第 44 卷 第 4 期 西 安 交 通 大 学 学 报 Vol. 44 №4
2010 年 4 月 JOURNAL OF XI'AN JIAOTONG UNIVERSITY Apr. 2010

高效的用户访问预测新算法

冯少荣，张东站
（厦门大学信息科学与技术学院，361005，福建厦门）

摘要：针对基于 Web 日志挖掘的用户访问预测经典算法的不足，提出了基于 Markov 链和关联规则的预测算法(MAPA)．使用二阶 Markov 链找到用户下一步或将来可能访问的页面集，生成预测候选集；使用二项关联规则从正向和反向 2 个角度修正 Markov 的预测结果，从而生成最后的预测页面．通过引入用户反馈机制，提出了带反馈的 Markov 预测算法(MPAF)，即在预测过程中逐步构造历史预测树，把历史预测信息保存到历史预测树中，并根据用户的反馈来判断预测的正确性．在预测过程中，用二阶 Markov 预测算法生成预测候选集，再利用历史预测信息动态地调整预测算法，从而生成预测页面．理论分析证明，这 2 种预测算法具有线性时间复杂度的预测效率．实验结果表明，MAPA 和 MPAF 在预测准确率上平均提高 5％和 10％．

关键词：数据挖掘；Web 日志挖掘；访问预测；Markov 预测；关联规则
中图分类号：TP311 **文献标志码**：A **文章编号**：0253-987X(2010)04-0028-06

Two New Efficient Algorithms to User Access Prediction

FENG Shaorong，ZHANG Dongzhan
(School of Information Science and Technology，Xiamen University，Xiamen，Fujian 361005，China)

Abstract：A Markov chain and association rule prediction algorithm (MAPA) is proposed to deal with shortcomings of existing algorithms on user access prediction based on web log mining. The algorithm uses the second-order Markov chain to find the pages which users may visit in either the next step or future，so as to generate the candidate prediction page set. Then the two-item association rules are used to correct the prediction result from the forward and the reverse perspectives to get the last prediction page. The algorithm integrates the advantages of both the Markov chain and the association rule well. A Markov prediction algorithm with feedback (MPAF) is proposed by introducing user feedback mechanism. The algorithm creates a history prediction tree (HPT) step by step during the prediction process，saves the history prediction information into HPT，and determines whether the prediction is correct according the user's feedback. The algorithm generates the candidate prediction page set using the second order Markov prediction algorithm at first，and then the last prediction page is generated by dynamically adjusting the prediction algorithm according the historical prediction information. Theoretical analyses show that both the prediction algorithms have linear time complexity. Experimental results show that the average prediction accuracy of MAPA and MPAF is increased by 5％ and 10％，respectively.

Keywords：data mining；Web log mining；access prediction；Markov prediction；association rule

当前，用户访问预测的研究主要分为以下 3 类：①利用关联规则预测用户的下一个访问页面[1]，结合粗糙集理论与关联规则的预测方法[2]；②基于 Markov 模型的用户访问预测算法[3]和基于用户分

收稿日期：2009-09-07． 作者简介：冯少荣(1964－)，男，博士，副教授． 基金项目：国家自然科学基金资助项目(50604012)．

类的多 Markov 链预测算法[4]；③基于点击流树的预测算法[5].

本文根据以上用户访问预测过程中存在的问题，提出了基于 Markov 链和关联规则的用户访问预测算法及带反馈的 Markov 预测算法，并从理论上进行了分析，在实验中验证了算法的预测准确性.

1　基于 Markov 链和关联规则的预测

1.1　基于 Markov 链和关联规则的预测算法

二阶 Markov 链预测算法（MPA）（算法 1）如图 1 所示.

输入：用户的当前访问序列$(\dots, x_i, x_j)$
输出：预测页面或预测页面集
1. 根据当前访问序列构造$v(t)v(t-1)$.
2. 计算$\boldsymbol{x}(t+1)=\alpha_1 v(t)\boldsymbol{p}(1)+\alpha_2 v(t-1)\boldsymbol{p}(2)$，$\alpha_1+\alpha_2=1$.
3. 返回向量$\boldsymbol{x}$(t+1)中值最大的维对应的页面（或$\boldsymbol{x}(t+1)$中值大于某阈值的维对应的页面集），最后从$\boldsymbol{x}(t+1)$中选取概率最大的页面作为预测结果.

