

基于 Web 访问信息的用户兴趣迁移模式的研究

马力¹ 谭薇² 李培²

(西安邮电学院信息中心 西安 710061)¹ (西安邮电学院计算机学院 西安 710061)²

摘要 针对用户浏览网页的兴趣会随时间而变化这一现象,设计了一种网络用户兴趣迁移模式的挖掘模型。把用户的访问兴趣通过隐马尔可夫模型抽象成一种时间序列,以此反映用户兴趣的序列性,进而利用 GSP 算法从用户兴趣序列中挖掘出用户兴趣的迁移模式。实验证明该方法是有效的,从时间属性上更深层次地描述了用户兴趣的变化情况。

关键词 兴趣迁移模式,隐马尔可夫模型,序列模式挖掘

中图法分类号 TP393

文献标识码 A

Research of User Interest Drift Pattern Based on Web Access Information

MA Li¹ TAN Wei² LI Pei²

(Information Center, Xi'an Institute of Posts and Telecommunications, Xi'an 710061, China)¹

(School of Computer Science and Technology, Xi'an Institute of Posts and Telecommunications, Xi'an 710061, China)²

Abstract Focused on the phenomenon that the visited interest of user would change with time, a model of mining user's interest drift pattern was proposed. Abstracted the visited interest of user into a time sequence with the method of hidden Markov, which was used to reflect the sequential features of the user interest. Used GSP algorithm to mine the drift of user's visited interest pattern from the interested sequence of user. At last, verified the feasibility of the given model using the simulation experiments, and the model could give a deeper description of the draft of the visited interest on the attribute of time.

Keywords User interest drift pattern, Hidden markov model, Sequential pattern mining

1 引言

互联网的服务模式正在逐渐向分布式、主动式、个性化等方向演进,其中个性化服务要求人们对用户上网行为模式进行充分的理解。个性化服务技术的出现在一定程度上解决了 Internet 中信息海量增长与用户获取信息手段相对简单之间的矛盾。用户兴趣模型作为个性化推荐服务的基础,仍有许多未解决的问题,依旧是人们研究的热点问题之一。

最近几年来,用户建模技术作为个性化服务中的基础研究工作,发展愈来愈趋向成熟。针对由于用户兴趣模型无法准确描述用户发生迁移后的当前兴趣所造成的误差,以及根据模型造成的推荐误差,研究者们对此开展了一些学术研究和探索。目前,这些有关迁移模型的研究基本上可以分为对兴趣遗忘和时间窗的两类研究。Grabtree 和 Soltysiak^[1]采用时间窗方法来解决用户兴趣漂移问题,即只采用最近浏览的 1 个样本训练用户模型,1 为时间窗的大小。Widmer 和 Kuhl^[2]在时间窗方法的基础上进行了改进,使窗口的大小能够随着系统预测精度的变化而自动调整,这样训练出来的用户模型可以适应用户兴趣的不规则漂移。Maloof 和 Michals-

ki^[3]采用一种遗忘机制来衰减样本,每个样本赋予一个年龄,年龄随着时间的推移而增长,当年龄超过某一阈值时,样本将被完全遗忘,只有未被遗忘的样本才用来训练用户模型。这种对样本的时间遗忘机制使训练出来的用户模型能够反映用户兴趣的漂移。Koychev 和 Schwab^[4]分析了用户兴趣的漂移,改进了 Maloof 和 Michalski 的遗忘机制,提出用户模型的渐进遗忘法,以适应用户兴趣的漂移。Ma Shanle, Li Xue, Ding Yi^[5]等人采用引入时间敏感度函数的方法在理论和实验中实现了一种新颖的基于兴趣迁移的示范系统,克服了用户兴趣偏移的问题。上述方法从兴趣迁移的角度出发,改进了原有的用户模型,通过实验发现此类基于迁移的用户兴趣模型基本能够适应变化快速多样的用户兴趣。另外,上述方法中的个性化用户模型主要是基于用户的浏览内容来描述用户兴趣的,较少考虑用户的浏览行为。从一个较长时期上看,一个人的兴趣是在不断变化着的。但无论其兴趣改变与否,一个人的兴趣必定也能够通过他的浏览行为表现出来。用户对不同页面感兴趣的程度不同,其浏览行为在很大程度上体现出用户的真实兴趣。通过对他的浏览行为进行分析,可以了解用户的兴趣偏好。结合分析用户浏览内容和浏览行为等

