第42卷 第2期 西 安 交 通 大 学 学 报 Vol.42 №2
2008年2月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Feb. 2008

一种基于网页元数据的用户访问行为建模方法

杜瑾[1,2]，刘均[1,2]，郑庆华[1,2]，丁娇[1,2]，龚智勇[1,2]，韩殿哲[3]
（1.西安交通大学电子与信息工程学院，710049，西安；2.陕西省天地网技术重点实验室，710049，西安；
3.西安铁路信号工厂，710048，西安）

摘要：针对现有行为建模方法难以描述行为语义的问题，提出了一种分层次的用户行为元模型以及一种基于页面元数据的Web用户行为建模方法.该方法从URL的访问、活动、事务3个层次建立Web用户的行为模型，并对页面元数据获取以及在URL的访问、行为、事务之间转化等问题进行了说明.方法及模型不仅描述了用户访问序列信息，还增加了访问内容的局部主题和关键词等信息，为进一步获取Web用户的行为语义特征奠定了很好的基础.通过西安交通大学的Web教学系统验证表明，利用所提方法获得的序列划分准确率达86%以上.

关键词：行为建模；元数据；行为元模型
中图分类号：TP31 **文献标志码**：A **文章编号**：0253-987X(2008)02-0152-04

Method for User Browsing Behavior Modeling Based on Web Page Metadata

DU Jin[1,2], LIU Jun[1,2], ZHENG Qinghua[1,2], DING Jiao[1,2], GONG Zhiyong[1,2] HAN Dianzhe[3]
(1. School of Electronics and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China; 2. Shaanxi Key Laboratory of Satellite-Terrestrial Network Tech R & D, Xi′an 710049, China; 3. Xi′an Railway Signal Factory, Xi′an 710048, China)

Abstract: A hierarchical meta-model of user behaviors and a method for user behavior modeling based on the metadata in pages were proposed to overcome the difficulty of the existing methods in describing the semantics of behavior. In the method, the behavior model of web users is built with three levels, i.e., URL accesses, activities and sessions. Furthermore, both the abstraction of metadata in pages and the transformation among URL accesses, activities and sessions are specified. The behavior models generated by the method can describe not only the information of user accessing sequence, but also the additional information including local themes of pages content and keywords. The information gives the foundation for acquiring the semantic characteristics of user behaviors. Tests performed in web-based distance learning system of Xi′an Jiaotong University show that the accuracy of activity partition by the proposed method exceeds 86%.

Keywords: behavior modeling; metadata; behavior meta-model

Web用户行为建模及特征获取研究主要集中在Web数据挖掘领域，其数据来源是用户日志中的URL请求、页面间链接的拓扑结构、注册用户特征等.用户行为模式可解决页面自动导航、Web应用系统性能和页面重要性评价等问题.例如，从日志文件来发现以访问路径和频繁访问页面组形式的用户行为特征[1]，通过隐半马尔科夫模型来描述用户浏览行为[2]等.

目前，基于Web日志的行为建模方法只能描述用户在页面间的游走过程，无法描述用户行为内在

收稿日期：2007-05-24. 作者简介：杜瑾(1974—)，男，博士生；郑庆华(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60373105，60473136，60633020)；教育部高等学校博士学科点专项科研基金资助项目(20040698028)；国家科技攻关“十一五”专项经费资助项目(2006BAH02A24).

的语义信息，如用户的注册、登录、检索等具体行为含义. 事实上，在用户与 Web 应用系统交互的过程中，影响用户行为取向的关键因素是网页中蕴含的语义信息，不是页面间的链接关系. 因此，基于 Web 日志所获得的行为特征还难以作为个性化、自适应服务的依据.

针对上述问题，本文提出了一种基于分层结构的用户行为元模型，该元模型从“URL 访问-活动-事务”3 个层次定义了行为模型的框架结构，同时结合元模型进一步提出了基于页面元数据的 Web 用户行为建模方法，并对页面元数据获取以及 URL 的访问、行为、事务之间的转化等问题进行了说明. 行为建模方法不仅描述了用户访问序列信息，还增加了访问内容的局部主题和关键词等信息.

