doi:10.3969/j.issn.1000-436x.2014.09.011

融合停留时间的隐 Markov 个性化推荐模型

刘胜宗[1]，樊晓平[1,2]，廖志芳[3]，胡佳[1]

(1. 中南大学 信息科学与工程学院，湖南 长沙 410075；2. 湖南财政经济学院 网络化系统研究所，湖南 长沙 410205；
3. 中南大学 软件学院，湖南 长沙 410075)

摘 要：静态模型在推荐系统中往往将用户的兴趣偏好看作是固定不变的，而在一定程度上与实际并不符合。为此，基于隐 Markov 动态模型提出一种融合停留时间的类时齐隐 Markov 个性化推荐模型（ctqHMM）。该模型用隐含状态变量的转移来模拟 Web 用户的兴趣变迁，并用停留时间来描述用户对某一偏好感兴趣的程度和所推荐页面的重要性。然后，提出一种基于该模型平稳分布的用户聚类方法，并将其用于推荐系统中。在真实的 Web 服务器访问记录数据上的实验证明，类时齐隐 Markov 模型具有更好的推荐性能。

关键词：Web 挖掘；类时齐隐 Markov 模型；平稳分布；用户聚类；个性化推荐；HMM

中图分类号：TN911.23 **文献标识码**：A **文章编号**：1000-436X(2014)09-0112-10

Hidden Markov model fused with staying time for personalized recommendation

LIU Sheng-zong[1], FAN Xiao-ping[1,2], LIAO Zhi-fang[3], HU Jia[1]

(1. School of Information Science and Engineering, Central South University, Changsha 410075, China;
2. Laboratory of Networked Systems, Hunan University of Finance and Economics, Changsha 410205, China;
3. School of Software, Central South University, Changsha 410075, China)

Abstract: Static model in the recommendation system often regards the user's interest as changeless, which is inconsistent with the actual to a certain extent. With regards to this, a hidden Markov model fused with staying time for personalized recommendation (ctqHMM) based on the HMM dynamic model is proposed. The proposed model employs the transfer of the implicit state variables to simulate the changes of Web users' interests, and uses staying time to describe the level of interest to the specific preference and the importance of the recommended pages. Then, a user's clustering method based on the stationary distribution of the ctqHMM is also proposed and applied into the recommending systems. Experiment results on real Web server access log data show the encouraging performance of the proposed method over the state-of-the-arts.

Key words: Web mining; classified time homogeneous hidden Markov model; stationary distribution; user clustering; personalized recommendation; HMM

1 引言

个性化推荐系统用来发掘 Web 用户感兴趣的信息，帮助用户从海量的信息中找到他们喜欢的但并未浏览或标注过的资源，解决信息过载问题[1~3]。目前大部分的推荐系统研究将用户的偏好看作固定不变的静态模式，然而在现实环境中，用户偏好会随时间变化。有研究表明，用户的偏好在新事物的刺激下随着个人兴趣的演化而改变[2]。而传统的推荐系统能够识别用户的偏好，但不能跟踪偏好的

收稿日期：2014-01-08；**修回日期**：2014-02-06

基金项目：国家科技支撑计划基金资助项目(2012BAH08B00)；国家自然科学基金资助项目(61073105)；湖南省自然科学基金资助项目(12JJ3074)；湖南省科技支撑计划基金资助项目(2012GK4006)

Foundation Items: The National Key Technology R&D Program of China(2012BAH08B00); The National Natural Science Foundation of China(61073105); The Natural Science Foundation of Hunan Province (12JJ3074); Key Technology R&D Program of Hunan Province (2012GK4006)

转移，即静态偏好模型具有局限性，因此，采用动态偏好模型有助于改善这种状况。

静态模型存在数据稀疏问题，而在很多 Web 系统中，用户访问记录可以通过服务器日志获取，充分利用这些信息可以缓解稀疏问题[1,4~6]。文献[7]提出一阶 Markov 模型预测用户的后续访问页面，该模型在无噪数据中取得较好的精度；文献[1]结合关联规则挖掘改进了传统 Markov 链预测方法，降低了全阶 Markov 模型的时间复杂度；文献[8]提出多 Markov 链并对不同类别的用户选用对应的 Markov 链进行预测；文献[9]利用半马氏过程对网页进行排序，取得了很好的效果，但该方法并没挖掘用户的个性化需求。

为了跟踪隐藏在用户浏览页面序列[4,10,11]及停留时间分布中的用户偏好转移过程，本文提出分类时间齐次隐 Markov 推荐模型（ctqHMM），简称类时齐隐 Markov 模型，该模型将停留时间融合到隐 Markov 链中，并使用其平稳分布来表征用户的兴趣特征，该模型包括学习和预测 2 个过程，分别对应用户聚类和个性化推荐。实验验证该方法具有更好的推荐效果。

2　类隐 Markov 链偏好分析模型

由于用户的浏览行为一般是在自身偏好驱使下进行，因此用户的偏好往往隐藏在用户的页面访问序列中[2,7,12]。

定义 1　设 v 表示某用户在任一时刻的偏好，那么所有时刻该用户的访问偏好构成用户的访问偏好集 $V=\{v_1,v_2,\cdots,v_N\}$。

用户的访问偏好又称为用户访问兴趣概念，表示用户所感兴趣的某类主题，在具体分析过程中，偏好往往用资源所属的类别或者相应的标签[13~15]来对应。

定义 2　Web 访问图是偏好集 V 的偏好分布模型，用有向图 $PG=\{Page,Edge\}$ 表示，其中，$Page=\{page_1,page_2,\cdots,page_I\}$ 为页面集合，$Edge=\{edge_1,edge_2,\cdots,edge_J\}$ 表示 $page$ 间跳转关系集合，每个偏好对应于若干个不同的页面，每个页面也可以放置不同的偏好。用户在会话期内对页面的请求序列称为页面访问序列，将该序列根据观察概率矩阵进行转换，可得到用户的偏好序列。在访问过程中，若偏好从 $v_i(v_i\in V)$ 变为 $v_j(v_j\in V,v_i\neq v_j)$，则称用户偏好发生了转移。

隐 Markov 链用户偏好分析模型认为用户的页面访问序列和偏好序列形成一个特殊的隐 Markov 链。然而有些用户的浏览行为相似，因此根据用户分类假设理论[7,16,17]，可以分别创建与各用户分类相应的隐 Markov 链浏览子模型，所有子模型的集合组成了类隐 Markov 模型。

