doi:10.3969/j.issn.1005-3026.2016.06.004

# 基于马尔科夫模型和贝叶斯定理的Web用户浏览行为预测模型

毕 猛[1,2], 侯 林[1], 倪 盼[3], 周福才[1]
(1. 东北大学 软件学院, 辽宁 沈阳 110169; 2. 沈阳工业大学 管理学院, 辽宁 沈阳 110023;
3. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110819)

**摘 要**: 对用户的Web浏览行为进行分析,既可以使用户减少等待时间,同时也能减轻网络负载. 依据Web网站的层次结构特点,首先设计了基于Hash表的反向索引结构来提高数据的预处理速度. 在此基础上,利用分层思想构建了基于马尔科夫模型和贝叶斯定理的Web用户浏览行为预测模型. 给出了模型的设计思想、相关定义、模型框架以及模型中所涉及的关键构建方法等. 最后,对模型进行了实验分析,结果表明在适当的预测准确率前提下,模型能够有效减少在预测时所需的候选网页数量,并大幅提升预测效率.

**关 键 词**: Web站点;用户浏览行为预测;马尔科夫模型;贝叶斯定理
**中图分类号**: TP 309 **文献标志码**: A **文章编号**: 1005-3026(2016)06-0775-06

# Users' Web Browsing Behavior Prediction Model Based on Markov Model and Bayesian Theorem

*BI Meng*[1,2], *HOU Lin*[1], *NI Pan*[3], *ZHOU Fu-cai*[1]
(1. Software College, Northeastern University, Shenyang 110169, China; 2. Management College, Shenyang University of Technology, Shenyang 110023, China; 3. School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: ZHOU Fu-cai, E-mail: fczhou@mail.neu.edu.cn)

**Abstract**: According to the novel aspect of natural hierarchical property of Web site, the inverted index structure was proposed based on Hash table (IIS-HT) to promote the speed of data preprocessing. Based on IIS-HT, a prediction model was also proposed which was based on statistics to predict users' browsing behavior. The design idea, definition, framework and key construction methods of the model were also given. Finally, the proposed model was tested with real data. The experimental results show that the model and prediction algorithm could reduce the scope of candidate pages and improve the speed of prediction with adequate accuracy.

**Key words**: Web site; users' browsing behavior prediction; Markov model; Bayesian theorem

Web挖掘技术[1]因为具有改善网站服务质量、提高用户满意度以及提供个人化服务等特点,而受到企业界和学术界的日益关注. 目前的Web用户浏览情况挖掘就是对网站的日志记录进行分析,从而提取用户浏览网站的特征信息[2-3]. 目前已有一些学者提出了针对Web用户浏览行为分析的数学模型. 文献[4]利用马尔科夫模型来分析预测用户下一个可能浏览的网页. 文献[5]提出了两个模型来研究Web浏览行为的特征化分析,以Morse's模型及马尔科夫模型分别对网页访问及使用的情况予以模型化分析. 文献[6-7]利用高阶马尔科夫(Markov)模型来预测用户浏览行为,然而,当所使用的阶数 $n$ 越高,计算代价也越高. 文献[8]提出了一种使用贝叶斯(Bayesian)分类算法和领域本体过滤中文网页的方法,该方法可以区分相同领域中的不同网页并

收稿日期: 2015-04-09
基金项目: 国家科技重大专项(2013ZX03002006);辽宁省科技攻关项目(2013217004);辽宁省博士启动基金资助项目(20141012);沈阳市科技计划项目(F14-231-1-08);中央高校基本科研业务费专项资金资助项目(N130317002).
作者简介: 毕 猛(1982-),男,辽宁沈阳人,东北大学博士研究生,沈阳工业大学工程师;周福才(1964-),男,吉林长春人,东北大学教授,博士生导师.

可兼顾网页过滤的实时性. 虽然上述研究在 Web 用户浏览行为分析和预测方面取得了一些成果,但是也存在计算代价高、数据预处理时间长、预测准确率低等问题.