1　基于网页元数据的行为元模型

1.1　网页元数据定义

网页元数据可分为描述性元数据(Descriptive Metadata, DM)与结构性元数据(Structural Metadata, SM)2 种类型[3]，其中描述性元数据主要包括网页标题以及由状态变量构成的二元组，可定义为

$$d_{DM}=(T,\{(V,a_{RW})\}) \tag{1}$$

式中：T 表示标题；V 表示变量；a_{RW} 表示读写属性. 其中，$\{(V,a_{RW})\}$ 是变量 V 及对应的读写属性构成的二元组集合. 结构性元数据主要包括页面中的超链接以及对应的标题，可定义为

$$d_{SM}=(\{u_1,u_2,\cdots,u_n\}) \tag{2}$$

式中：u_i 表示第 i 个页面访问路径. ({log in. jsp? user & pass,/registration. htm,…})是一个结构性网页元数据的实例，表示当前页面中包含“login. jsp? user & pass”与“registration. htm”2 个超链接.

1.2　行为元模型

基于页面元数据的定义，对行为建模的相关概念定义如下.

定义 1　根据勒温的理论[4]，行为可定义为主体在其内在因素或内在环境的影响下，与外部环境之间的相互作用.

行为具有目的性，是行为主体根据其内在因素中的目的实施的，并且会反作用于内在因素与外部环境，同时行为也具有分层特性，可以分解为粒度更小的子行为，也能够组合成粒度更大的行为.

定义 2　行为模型是指对特定主体在特定时刻或时间段的行为，以及行为之间的关系、涉及环境因素的形式化描述.

定义 3　行为元模型是对行为模型在结构与语义上的定义，是用户行为上一层的抽象，它与行为模型是类与实例的关系.

Web 用户的行为是指，用户与 Web 应用系统之间的交互行为. 结合行为的分层特性，本文提出了一种 3 层框架的行为元模型，描述如下.

(1)URL 访问. 它处于行为模型的最底层，对应于 Web 日志中的 URL 请求，可用元组(ID_{User}，$ID_{Request}$，t，t_D，M，u，S_{State})来描述，描述信息中不涉及行为的语义信息，其中 ID_{User} 惟一标示用户的 ID，$ID_{Request}$ 惟一标示某个 URL 的请求，t 为当前 URL 的请求时间，t_D 为浏览时长，M 表示请求方法，u 表示页面访问路径，S_{State} 是一个由变量、变量值构成的二元组集合，用来描述外部环境状态以及当前行为对环境状态的影响(u 中所传递的参数变量以及变量值). 该层行为是通过对 Web 日志分析获得的.

(2)活动(Activity). 它处于行为模型的中间层，是指与应用系统特定功能相对应的用户行为，如用户注册、登录、检索等，同时还描述用户访问序列的局部语义信息，如主题、标题、关键词、超链接等. 该层行为可用元组(ID_{User}，$ID_{Activity}$，$N_{Activity}$，t，t_D，S_{State})来描述，其中 $ID_{Activity}$ 惟一标示某个活动的 ID，$N_{Activity}$ 表示当前活动的名称，S_{State} 包含着与当前活动有关的局部语义信息变量.

(3)事务(Session). 它是行为模型的最上层，是指完成特定目标的一系列活动. 事务可用元组(ID_{User}，$ID_{Session}$，t，t_D，S_{State}，S_{Aid})来描述，其中 $ID_{Session}$ 惟一标示当前事务，S_{Aid} 表示本事务中包含活动序列的所有活动的 ID 集合.

2　行为建模方法

结合第 1 节中的行为元模型，本文提出了一种基于页面元数据的用户行为建模方法，其机理如图 1 所示.

