

基于隐反馈的类时齐 Markov 推荐模型

刘胜宗¹, 廖志芳¹, 胡 佳¹, 樊晓平^{1,2}

(1. 中南大学信息科学与工程学院/中南大学软件学院, 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)

摘 要: 传统 Markov 链模型在用户浏览行为预测方面体现出较好的性能,但不能很好的体现出用户的兴趣度和所推荐的页面的重要性,因此本文提出类时齐 Markov 模型.该模型给不同的类别用户单独创建时齐 Markov 模型,并用时齐 Markov 模型的平稳分布表征用户的访问兴趣和页面的重要程度.本文进而提出了基于隐反馈的类时齐 Markov 推荐模型,在真实的 WEB 服务器日志数据上的实验证明,类时齐 Markov 模型具有更好的推荐性能.

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

中图分类号: TN911.23 **文献标识码:** A **文章编号:** 0372-2112 (2014)04-0703-08

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.3969/j.issn.0372-2112.2014.04.013

Classified Time Homogeneous Markov Model for Recommendation Based on Implicit Feedback

LIU Sheng-zong¹, LIAO Zhi-fang¹, HU Jia¹, FAN Xiao-ping^{1,2}

(1. School of Information Science and Engineering / School of Software, Central South University, Changsha, Hunan 410075, China;

2. Laboratory of Networked Systems, Hunan University of Finance and Economics, Changsha, Hunan 410205, China)

Abstract: Markov chain model shows good performance in the user browsing behavior predictions. But it does not work well in reflecting user's interestingness and the importance of the recommended pages. Therefore, this paper proposes classified time homogeneous Markov model. The proposed model create a time homogeneous Markov model separately for every different category of users and use the stationary distribution of the time homogeneous Markov model to characterize users' access interest and pages' importance. Then this paper puts forward a classified time homogeneous Markov model for recommendation based on implicit feedback. The results of experiment with some real WEB server log data show that the proposed model and algorithm have more perfect performance.

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

1 引言

用户浏览网页时的相关反馈包括显性和隐性反馈^[1].显性反馈指系统通过问卷或评分等方式让用户对目标资源给出倾向性的反馈信息^[2];而隐性反馈则是系统通过其他后台机制(如服务器日志)获取到的反馈信息^[3,4].目前大部分推荐系统主要是使用显性反馈的方式,即需要用户对每次查看的对象都进行评分或相关操作^[5],通过这种方式来获取评价方法存在缺点:对用户不友好,大多用户不愿意评分,导致评分数据更加稀疏,使得基于评分方式推荐的算法适应性降低.而用户在浏

览时,会产生大量的隐性反馈信息,通过从网络日志中获取并分析这些信息可知,用户在网上浏览的行为基本可分为:阅读网页和网页跳转^[6-9].阅读网页是用户获取信息的直接方式,而体现某一网页对用户的兴趣度则可以通过“使用者的停留时间”体现出来^[8],对隐性反馈的可靠性研究表名,用户在某个页面上的停留时间和用户对该页面的评价近似是正相关的^[4,8].

传统的 Markov 链用户浏览预测模型^[10]将用户的浏览过程抽象为特殊的随机过程,并用转移概率矩阵来描述用户的浏览特征,并在此基础上对用户的后续浏览路径进行预测,这种方法是可行的,后来的研究者针对

该方法进行了改进,文献[11]提出了多 Markov 链的方法对同类别的用户建立一条 Markov 链,预测时,先判断当前用户的类别,再选用对应的 Markov 链进行预测;文献[12]利用多阶 Markov 链改进了用户分类方法,提出了 web 用户聚类方法;文献[1]结合关联规则挖掘的方式改进了传统 Markov 链预测方法,提出了一个新的两层预测框架下的样本分类算法 EC,该算法在不影响准确度的前提下,降低了全阶 Markov 模型的时间复杂度。然而这些方法仅仅是根据用户在网页间的转移概率来进行预测,对隐反馈信息的利用不够充分,而用户的兴趣度不仅仅体现在页面浏览序列中,也跟用户在某页面的停留时间、页面的重要性以及用户当前的兴趣与页面的主题等信息有很大的关联性^[13]。

据此,本文提出了基于隐反馈的类时齐 Markov 推荐模型,先将用户按浏览序列进行聚类,然后给每个类别的用户建立起对应的时齐 Markov 模型,推荐时,先根据目标用户的已有浏览序列判断出该用户的类别,再根据相应类别的时齐 Markov 模型的平稳分布,选取概率最大的前 N 个资源推荐给该用户,通过实验验证,该方法具有更好的推荐效果。