$\boldsymbol{p}(1)$表示一阶概率转移矩阵；$\boldsymbol{p}(2)$表示二阶概率转移矩阵；
α_1 为一阶转移矩阵的权值；α_2 为二阶转移矩阵的权值

图 1　二阶 Markov 链预测算法

命题 1　算法 1 的时间复杂度是 $O(n)$，其中 n 是 Web 站点的页面数.

证明 1　分析算法 1 可知，在算法第 2 步的计算中需要扫描 $\boldsymbol{p}(1)$的第 j 维 $p_j(1)$和 $\boldsymbol{p}(2)$的第 i 维 $p_i(2)$，由于 $\boldsymbol{p}(1)$和$\boldsymbol{p}(2)$是 n 维矩阵，故算法时间复杂度是 $O(n)$.

二项关联规则挖掘算法（算法 2）只需要对数据库进行一轮扫描，算法描述如图 2 所示.

输入：用户会话集S
输出：关联规则矩阵$\boldsymbol{R}$
1. while $(s_i \in S)$ //对于每一个会话 s_i
2. { while $(x_u \in s_i)$ //x_u初始为s_i的最后页面
3. 　{ for each $(x_v \in s_i - \{x_u, \dots\})$ //$s_i - \{x_u, \dots\}$ 表示在 x_u之前的页面
4. 　　{ $R_{x_v x_u}$++； //矩阵 R 中对应的数值加1
5. 　　　$x_v = x_v$.previous； //取页面 x_v前一个页面
6. 　　}
7. 　　R_{x_u}++；
8. 　$x_u = x_u$.previous；
9. }}
10. for each (x_u, x_v)
11. {if $((R_{x_v x_u}/|S|)$ <min_sup) //如果支持度小于阈值 min_sup
12. 　　　$R_{x_v x_u}=0$；
13. 　else $R_{x_v x_u}=R_{x_v x_u}/R_{x_v}$； //计算可信度
14.}

图 2　基于矩阵的二项关联规则挖掘算法

命题 2　算法 2 的时间复杂度是 $O(\lambda^2|S|)$，空间复杂度是 $O(n^2)$，其中$|S|$表示用户会话集（训练数据集）的大小.

证明 2　分析算法 2 可知，第 1 行的循环执行次数为$|S|$，第 2 行的循环执行次数$|s_i|$，第 3 行的循环针对会话 s_i 中的每个页面分别执行的次数为 1，2，…，$|s_i|$. $|s_i|$为用户访问序列的长度，服从泊松分布，故它的数学期望值是 λ. 第 3 行循环平均执行次数为$(1+\lambda)/2$，因此该算法执行的总时间为 $\lambda(1+\lambda)|S|/2$，算法 2 的时间复杂度是 $O(\lambda^2|S|)$. 执行算法 2 需要一个空间来存储一个 n 维关联矩阵，所以该算法的空间复杂度是 $O(n^2)$.

基于 Markov 链和关联规则的预测算法（MA-PA）（算法 3）如图 3 所示.

输入：当前访问序列$v=(t_1, t_2, \dots t)$，阈值 min_rule 和 max_rule
输出：预测页面
1. 利用二阶Markov 预测算法计算出用户有可能访问的页面子集 m 和对应的 markoy 预测概率 p_m；
2. for each $(r_i \in m)$
3. {predicProb $(r_i)=p_m(r_i)$； //初始化预测概率
4. 　for each $(P_j \in v)$
5. 　　{ conf $(P_j \rightarrow r_i)=R_{p_j r_i}$；} //从关联矩阵 $\boldsymbol{R}$ 中查找相应规则的可信度
6. 　　if (for each $P_j \in v$, conf $(P_j \rightarrow r_i)$<min_rule)
7. 　　{ remove r_i from m } //从 m 中删除 r_i
8. 　　for each $(P_j \in v)$
9. 　　{ if (conf $(P_j \rightarrow r_i)$>max_rule)
10. 　　　{ predictProb (r_i)+=$\lambda\omega_j$conf $(P_j \rightarrow r_i)$；} //计算预测概率
11. 　　}}
12. 返回predictProb中最大的页面

图 3　基于 Markov 链和关联规则的预测算法

命题 3　算法 3 的时间复杂度是 $O(n)$.