定义 3　类隐 Markov 链用户浏览模型用五元组 $cHMM=(X,Y,K,P(C),HMC)$ 表示，其中，X 称为状态序列，取值于用户偏好集 $V=\{v_1,v_2,\cdots,v_N\}$，每个 $v_i(v_i\in V)$ 表示用户的偏好，对应于模型的一个状态，状态序列即用户的访问偏好序列；Y 取值于资源集 $Page=\{page_1,page_2,\cdots,page_J\}$，称为观察序列，即用户的页面访问序列；$K$ 表示用户类别的数目；$C=\{c_1,c_2,\cdots,c_k\}$ 表示用户类别，$P(C)$ 表示 C 的概率分布。$HMC=\{hmc_1,hmc_2,\cdots,hmc_K\}$，其中，$hmc_k$ 为描述类别为 c_k 的用户的浏览特征的隐 Markov 链。hmc_k 可以表示为一个三元组：$hmc_k=(\boldsymbol{\pi}^k,\boldsymbol{A}^k,\boldsymbol{B}^k)$，其中，$\boldsymbol{A}^k$ 为隐 Markov 链 hmc_k 的状态转移概率矩阵，$\boldsymbol{A}^k=(a_{ij}^k)=P[X_{t+1}=v_j\mid X_t=v_i,C=c_k]$；每项 a_{ij}^k 表示类别为 c_k 的用户由状态 v_i 跳转到状态 v_j 的概率；$\boldsymbol{B}^k$ 为观察概率矩阵，$\boldsymbol{B}^k=(b_{jl}^k)=P[Y_t=page_l\mid X_t=v_j,C=c_k]$，每项 b_{jl}^k 表示在一个偏好 v_j 上，类别为 c_k 的用户访问页面 $page_l$ 的概率；$\boldsymbol{\pi}^k$ 为 X 的初始分布，其中，每一项 $\pi_j^k=P(X_1=v_j\mid C=c_k)$。对于以上的 i、j、k、l 均有 $1\leqslant k\leqslant K,1\leqslant i,j\leqslant N,1\leqslant l\leqslant M$。

在类隐 Markov 模型中，设 y_t 为用户在 t 时刻访问的页面资源（即 $page_t$），那么页面访问序列可表示为 $(y_1,y_2,\cdots,y_t)$，根据定义，该序列出现概率为

$$P(y_1,y_2,\cdots,y_t)=\sum_{k=1}^{K}P(y_1,y_2,\cdots,y_t\mid C=c_k)P(C=c_k) \tag{1}$$

其中，$P(y_1,y_2,\cdots,y_t\mid C=c_k)$ 表示在用户类别为 c_k 的隐 Markov 模型 hmc_k 中，页面访问序列 $Y(y_1,y_2,\cdots,y_t)$ 出现的概率，计算为

$$P(y_1,y_2,\cdots,y_t\mid C=c_k)=\sum_{\boldsymbol{X}}P(y_1,y_2,\cdots,y_t,\boldsymbol{X}\mid hmc_k) \tag{2}$$

$$\begin{aligned}&P(y_1,y_2,\cdots,y_t,\boldsymbol{X}\mid hmc_k)\\&=\boldsymbol{\pi}^k(x_1\mid C=c_k)\boldsymbol{B}^k(y_1\mid x_1,C=c_k)\cdot\\&\prod_{i=1}^{t-1}\boldsymbol{A}^k(x_{i+1}\mid x_i,C=c_k)\boldsymbol{B}^k(y_{i+1}\mid x_{i+1},C=c_k)\end{aligned} \tag{3}$$

其中，向量 $\boldsymbol{X}$ 为观察序列 Y 所对应的隐含状态序列，$\boldsymbol{\pi}^k(x_1 \mid C=c_k)$ 、 $\boldsymbol{A}^k(x_{i+1} \mid x_i, C=c_k)$ 、 $\boldsymbol{B}^k(y_i \mid x_i, C=c_k)$ 可以从模型 hmc_k 的初始分布 $\boldsymbol{\pi}_1^k$ 和转移概率矩阵及观察矩阵获取。

3 融合停留时间的类时齐隐 Markov 模型

在传统模型中，用户在页面上的停留时间被看作常量。本文则将用户浏览过程分解为 2 个过程，跳转过程（页面跳转和偏好转移）和用户停留过程(页面停留和偏好暂驻)[9]。跳转过程采用嵌入隐 Markov 过程来描述，而停留过程采用停留时间分布描述。正常情况下，用户对页面的兴趣程度是影响用户在页面上的停留时长的主要因素。所有用户访问某页面的停留时间总和越长，表示该页面的关注度就越大以及用户对此页面所包含的偏好的兴趣程度就越高，因此，这种页面及相应的偏好也就占有越高的推荐度。

引入页面停留时间后，可以将用户的浏览行为过程 $\{Z_t, t \geqslant 0\}$ 表示为 $\{Z=(X_n, Y_n, S_n),\ n \geqslant 0,\ n \in N\}$，$X_n$ 表示偏好转移过程（隐含状态），Y_n 表示页面跳转过程（可观测的输出），S_n 表示用户在页面上的停留时间过程，X_n 和 Y_n 过程可以用隐 Markov 过程来描述，而停留时间过程并不满足 Markov 性，所以称过程 $\{X_n, Y_n\}$ 为过程 $\{Z_t\}$ 的嵌入隐 Markov 过程。

而对于随机变量 S_n，假设 S_n 只依赖于当前所处的页面 Y_n，与其他的页面无关，那么过程 Z_t 则可以视为时间齐次的隐 Markov 过程[9]。此时，停留时间由当前用户所处页面的属性决定，模型考虑影响停留时间的主要因素，忽略了其他影响。根据用户在页面上的停留时间分布 S_n，可以求得对应的隐含状态的停留时间分布 S_n'。

定义 4 类时齐隐 Markov 模型表示为五元组 $ctqHMM=(X, Y, K, P(C), EHMC)$，其中，$X$,$Y$,$K$,$C$, $P(C)$ 同 *cHMM* 模型中的定义一样。而 *EHMC* 为时齐隐 Markov 链的集合，即 $EHMC=\{ehmc_1, ehmc_2, \cdots, ehmc_K\}$，其中，$ehmc_k$ 为描述类别为 c_k 的用户浏览特征的时齐隐 Markov 链。每个时齐隐 Markov 链 $ehmc_k$ 可以表示为一个五元组：$ehmc_k=(\boldsymbol{\pi}^k, \boldsymbol{A}^k, \boldsymbol{B}^k, \boldsymbol{T}_X^k, \boldsymbol{T}_Y^k)$，其中，$\boldsymbol{A}^k$，$\boldsymbol{B}^k$，$\boldsymbol{\pi}^k$ 的定义同 *HMC* 中 hmc_k 定义一致。$\boldsymbol{T}_X^k$ 表示隐含状态上的停留时间分布，对应于 S_n'，其中，每一项 $T_X^k(i)$ 表示类别为 c_k 的用户在状态 v_i 上的停留时间的总和。$\boldsymbol{T}_Y^k$ 表示类别为 c_k 的用户在资源页面上的停留时间分布，对应于 S_n，其中每一项 $T_Y^k(l)$ 表示 c_k 类别用户在资源页面 p_l 上的停留时间的总和。同样有 $1 \leqslant k \leqslant K, 1 \leqslant i, j \leqslant N, 1 \leqslant l \leqslant M$。

