# 基于用户浏览行为分析的用户兴趣获取

尹春晖,邓 伟
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

**摘 要**:从用户的浏览行为可以反映用户的兴趣出发,分析了用户的浏览行为与兴趣之间的关系,提出了五种用户最小浏览行为组合,并在此基础上对其中三种行为进行转化,得到影响用户兴趣的关键的两种行为,并给出这两种行为与用户兴趣度之间的定量关系。结合对网页内容的挖掘,获取用户的兴趣。通过实验对文中的研究结果进行验证,实验结果证明,所采用的方法是合理和有效的,分析出的用户兴趣基本上可以正确反映用户的实际兴趣。

**关键词**:用户浏览行为;用户兴趣度;钩子
**中图分类号**:TP18 **文献标识码**:A **文章编号**:1673-629X(2008)05-0037-03

# Extracting User Interests Based on Analysis of User Behaviors

YIN Chun-hui, DENG Wei
(College of Computer Science and Technology, University of Soochow, Suzhou 215006, China)

**Abstract**: Based on the users' browsing behaviors can reflect users' interests. Analyses relations of users' browsing behaviors and interests. A minimum combination of five browsing behaviors is proposed, based on that, transforms three behaviors of this five, two pivotal behaviors influence users' interests are summarized. Combine with that, quantitative relations of this two behaviors and user interest degree is proposed. Combined with web page mining, users' interests are extracted. Verifying the researching results, experiments results prove that, the researching method is reasonable and effective. Users' interests analyzed can reflect users' real interests.

**Key words**: users' browsing behaviors; user interest degree; hook

## 0 引 言

基于用户兴趣的个性化服务[1]正逐渐成为学术界和商业应用中热门的研究方向。为了更好地实现个性化服务,必须不断学习用户的兴趣,并适时地更新。建立用户兴趣模型是解决这一问题的最佳方法。获取用户的兴趣又是建立用户兴趣模型的首要问题。在国内外对获取用户兴趣的研究中[2],重点都放在通过对用户浏览内容的分析来获取用户兴趣。文中提出一种兴趣获取的方法,通过对用户浏览行为的分析,根据浏览行为与用户兴趣之间的定量关系评价用户对网页的兴趣程度,结合对网页内容的挖掘,获得用户的兴趣。

## 1 浏览行为分析

### 1.1 浏览行为与用户心理

从心理学角度[3]来讲,人的行为可以反映人的兴趣和目的。在日常生活中,人们经常通过"察言观色"来推测一个人的兴趣和意图。同样的道理,在用户访问网络的过程中,也可以通过用户的浏览行为来推测用户的兴趣。比如用户保存或打印一个页面,可以推测用户对该页面感兴趣;用户频繁访问一个页面,可以推测用户对该页面感兴趣;用户在某页面上驻留时间较长,可以推测用户较为认真地阅读了该页面,对该页面有着比较浓厚的兴趣;用户下拉滚动条的时间较长,可以推测用户浏览了整个页面,对页面较有兴趣;用户按了 PAGE DOWN 键且时间较长或次数较多,可以推测用户对页面感兴趣。

### 1.2 浏览行为分类

浏览行为大致可以分为以下几类[4]:

(1)标记行为:增加书签、删除书签、保存页面、打印页面等。

(2)操作行为:复制、粘贴、剪切、拉动滚动条、点击链接等。

(3)重复行为:重复访问同一个页面等。

由分析可知,除了保存页面、打印页面、增加书签外,其他的行为都可转化为访问同一页面的次数和在

**收稿日期**:2007-08-25
**基金项目**:国家自然科学基金资助项目(60572074)
**作者简介**:尹春晖(1984-),男,江苏宿迁人,硕士研究生,研究方向为智能化信息处理、搜索引擎;邓 伟,副教授,硕士生导师,研究方向为智能化信息处理、神经网络、模式识别、语音信号处理。

页面上的驻留时间。所以,以下几个行为构成用户兴趣度估计的最小浏览行为组合:

·保存页面;

·打印页面;

·将页面存于书签中;

·访问同一页面的次数(重复访问);

·在页面上的驻留时间(复制、粘贴、拉动滚动条等)。

### 1.3 兴趣度的计算

网页兴趣度[5]是指用户对一个网页内容的感兴趣程度,采用0~1间的实数表示,0表示无兴趣,1表示最大兴趣。显然,用户兴趣与所浏览网页的兴趣度是密切相关的。把对网页 $L$ 的五种最小浏览行为表示为:保存页面($S(L)$)、打印页面($P(L)$)、将页面存于书签($B(L)$)、重复访问的次数($R(L)$)、在页面上的驻留时间($T(L)$)。这些行为在不同程度上反映了用户对页面 $L$ 的兴趣,为了区分这些行为所表示的不同兴趣程度[6],为每一个行为 $v$ 赋予一个权值 $Cv$,这样从用户这些行为推理得出的该用户对页面 $L$ 的感兴趣程度可以用下面公式来计算:

$$I(L) = \sum_{v \in F} Cv fv(L) \tag{1}$$

其中 $\sum_{v \in F} Cv = 1, F = \{S(L), P(L), B(L), R(L), T(L)\}$。$Cv$ 是分配给行为 $v$ 的权值;$fv(L)$ 是一个二值函数,如果用户对页面 $L$ 有行为 $v$,其函数值为1,否则为0。