证明 3　根据命题 1 可知，算法 3 第 1 行的时间复杂度是 $O(n)$，第 2 行的循环执行次数为$|m|$，第 4、第 6、第 8 行的循环执行次数均为$|v|$. 其中：m 是 Web 站点页面的子集，其大小由 Markov 预测算法调控，一般均远远小于 n；$|v|$是用户访问的长度，是服从泊松分布的随机变量，故它的数学期望值是 λ. 所以，算法 3 的时间复杂度为 $O(n+\lambda|m|)=O(n)$.

1.2　实验结果及分析

1.2.1　*实验数据*　实验数据来源于 1998 年 2 月微软服务器（www.microsoft.com）用户的一周访问日志（http://kdd.ics.uci.edu/databases/msweb/msweb.html），实验训练数据和测试数据的页面数及会话数如表 1 所示.

定义1　预测准确率定义如下：①1-预测准确率，即

表 1　实验数据

数据	页面数	会话数	$\|v\|\geqslant 2$ 的会话数
训练	294	32 711	26 168
测试	294	5 000	3 452

预测的页面在下一步被访问的比率；②x-预测准确率，即预测的页面在将来（从当前至会话结束）被访问的比率.

1.2.2　基于 Markov 链和关联规则的预测分析　图 4 是 MAPA 和 MPA 的最大 1-预测准确率，图 5 是 MPA 和 MAPA 最大 x-预测准确率. 在 MAPA 中，$\alpha_1=0.5$. 经过反复实验，当最小规则可信度阈值（min_rule）取 10%、最大规则可信度阈值（max_rule）取 20%、β_1（Markov 的预测权值）取 0.9 时，β_2（关联规则的预测权值 $\beta_2=1-\beta_1$）为 0.1，MAPA 的 1-预测准确率最好；β_1，$\beta_2=0.5$ 时（其余参数不变），MAPA 的 x-预测准确率最好.

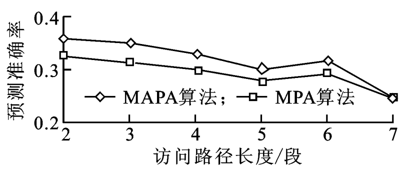

图 4　MPA 和 MAPA 最大 1-预测准确率

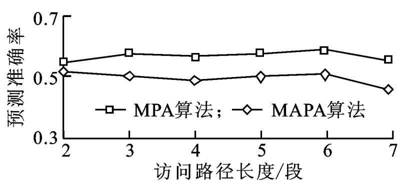

图 5　MPA 和 MAPA 最大 x-预测准确率

在 MPA 中，无论是 1-预测准确率还是 x-预测准确率，都是 $\alpha_1=0.5$ 时（代表一阶转移概率和二阶转移概率的权值相等）最好，也就是说当前访问页面和上一时刻的访问页面在 Markov 预测过程中的作用是等价的. 然而，在 MAPA 中，当 $\beta_2=0.1$（$\beta_1=0.9$）时，1-预测准确率最好，而当 $\beta_2=0.5$（$\beta_1=0.5$）时，x-预测准确率最好. β_2 为 0.1 说明关联规则修正对结果的影响较小，而 β_2 为 0.5 说明关联规则修正对结果影响较大. 根据经验可知，在一步转移过程中，当前已访问页面和下一访问页面之间的 Markov 性更好，在已访问页面和将来的访问页面之间关联关系更密切.

上述实验结果证明，在修正了关联规则以后，算法的预测准确率均有所提高，x-预测准确率的提高尤为明显. 由于用户访问 Web 站点的行为具有 Markov 性，而用户访问的页面之间具有关联性，所以用 Markov 链和关联规则相结合的算法吻合了用户的行为特征.

1.2.3　实验结果与点击流树算法对比　Web 日志挖掘是一个较新的研究领域，没有经典的参照算法和公认的实验测试数据. 预测算法应用于不同的数据集，预测准确率则有一定的差异，因此实验结果的对比并没有直接的参照意义，但可以作为一个侧面来体现算法的优越性. 本文引用了文献[6]的实验结果，在文献[6]中，准确率的计算和本文的 1-预测准确率相同，但是文献[6]推荐了 3 个页面，而本文只推荐了一个页面. 为了便于比较，本文允许预测算法推荐 3 个页面.

在实验测试数据中，本文的会话长度（单位：段）大于等于 8 的会话数较少，只有 295 个，这些会话的路径模式相似，故缺乏一般性和统计意义. 表 2 是预测算法推荐的 3 个页面最大 1-预测准确率. 如果忽略会话长度大于等于 8 的会话数，那么平均预测准确率可以达到 0.559（表 2 中前 5 项的平均值）. 表 3 是本文实验和文献[6]的对比.