类时齐隐 Markov 模型中内部概率的依赖关系可以表示为包含 2 个隐变量的贝叶斯网络，一个隐变量为类别变量 C，一个为隐含状态变量 X，如图 1 所示，其中第 1 层的节点 C 表示类别，第 2 层表示类时齐隐 Markov 链的兴趣转移部分，第 3 层表示页面浏览序列跳转部分，有向边表示条件依赖关系。

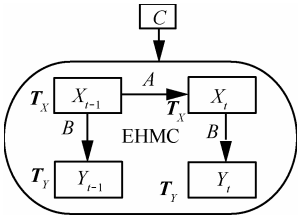


图 1 类时齐隐 Markov 模型的贝叶斯网络结构

式(1)~式(3)在 *ctqHMM* 模型中同样适用。用户在隐含状态上的停留时间通过观察概率表现为用户在对应的资源页面的停留时间。根据图 1 中第 2 层和第 3 层的关系可以得到

$$T_Y^k(l)=\sum_{i=1}^{N} T_X^k(i) B^k(y_l \mid x_i, C=c_k) \tag{4}$$

其中，B^k 表示观察概率，由于 $\boldsymbol{T}_Y^k$，$\boldsymbol{T}_X^k$ 均可看作行向量，它们的每个分量分别表示用户在隐含状态(y_l)和资源页面(x_i)的停留时间，因此式(4)也可表示为

$$\boldsymbol{T}_Y^k=\boldsymbol{T}_X^k \boldsymbol{B}^k \tag{5}$$

式(4)、式(5)给出了用户在隐含状态和资源页面上的停留时间之间的转换关系。

在时齐隐 Markov 链 *ehmc* 中，用户偏好程度用其平稳分布表示，而时齐隐 Markov 链的平稳分布由隐 Markov 过程的极限分布[18]以及停留时间分布共同决定。

停留时间分布是指用户在隐含状态上停留时间的分布情况。根据时间齐次的隐 Markov 过程的特性可知，在状态 i 上的停留时间 $T_X(i)$ 服从参数为 λ_i 的指数分布，其概率分布为

$$P(T_X(i) \leqslant t)=1-\mathrm{e}^{-\lambda_i t} \tag{6}$$

其中，λ_i 是由状态 i 决定的参数，而在状态 i 上停留时间的期望值 $\mu_i=1/\lambda_i$。

定义 5　如果存在概率分布 $\{\theta_i, i \geqslant 0\}$ 满足

$$\theta_i = \lim_{t\to\infty} \frac{TT}{t} \tag{7}$$

则称 $\theta = (\theta_i)_{i=1,2,\cdots,N}$ 为时齐隐 Markov 链的平稳分布，TT 表示[0,t]中状态 i 的累积停留时间。

而对于具体的某个类时齐隐 Markov 链 $ehmc_k$，其平稳分布可以根据对应嵌入隐 Markov 过程的极限分布和停留时间分布的参数求取。

$$\theta_i^k = \frac{\frac{\tilde{\theta}_i^k}{\lambda_i^k}}{\sum_{j=1}^{N_k} \frac{\tilde{\theta}_j^k}{\lambda_j^k}} \tag{8}$$

其中，$\tilde{\theta}^k = (\tilde{\theta}_i^k)_{i=1,2,\cdots,N_k}$ 表示用户类别 c_k 对应的嵌入隐 Markov 链唯一的极限分布，该值体现了隐含状态 i 所代表的兴趣偏好被用户访问的概率，λ_i^k 则表示对应类别下停留时间分布的参数，N_k 表示该类别中对应的隐含状态的数目。

在自身兴趣偏好的驱使下，用户对所有偏好的兴趣度都会收敛到过程 $\{Z_t, t \geqslant 0\}$ 的平稳分布 θ。这一平稳分布反应了所有偏好在该用户群体中的受欢迎程度。

4　类时齐隐 Markov 模型参数学习（基于平稳分布用户聚类）

在类时齐隐 Markov 模型中，需要学习以下几个参数：

1) 用户的类别数 K；

2) 任意一个用户属于类别 c_k 的概率 $P(C = c_k)$；

3) 各类时齐隐 Markov 链的初始分布、转移矩阵、观察矩阵、偏好停留时间分布。

可以看出，类时齐隐 Markov 模型的训练过程就是用户聚类以及各类别用户对应的时齐隐 Markov 链模型的参数学习过程。

设某个聚类结果将训练数据集 $D = \{d1, d2, \cdots, di, \cdots, dm\}$ 分成了 K 类，且第 k 类所包含的用户浏览序列的数目为 m_k，即

$$D = \bigcup_{k=1}^{K} D_k, m = \sum_{k=1}^{K} m_k \tag{9}$$

聚类的初始状态时 $K = m, m_k = 1$，此时一个用户的浏览序列对应一个类别。

设 C 表示用户的类别，则 C 的分布为

$$P(C = c_k) = \frac{m_k}{m} \tag{10}$$

对于每个用户类别 c_k，$ehmc_k$ 就是建立在训练数据集 D_k 上的时齐隐 Markov 链模型，那么建立时齐隐 Markov 模型 $ehmc_k$ 需要确定 5 个参数：$\boldsymbol{A}^k$、$\boldsymbol{B}^k$、$\boldsymbol{\pi}^k$、$\boldsymbol{T}_X^k$、$\boldsymbol{T}_Y^k$。

对于 $\boldsymbol{A}^k$、$\boldsymbol{B}^k$、$\boldsymbol{\pi}^k$，这里采用基于 EM 算法的 Baum-Welch 算法进行训练：首先，根据训练集的特性按照经验给出初始值 $\boldsymbol{\pi}^k$；然后利用训练子集 $D_k = \{Y_1^k, Y_2^k, \cdots, Y_{m_k}^k\}$ 不断迭代，并在迭代过程中更新初始模型，使每次迭代让似然函数 $P(D_k \mid ehmc_k)$ 朝着局部最大方向变化，以保证得到该似然函数最大的模型[18]。