针对上述问题,依据 Web 网站的分层结构特点[9]来进行用户浏览行为预测. 首先,提出了基于 Hash 表的反向索引结构(inverted index structure based on Hash table,IIS - HT),并通过 IIS - HT 来提升数据预处理效率. 然后,针对 Web 网站的分层特性,并结合马尔科夫模型与贝叶斯定理,提出了基于马尔科夫模型与贝叶斯定理的分层预测模型(hierarchical prediction model based on Markov model and Bayesian theorem, HPM - MMBT). 在 HPM - MMBT 中,将用户的浏览行为依据网站的分层特点分为网页类别(分层 - 1)和网页(分层 - 2). 在网站中的每个不同的类别都会包含许多网页,而网页类别的数量一定小于网页的数量,所以相较各网页类别之间的转换概率相对稳定,因此在分层 - 1 可以利用马尔科夫模型过滤出用户最可能浏览的网页类别. 当可能的网页类别被过滤出之后,在分层 - 2 以贝叶斯定理作为预测方法,从而找出用户最有可能浏览的网页. 最后,利用网站日志文档[10]对模型进行了相关测试.

## 1 基于 IIS - HT 的数据预处理方法

在数据预处理阶段,可以利用反向索引结构(如图 1 所示)进行网站日志数据的预处理. 在反向索引结构中,每读取一条日志记录,就会获得一个用户的 IP 地址,将该 IP 地址作为搜索关键字,在用户列表(user list)中查找与该 IP 地址对应的位置,然后将该网页信息加入到网页信息列表(page list)中. 显然,反向索引结构易于实现;但是,当网站日志记录很大的时候,用户列表将会变得很长,进而导致查询和对比开销变得极为庞大.

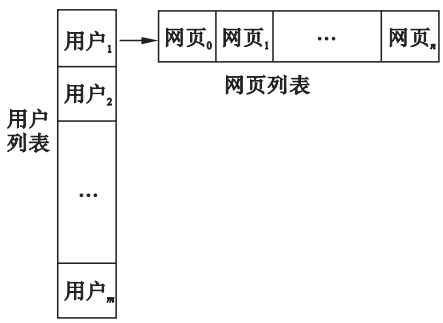


图 1 反向索引结构

Fig. 1 Inverted index structure

为了解决上述问题,将 Hash 链表和反向索引结构结合,设计了基于 Hash 表的反向索引结构,如图 2 所示. IIS - HT 包括两部分:第一部分是由用户列表和与之连接的 Hash 表组成,每读取一条日志记录时,就会获得一个用户的 IP 地址,将该 IP 地址的每一个字符转换为 ACII 码并累加求和,并对该求和结果进行模运算,最后将运算结果作为 Hash 表的索引值;在相对应的 Hash 索引值之下找到用户列表,然后进入 IIS - HT 的第二个组成部分,即反向索引. 在该反向索引结果中,搜索和比较用户 IP,如果该 IP 已经存在于用户列表中,则在网页列表中增加该网页信息,否则将该 IP 地址加入到用户列表中.

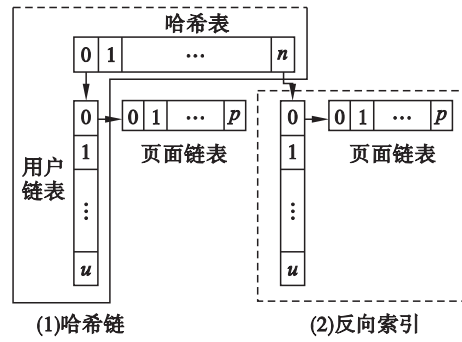


图 2 基于 Hash 表的反向索引结构

Fig. 2 Inverted index structure based on Hash table

## 2 HPM - MMBT 模型设计

### 2.1 设计思想

结合马尔科模型和贝叶斯定理,本文所设计的面向 Web 用户浏览行为的预测模型架构如图 3 所示. 其中 $S$ 是网页相似度矩阵,$P$ 是马尔科夫转换矩阵,$R$ 是相关性矩阵.

首先,依据 Web 网站的分层属性,将其分为两类:即网页类别(如某个网站的新闻类网页、娱乐网页等类别)和网页. 本文将网页类别定义为分层 - 1,网页定义为分层 - 2. 在网站中,不同的网页类别会包含很多不同的网页,但是网页类别的数量是明显小于网页数据量的. 因此,网页类别之间的转换是相对稳定的,所以在分层 - 1 可以利用马尔科夫模型对用户可能浏览的网页类别进行过滤. 依据 $t-1$ 和 $t-2$ 时刻用户浏览网页类别的情况,就可以对 $t$ 时刻用户浏览网页类别的情况进行预测. 在预测出用户在 $t$ 时刻可能浏览的网页类别后,在分层 - 2 可以利用贝叶斯定理并结合用户在 $t-1$ 时刻的状态来预测用户在 $t$ 时刻可能浏览的网页.