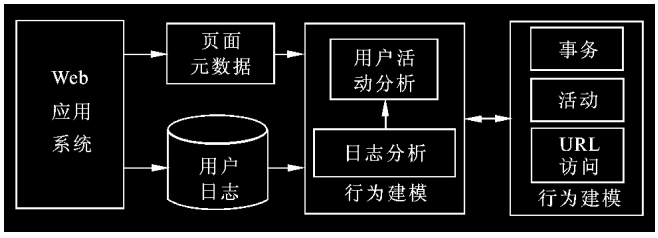

图 1　基于页面元数据的用户行为建模机理

2.1　日志分析

Web 日志格式的局限性导致用户 ID 无法区

分,且地址转换(NAT)网关与代理的客户端在 Web 日志中记录的 IP 地址、浏览器类型、操作系统类型等内容完全一致的概率比较大,因此日志分析应识别出 Web 日志中的每个用户,并获得完整的访问序列.

现有 Web 应用系统一般采用基于环境变量的动态页面技术,其中描述环境状态的环境变量值在导航过程中应相对稳定.此外,造成环境变量变化的页面,其 URL 请求的日志不会因代理缓冲(Cache)机制而丢失.结合这些特点,本文给出了页面距离的概念,并提出了 2 个用于区分用户规则及日志分析的方法.

定义 4 页面相对距离,即页面 p_j 相对 p_i 的距离,是指 p_i 通过页面间的导航链接到达 p_j 所需的最少步骤,记为 $D_R(p_i,p_j)$.特别地,$D_R(p_i,p_i)=0$;若 p_i 无法通过超链接到达 p_j,$D_R(p_i,p_j)=\infty$;若 $p_j\in p_i.\text{metadata.url}$,$D_R(p_i,p_j)=1$,即页面间存在直接链接关系,其中 $p_i.\text{metadata.url}$ 是指页面 p_i 中超链接的集合.

规则 1 设 r_i 为 Web 日志 L 中的一个 URL 请求,$R\subset L$ 为与 r_i 时间间隔小于特定阈值 t_0 且页面距离小于特定阈值 l_0 的后序 URL 请求的集合,即 $R=\{r\mid 1\leqslant D_R(r_i.\text{url},r.\text{url})<l_0\wedge(|r.t-r_i.t|)<t_0\wedge r\in L\}$,对于任何 $r\in R$,若 $r_i.S_{\text{State}}=r.S_{\text{State}}$,则 $r_i.\text{ID}_{\text{User}}=r.\text{ID}_{\text{User}}$.

规则 1 的含义是:若环境状态不发生变化,时间与相对距离接近的 2 个 URL 请求来自同一用户.

规则 2 设 r_i 为 L 中的一个 URL 请求,若不存在 $r\in R$,使得 $r_i.S_{\text{State}}=r.S_{\text{State}}$,则当 r_j 满足条件① $r_j\in R$,② $D_R(r_i.\text{url},r_j.\text{url})=1$,③ 设 $S_{\text{Change}}=r.S_{\text{State}}-r_i.S_{\text{State}}$,$S_{\text{Change}}\neq\varnothing$并且构成 S_{Change}的二元组元素的变量属性在页面 $r_i.\text{url}$ 的元数据中是可写的,则等式 $r_i.\text{ID}_{\text{User}}=r_j.\text{ID}_{\text{User}}$成立.

规则 2 的含义是:若环境状态发生变化,而其中的 2 个 URL 请求的页面具有直接链接关系且发生变化的环境变量在前一个 URL 对应页面中是可写的,则这 2 个请求来自同一用户.

基于规则 1、规则 2 的日志分析方法正如图 2 所示.

每个分组中的每一类 URL 请求对应于某个用户,若某分组中只有一类,则该分组对应于某个用户.利用日志分析方法可将 Web 日志的访问请求按用户进行分组,并生成(ID_{User},$\text{ID}_{\text{Request}}$,$t$,$t_D$,$M$,$u$,$S_{\text{State}}$)的 URL 访问序列.