设模型 $ehmc_k$ 对应 $D_k = \{Y_1^k, Y_2^k, \cdots, Y_{m_k}^k\}$ 数据集，其中，$Y_l^k = (y_1^l, y_2^l, \cdots, y_{t_l}^l)$，$1 \leqslant l \leqslant m_k$，那么 t 时刻用户 l 处于偏好 v_i 且 t+1 时刻处于 v_j 的概率为

$$\Lambda_t^l(i,j) = P(x_t = v_i, x_{t+1} = v_j \mid Y_l^k, C = c_k) \tag{11}$$

由式(11)可知，用户在 t 时刻处于偏好 v_i 的概率为

$$\theta_t^l(i) = P(x_t = v_i \mid Y_l^k, C = c_k) = \sum_{j=1}^{N} \Lambda_t^l(i,j) \tag{12}$$

在 Baum-Welch 算法的迭代过程中，参数采用式(13)~式(15)进行更新。

$$\tilde{\pi}_j^k = \frac{\sum_{l=1}^{m_k} \theta_t^l(j)}{\sum_{l=1}^{m_k} \sum_{j=1}^{N} \theta_t^l(j)} \tag{13}$$

$$\tilde{a}_{ij}^k = \frac{\sum_{l=1}^{m_k} \sum_{t=1}^{t_l - 1} \Lambda_t^l(i,j)}{\sum_{l=1}^{m_k} \sum_{t=1}^{t_l - 1} \theta_t^l(j)} \tag{14}$$

$$\tilde{b}_{ij}^k = \frac{\sum_{l=1}^{m_k} \sum_{t=1 \wedge y_{tl}^l = v_l}^{t_l} \theta_t^l(i,j)}{\sum_{l=1}^{m_k} \sum_{t=1}^{t_l} \theta_t^l(j)} \tag{15}$$

经过多次迭代之后，将学习得到使似然函数 $P(D_k \mid ehmc_k)$ 最大的模型。

对于停留时间 $\boldsymbol{T}_X^k$、$\boldsymbol{T}_Y^k$，根据式(4)和式(5)可知，在观察概率矩阵已求解的情况下，这两者是相互可以转换的，那么这里只给出 $\boldsymbol{T}_X^k$ 的估计方法。影响停留时间的因素包括但不限于：页面加载速度、资源大小、用户暂时离开而延长浏览时间等[6]。由于这

些干扰因素的存在，因此在估计停留时间时，采用下面的去噪方法，以获取停留时间 λ^k（表示 c_k 类别下用户偏好的停留时间分布，在不影响理解的情况下，用 λ 表示）的无偏估计。

在用户类别 c_k 下，设 $T_X^k(i)$ 为在偏好 i 上真实停留时间（用 T_i 表示），S_i 为停留时间的观测值，U 为噪声，假设 U 和 T_i 是相互独立的，那么认为 S_i 是 T_i 和 U 的联合

$$S_i = U + T_i \tag{16}$$

其中，T_i 服从指数分布，由于卡方分布经常用来模拟非负值的噪声分布，因此这里设 U 服从 $\chi^2(r)$ 分布，其自由度为 r。

下面根据该联合分布来估计停留时间的参数 λ_i。

设 $\mathrm{E}(S_i)$，$\mathrm{Var}(S_i)$ 分别为 S_i 的期望和方差。而在指数分布 T_i 中，期望和方差分别为 $1/\lambda_i$ 和 $1/\lambda_i^2$，在卡方分布 U 中，期望和方差分别为 r 和 $2r$。又由于 T_i 和 U 相互独立，则有

$$\begin{cases} \mathrm{E}(S_i) = \mathrm{E}(U + T_i) = r + \dfrac{1}{\lambda_i} \\ \mathrm{Var}(S_i) = \mathrm{Var}(U + T_i) = 2r + \dfrac{1}{\lambda_i^2} \end{cases} \tag{17}$$

若给定观测值样本，$\mathrm{E}(S_i)$ 用样本均值 $\overline{S}_i$ 代替，$\mathrm{Var}(S_i)$ 用样本方差 S_i^2 表示，则有

$$\begin{cases} \overline{S}_i = r + \dfrac{1}{\lambda_i} \\ S_i^2 = 2r + \dfrac{1}{\lambda_i^2} \end{cases} \tag{18}$$

对这个方程组进行变形可得

$$\begin{cases} \lambda_i = \dfrac{1}{1 \pm \sqrt{S_i^2 - 2\overline{S}_i + 1}}, \lambda_i > 0 \\ r = (\overline{S}_i - 1) \mp \sqrt{S_i^2 - 2\overline{S}_i + 1}, r > 0 \end{cases} \tag{19}$$

由于实际中样本数据的稀疏性，将该方程组求解问题转化为优化问题，优化的目标函数是方程组(18)中 3 个方程中 r 的差距

$$\min_{\lambda_i} \left(\left(\overline{S}_i - \frac{1}{\lambda_i} \right) - \frac{1}{2} \left(S_i^2 - \frac{1}{\lambda_i^2} \right) \right)^2, \quad \lambda_i > 0 \tag{20}$$

目标函数式(20)表示需要求取使 r 的偏差最小的 λ_i，该优化问题可以通过梯度下降法求解。

根据时齐隐 Markov 模型的平稳分布的特性，将 m 个用户访问序列转化成 m 个时齐隐 Markov 链，然后通过对这 m 个链进行聚类，完成用户聚类。这里，聚类要解决 3 个问题：1) 定义合适的聚类相异度；2) 聚类的合并；3) 聚类结果评价准则函数。

1) 聚类相异度的定义

用户间的相异度是通过用户的浏览行为差异来衡量的，本文采用时齐隐 Markov 链的平稳分布的差异来衡量用户之间的相异度，而平稳分布是离散型的分布，在离散分布领域，一般选取 Kullback-Liebler（KL）距离来描述 2 个离散分布之间的相异度。2 个分布之间的 KL 距离越小，Markov 链的动态特征就越相近。

设有 2 个离散分布 $p = (p_1, \cdots, p_n, \cdots)$，$q = (q_1, \cdots, q_n, \cdots)$，则 p 关于 q 的 KL 距离（$KLDist$）定义如下

$$KLDist(p, q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{21}$$

由于 KL 距离具有方向性，该定义不满足对称性质，因此需要将 2 个时齐隐 Markov 链 $ehmc_k, ehmc_l$ 的距离（相异度）定义为

$$Disim(\lambda^k, \lambda^l) = \frac{1}{2}(KLDist(\lambda^k, \lambda^l) + KLDist(\lambda^l, \lambda^k)) \tag{22}$$