表 2　预测算法推荐的 3 个页面的最大 1-预测准确率

会话长度/段	2	3	4	5	6	7	准确率平均值
预测准确率	0.570	0.574	0.565	0.564	0.520	0.457	0.542

表 3　本文算法与点击流树算法（CST）实验结果对比

算法	数据集	平均预测准确率	一次预测的平均时间/ms
MAPA	Msweb	0.025	0.559
CST	ClarkNet	0.080	0.551

虽然不同的算法作用于不同的数据集，预测准确率不具备直接的可比性，但是可以从一个侧面体现出算法的预测准确性. 一次预测的平均时间在忽略了 CPU 的频率以后是具备可比性的. 本文的实验是在奔腾 4、CPU 频率为 2.93 GHz、内存为 1 GB、操作系统为 Windows XP 及 Java 的环境上运行的，程序未经优化. 文献[6]的实验是在奔腾 4、CPU 频率为 2.4 GHz、内存为 552 MB、操作系统为 Windows XP 及 Java 的运行环境上运行的. 从实验

环境来看，本文算法是一次性将实验训练数据和测试数据读入内存，内存的开销在 100 MB 左右，因此程序运行以后不需要从硬盘读数据，内存对计算时间无影响. 扣除 CPU 频率对计算时间的影响，MAPA 的预测效率是 CST（折算 CPU 频率后，平均预测时间为 0.065 ms）的 2.5 倍.

基于 Markov 链和关联规则的预测算法只要维护 3 个矩阵（一阶转移概率矩阵、二阶转移概率矩阵和关联矩阵），这些矩阵不会随着 Web 日志的增多而增大. 相反，Web 日志数据量越大，这 3 个矩阵所保存的知识就越多，预测准确率越高. 预测的时间复杂度仅与这 3 个矩阵的维数（Web 站点页面的数量）有关系，不会受 Web 日志数据量的影响.

2　带反馈的 Markov 预测算法

用户访问预测的经典算法有一个共同的缺点，就是不考虑用户反馈，无法判断预测的结果是否正确，也无法动态地调整预测算法. 本文引入了用户反馈[7]机制并保存了一定量的历史预测记录 ξ，根据用户反馈来判断预测是否正确，从而建立一张包含预测准确率的历史预测树（HPT）. 预测过程中，可以直接使用正确率高的历史预测结果作为当前的预测值，或者把历史预测准确率作为当前预测的计算参数，动态地调整 Markov 预测算法，从而提高预测准确率和预测效率.

2.1　历史预测树

历史预测树构造算法（算法 4）如图 6 所示.

```
输入：用户当前访问序列v=(t1, t2, …, ts)，预测页面 t
输出：HPT
1. cnod = root； // 当前节点指向根节点
2. for each (t in v)  // 如果t在v中
3. { if (t in cnode.childList) 如果 t是当前节点的子节点
4.    { cnode = node which node.t = t in cnode.childList ;}
5.    else
6.    { newNode.t = t； // 新建立一个节点
7.       newNode.parent = cnode ;
8.       cnode.childList ∪= newNode ;
9.       cnode = newNode ;
10.      建立newNode 与包含 t 的节点进行链接 ;}
11 }
12. 把预测记录 η(0, t, 0 currentTime) 插入 cnode.hrList ;
```

图 6　历史预测树构造算法

2.2　带反馈的 Markov 预测算法（MPAF）

带反馈的 Markov 预测算法（MPAF）通过引入反馈机制建立了历史预测树，利用历史预测信息动态地修正了 Markov 预测算法的预测结果（简称为 Markov 的预测结果），从而保证低阶 Markov 的预测准确率. 该算法首先从 HPT 中查询当前会话的历史预测信息，如果 HPT 中存在预测正确率大于指定阈值（T_{good}）的预测页面，则直接把该页面作为当前会话的预测结果；否则，从 HPT 中找出当前会话的历史预测结果和预测正确率，从正向和反向来修正 Markov 的预测结果. 用户反馈是建立 HPT 的基础，反馈过程根据用户下一步访问页面来判断当前预测的正确性.

如果在查询过程中未找到预测页面，返回 ξ，并利用历史预测准确率从正向和反向来修正 Markov 的预测结果. 假设当前用户已访问的页面序列 $v=\{t_1, t_2, \cdots, t_s\}$，Markov 的预测结果集 $m=\{r_1, r_2, \cdots, r_t\}$，对应的 Markov 预测概率 $p_m=\{p_m(r_1), p_m(r_2), \cdots, p_m(r_t)\}$.