输入:Web日志L
输出:序列集合$s_{\text{Serial}}=\{s_{\text{Serial1}},s_{\text{Serial2}},\cdots,s_{\text{Serial}n}\}$
过程:$S_{\text{Group}}=\text{GroupedBy}(a_{\text{IP}},a_{\text{Browser}},a_{\text{OS}},L)$
//根据日志中的客户端IP地址、浏览器类型、操作系统类型等属性对URL请求进行分组,而每个组中的URL请求具有3个相同的属性
$m=|S_{\text{Group}}|$ //m为分组的个数
for $i=1$ to m do
{$S_{\text{RoughSerial}i}=R_1(S_{\text{Group}i})$
//对每个分组中的所有符合规则1(R_1)的URL请求按照页面相对距离从小到大的顺序进行归类,直至处理所有符合R_1的URL请求
$s=|S_{\text{RoughSerial}i}|$
for $j=1$ to s do
{$U_{\text{RepaireURL}}(S_{\text{RoughSerial}ij})$ //根据页面的链接关系填充丢失的URL请求
$S_{\text{Serial}i}=R_2(S_{\text{RoughSerial}i})$
//对每个分组中的所有符合R_2的URL请求进行归类
}
}
$s_{\text{Serial}}=\cup s_{\text{Serial}i}$

图 2 日志分析方法

2.2 活动分析

活动分析是指将特定用户的 URL 访问序列转化为能够描述行为语义的活动序列,主要依据页面的聚类特性,这种特性体现在链接关系与页面内容 2 个方面.根据这些特性,结合在页面主题信息采集中的 Sibling、Pagerank 等思想,可以得到如下规则,以用于判定页面是否属于同一主题.

规则 3 存在直接或间接链接关系的页面可能属于同一主题;页面间相对距离越小,则属于同一主题的可能性越大;页面互相之间存在链接关系,则属于同一主题的可能性最大[5].

规则 4 内容(特别是标题)相同或相似的页面很可能属于同一主题[6].

综合规则 3、规则 4,活动分析的基本思路可归纳为:首先根据页面的超链接、标题以及变量这 3 种元数据信息对页面进行聚类,其次根据聚类结果将 URL 访问序列划分为活动序列,并为每个活动设置标题.

定义 5 页面绝对距离

$$D_A(p_i,p_j)=\frac{D_R(p_i,p_j)D_R(p_j,p_i)}{D_R(p_i,p_j)+D_R(p_j,p_i)} \tag{3}$$

显然,$D_A(p_i,p_j)\leqslant\min(D_R(p_i,p_j),D_R(p_j,p_i))$.

定义 6 设 $S(p_i,p_j)$是在标题、变量构成的向量空间中获得的页面 p_j 与 p_i 的相似度[6-7],则相似距离 $D_S(p_i,p_j)=D_A(p_i,p_j)S(p_i,p_j)$.

由规则 3、规则 4 知，$D_S(p_i,p_j)$越小，页面 p_j 与 p_i 属于同一主题的可能性就越大. 基于此，采用 AGNES 聚类方法可对页面进行聚类.

设 URL 访问序列为 $r_1,r_2,\cdots,r_i,\cdots,r_m$，其对应的页面为 $p_1,p_2,\cdots,p_i,\cdots,p_m$，则根据网站内所有页面的聚类结果提出的活动分析方法的步骤如图 3 所示.

步骤1：若寻找到满足如下条件的子序列$p_j,\ldots,p_i(i\geqslant j)$，即$p_j,\ldots,p_i$属于同一类且$p_{j-1}$与$p_{i+1}$都不属于该类，就将$r_j,\ldots,r_i$合并为一个活动.

步骤2：若寻找到$p_j,\ldots,p_i$的中心，中心页面p_j应满足，对于任意页面

$P_k(j\leqslant k\leqslant i)$，$\sum_{x=j}^{i} D_S(p_x,p_l)\leqslant\sum_{x=j}^{i} D_S(p_x,p_k)$

步骤3：$A_C.N_{\text{Activty}}=p_l.\text{title}$，$A_C$ 为当前活动

$A_C.t_D=r_i.t-r_j.t$；

$A_C.S_{\text{State}}=r_i.S_{\text{State}}$.