文献[7]中指出兴趣度 $d(P) = a * t(P) + b * v(P) + c$,其中 $t(P)$ 是网页 $P$ 的浏览的时间,$v(P)$ 是拉动滚动条的次数。但是笔者认为,$v(P)$ 已经被包含在 $t(P)$ 之中。因为如果网页的浏览时间长,就会相应地造成拉动滚动条的次数也会多。反之,拉动滚动条的次数多,网页的浏览时间也会相应的长。为了比较准确地找到网页兴趣度与用户浏览行为之间的关系,对用户浏览行为作了进一步的分析,前面已经指出,用户的最小浏览行为有五种。但是,进一步分析后可以发现:1) 复制、粘贴、拉动滚动条等行为必定增加网页浏览时间;2) 保存的页面、标记书签的页面,以后会被用户多次调出来重新浏览,这体现为访问次数。所以,能够揭示用户对网页 $L$ 兴趣度的关键的两种行为是:在网页上的浏览时间 $t(L)$(简称 $T$ 行为)和对网页的访问次数 $r(L)$(简称 $R$ 行为)。为了找到 $T$、$R$ 行为与网页兴趣度之间的定量关系[8],必须进行详细的分析和实验。实验分析后得到兴趣度量化估算公式:

$$I(L) = a * t(L) + b * r(L) + c \tag{2}$$

其中 $a$, $b$, $c$ 是与 $t(L)$ 和 $r(L)$ 无关的未知系数。

### 1.4 兴趣的表示

前面已经通过对用户浏览行为的分析获得用户对网页的兴趣度。但是,如何表示用户本身的兴趣,才是建立用户兴趣模型要解决的关键的问题。为了得到用户的兴趣[9],必须先对用户浏览的页面进行聚类分析,得到基于用户兴趣类的页面簇。在此基础上,对兴趣子类的权值高的特征向量集合值进行概化。例如,某一兴趣子类的权值较高的特征向量集合为{足球,NBA,意甲,中超,……},则可以用"体育"这个关键词来概化该兴趣子类。每个用户的兴趣就可以用若干个关键词来简要地描述。用户兴趣子类的权值如何确定呢?兴趣子类的权值可以采用页面的兴趣度来进行量化。兴趣子类 $K_i$ 的兴趣度权值计算公式为:

$$\text{InterestDegree}(K_i) = \sum_{j=1}^{n} L_j \tag{3}$$

其中,$L_j$ 为兴趣子类 $K_i$ 中任一个页面的兴趣度,$n$ 为兴趣子类 $K_i$ 中的页面数目。

当得到用户的兴趣子类及其权值之后,便可以采用二层树状结构模型来表示用户兴趣。用一组关键字 $(K_1, K_2, \cdots, K_m)$ 来表示用户的 $m$ 个兴趣类。每一兴趣类都可以通过计算得到其权值 $\text{InterestDegree}(K_i)$(下面用 $\text{ID}(K_i)$ 表示)。因此,用户的兴趣可以表示为 $((K_1, \text{ID}(K_1)), (K_2, \text{ID}(K_2)), \cdots, (K_m, \text{ID}(K_m)))$ 的加权矢量格式。

## 2 行为数据采集

### 2.1 客户端远程 Agent 方法

用户行为数据的采集是获取用户兴趣的基础。在客户端采用远程 Agent 的方法来获取用户行为数据,是一种比较良好的方法。所谓"客户端远程 Agent"就是运用 Applet 技术在客户端实现用户浏览行为的获取。当用户第一次访问站点中的某个页面时,Applet 程序被下载到本地。Applet 获取页面的装载时间和离开时间,将用户浏览时间发送给 Web 服务器,在服务器端监视并获取远程 Agent 送回的用户浏览行为信息。但是必须指出的是,Applet 在第一次加载的时候,它会产生一些额外的时间和系统资源开销。另外,这种方法只能收集单个站点、单个用户的浏览行为。为了解决这些问题,将采用钩子函数,它能较好地捕获用户的浏览行为。

### 2.2 钩子函数

钩子(hook)[10]是 Windows 系统中的一种特殊的消息处理机制。钩子机制允许应用程序截获处理 Windows 消息或特定事件,与 DOS 中断截获处理机制有类似之处。钩子是 Windows 消息处理机制的一个