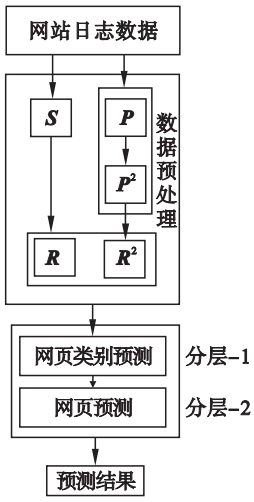


图 3　HPM – MMBT 模型
Fig. 3　HPM-MMBT model

## 2.2　相关定义

定义 1 是本文用到的相关定义的描述.

**定义 1**

$$\text{hierarchy}=\begin{cases}\text{depth}, & \text{depth}\leqslant\text{hierarchy};\\ \text{hierarchy}, & \text{else}.\end{cases}$$

depth 表示网页类别的深度(即网页类别目录的深度),hierarchy 表示 depth 的观测值,即那一层网页类别需要被观测. 如果 depth 小于等于 hierarchy,则 hierarchy 等于 depth;否则,hierarchy 为设定值.

对于某用户的日志记录集合 $R=\{\text{Record}_1,\text{Record}_2,\cdots,\text{Record}_i,\cdots,\text{Record}_m\}$,它保存了 $m$ 条用户访问记录,每一个访问记录 Record 都是一个访问序列,该访问序列包含了 $n$ 个被用户依时间顺序浏览的网页,即 $\text{Record}_i=\{\text{page}_{i1},\text{page}_{i2},\cdots,\text{page}_{in}\}$. $\text{Record}_i$ 代表了用户 $i$ 的浏览网页的情形. 当 hierarchy 的值为 1 时,根据 $\text{Record}_i$ 中各网页所属的类别,给出 $\text{Record}_i$ 的定义:

**定义 2**

$\text{Record}_i=\{(\text{category}_{i1},\text{page}_{i1}),(\text{category}_{i2},\text{page}_{i2}),\cdots,(\text{category}_{in},\text{page}_{in})\}$.

$\text{Record}_i$ 表示用户浏览网页类别和具体网页的情况,$\text{category}_{il}$ 表示第 $l$ 个网页的类别.

## 2.3　模型构建

### 2.3.1　网页相似度矩阵的建立

首先,需要获得用户浏览网站的日志文件. 之后,当 hierarchy $=l$ 时,假设有 $k$ 个网页类别. 因此,当用户在 $l$ 层网页类别浏览时,可以从 $m$ 个用户浏览记录中获得 $m\times k$ 矩阵. 在该矩阵中,行向量为 $V_i^{\text{col}}=<C_{1,i},C_{2i},\cdots,C_{hi},\cdots,C_{mi}>$. $V_i^{\text{col}}$ 表示网页类别 $i$ 被第 $m$ 个用户浏览过. 如果 $V_i^{\text{col}}$ 中网页类 $i$ 被第 $h$ 个用户浏览过,则 $C_{hi}=1$,否则 $C_{hi}=0$.

本文利用集合相似度式(1)和欧几里得距离式(2)来计算两个不同网页类别的相似度.

$$\text{SetSim}(V_i^{\text{col}},V_j^{\text{col}})=\frac{|V_i^{\text{col}}\cap V_j^{\text{col}}|}{|V_i^{\text{col}}\cap V_j^{\text{col}}|},\quad(1)$$

$$D(V_i^{\text{col}},V_j^{\text{col}})=\sqrt{\sum_{k=1}^{m}(V_{ki}^{\text{col}},V_{kj}^{\text{col}})^2}.\quad(2)$$

将式(2)进行标准化,得到式(3):

$$\text{ND}(V_i^{\text{col}},V_j^{\text{col}})=1-\sqrt{\frac{\sum_{k=1}^{m}(V_{ki}^{\text{col}},V_{kj}^{\text{col}})^2}{m}}.\quad(3)$$

为式(1)和(3)赋予不同权重值,得到式(4):

$$S(V_i^{\text{col}},V_j^{\text{col}})=\text{SetSim}(V_i^{\text{col}},V_j^{\text{col}})\cdot W_{\text{SS}}+\text{ND}(V_i^{\text{col}},V_j^{\text{col}})\cdot W_D.\quad(4)$$

其中,$W_{\text{SS}}+W_D=1$,$W_{\text{SS}},W_D>0$.