其中，λ^k 为 $ehmc_k$ 的平稳分布，λ^l 为 $ehmc_l$ 的平稳分布。

2) 聚类的合并更新

用户的聚类可以视为对应的 2 个时齐隐 Markov 模型进行聚类，那么设 $ehmc_k$、$ehmc_l$ 模型分别描述了访问序列集合 D_k 和 D_l，那么合并这 2 个模型，就是将 D_l 合并到 D_k，然后利用更新后的训练子集 D_k 对 $ehmc_k$ 模型进行更新学习，最后返回更新后的结果。

3) 聚类结果评价准则函数

在 ctqHMM 中，由于分类事先未知，因此模型可以表示为含有隐变量的贝叶斯网络[8]，如图 2 所示。

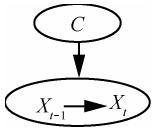


图 2 类时齐隐 Markov 链分类模型的网络结构

节点 C 表示类别，下面表示状态变化，有向边表示条件依赖关系，可以发现，在图 2 中只考虑图 1 中的上半部分，并不需要用到观察值。根据贝叶斯网络

学习理论，一个贝叶斯网络模型 M 的优劣取决于它对于学习数据集 D 的后验概率 $P(M|D)$，$P(M|D)$ 越大，该模型 M 就越优，反之亦然。因此本文将 $P(M|D)$ 作为评价聚类结果的准则函数，记为 F。

根据贝叶斯公式有

$$F = P(M|D) = \frac{P(M)P(D|M)}{P(D)} \tag{23}$$

其中，$P(D)$ 是数据集 D 的边际，不随聚类结果而变化，可看作常数；$P(M)$ 表示模型 M 的先验，可视为均匀分布[8]，这样 $P(M)$ 也可看作常数；从而对后验概率的计算转化成了对 $P(D|M)$ 的计算。在贝叶斯网络理论中，$P(D|M)$ 称为模型 M 的似然函数。

$$P(D|M) = L(D,C)L(D,X_{t-1},X_t) \tag{24}$$

其中，$L(D,C)$ 表示对于节点 C 的似然函数，$L(D,X_{t-1},X_t)$ 表示对于节点 (X_{t-1},X_t) 的似然函数，可分别用式(25)及式(26)计算[8]。

$$L(D,C) = \frac{\Gamma\left(\frac{1}{N}\right)}{\Gamma\left(\frac{1}{N}+m\right)}\prod_{k=1}^{m}\frac{\Gamma\left(\frac{1}{N_k}+m_k\right)}{\Gamma\left(\frac{1}{N_k}\right)} \tag{25}$$

$$L(D,X_{t-1},X_t) = \prod_{k=1}^{K}\prod_{i=1}^{n}\frac{\Gamma\left(\frac{1}{N_{k_i}}\right)}{\Gamma\left(\frac{1}{N_{k_i}}+S_{k_i}\right)}\prod_{j=1}^{n}\frac{\Gamma\left(\frac{1}{N_{k_{ij}}}+S_{k_{ij}}\right)}{\Gamma\left(\frac{1}{N_{k_{ij}}}\right)} \tag{26}$$

其中，N_k 为 D_k 中包含的隐含状态的个数，N_{k_i} 和 S_{k_i} 均表示 D_k 中隐含状态 i 出现的次数，$N_{k_{ij}}$ 和 $S_{k_{ij}}$ 表示 D_k 中所有用户浏览偏好序列中，隐含状态对 (x_i,x_j) 出现的次数。

因此，利用式(23)~式(26)可以计算出任意聚类结果所确定的贝叶斯网络的后验概率，其中后验概率最大的聚类结果为最优。

5　基于 ctqHMM 的个性化推荐

基于 ctqHMM 进行个性化推荐首先需要对用户的类别进行判定。根据贝叶斯公式，访问序列为 $(y_1,y_2,\cdots,y_t)$ 的用户属于类别 c_k 的概率为

$$\begin{aligned} &P(C=c_k \mid y_1,y_2,\cdots,y_t) \\ &= \frac{P(y_1,y_2,\cdots,y_k \mid C=c_k)P(C=c_k)}{P(y_1,y_2,\cdots,y_t)} \end{aligned} \tag{27}$$

其中，分母 $P(y_1,y_2,\cdots,y_t)$ 为边际概率，对于不同的聚类结果，该值保持不变，因此用户属于 c_k 的概率和分子是正相关的，根据贝叶斯判定规则有：

$$\begin{aligned} &P(y_1,y_2,\cdots,y_l \mid C=c_k)P(C=c_k) \\ &= \max_{j=1,2,\cdots,K}(P(y_1,y_2,\cdots,y_l \mid C=c_j)P(C=c_j)) \end{aligned}$$

则用户的类别为 c_k。

确定用户类别为 c_k 后，接下来需要通过其页面访问序列挖掘用户的访问偏好，根据 c_k 类别对应的时齐隐 Markov 模型 $ehmc_k$ 的参数 $\boldsymbol{A}^k$、$\boldsymbol{B}^k$、$\boldsymbol{\pi}^k$、$\boldsymbol{T}_X^k$ 以及观察序列 $(y_1,y_2,\cdots,y_r,\cdots,y_t)$，可以采用维特比（Viterbi）算法估计隐含状态序列（用户偏好序列）的最佳值来挖掘用户页面访问序列中的兴趣偏好。

设隐含状态转移序列 $(\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_t)$ 是根据维特比算法求取的最佳状态转移序列，计算状态转移序列中出现次数最多的状态 ω（若次数最多的状态不止一个，取离时刻 t 最近的状态），用向量 $\boldsymbol{V}$ 表示预测向量，其中每一个分量指用户在第 t+1 时刻访问某页面的概率，即

$$\boldsymbol{V} = [P(y_{t+1}=p_1),P(y_{t+1}=p_2),\cdots,P(y_{t+1}=p_M)] \tag{28}$$

那么可以根据式(29)对用户在 t+1 时刻访问页面进行预测

$$\boldsymbol{V} = [b_{\omega l}^k],\ 1 \leqslant l \leqslant M \tag{29}$$

在向量 $\boldsymbol{V}$ 中，概率值最大的分量对应的页面即为用户在 t+1 时刻最可能访问的页面，根据 top-N 的原则，系统选取概率值排在前 N 大的页面作为推荐页面，推荐给用户。

