$、$M_s(g)(h)$。

3.2　讨　论

从两个方面来分析实验结果：首先，结果存在一定的逻辑关系。M_crs与M_cre在M_cr的基础上增加了限制条件，因而其实验结果的值不会高于M_cr的结果。同样，M_crse不会获得比M_crs和M_cre更好的结果，M_s也不会获得比M_o更好的结果。其次，图1表明h1-30和h2-10的性能比较好，并且结果相近。对于考虑会话起点与终点的估测方法M_crs、M_cre、M_crse而言，h2-10的性能更

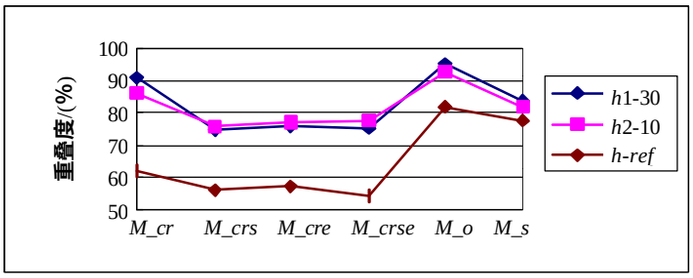

图1　六种评测方法对三种启发式方法准确度的评分

好。这可能是因为实验数据包括大量的一连串持续时间非常短的会话。*h*1-30将这些会话都归入一个构造会话中，而*h*2-10能够根据页面停留时间将两个真实会话进行正确的划分。但是对于页面停留时间很短的真实会话而言，*h*2-10也可能会发生推测错误，因而对于*M_cr*而言，*h*1-30具有较好地性能。*h-ref*方法的评分相对较低。这是因为：1) *h-ref*方法将真实会话中所包含的多个时间重叠的不同浏览窗口的活动划分到各个不同的构造会话中去。但是这类情况在实验数据中并不多，只有3%的真实会话包括并行活动。2) *h-ref*中设定无法确定引用页的页面停留时间为10 min，而实验数据中的此类页面的停留时间很长，导致识别方法将页面划分至新的构造会话中去。而且，一旦一个页面请求被错误划分，后续的页面也将被错误划分。在今后的工作中，可以考虑如何更好地将基于引用页的启发式方法与基于时间的启发式方法结合起来。

总之，选择30 min作为会话最长持续时间的启发式方法的识别结果比较准确，非常适合本文所调查的这个站点。

4　结 束 语

本文根据多种数据挖掘应用的需要提出了一套评测系统，用于比较各种启发式方法的性能。实验结果显示了每个启发式识别方法的参数的影响作用，并使用满足不同应用需求的评测方法来分析各种启发式识别方法的性能。从而，Web挖掘分析可以选择最合适的启发式方法来完成数据预处理阶段的会话识别处理，并在此基础上进行进一步的数据挖掘工作。

今后的一个研究方向是缓存对会话识别的影响，它包括估测每种启发式方法的错误对后继挖掘分析结果的影响。即，对所开发出的模式的支持度与准确度的估测。我们还将进一步完善用于选择数据预处理启发式方法的机制，从而在应用启发式方法对真实数据分析之前就能够得到比较出各种启发式会话识别方法的性能差异。

参 考 文 献

1　Cooley R, Mobasher B, Srivastava J. Data preparation for mining world wide web browsing patterns. Journal of Knowledeg and Information Systems, 1999, 1(1): 135-139

2　World wide web committee web usage characterization activity. W3C Working Draft:Web Characterization Terminology Definitions Sheet.www.w3.org/1999/05/WCA-terms, 1999

3　Catledge R, Pitkow J. Characterizing browsing behaviors on the world wide web. Computer Networks and ISDN Systems, 1995, 27, 86-92

4　Borges J, Levene M. Data mining of user navigation patterns. KDD' 99 workshop on web usage analysis and user profiling WEBKDD' 99, 1999: 180-186

5　Spiliopoulou M, Faulstich C L. WUM: a tool for web utilization analysis. In extend version of Proc. EDBT Workshop WebDB' 98, LNCS 1590, 1999: 184-203