在获得 2 个网页类别的相似度后,可以得到相似度矩阵 $\boldsymbol{S}$:

$$\boldsymbol{S}=\begin{bmatrix}S_{11}&\cdots&S_{1k}\\ \square&&\square\\ S_{k1}&\cdots&S_{kk}\end{bmatrix}.$$

$S_{ij}$ 是通过式(4)计算得出的.

### 2.3.2　马尔科夫转换矩阵的建立

与相似度矩阵 $\boldsymbol{S}$ 的建立方法相类似,同样是利用用户的浏览行为日志记录来建立一阶和二阶马尔科夫转换矩阵.

首先建立一阶马尔科夫转换矩阵,如式(5)所示. 二阶以及 $n$ 阶马尔科夫转换矩阵可以通过一阶马尔科夫转换矩阵得到.

$$\boldsymbol{P}=\begin{pmatrix}P_{11}&\cdots&P_{1k}\\ \square&&\square\\ P_{k1}&\cdots&P_{kk}\end{pmatrix},P_{ij}=\frac{\text{Number}(i,j)}{\sum_{j=1}^{k}\text{TotalNumber}(i,j)}.\quad(5)$$

在矩阵 $\boldsymbol{P}$ 中,元素 $P_{ij}$ 代表从一个网页类别转向另一个网页类别的概率(例如,从娱乐类网页转向体育类网页的概率). 在式(5)中,分母表示从网页类别 $i$ 转换到全部 $k$ 个网页类别的总次数,分子表示所有从网页类别 $i$ 转换到网页类别 $j$ 的总次数.

### 2.3.3　相关性矩阵的建立

将相似度矩阵 $\boldsymbol{S}$ 和 $n$ 阶马尔科夫转换矩阵 $\boldsymbol{P}^n$ 中的对应位置相乘就可以获得一个相关性矩阵 $\boldsymbol{R}^n$,如式(6)所示.

$$\boldsymbol{R}^n=\begin{bmatrix}R_{11}^n&\cdots&R_{1k}^n\\ \square&&\square\\ R_{k1}^n&\cdots&R_{kk}^n\end{bmatrix},R_{ij}^n=S_{ij}\cdot P_{ij}^n.\quad(6)$$

在相关性矩阵 $\boldsymbol{R}^n$ 中每一个元素都表示两个网页类别直接的相关程度.

2.3.4 用户浏览行为预测方法

依据 2.1 节的分层预测模型,在分层 -1 可以过滤出用户下一步可能访问的网页类别集合;在分层 -2,可以利用贝叶斯定理对网页类别集合中的网页进行预测,并获得可能被用户浏览的网页结合.

为了减少预测过程中所花费的代价开销,依据 2.1 节提出的分层预测模型框架,从分层 -1 和分层 -2 进行分类预测.

1) 分层 -1 中的网页类别预测. 在分层 -1 中,主要是对用户浏览的网页类别进行预测. 如图 4 所示,利用网页类别集合 $C_{t-1}$ 和 $C_{t-2}$,可以得到 $C_t$ 的预测集合 $\theta$. 在分层 -2 中只对预测集合 $\theta$ 中的网页进行预测,进而减少分层 -2 中的相关代价开销.

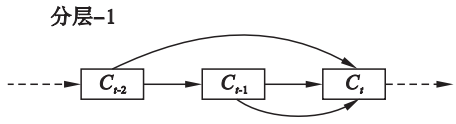


图 4 分层 -1 中的网页类别预测

Fig. 4 Category prediction in level one

在分层 -1 中,可以通过相关矩阵获得可能的网页类别集合. $\boldsymbol{R}^n_{C_{t-n}}$ 是相关矩阵 $\boldsymbol{R}$ 的 $C_{t-n}$ 列的列向量 $< C_{t-n,1}, C_{t-n,2}, \cdots, C_{t-n,k} >$. 当选择 $n=1$ 和 $n=2$ 时,可得 $\boldsymbol{R}^1_{C_{t-1}}$ 和 $\boldsymbol{R}^2_{C_{t-2}}$. 取出 top $-r_1$ 个网页类别构成集合 $\theta$.