带反馈的 Markov 预测算法（算法 5）如图 7 所示.

```
输入：用户当前访问序列v=(t1, t2, …, ts)，历史预测记录
        (HRS)ξ，阈值 min_p
输出：预测页面 t
1. 利用二阶 Markov 预测算法计算出用户有可能访问的页面集 m
      和对应的 Markov 预测概率 pm
2. for each (η in ξ)
3. { if ((η.y / (η.x+η.y))<min_t &&η.t in m)
4.      {     remove η.t from m;
5.            remove η from ξ
6.      }
7. }
8.  for each (t in m)
9.  { if (t in ξ.t)  // 如果 t 在 ξ.t 中
10.         predictProb (t) = pm (t)+λhpp(t)
11 . eles  predictProb(ri) = pm (ri)+λpm(ri)
12. }
13. 返回predictProb 值最大的页面
```

图 7　带反馈的 Markov 预测算法

命题 4　算法 5 的时间复杂度是 $O(n)$.

证明 4　根据命题 1 可知，算法 5 第 1 行的时间复杂度是 $O(n)$，第 2 行的循环执行次数为 $|\xi|$，第 8 行的循环执行次数为 $|m||\xi|$. 其中：m 是 Markov 预测算法计算出用户有可能访问的页面集，它远远小于 n；ξ 是和当前会话对应的历史预测记录集，一般也远远小于 n，所以 $|m||\xi|<n$，故算法 5 的时间复杂度是 $O(n)$.

带反馈的 Markov 预测算法的在线部分的时间复杂度是 $O(n)+O(\lambda^2|S_{\mathrm{page}}|)$，其中 $|S_{\mathrm{page}}|$ 是在 HPT 中页面为 t_s 的处于叶子节点的个数，λ 是用户访问路径长度的数学期望值. 因此，只要把 HPT 的规模控制在适当的范围内，定时裁剪 HPT 中对预测指导意义不大的分支即可. 所以，带反馈的 Markov 预测算法具有很好的预测效率.

用户反馈是建立 HPT 的基础，反馈过程是根据用户的下一步访问页面来判断当前预测的正确性，如果判断正确，则对应 ξ 的正确字段(y)加 1，否则错误字段(x)加 1.

2.3 实验结果分析

2.3.1 实验数据 实验数据和预测准确率的定义同 1.2.1 节. 先建立 HPT，为了与原测试数据(1.2.1 节中定义的)不重复，从实验训练数据中随机抽取了 10 954 个会话来建立 HPT 的训练数据，再在已有 HPT 的基础上对原测试数据进行测试.

2.3.2 MPAF 的实验分析 经过多次的实验观察发现，MPAF 在 T_{good}=0.8、阈值 min_ p 为 0.1 时，预测性能较好，结果如表 4、表 5 所示. T_{good}=0.8 表示，对于当前会话，如果在 ξ 中存在预测准确率大于 0.8 的预测结果，就把该结果作为当前的预测结果，无需调用预测算法. min_ p 为 0.1 表示，对于当前会话，如果在 ξ 中存在预测准确率低于 0.1 的预测结果，并且该结果在当前预测的候选集中，就将其从候选集中删除.

表 4 T_{good}=0.8 时 MPAF 获得预测结果的次数

预测总次数	获得预测结果的次数	预测率
6 051	1 413	0.234

表 5 min_ p 为 0.1 时 MPAF 从 Markov 候选集中删除预测结果的次数

候选集中包含的总页面数	预测结果的删除次数	删除次数的比率
22 760	3 559	0.156

MPA、MAPA 和 MPAF 最大 1-预测准确率比较如图 8 所. 从图中可看出，当 T_{good}=0.8、min_ p 为 0.1、历史预测的权值 β 为 0.5 时，MPAF 的最大 1-预测准确率比较高.

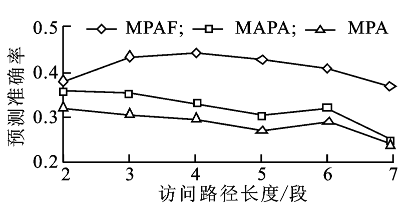

图 8 最大 1-预测准确率比较

从理论上讲，如果不考虑 HPT 的规模，那么随着历史预测次数的增多，HPT 提供的参考意义更大，因此可进一步提高预测准确率. 为了限制 HPT 的大小，本文在实验中裁剪了历史预测次数小于 5 的分支和预测正确率在区间[0.4，0.6]的分支.