2 传统 Markov 链用户浏览预测模型

用户浏览过程:设 Z_t 表示用户在 t 时刻访问的页面资源,状态空间 V 是包含所有页面资源的有限离散空间,设 $|V| = N$,那么过程 $\{Z_t, t \geq 0\}$ 包含了用户浏览的所有行为特征信息,该时间参数连续而状态参数离散的随机过程就是用户浏览过程^[13,14]。

基于 Markov 链预测模型将所有用户在 WEB 页面上的浏览过程构建成 Markov 单链模型。

定义 1 单 Markov 链模型^[11]

单 Markov 链模型由三元组 $SMC(X, TR, \lambda)$ 表示,其中: X 表示离散的随机变量,取值范围为 $\{x_1, x_2, x_3, \dots, x_n\}$,其中每个 x_i 称为模型的状态,它对应于 Web 中的一个页面; TR 是转移概率矩阵,其每一项 $TR_{ij} = P(X_t = x_j | X_{t-1} = x_i)$ 表示由状态 x_i 转移到状态 x_j 的概率; λ 为初始状态分布概率向量,其每一项表示为 $\lambda_i = P(x_{t_0} = x_i)$ 。

转移矩阵和初始分布等参数可由相应算法学习得到,然后可以根据初始状态分布和转移概率矩阵预测任意时间段的状态分布^[11]。

该模型将停留时间看作常量,优点是简单,易于计算,但也有缺点,如:用户因误点击而快速离开页面,单链模型将这种情况视为正常的,这会影响模型的推荐精度。为此,本文结合 Browserank^[13]中的时齐 Markov 模型提出了类时齐 Markov 模型。

3 类时齐 Markov 模型

引入页面停留时间因素后,用户浏览过程 $\{Z_t, t \geq 0\}$ 可表示为 $\{Z = (X_n, Y_n), n \geq 0, n \in N\}$, X_n 表示纯跳过程, Y_n 表示用户在页面上的停留时间过程,纯跳过程满足 Markov 性,而停留时间过程并不满足 Markov 性,在随机过程领域中,称过程 $\{X_n\}$ 为过程 $\{Z_t\}$ 的嵌入 Markov 链,记为 EMC。

对于 Y_n ,假设 Y_n 只依赖于当前所处的状态 X_n ,与其他的状态无关,那么过程 Z_t 则可以视为时间齐次的 Markov 过程。此时,用户在当前页面上的停留时间由当前页面本身决定,与其他的因素无关,也就是该模型只考虑影响停留时间的主要因素,其他因素视为噪声。

定义 2 时齐 Markov 模型

时齐 Markov 模型表示为一个三元组 $qm(X, T, EMC)$ 。其中 X 是一个离散随机变量,其值域是问题域中所有状态的集合, T 表示状态停留时间分布, EMC 表示嵌入链。

在上面的定义中, T 对应于 Z_t 过程中的子过程 Y_n ,而 EMC 则描述的是子过程 X_n 。在 Z 中, X_n 过程是纯跳过程,因此可以将 $\{X_n\}$ 看作离散参数齐次 Markov 链,它的一步转移概率矩阵 $TR(m, 1)$ 与时间 m 无关,即:

$$\begin{aligned} TR_{ij}(m, 1) &= P\{X_m = j | X_{m-1} = i\} \\ &= P\{X_1 = j | X_0 = i\} \quad i, j \in N \end{aligned} \quad (1)$$

通过观察和分析大量用户的浏览行为可发现,有些用户的浏览行为呈现出相似的特点。据此本文将用户先聚类,然后给每类用户分别建立起时齐 Markov 模型,最后再进行推荐。基于这种思想建立起来含有多个时齐 Markov 过程的模型叫做类时齐 Markov 模型。可知时齐 Markov 模型是将所有用户看作一个类别的类时齐 Markov 模型。

定义 3 类时齐 Markov 模型

类时齐 Markov 模型可以表示为一个五元组 $CQM(X, K, C, P(C), QM)$,其中 X 是一个离散随机变量,值域为 $\{x_1, x_2, x_3, \dots, x_n\}$,每个 x_i 对应于模型的一个状态即网页资源; K 表示模型中所包含的用户类别的数目; $C = \{c_1, c_2, \dots, c_k\}$ 表示用户类别, $P(C)$ 表示 C 的概率分布; $QM = \{qm_1, qm_2, \dots, qm_k\}$ 为所有用户类别对应的 qm 的集合,其中 qm_i 表示类别为 c_i 的用户群的时齐 Markov 模型。

在类时齐 Markov 模型(CQM)中,每个类别对应的时齐 Markov 模型(qm)都是相互独立的,因此下面对 CQM 的讨论是针对某一个用户类别对应的 qm 而展开。