假设一个 Web 站点有 3 个网页类别,且 $n=1$ 和 $n=2$ 时,则可以得到 $3\times 3$ 的相关矩阵 $\boldsymbol{R}^1$ 和 $\boldsymbol{R}^2$. 当 $C_{t-1}=C_2$ 和 $C_{t-2}=C_1$ 时,可以得到列向量 $\boldsymbol{R}^1_{C_2} = <0.13, 0.18, 0.27>$ 和 $\boldsymbol{R}^2_{C_1} = <0.16, 0.23, 0.12>$. 当取出前 top $-r_1$ ($r_1=2$) 个相关网页类别时,只有 $\boldsymbol{R}^1$ 中的 $C_3$ (0.27) 和 $\boldsymbol{R}^2$ 中 $C_2$ (0.23) 满足要求. 因此,在分层 -1 中的预测网页类别集合 $\theta = \{C_2, C_3\}$.

$$\boldsymbol{R}^1 = \begin{bmatrix} 0.25 & 0.15 & 0.34 \\ 0.13 & 0.18 & 0.27 \\ 0.22 & 0.16 & 0.26 \end{bmatrix}, \boldsymbol{R}^2 = \begin{bmatrix} 0.16 & 0.23 & 0.12 \\ 0.24 & 0.18 & 0.12 \\ 0.10 & 0.11 & 0.21 \end{bmatrix}$$

2) 分层 -2 中的网页预测. 如图 5 所示,在分层 -2 中,利用贝叶斯定理计算下一个可能浏览的网页 $\text{page}_b$,并获得最可能被用户浏览的前 top $-r_2$ 页面. 预测的网页集合 $\tau$ 将作为分层 -2 的最终的输出结果.

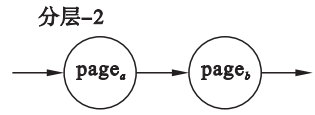


图 5 分层 -2 中的网页预测

Fig. 5 Page prediction in level two

由于在分层 -1 中已经完成了网页类别的预测工作,因此 $\text{page}_b$ 一定属于网页类别集合 $\theta$. 下面,定义 $P_\theta$ 作为所有可能的候选网页集合.

$P_\theta = \{\text{page}_{b_j} | \text{category of page}_{b_j} \text{must belong to predicted category set } \theta, 1 \leqslant j \leqslant r\}$. 将 $P_\theta$ 代入贝叶斯定理中计算,得到式(7):

$$P(\text{page}_{b_i} | \text{page}_a) = \frac{P(\text{page}_a | \text{page}_{b_i}) P(\text{page}_{b_i})}{P(\text{page}_a)} = \frac{P(\text{page}_a | \text{page}_{b_i}) P(\text{page}_{b_i})}{\sum_{j=1}^{r} P(\text{page}_a | \text{page}_{b_j}) \text{page}_{b_j}}. \quad (7)$$

同时,还可以得到候选网页集合 $\tau$:

$\tau = \{\text{page}_{b_j} | \text{page}_{b_j}$ 是 $P_\theta$ 中通过贝叶斯定理计算得到的前 top $-r_2$ 个网页$\}$. 因为已有的网页类别集合 $\theta = \{C_2, C_3\}$,因此在分层 -2 中,只需要考虑这 2 个网页类别集合中的网页.

在分层 -2 中,候选网页结合 $P_\theta$ 为 $P_\theta = \{\text{page}_{C_2,1}, \text{page}_{C_2,2}, \text{page}_{C_2,3}, \text{page}_{C_3,1}, \text{page}_{C_3,2}\}$.

## 3 实验分析

### 3.1 测试数据

选择 1995 年 7 月 NASA 网站的日志文件[10] 作为测试数据,该日志文件大小为 205.2 MB,经过预处理后,该日志文件的具体统计信息情况如表 3 所示.

表 1 日志文件信息

Table 1 Information of log file

| 项目 | NASA_Jul |
|---|---|
| 用户个数 | 24 787 |
| Session 个数 | 51 681 |
| 浏览的平均长度 | 2.72 |

依据实际中的不同应用情况,本文将上述测试数据集分为两类:大数据集和小数据集;其中大数据集合包含 1 000 条日志记录,小数据集合包含 100 条日志记录.

在本文中,使用预测准确率 (prediction accuracy,PA) 来评估模型的实际应用效果. PA 的计算公式为

$$PA = \frac{cache_{access}}{request_{all}}. \quad (8)$$

### 3.2　测试结果分析

分别从大数据集合和小数据集合两个方面，对 HPM – MMBT 的预测准确率进行测试，并与仅采用 Markov 方法和仅采用 Bayesian 定理的方法进行对比分析. 设定 top – $r_1$ = 10，top – $r_2$ = 10.

1）采用小数据集测试时的预测准确率分析. 如表 2 所示，当 hierarchy = 1 ~ 2 时，HPM – MMBT 的预测准确率 PA 比 Bayesian 方案和一阶 Markov 方案略高. 当 hierarchy = 3 ~ 5，HPM – MMBT 的预测准确率 PA 比 Bayesian 方案和一阶 Markov 方案略低.