基于类时齐隐 Markov 模型的个性化推荐一般框架如下。

输入：训练数据集中所有用户的浏览序列集合 D，目标用户 U，候选推荐资源数目 N

输出：推荐给用户 U 的候选资源列表 List

处理过程如下所述。

聚类学习过程

1) 初始化，将 D 中的每个用户看作一个用户类别，并根据 Baum-Welch 算法结合式(10)~式(20)计算出每个用户类别对应的模型参数。

2) 根据式(22)计算出每 2 个用户类别间的相异度，并从小到大排列成队列 *Queue*（初始聚类结果）。

3) 利用式(23)~式(26)对初始的聚类结果计算准则函数值 F_{old}。令 F_{new}=F_{old}。

4) 循环生成聚类结果

while F_{new}>=F_{old}

F_{new}=F_{old}

For i=0 to Length(*Queue*)

将 *Queue*(*i*)对应的这 2 个类别进行临时合并生成新类别，并将新类别替换原来的 2 个用户类别生成待定聚类结果。

对待定聚类结果根据式(23)~式(26)计算准则函数值 P_{new}。

If F_{new}>F_{old} Then

将 *Queue*(*i*)对应的2个类别正式合并，且加入待聚类用户类别集合中生成候选聚类结果；

对候选聚类结果重新排序得到新的相似度队列 *Queue*。

Break；

End If

End For

End While

5) 得到最终候选聚类结果输出，供后续流程使用。

判断用户 *U* 类别过程

6) 根据贝叶斯判定规则判定目标用户 U 的类别。

产生推荐列表过程

7) 根据待测用户观察序列 $(y_1,y_2,\cdots,y_r,\cdots,y_t)$ 及其所属类别对应的时齐隐 Markov 模型，采用维特比算法估计隐含状态序列的最佳值 $(\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_t)$。

8) 计算 $(\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_t)$ 中出现次数最多的状态 ω，利用式(28)、式(29)计算出用户 U 在 t+1 时刻分别访问各页面的概率值，并选取前 N 大值所对应的页面加入推荐页面列表 List 中。

9) 输出列表 List，结束。

6 实验分析

6.1 数据集

本文选取了 2 组实验数据集，第 1 组数据是来自 UCI 通用数据集中匿名的微软网站数据集（数据 1），第 2 组数据采用某电子商务网站的点击数据（数据 2）。

如表 1 所示，数据 1 记录了 Microsoft 站点一周的访问日志数据，该数据里包含了 32 710 个用户对 5 771 个页面发起的 127 536 次 HTTP 请求，每个用户平均访问的页面数是 4（127 536/32 710），根据用户会话所访问的页面信息内容，在该数据集上定义了 13 个偏好：Windows 下载、Office 下载、IE 下载、Windows 升级、Office 升级、IE 升级、Windows 帮助、Office 帮助、IE 帮助、Windows 技术资讯、Office 技术资讯、IE 技术资讯、Microsoft 新闻。

数据 2 来自于国内某统计服务供应商提供的点击数据（click_through_data），该点击数据文件记录了每个顾客的 HTTP 请求信息。数据 2 选取了 2012 年 1 月 7 日 00:00 到 24:00 这个时间段的 1 379 746 次请求，涉及到 246 971 个页面，300 930 个用户的浏览序列，每个用户平均访问的页面数约为 5 (1 379 746/ 300 930)，根据用户会话所访问的商品信息内容，在该数据集上定义了女装、男装、鞋类、箱包、配饰、运动户外、珠宝手表、数码、家电、办公等 100 多个兴趣偏好。

表 1 实验数据集情况

数据	用户数目	页面数目	请求数目
数据 1	32 710	5 771	127 536
数据 2	300 930	246 971	1 379 746

6.2 用户停留时间分布的验证

本实验对样本中的页面停留时间分布进行验证，结果如图 3 和图 4 所示。图中横坐标表示停留时间的长度（单位为 s），本文选取前 30 min 的分布情况；纵坐标表示某个长度的停留时间出现的次数的自然对数值。从前面的分析知道，如果停留时间观察样本精确地服从指数分布，那么图中的自然对数分布应该是一条直线，而实际情况是各数据集的停留时间分布曲线并不是直线，数据 1 是近

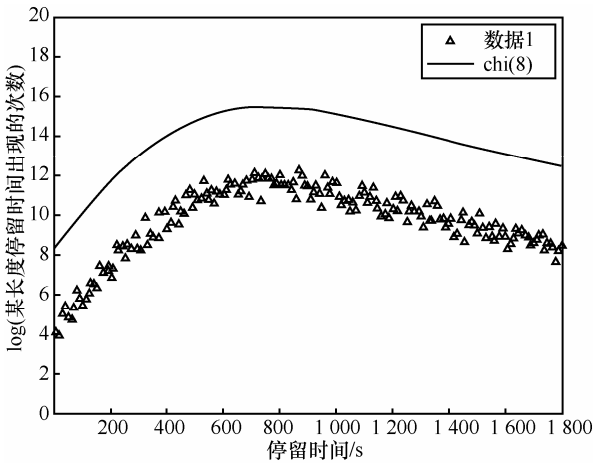


图 3 数据 1 的停留时间观测值的对数分布

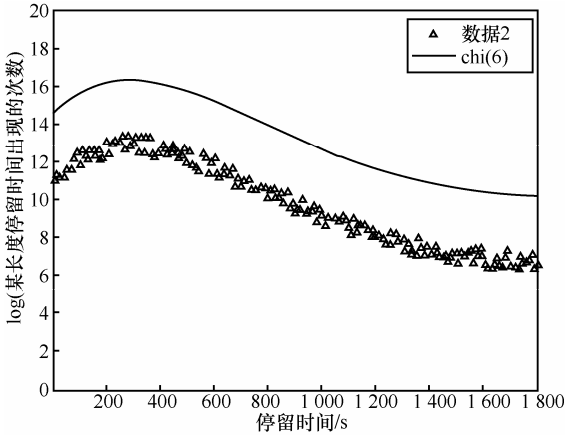


图 4　数据 2 的停留时间观测值的对数分布

似符合自由度为 8 的卡方分布，而数据 2 是近似符合自由度为 6 的卡方分布。这组实验说明，停留时间的观测值并不严格的遵守指数分布，而是掺杂了服从卡方分布的噪声数据。