平台。钩子程序是一个应用程序定义的回调函数,它只能定义为普通的C函数。

钩子上可以挂接多个回调函数以构成一个钩子函数链。钩子函数可根据各自的功能对消息进行监视、修改和控制等。当钩子所监视的消息到达后,Windows调用链表中的第一个钩子程序,在目标窗口处理函数之前处理它,然后交还控制权或将消息传递给下一个钩子函数以致最终达到窗口函数。一个钩子处理一种类型的消息。某些类型的钩子只能在系统范围内设置处理子程,其余类型的钩子还可以在特定的线程中设置。所以,可以在系统中安装自定义的钩子,监视系统中特定事件的发生,完成特定的功能,比如截获键盘、鼠标的输入、屏幕取词等等,以获取用户的浏览行为。

## 3 实验分析

在本地机器上的D盘的根目录Test(D:\Test)下,保存着不同主题内容的20个网页。

实验让用户随意浏览Test目录下这不同主题内容的20个网页,利用钩子函数获取用户的浏览行为数据,并对每个网页先给出预估计的兴趣度。再由兴趣度经验计算公式 $I(L) = 0.112 * t(L) + 0.071 * r(L) + 95.883$ 计算出对各个网页的兴趣度,并将其与预估兴趣度进行比较(见表1)。

表1 计算兴趣度与预估兴趣度

| 计算兴趣度 | 预估兴趣度 | 比值 |
|---|---|---|
| 0.6142 | 0.65 | 0.94 |
| 0.6810 | 0.8 | 0.85 |
| 0.9355 | 0.95 | 0.98 |
| 0.2672 | 0.2 | 0.75 |
| 0.6012 | 0.65 | 0.92 |
| 0.2745 | 0.25 | 0.91 |
| 0.3615 | 0.4 | 0.90 |
| 0.3827 | 0.45 | 0.85 |
| 0.6618 | 0.4 | 0.60 |
| 0.6182 | 0.65 | 0.95 |
| 0.6811 | 0.6 | 0.88 |
| 0.7143 | 0.6 | 0.83 |
| 0.5332 | 0.5 | 0.93 |
| 0.2885 | 0.3 | 0.96 |
| 0.3813 | 0.25 | 0.65 |
| 0.2358 | 0.15 | 0.63 |
| 0.2258 | 0.2 | 0.88 |
| 0.1215 | 0.1 | 0.82 |
| 0.7251 | 0.85 | 0.85 |
| 0.8098 | 0.7 | 0.86 |

从表中可以看出,计算兴趣度和预估兴趣度两者的比值平均达到0.847,最高的达到0.98,最低也达到0.60,三分之二以上达到0.8。这说明了计算所得的网页兴趣度和实际的兴趣度是非常接近的,进一步验证了文中的研究结果兴趣度量化估算公式是合理和有效的。

## 4 结 语

对用户的浏览行为进行了分析,在此基础上,提出一种用户浏览行为与用户网页兴趣度之间的定量关系。并通过实验验证了这种定量关系是合理和有效的。

获取用户兴趣只是建立用户兴趣模型的第一步,如何使用合适的数据结构来存储用户兴趣,如何建立和更新兴趣模型将是下一步的研究内容。

**参考文献:**


[1] 谭 琼,李晓黎,史忠植. 一种实现搜索引擎个性化服务的方法[J]. 计算机科学,2002,29(1):23-25.

[2] Rucker J, Polanco M J. Siteseer: Personalized Navigation for the Web[J]. Communication of the ACM, 1997,40(3):50-55.

[3] 金海金. 基于用户行为及语义相关实时更新的用户兴趣模型[D]. 成都:西南师范大学,2005.

[4] 孙铁利,杨凤芹. 根据用户隐式反馈建立和更新用户兴趣模型[J]. 东北大学学报:自然科学版,2003,35(3):99-104.

[5] Seo Y, Zhang B. Learning user's preferences by analyzing Web-browsing behaviors[J]. Artificial Intelligence, 2001,15(6):381-387.

[6] Srivastava J, Cooley R, Deshpande M, et al. WebUsage Mining: Discovery and applications of Usage Patterns from Web Data[J]. SIGKDD Explorations, 2000,2(1):45-47.

[7] 付关友,朱征宇. 个性化服务中基于行为分析的用户兴趣建模[J]. 计算机工程与科学,2005,27(12):76-78.

[8] Goecks J, Shavlik J. Learning Users' Interests by Unobtrusively Observing Their Normal Behavior[C]// In Proceedings of the ACM Intelligent User Interfaces Conference(IUI) [s.l.]:[s.n.], 2000.

[9] 赵银春. 基于Web浏览内容和行为相集合的用户兴趣挖掘[J]. 计算机工程,2005,31(12):93-94.

[10] Jeffrey. Windows核心编程[M]. 王建华,张焕生,侯丽坤,等译. 北京:机械工业出版社,2005:463-549.