在时齐 Markov 模型(qm)中,网页资源的重要性由该模型的平稳分布描述。而 qm 的平稳分布由嵌入链的极限分布和停留时间分布共同决定。

在时齐 Markov 模型中,嵌入链 $\{X_n, n \geq 0\}$ 是不可约的,因此该嵌入链存在而且仅存在唯一的极限分布^[13],表示各状态被访问的可能性。

而停留时间分布是指用户在页面上停留时间的分布情况。根据时齐 Markov 过程的特性可知,在页面 i 上的停留时间 T_i 服从参数为 λ_i 的指数分布,其概率分布为:

$$P(T_i \leq t) = 1 - e^{-\lambda_i t} \quad (2)$$

其中 λ_i 是由网页 i 决定的参数。

定义 4 时齐 Markov 模型平稳分布

如果存在概率分布 $\{\pi_i, i \geq 0\}$ 满足:

$$\pi_i = \lim_{t \rightarrow \infty} \frac{TT_i}{t} \quad (3)$$

则称 $\pi = (\pi_i)_{i=1,2,\dots,N}$ 为时齐 Markov 模型的平稳分布,其中 TT_i 表示 $[0, t]$ 中过程停留在资源 i 的累积时间。

平稳分布 π_i 越大,说明网页 i 被访问的频率大、停留时间长、受关注度高,越应该推荐给用户。

而时齐 Markov 模型的平稳分布可以根据式(4)求得:

$$\pi_i = \frac{\frac{\bar{\pi}_i}{\lambda_i}}{\sum_{j=1}^N \frac{\bar{\pi}_j}{\lambda_j}} \quad (4)$$

$\bar{\pi} = (\bar{\pi}_i)_{i=1,2,\dots,N}$ 表示该模型中嵌入链唯一的极限分布, λ_i 表示停留时间分布的参数。时齐 Markov 模型的平稳分布反应了所有网页的平均停留时间分布。

嵌入链的极限分布 $\bar{\pi}$, 以及停留时间分布的参数 λ 需要通过对训练集的监督学习进行参数估计。

首先讨论嵌入链平稳分布 $\bar{\pi}$ 的估计方法。嵌入链 EMC 就是建立在训练数据集 D 上的单 Markov 链模型,同时设 N 表示 D 中状态的个数,嵌入链的转移矩阵 TR 的每一项由式(5)求得:

$$TR_{ij} = \begin{cases} \alpha \frac{S_{ij}}{\sum_{j=1}^N S_{ij}} + (1 - \alpha) \frac{1}{N}, & \text{if } \sum_{j=1}^N S_{ij} \neq 0 \\ \frac{1}{N}, & \text{else} \end{cases} \quad (5)$$

S_{ij} 表示 D 中所有用户浏览序列中,状态对 (x_i, x_j) 出现的次数。

嵌入链的平稳分布 $\bar{\pi}$ 用幂迭代法进行求解,通过对嵌入链的转移矩阵 TR 进行多次幂迭代运算,直到目标函数的值小于事先设定好的阈值。

然后讨论停留时间 λ 的估计方法。由于停留时间的观测值和真实值存在误差(噪声),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 的无偏估计。

设 T_i 为在页面 i 上真实停留时间, 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,那么可以认为 S_i 是 T_i 和 U 的联合分布:

$$S_i = U + T_i \quad (6)$$

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,因此这里假设 U 服从 $\chi^2(r)$ 分布,其自由度为 r 。

设 $E(S_i)$, $\text{Var}(S_i)$ 分别为 S_i 的期望和方差。由于 T_i 和 U 相互独立,则有:

$$\begin{cases} E(S_i) = E(U + T_i) = r + \frac{1}{\lambda_i} \\ \text{Var}(S_i) = \text{Var}(U + T_i) = 2r + \frac{1}{\lambda_i^2} \end{cases} \quad (7)$$

在给定观测值样本之后, $E(S_i)$ 将用样本均值 \bar{S}_i 代替,而 $\text{Var}(S_i)$ 用样本方差 S_i^2 表示,则有:

$$\begin{cases} \bar{S}_i = r + \frac{1}{\lambda_i} \\ S_i^2 = 2r + \frac{1}{\lambda_i^2} \end{cases} \quad (8)$$

在实际中,由于样本数据稀疏,该方程组无唯一解,因此,将该方程组求解问题转化为求优解问题,目标函数是方程组(8)中两个方程中 r 的差

$$\min_{\lambda_i} \left(\left(\bar{S}_i - \frac{1}{\lambda_i} \right) - \frac{1}{2} \left(S_i^2 - \frac{1}{\lambda_i^2} \right) \right)^2 \quad \lambda_i > 0 \quad (9)$$