6.3　浏览序列长度 *L* 对聚类影响实验

本实验比较本文模型的聚类和传统 Markov 链模型的聚类的准确性，同时检验训练集中用户浏览序列长度 *L* 对聚类结果的影响。采取和文献[4]同样的实验方式，首先从 2 个数据集中各抽取 8 个测试用例，每个用例的用户数目在 80~100 之间，页面数在 50~80 之间，并人为地控制用户浏览序列的长度，数据 1 的测试用例中用户浏览序列的平均长度分别为：3.56、6.37、8.41、9.22、11.68、13.78、15.63、16.91，数据 2 的分别为：2.92、5.66、7.48、8.69、10.42、13.67、15.92、17.83。首先，采用本文算法和传统 Markov 聚类算法进行聚类处理，得到聚类结果，然后，手工分析其结果，计算出相应的准确率，实验结果如表 2 所示。可看出，*L* 较小时，两者之间的准确率都较低，且相差不大，当 *L* 逐渐增大时，两者的准确度明显增加，到了一定范围时，递增趋势减缓，这说明用户浏览序列越长，能够反应出用户更多的偏好信息，有利于聚类结果；同时可以看到 *L* 增大时，本文模型的聚类算法优于传统 Markov 链聚类算法，这是因为当用户浏览序列增加时，用户访问兴趣偏好发生转移的概率增加，传统 Markov 链聚类算法是一个静态模型，不能有效跟踪用户偏好转移情况，而本文模型能较好地对用户兴趣迁移动态过程的特征进行建模。可以发现，当 *L* 的平均值大于 9 时，聚类准确率保持在 75%以上，且增长趋势变缓，因此后续实验将 *L* 控制在 9 以上。

6.4　*L* 对推荐结果的影响

本实验检验目标用户的浏览序列长度对推荐准确率的影响，准确率表示用户对系统所推荐的资源感兴趣的概率[19]。先从数据 1 和数据 2 中浏览序列长度大于 9 的数据子集中分别随机抽取 100、300 和 500 个用户，然后将每个实验数据中的浏览序列集进一步划分为训练集和测试集，其中，训练集占 90%。其中，数据 1 中抽取的 3 个实验数据的平均浏览序列长度分别为 11.32、11.94、11.85，数据 2 的则为 13.64、13.63、13.79。实验通过训练集训练模型，然后对测试集进行推荐，根据 top-*N* 的原则取排前 20 的推荐结果并求取准确率，实验结果如表 3 所示。随着目标用户浏览序列长度增加，准确率逐步提高，因为对目标用户的推荐是基于目标用户历史浏览进行的，当序列过短时，获取的信息过少，导致分类错误的可能性大，影响到用户兴趣挖掘从而降低准确率，随着序列长度增加，分类正确的可能性就增加，从而提高了预测精度。同时对比不同用户数目时的实验结果可以发现，L_{test} 固定时，

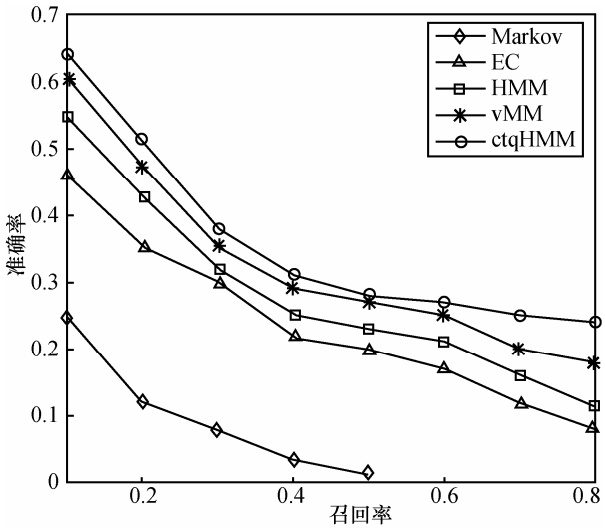
表 2　　**用户浏览序列长度对聚类准确率的影响**

用例编号	*L* 平均值(数据 1)	数据 1 的准确率		*L* 平均值(数据 2)	数据 2 的准确率	
		本文聚类方法	传统 Markov 聚类		本文聚类方法	传统 Markov 聚类
1	3.56	0.226 1	0.224 2	2.92	0.216 2	0.217 1
2	6.37	0.531 2	0.429 3	5.66	0.520 3	0.417 4
3	8.41	0.649 4	0.531 6	7.48	0.632 7	0.525 8
4	9.22	0.768 5	0.654 7	8.69	0.757 6	0.651 9
5	11.68	0.824 3	0.696 5	10.42	0.811 9	0.702 2
6	13.78	0.838 7	0.756 4	13.67	0.833 4	0.739 2
7	15.63	0.847 1	0.808 2	15.92	0.849 5	0.811 2
8	16.91	0.869 1	0.816 1	17.83	0.871 2	0.824 1

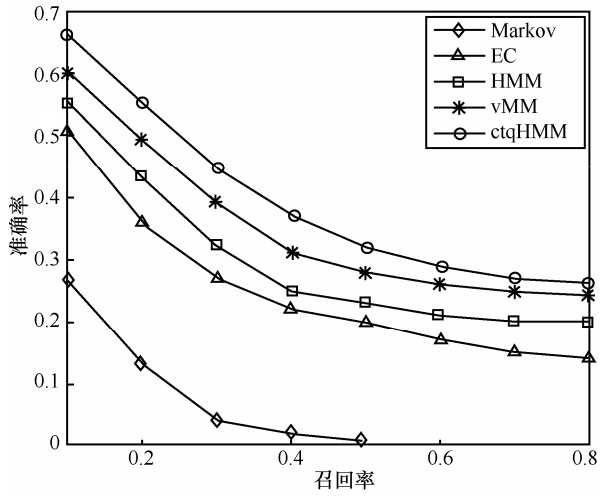
表 3　　**目标用户浏览序列长度对推荐准确率的影响**

目标用户浏览序列长度 L_{test}	数据 1 的准确率			数据 2 的准确率		
	100 用户 (L_{train}=11.32)	300 用户 (L_{train} =11.94)	500 用户 (L_{train} =11.85)	100 用户 (L_{train} =13.64)	300 用户 (L_{train} =13.63)	500 用户 (L_{train} =13.79)
1	0.225 1	0.212 5	0.234 1	0.209 6	0.224 1	0.231 4
3	0.325 6	0.332 1	0.339 8	0.335 5	0.374 5	0.342 9
5	0.402 1	0.416 8	0.417 5	0.411 3	0.420 3	0.429 8
7	0.532 6	0.551 2	0.556 8	0.565 9	0.570 1	0.586 3
9	0.607 7	0.623 5	0.634 7	0.627 5	0.632 1	0.639 8
11	0.610 8	0.632 1	0.645 5	0.638 7	0.643 3	0.653 9
13	0.620 3	0.642 2	0.658 4	0.647 2	0.657 4	0.662 5