MPA、MAPA 和 MPAF 最大 x-预测准确率比较如图 9 所示. 从图中可以看出，当 T_{good}=0.8、min_ p 为 0.1、历史预测的权值 β 为 0.1 时，MAPA 最大 x-预测准确率比较高.

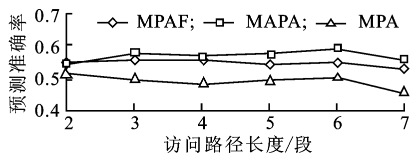

图 9 最大 x-预测准确率比较

MPAF 的 x-预测准确率略低于 MAPA 的原因是，在 MPAF 中，反馈是基于用户的下一个访问，即在 HPT 中，ξ 保存的是 1-预测准确率，所以 MPAF 对于多步转移的预测效果较差. 当然，反馈也可以基于多步访问，也就是说，可以让 HPT 中的 ξ 保存 x-预测准确率，那么 1-预测准确率就会下降. 如果让 HPT 同时保存 1-预测准确率和 x-预测准确率，就会增加预测的计算复杂度，因此可以根据需要，建立 2 颗 HPT 来分别保存 1-预测准确率和 x-预测准确率.

3 结 论

通过对 Web 日志进行挖掘，可以帮助站点管理者发现用户访问页面的行为规律，理解用户的行为意图，预测用户将来的访问页面，并把预测的页面提前发送给用户，从而改善用户的访问效率，为用户提供个性化服务[8]. 在 Internet 中，应将用户从被动的寻找信息转化为主动感知浏览者的需求，同时利用预测的页面优化代理服务器或 Web 服务器的缓存置换策略，才能真正改善服务器的性能和设计.

本文分析了已有的基于 Web 日志挖掘的用户访问预测算法的优缺点，提出了基于 Markov 链和关联规则的预测算法和带反馈的 Markov 预测算法. 理论分析和实验结果表明，本文的 2 个预测算法在预测准确率上有明显提高. 为了进一步提高预测准确率，可以采用高效的处理方法对 Web 日志进行预处理，同时采用 Web 内容挖掘和 Web 日志挖掘相结合的办法[9]解决基于 Web 日志挖掘的用户访问预测的滞后性问题. 这些研究将是本文下一步的工作.

参考文献：

[1] GÉRY M, HADDAD H. Evaluation of web usage mining approaches for user's next request prediction [C] // Proceedings of the 5th ACM International Workshop on Web Information and Data Management. New York, USA: ACM, 2003: 74-81.

[2] ZHANG Zhili, SHI Lei, GUO Shen, et al. Appling association rule to Web prediction [C] // Proceedings of the 1st International Multi-Symposiums on Computer and Computational Sciences. Los Alamitos, CA, USA: IEEE Computer Society, 2006: 522-527.

[3] ZUKERMAN I, ALBRECHT D W, NICHOLSON A E. Predicting user's requests on the WWW[C] // Proceedings of the 7th International Conference on User Modeling. Berlin, Germany: Springer, 1999: 275-284.

[4] 刑永康，马少平. 多 Markov 链用户浏览预测模型[J]. 计算机学报，2003，26(11)：1510-1517.
XING Yongkang, MA Shaoping. Modeling user navigation sequences based on multi-Markov chains[J]. Chinese Journal of Computers, 2003, 26(11): 1510-1517.

[5] GÜNDÜZ S, ÖZSU M T. A web page prediction model based on click-stream tree representation of user behavior [C] // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2003: 535-540.

[6] ÖGÜDÜCÜ S G, ÖZSU M T. Incremental click-stream tree model: learning from new users for web page prediction [J]. Distributed and Parallel Databases, 2006, 19(1): 5-27.

[7] HUANG Yinfu, HSU J M. Mining web logs to improve hit ratios of prefetching and caching [J]. Knowledge-Based Systems, 2008, 21(1): 62-69.

[8] SUTHEERA P, HIDEKAZU T. Mining web logs for a personalized recommender system[J]. Joho Shori Gakkai Zenkoku Taikai Koen Ronbunshu, 2005, 67(3): 19-20.

[9] LIU Haibin, KEŠELJ V. Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests [J]. Data & Knowledge Engineering, 2007, 61(2): 304-330.

（编辑　苗凌）