目标函数(9)表示需要求得 r 的偏差最小的 λ_i ,这可以通过梯度下降法求解。

而类时齐 Markov 模型(CQM),则是基于用户分类的思想,将具有相似浏览行为的用户组成相同的用户类别,分别给每个用户类别建立起对应的时齐 Markov 模型(qm),并对这些 qm 进行训练学习,最后求出各个 qm 的平稳分布,从而获取各类用户中最受欢迎的页面资源,作为推荐对象。

4 嵌入链的聚类

本文将 m 个用户访问序列转化成 m 个时齐 Markov 模型的嵌入链,然后通过对这 m 个嵌入链进行聚类,完成用户聚类。

基于用户浏览行为的聚类要解决:(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,而用户的浏览行为在前面已经模型化为 Markov 链,因此这里只要定义 Markov 链之间的相异度。表征

Markov 链动态特性的转移矩阵是离散型分布,在离散分布领域,常选取 Kullback-Liebler(KL)距离来描述转移矩阵之间的相异度.两个转移矩阵之间的 KL 距离越小,说明这两个分布差别就越小,Markov 链的动态特征就越相近.

对于任意的两个转移矩阵 $\mathbf{TR}_k, \mathbf{TR}_l$, 它们的第 i 行分别为: $p_{k_{ij}}, p_{l_{ij}} | j = 1, 2, \dots, n$. 这两行之间的 KL 距离则表示为:

$$\text{KLDist}(p_{k_i}, p_{l_i}) = \sum_{j=1}^n p_{k_{ij}} \lg \frac{p_{k_{ij}}}{p_{l_{ij}}} \quad (10)$$

因此,这两个转移概率矩阵 $\mathbf{TR}_k, \mathbf{TR}_l$ 之间的相异度用所有行分布的 KL 距离的均值来表示:

$$\text{Disim}(\mathbf{TR}_k, \mathbf{TR}_l) = \frac{1}{n} \sum_{i=1}^n \text{KLDist}(p_{k_i}, p_{l_i}) \quad (11)$$

KL 距离具有方向性,为了满足距离对称,将两个嵌入的 Markov 链 $\text{EMC}^k, \text{EMC}^l$ 的距离定义为:

$$\begin{aligned} \text{Disim}(\text{EMC}^k, \text{EMC}^l) \\ = \frac{1}{2} (\text{Disim}(\mathbf{TR}_k, \mathbf{TR}_l) + \text{Disim}(\mathbf{TR}_l, \mathbf{TR}_k)) \end{aligned} \quad (12)$$

(2) 聚类的合并

聚类的合并是基于各类别对应的嵌入链之间的相异度.开始时,将每个用户的访问序列当作一条 Markov 链,然后计算每条链之间的距离,选取距离最小的两条链进行合并成一个新的聚类,并将该新的聚类和其他待聚类的链重新进行距离计算.合并两个代表 Markov 链动态特征的转移概率矩阵时,首先将转移矩阵的状态合并,生成新的转移概率矩阵,然后计算转移矩阵中每个概率值.

假设训练数据集 $D = \{d_1, d_2, \dots, d_i, \dots, d_m\}$ 被分成了 K 类,且第 k 类所包含的用户浏览序列的数目为 m_k , 即:

$$D = \bigcup_{k=1}^K D_k, \quad m = \sum_{k=1}^K m_k \quad (13)$$

聚类的初始状态可以看成特殊的聚类结果,即:一个用户的浏览序列对应一个类别,此时 $K = m, m_k = 1$.

设 C 表示用户的类别,则 C 的分布为:

$$P(C = c_k) = \frac{m_k}{m} \quad (14)$$

对于每个用户类别 c_k ,其嵌入链模型 EMC^k 就是建立在训练数据集 D_k 上的单 Markov 链模型,同时设 D_k 包含的状态的个数为 N_k ,可以利用下面的方法来计算对应的转移矩阵 \mathbf{TR}^k 的每一项:

$$\mathbf{TR}_{ij}^k = \begin{cases} \alpha \frac{S_{k_{ij}}}{\sum_{j=1}^{N_k} S_{k_{ij}}} + (1 - \alpha) \frac{1}{N_k}, & \text{if } \sum_{j=1}^{N_k} S_{k_{ij}} \neq 0 \\ \frac{1}{N_k}, & \text{else} \end{cases} \quad (15)$$

式(15)中, $S_{k_{ij}}$ 表示 D_k 中所有用户浏览序列中,状态对 (x_i, x_j) 出现的次数. α 为归一化因子,而对于两个嵌入链 EMC^k 和 EMC^l ,合并后的嵌入链 EMC^{k+l} 的转移矩阵的元素计算方法为:

$$\mathbf{TR}_{ij}^{k+l} = \begin{cases} \alpha \frac{S_{(k+l)_{ij}}}{\sum_{j=1}^{N_{k+l}} S_{(k+l)_{ij}}} + (1 - \alpha) \frac{1}{N_{k+l}}, & \text{if } \sum_{j=1}^{N_{k+l}} S_{(k+l)_{ij}} \neq 0 \\ \frac{1}{N_{k+l}}, & \text{else} \end{cases} \quad (16)$$

其中 $D_{(k+l)} = D_k \cup D_l$,其余参数参照式(15).