当用户数目增加，训练集的 L_{train} 保持基本一致时，同样可以提高准确率，因为训练的用户数目越多，越容易找到适合目标用户的推荐对象。



(a) 数据 1(500 用户, L_{train} = 11.85, L_{test} = 13)



(b) 数据 2(500 用户, L_{train} = 13.79, L_{test} = 13)

图 5　ctqHMM 算法在不同数据集上的效果对比

6.5　ctqHMM 算法推荐效果比较

本实验中，比较了 ctqHMM 算法和一些常见的基于随机过程推荐算法的有效性，包括传统类 Markov 链法[8]、EC 法[1]、隐 Markov 链 HMM[2]以及变阶 Markov 链 vMM[7]，并用召回率—准确率评价准则[19]进行评价，实验设置：从数据 1 中抽取 500 个用户，训练集平均长度为 11.85，目标用户浏览序列长度为 13；从数据 2 中抽取 500 个用户，训练集平均长度为 13.79，目标用户浏览序列长度为 13。图 5 显示了在数据 1 和数据 2 上所有方法的召回率—准确率曲线。图 5 中，ctqHMM 算法比其他方法有更好的 precision-recall 曲线，因此推荐的效果更好，比起参照方法中效果最好的 vMM 算法，ctqHMM 算法提升了将近 2%~6%。

7　结束语

本文通过对用户浏览网页所产生的日志进行分析，对用户的浏览兴趣偏好进行了动态建模，并对跳转关系、停留时间等因素进行综合分析，结合传统的 Markov 预测模型，提出了新的类时齐隐 Markov 模型，并将该模型用于 Web 系统中资源推荐，并在实际的数据集上和相关参照方法进行了比较实验，实验结果表明本文提出的模型在推荐准确率以及召回率指标上比其他算法的预测性能更好。

本文为动态跟踪用户的偏好转移提供了一种思路，研究中也存在值得改进的地方，模型目前只适用于资源类型单一的网页级别分析，而难以适用于整个互联网范畴的站点级别分析，而且并未考虑影响停留时间的其他因素等，这些问题都将成为下一步研究的重点。

参考文献：

[1] AWAD M, ISSA K. Prediction of user's web-browsing behavior:

application of markov model[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, 42(4):1131-1142.

[2] SAHOO N, VIR S, MUKHOPADHYAY T. A hidden Markov model for collaborative filtering[J]. MIS Quarterly-Management Information Systems, 2012, 36(4):13-29.

[3] 许海玲,吴潇等.互联网推荐系统比较研究[J]. 软件学报, 2009, 20(2): 350-362.
XU H L, WU X, *et al*. Comparison study of Internet recommendation system[J]. Journal of Software, 2009,20(2):350-362..

[4] BHAWNA N, SURESH J. Generating a new model for predicting the next accessed web page in web usage mining[A]. Proceeding in 3rd International Conference on Emerging Trends in Engineering and Technology[C]. India, 2010. 485-490.

[5] MIKA P. Ontologies are us: a unified model of social networks and semantics[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2007, 5(1): 5-15.

[6] OARD DW, KIM J. Implicit feedback for recommender systems[A]. Proceeding of the AAAI Workshop on Recommender Systems[C]. USA, 1998.81-83.

[7] BORGES J, LEVENE M. Data mining of user navigation patterns[A]. Proceedings of the 1999 KDD Workshop on Web Mining[C]. San Diego California, 1999. 92-111.

[8] 邢永康, 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 Y K, MA S P. Modeling user navigation sequences based on multi-markov chains[J]. Chinese Journal of Computers,2003, 26(11): 1510-1517.

[9] LIU Y T, LIU T Y, GAO B, *et al*. Browserank: letting web users vote for page importance[A]. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. Singapore, 2008. 451-458.

[10] JOSEPH A, JOHN R. Recommender systems: from algorithms to user experience[J]. User Model User-Adap Inter, 2012, 22:101-123.

[11] LICHTENWALTER R N, LUSSIER J T, CHAWLA N V. New perspectives and methods in link prediction[A]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining[C]. USA, 2010.243-252.

[12] KNIJNENBURG B P, WILLEMSEN M C, GANTNER Z, *et al*. Explaining the user experience of recommender systems[J]. User Model User-Adap Inter, 2012, 22: 441-504.

[13] 刘凯鹏, 方滨兴. 一种基于社会性标注的网页排序算法[J]. 计算机学报, 2010, 33(6):1014-1023.
LIU K P, FANG B X. A novel page ranking algorethm based on social annotations[J]. Chinese Journal of Computers,2010, 33(6):1014-1023.

[14] HOTHO A, JASCHKE R, SCHMITZ C, *et al*. Information Retrieval in Folksonomies: Search and Ranking[M]. Berlin Heidelberg, Springer, 2006.411-426.

[15] SONG G H, SUN S T, FAN W. Applying user interest on item-based recommender system[A]. Proceedings of the 5th International Joint Conference on Computational Sciences and Optimization (CSO)[C]. Harbin, China, 2012.635-638.

[16] 边肇祺, 张学工等. 模式识别(第二版)[M]. 北京: 清华大学出版社, 2000.
BIAN Z Q, ZHANG X G, *et al*. Pattern Recognition(2)[M]. Beijing: Tsinghua University Press, 2000.

[17] WAN M, JONSSON A, WANG C, *et al*. A random indexing approach for web user clustering and web prefetching[J]. New Frontiers in Applied Data Mining Lecture Notes in Computer Science, 2012, 7104: 40-52.

[18] 孙荣恒. 随机过程及其应用[M]. 北京: 清华大学出版社, 2004.
SUN R H. Stochastic Process and Application[M]. Beijing: Tsinghua university press,2004.

[19] 朱郁筱,吕琳媛. 推荐系统评价指标综述[J].电子科技大学学报, 2012, 2(41): 163-175.
ZHU Y X, LV L Y. Evaluation metrics for recommender systems[J]. Journal of University of Electronic Science and Technology of China,2012, 2(41): 163-175.

作者简介：

刘胜宗（1986-），男，湖南邵阳人，中南大学博士生，主要研究方向为数据挖掘、智能信息处理、推荐系统。

樊晓平[通信作者]（1961-），男，浙江绍兴人，博士，中南大学教授、博士生导师，湖南财政经济学院副院长，主要研究方向为机器人控制、智能控制、无线传感器网络、智能交通系统。E-mail:xpfan@mail.csu.edu.cn。

廖志芳（1968-），女，湖南长沙人，博士，中南大学副教授，主要研究方向为数据挖掘、推荐系统等。

胡佳（1985-），男，湖南岳阳人，中南大学博士生，主要研究方向为无线传感器网络、压缩感知。