步骤4：重复步骤1～步骤3，直到所有的URL访问行为都被合并到某一个活动中.

图 3　活动分析方法

事务分析是从特定用户的一次访问序列中提取出的，为完成某特定目标而进行的活动序列. 通常情况下，用户的访问行为是随机的，有可能不带任何目标，也可能有一个或多个目标且各目标的访问活动交叉进行. 为了方便起见，本文将用户一次访问序列中的所有活动的集合作为一次事务.

3　实　验

在西安交通大学网络教育学院的教学网站平台上，对行为建模方法进行了实验.

第 1，收集相同时间段内来自同一代理服务器的所有访问日志；第 2，进行日志分析，获得每个学习者的 URL 访问序列；第 3，采用人工分析法生成活动序列；第 4，采用本文方法产生活动序列；第 5，对 2 种生成序列进行对比，以验证本文方法的有效性. 同时，采用简单匹配系数来描述用 2 种方法得到的活动序列的相似度. 活动序列 s_i、s_j 的相似度

$$d_{\text{sim}}(s_i,s_j)=\frac{A(s_i,s_j)+C(s_i,s_j)}{A(s_i,s_j)+B(s_i,s_j)+B(s_j,s_i)+C(s_i,s_j)} \quad (4)$$

式中：$A(s_i,s_j)$表示在 2 种活动序列中 2 个 URL 访问属于同一活动的次数；$B(s_i,s_j)$表示在活动序列 s_i 中 2 个 URL 访问属于某一活动，但不属于活动序列 s_j 对应活动的次数；$C(s_i,s_j)$表示在 2 种活动序列中 2 个 URL 访问均不属于同一活动的次数.

实验共提取出 258 名学生的 URL 访问序列，采用 2 种方法获取的序列匹配系数在最差的情况下为 0.864，也就是说与人工标记方法得到的活动序列进行对比，本文的活动分析方法的准确率在 86% 以上. 当一个活动包含的页面个数较多时，所获得的活动序列将划分得更加精确，这是因为页面样本数越多，页面聚类的精度就越高.

4　结　论

本文针对现有 Web 用户行为模型缺少描述行为语义能力的问题，提出了一种 3 层结构的 Web 用户页面访问行为元模型. 该元模型从 URL 的访问、活动、事务 3 个层次定义了行为的描述框架. 基于元模型，又进一步提出了一种基于页面元数据的 Web 用户行为建模方法，它根据 Web 日志中的 URL 请求以及页面的标题、变量、超链接等元数据信息可生成由不同粒度行为序列构成的行为模型，该模型可以很好地描述行为语义，具有精确的行为识别能力.

参考文献：

[1] SRINIVASAN S, KRISHNA V, HOLMES S. Web-log-driven business activity monitoring [J]. Computer, 2005,38(3):61-68.

[2] 谢逸，余顺争. 基于 Web 用户浏览行为的统计异常检测 [J]. 软件学报，2007，18(4):967-977.
XIE Yi, YU Shunzheng. Anomaly detection based on web users' browsing behaviors [J]. Journal of Software, 2007,18 (4):967-977.

[3] Digital Library Federation. Metadata encoding and transmission standard [EB/OL]. [2007-01-10]. http://www.loc.gov/standards/mets.

[4] 库尔特·勒温. 拓扑心理学原理 [M]. 高觉敷，译. 北京：商务印书馆，2003.

[5] HENZINGER M R. Hyperlink analysis for the Web [J]. IEEE Internet Computing, 2001, 5 (1):5-50.

[6] YANG Y, SLATTERY S, GHANI R. A study of approaches to hypertext categorization [J]. Journal of Intelligent Information Systems, 2002, 18(2/3):219-241.

[7] BEIL F, ESTER M, XU X. Frequent term-based text clustering [C] // Proceedings of 2002 International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002:436-442.

（编辑　苗凌）