(3) 评价聚类结果的准则函数

在 Markov 链分类模型中,由于分类事先未知,因此该分类可以看成隐变量,因此模型也可以表示为一个含有隐变量的贝叶斯网络^[11],如图 1 所示.

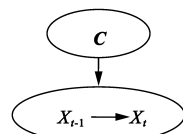


图1 Markov链分类模型的网络结构

上面的节点 C 表示类别,下面表示 Markov 链,有向边表示条件依赖关系.因此不同的聚类结果对应不同的贝叶斯网络模型.那么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.因此本文将概率 $P(M|D)$ 作为评价聚类结果的准则函数,记为 F .

根据贝叶斯公式有:

$$F = P(M|D) = \frac{P(M)P(D|M)}{P(D)} \quad (17)$$

其中 $P(D)$ 是学习数据的边际概率,不随聚类结果而变化;而 $P(M)$ 表示模型 M 的先验概率,由于无法确定模型分布,因此按照文献[11]的处理方式,将其看作均匀分布,这样 $P(M)$ 可以看作常量;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. Cooper 等推导出了具体的计算公式.

如图 1 所示,该贝叶斯网络有两个节点,所以

$$P(D|M) = L(D, C) L(D, X_{t-1}, X_t) \quad (18)$$

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,可分别用下面的公式计算^[11]:

$$L(D, C) = \frac{\Gamma(\frac{1}{N})}{\Gamma(\frac{1}{N} + m)} \prod_{k=1}^m \frac{\Gamma(\frac{1}{N_k} + m_k)}{\Gamma(\frac{1}{N_k})} \quad (19)$$

$$L(D, X_{t-1}, X_t) = \prod_{k=1}^K \prod_{i=1}^n \frac{\Gamma(\frac{1}{N_{k_i}})}{\Gamma(\frac{1}{N_{k_i}} + S_{k_i})} \cdot \prod_{j=1}^n \frac{\Gamma(\frac{1}{N_{k_{ij}}} + S_{k_{ij}})}{\Gamma(\frac{1}{N_{k_{ij}}})} \tag{20}$$

式中 N_{k_i} 和 S_{k_i} 均表示 D_k 中状态 i 出现的次数,其他参数含义和式(15)(16)中的相同。

利用式(17)~(20)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 F 。聚类过程中,反复进行迭代,直到后验概率不再增大,将其对应的聚类结果输出。

5 基于 CQM 的个性化推荐

类时齐 Markov 模型建立后,通过模型学习,估计出所有的参数后,便可以采用该模型来描述各类别的用户的浏览特征,继而对用户进行相关资源页面的推荐。推荐分为两个步骤:(1)判定用户类别;(2)TOP-N 推荐。

(1)判定用户类别

根据贝叶斯公式,访问序列为 (x_1, x_2, \cdots, x_t) 的用户属于类别 c_k 的概率为

$$P(C = c_k | x_1, x_2, \cdots, x_t) = \frac{P(x_1, x_2, \cdots, x_t | C = c_k) P(C = c_k)}{P(x_1, x_2, \cdots, x_t)} \tag{21}$$

式(21)中,分母 $P(x_1, x_2, \cdots, x_t)$ 表示序列的边际概率,对于不同的用户聚类结果,该值保持不变,因此用户属于 c_k 的概率和分子部分是正相关的,因此有如下的判定规则:

if $P(x_1, x_2, \cdots, x_t | C = c_k) P(C = c_k)$
= $\max_{j=1,2,\cdots,K} (P(x_1, x_2, \cdots, x_t | C = c_j) P(C = c_j))$
then 用户的类别为 c_k

这是基于最小错误率的贝叶斯判定规则,按这种规则得到的分类结果的错误率最小^[11]。

(2)TOP-N 推荐

确定用户的类别 c_k 后,该用户的推荐结果集就根据该用户类别对应的时齐 Markov 模型的平稳分布 π_k 进行获取,前面的讨论指出时齐 Markov 模型的平稳分