如表 3 所示，在分层 – 1 中获得了网页类别预测集合 $\theta$ 之后，随着网站层数的增加，在分层 – 2中的每一层待预测网页的数量远小于每一层的网页总数，特别是在层数 = 3 ~ 5 时，预测的集合 $P_\theta$ 中的网页数量分别仅占每一层网页总数量的 15.49%，8.33% 和 8.82%.

表 2　采用小数据集测试时的预测准确率

Table 2　Hit-ratio to small data　%

| 层数 | Bayesian | Markov | HPM-MMBT |
|---|---|---|---|
| 1 | 32.54 | 32.54 | 33.18 |
| 2 | 52.13 | 52.13 | 52.69 |
| 3 | 41.98 | 40.37 | 40.04 |
| 4 | 47.46 | 46.42 | 40.92 |
| 5 | 42.91 | 43.15 | 41.39 |
| 平均值 | 43.41 | 42.92 | 41.64 |

表 3　采用小数据集测试时的候选网页情况

Table 3　Candidate page set to small data

| 层数 | HPM – MMBT | 百分比/% |
|---|---|---|
| 1 | 544.82 | 75.25 |
| 2 | 392.53 | 54.14 |
| 3 | 112.93 | 15.49 |
| 4 | 60.31 | 8.33 |
| 5 | 63.78 | 8.82 |

从表 2，表 3 可知，HPM – MMBT 在保证一定预测准确率的情况下，可以减少预测范围，提高预测效率.

2）采用大数据集测试时的预测准确率分析. 如表 4 所示，当 hierarchy = 1 ~ 2 时，HPM – MMBT 的预测准确率 PA 比 Bayesian 方案和一阶 Markov 方案略高. 当 hierarchy = 3 ~ 5，HPM – MMBT 的预测准确率 PA 比 Bayesian 方案和一阶 Markov 方案略低.

表 4　采用大数据集测试时的预测准确率

Table 4　Hit-ratio to large data　%

| 层数 | Bayesian | Markov | HPM-MMBT |
|---|---|---|---|
| 1 | 62.42 | 62.67 | 63.12 |
| 2 | 68.54 | 68.41 | 68.96 |
| 3 | 65.83 | 66.12 | 63.35 |
| 4 | 64.84 | 64.59 | 61.89 |
| 5 | 69.24 | 69.73 | 63.96 |
| 平均值 | 66.17 | 66.31 | 64.26 |

如表 5 所示，在分层 – 1 中获得了网页类别预测集合 $\theta$ 之后，随着网站层数的增加，在分层 – 2中的每一层待预测网页的数量远小于每一层的网页总数，特别是在层数 = 3 ~ 5 时，预测的集合 $P_\theta$ 中的网页数量分别仅占每一层网页总数量的 15.49%，8.33% 和 8.82%. 特别是在层数 = 3 ~ 5 时，预测的集合 $P_\theta$ 中的网页数量分别仅占每一层网页总数量的 13.91%，11.49%，11.69%.

表 5　采用大数据集测试时的候选网页情况

Table 5　Candidate page set to large data

| 层数 | HPM – MMBT | 百分比/% |
|---|---|---|
| 1 | 588.72 | 81.21 |
| 2 | 392.68 | 54.16 |
| 3 | 101.47 | 13.91 |
| 4 | 83.24 | 11.49 |
| 5 | 84.51 | 11.69 |

通过上面的实验分析可知，在适当的预测准确率前提下，模型能够有效地减少在预测时所需的候选网页数量并大幅提升预测效率.

## 4　结　　语

本文首先给出了基于 IIS – HT 的数据预处理方法，并以此为基础构建了 HPM – MMBT 模型. 对 HPM – MMBT 模型进行实验分析，结果表明，该模型的预测准确率与 Bayesian 方案和一阶 Markov 方案基本一致，但是在预测时所需的候选网页数量减少，实现了大幅度提升预测效率的目的.

### 参考文献：

[1] Li M J，Yu X M，Ryu K H. MapReduce-based web mining for prediction of web-user navigation [J]. *Journal of Information Science*，2014，40(5)：557 – 567.

[2] Torres S D，Hiemstra D. Analysis of search and browsing behavior of young users on the web [J]. *ACM Transactions on the Web*，2014，8(2)：1 – 54.

（下转第 799 页）