布 π_k 描述的是平均停留时间的稳定分布。因此,首先将页面集合按平均停留时间由大到小排序,将排在前 N 位的页面资源推荐给当前用户。

6 实验结果及分析

为了检验本文提出的类时齐 Markov 模型 CQM 的有效性,本文选取了三组实验数据,如表 1 所示,数据一是 EPA WEB 服务器的日志信息,记录了一天时间里对服务器的 47748 次 HTTP 请求。在研究每个用户浏览过程时,根据本文提出的模型的要求,将会话期内来自相同 IP 地址的请求看作同一用户的请求,并按请求的时间先后顺序排序,进而将这些数据转化为 2266 个用户的浏览序列,涉及到 3086 个页面,每个用户平均访问的页面数是 21(47748/2266)。

数据二记录了 microsoft.com 站点一周内 32710 个用户对 5771 个页面发起的 127536 次 HTTP 请求,每个用户平均访问的页面数是 4(127536/32710)。

数据三来自于国内某统计服务供应商提供的点击数据(click_through_data),该数据记录了每个顾客的 HTTP 请求信息。数据三选取了一天内的 1379746 次请求信息,涉及到 246971 个页面,300930 个用户的浏览序列,每个用户平均访问的页面数约为 5(1379746/300930)。

表 1 实验数据集情况表

	用户数目	页面数目	请求数目
数据一	2266	3086	47748
数据二	32710	5771	127536
数据三	300930	246971	1379746

本文在这三个数据集上分别做了四组实验,第一组是验证停留时间分布的实验,第二组是推荐准确度对比实验,第三组是召回率指标对比实验,第四组则是验证聚类结果中各类别下元素个数对准确率的影响。

实验 1 验证了样本中的页面停留时间分布,结果如图 2、3、4 所示。图中横坐标表示停留时间的长度(单位为 s),本文选取长度在 0 到 1800 范围内;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,如果停留时间样本精确的服从指数分布,

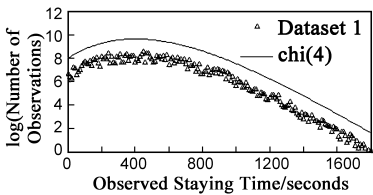


图2 数据一的停留时间观测样本的对数分布图

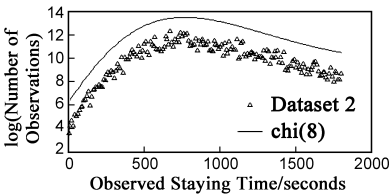


图3 数据二的停留时间观测样本的对数分布图

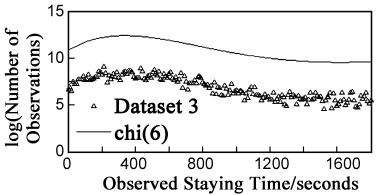


图4 数据三的停留时间观测样本的对数分布图

那么图中的自然对数分布应该是直线,但实际情况并非如此,各数据集的停留时间分布曲线并非直线,其中数据一是接近自由度为4的卡方分布,而数据二则是接近自由度为8的卡方分布,而数据三则是接近自由度为6的卡方分布,验证了前面关于停留时间分布的假设。

实验2比较了本文方法(CQM)和传统的 Markov 单链^[6]、多链模型^[11]、文献[1]的方法(EC)以及文献[12]的方法(vMM)等参照方法之间的平均推荐准确率^[15],随机选取各数据集中80%的用户的浏览序列作为训练集,用于模型的学习及聚类,其余20%作为测试集,并设 $N=20$ 。实验结果如图5所示,图中横坐标表示测试用户浏览序列的长度,纵坐标表示平均准确率。可以看出,在测试用户浏览序列长度较小的时候(以数据一的结果为例),浏览序列长度 L 小于3时,传统的单链模型和vMM方法以及EC方法的精确度比CQM及多链模型要高,准确度从高到低排列依次为:EC、vMM、单链模型、CQM、多链模型;而当浏览序列长度 L 介于3~6之间时,准确度从高到底排列依次为:EC、vMM、CQM、单链模型、多链模型;随着浏览序列长度增大到大于6时,准确度从高到底排列依次为:CQM、EC、vMM、多链模型、单链模型。产生这种现象的主要原因是:在CQM和

多链模型中涉及到的用户聚类,是基于用户历史浏览序列进行的,当序列过短时,获取的用户信息少,从而导致用户的分类不够准确,出现分类错误的可能性越大,随着用户浏览序列长度越来越大,分类正确的可能性就增加,从而提高了预测准确率;而EC和vMM方法在浏览序列短的时候等价于单链模型,因此保持了在浏览序列短时也有高准确率的优势,随着浏览序列增长,EC和vMM的多阶特性体现出来,因此一直高于单链模型;由于EC、vMM和多链模型只关注挖掘用户浏览序列并没有关注时间信息,因此到了一定的浏览序列长度之后,这些方法的精度就不及CQM。

对照图5中的(a)(b)(c),图5(a)中的准确率普遍比(b)和(c)中要高,即在相同的浏览序列长度的情况下,图5(a)中的准确率取值均大于(b)和(c)中对应的值,甚至超过了10个百分点,这是因为数据二和数据三相对来说比数据一要更加稀疏,这意味着数据二和数据三中存在大量短浏览序列的情况,从而影响了模型的预测准确性。尽管如此,在用户浏览序列长度大于5时,CQM的预测推荐准确率还是维持在50%以上,比起其他参照方法具有一定的优势。

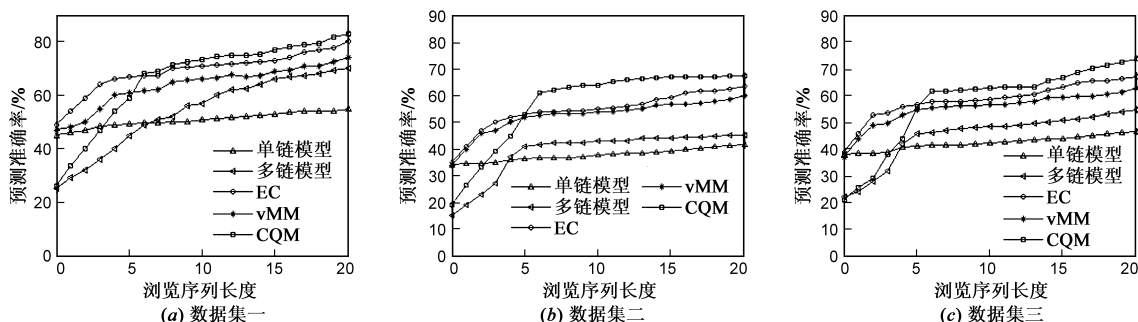


图5 本文模型和传统模型的平均准确率比较

实验3比较了各方法的平均推荐召回率^[15],相关设置和实验2一样。实验结果如图6所示,跟实验2类似,CQM和传统的多链模型在测试用户浏览序列长度较小的时候,推荐效果比传统的单链模型、EC以及

vMM要差,而当浏览序列长度超过了一定大小时,其召回率则超过了单链模型,同时CQM超过了EC和vMM;CQM和传统的多链模型在召回率这一指标上相比,前者的效果要好。

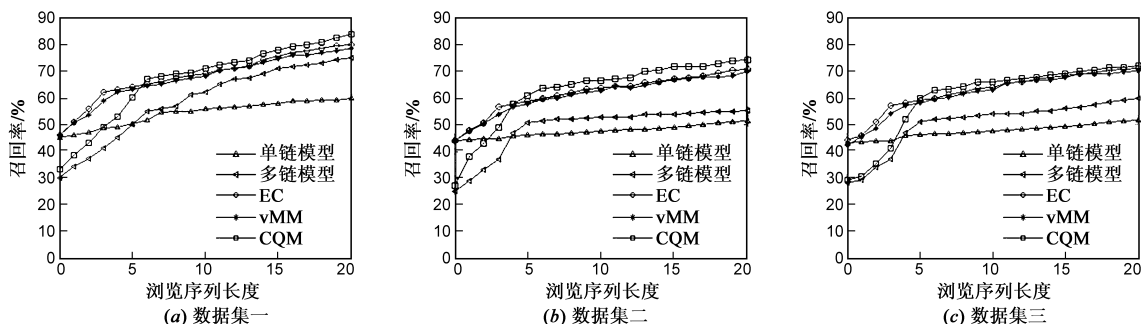


图6 本文模型和传统模型的平均召回率比较

实验 4 就是为了检验聚类结果中各个类别下的用户个数对模型推荐准确度的影响,在实验中,分别针对每个聚类的个数进行,实验结果如图 7 所示,从图中可以看出,当某一聚类中的元素个数偏低时,推荐准确度很低,随着个数增加,推荐准确率开始增长,在元素个数处于 10~100 之间时准确率的增加最为迅速,这就意味着随着类别中的元素个数的增加使得用户信息的获取也越充分,而提升到一定程度后,增长的趋势变缓,这是由于随着元素个数的增加,信息的冗余程度也越高,因此预测准确率的提升变得缓慢.由于在类别中的元素个数比较少的时候,准确率不是很理想,因此在实验 2 和 3 中计算的都是平均准确率,而不是单个类别下的推荐准确率.

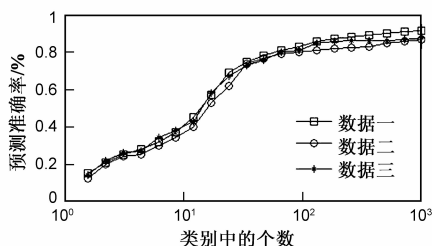


图7 聚类结果里每类元素个数对预测准确率的影响

7 结论

本文通过对用户的隐性反馈诸如跳转关系、停留时间等因素进行综合考虑,结合传统的 Markov 预测模型,提出了新的类时齐 Markov 模型,并将该模型用于 WEB 系统中资源推荐,最后在三个数据集上和其他相关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更好.

但是本文主要是基于用户的网页间的跳转行为和停留时间进行的预测和推荐,分析节点粒度是网页,而用户在网页间的随机游走受潜在的兴趣驱使,接下来可以将节点粒度提升到兴趣概念级别,并结合隐含概念挖掘技术进一步研究,分析用户的兴趣迁移过程.

参考文献

- [1] Awad M A, Khalil I. 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] 吴永辉, 王晓龙, 等. 基于主题的自适应在线网络热点发现方法及新闻推荐系统[J]. 电子学报, 2010, 38(11): 2620 - 2624.
Wu Y H, Wang X L, et al. Adaptive on-line web topic detection method for web news recommendation system[J]. Acta Electronica Sinica, 2010, 38(11): 2620 - 2624. (in Chinese)
- [3] 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.
- [4] 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. (in Chinese)
- [5] 黄世平, 黄晋, 等. 自动建立信任的防攻击推荐算法研究[J]. 电子学报, 2013, 41 (2): 382 - 387.
Huang S P, Huang J, et al. Anti-attack recommender algorithm based on automatic trust establishment[J]. Acta Electronica Sinica, 2013, 41(2): 382 - 387. (in Chinese)
- [6] Wan M, Jönsson A, et al. A random indexing approach for web user clustering and web prefetching[A]. Proceedings of the 15th International Conference on New Frontiers in Applied Data Mining [C]. Berlin, 2011. 40 - 52.
- [7] Bhawna N, Suresh J. Generating a new model for predicting the next accessed web page in web usage mining[A]. Proceedings of 3rd International Conference on Emerging Trends in Engineering and Technology[C]. India, 2010. 485 - 490.
- [8] Oard D W, Kim J. Implicit feedback for recommender systems[A]. Proceedings of the AAAI Workshop on Recommender Systems[C]. Wollongong, 1998. 81 - 83.
- [9] 张引, 张斌, 等. 面向自主意识的标签个性化推荐方法研究[J]. 电子学报, 2012, 40(12): 2353 - 2359.
Zhang Y, Zhang B, et al. Autonomy oriented personalized tag recommendation[J]. Acta Electronica Sinica, 2012, 40 (12): 2353 - 2359. (in Chinese)
- [10] Knijnenburg B P, Willemsen M C, et al. Explaining the user experience of recommender systems[J]. User Modeling and User-Adapted Interaction, 2012, 22(4-5): 441 - 504.
- [11] 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. (in Chinese)
- [12] 林文龙, 刘业政, 等. 基于混合隐 Markov 链浏览模型的 WEB 用户聚类与个性化推荐[J]. 情报学报, 2009, 28 (4): 557 - 564.
Lin W L, Liu Y Z, et al. Web user clustering and personalized recommendation based on mixtures of hidden Markov chain models[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28 (4): 557 - 564. (in Chinese)
- [13] Liu Y T, Liu T Y, et al. Browse Rank: 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.

- [14] Konstan J A, Riedl J. Recommender systems: from algorithms to user experience[J]. User Modeling and User-Adapted Interaction, 2012, 22(1-2): 101-123.
- [15] 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, Lu L Y. Evaluation metrics for recommender systems[J]. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 163-175. (in Chinese)

作者简介



刘胜宗 男, 1986 年出生, 湖南邵阳人. 中南大学信息科学与工程学院博士研究生. 研究领域为数据挖掘、推荐系统、智能信息处理.

E-mail: lshz179@163.com



廖志芳 女, 1968 年出生, 湖南长沙人, 中南大学软件学院副教授、硕士生导师. 研究领域为数据挖掘、推荐系统等.



樊晓平(通信作者) 男, 1961 年出生, 浙江绍兴人. 中南大学信息科学与工程学院教授、博士生导师, 湖南财政经济学院网络化系统研究所所长. 研究领域为无线传感器网络、网络化系统控制、智能信息处理、智能交通系统等.

E-mail: xpfan@mail.csu.edu.cn