到稿日期:2010-06-18 返修日期:2010-11-25 本文受陕西省自然科学基金资助项目(SJ08-ZT15),陕西省教育厅专项科研计划项目(08JK425)资助。

马力(1961-),男,教授,硕士生导师,主要研究方向为人工智能;谭薇(1982-),女,硕士生,主要研究方向为智能信息处理, E-mail: jellytw@163.com(通信作者);李培(1980-),女,讲师,主要研究方向为网络安全、智能信息处理。

相关信息,才能更为精确地描述用户兴趣。

2 模型和定义

本文结合 Web 使用挖掘和内容挖掘的思想,提出了一种用户兴趣迁移模式挖掘的模型,以发现用户兴趣的迁移模式,反映用户的访问偏好。图 1 显示了用户兴趣迁移模式的挖掘模型。首先,从 Web 访问信息中得到用户访问序列;然后,用一组主题矢量来表示页面,进行页面聚类;继而,将用户访问序列同用户兴趣相结合,形成用户兴趣序列库,供迁移模式挖掘算法使用;最后,从用户兴趣序列库中分析得到用户兴趣迁移模式。

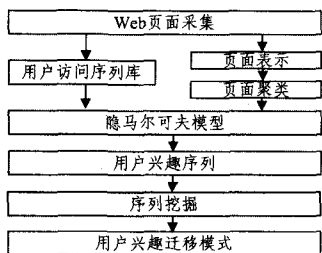


图1 用户兴趣迁移模式挖掘模型

2.1 模型中的基本概念

定义 1(访问序列) 我们将用户访问某个感兴趣页面的请求称为一个事件,因此采集的网页记录可以看成是由某个用户的若干事件组成的,事件按发生时间的升序排列,称为访问序列。

定义 2(用户兴趣序列) 给定一个用户访问序列,如果将这些序列中的每个页面替换成它相应的主要内容表示,连续相同的内容合并,那么替换后的这个序列就是用户兴趣序列。

定义 3(用户兴趣序列迁移模式) 最频繁访问的用户兴趣序列就是用户兴趣的序列迁移模式。

2.2 模型的分析

页面采集工作是指将用户浏览过的 Web 页面收集起来,清除一些无用数据,只保留用户感兴趣的内容页面,作为数据挖掘的数据源。本文在进行网页采集时,加入时间属性,即当采集到一篇用户喜欢的页面文档后,把其采集到该页面的时间作为页面的一个属性,称为页面加入时间。页面加入时间以绝对时间表示,比如从 2010 年 1 月 1 日零点到当前时间的秒数。

页面表示工作是指将网页文本采用向量空间模型 VSM (Vector Space Model)表示。这个过程包括文本预处理工作、特征提取工作、特征词权值计算工作。

页面聚类工作是采用课题组改进的基于后缀树的聚类算法对采集到的页面进行页面文本聚类,使得相似的页面文本组成簇。聚类结果中每一篇页面文档可能会属于多个类别,每一个类别又会包含至少一篇页面文档。

用户访问序列库构造工作是指将采集到的所有页面按其时间属性的升序排列,根据指定的时间周期进行分割,构成访问序列集。从开始采集数据的那一刻开始计时,每隔一个时间周期内的访问页面放在同一条序列中。

引入隐马尔可夫模型是为了完成用户访问序列同用户兴趣类相结合的工作,即将访问序列中的每个页面替换成它可能属于的相应的兴趣类,由用户访问序列得到用户兴趣序列。

序列模式挖掘工作是指从用户兴趣序列数据库中挖掘频繁出现的子序列,这种频繁子序列被定义为用户兴趣迁移序列。

3 基于隐马尔可夫模型的用户兴趣序列发现

本文提到的用户兴趣是用类别来表示的,用户兴趣类别序列就反映了用户兴趣间的转移关系,将用户兴趣类别序列定义为用户兴趣序列。一条用户兴趣序列可以按下式进行描述:

$Q = \langle (e_1, \text{cluster}, e_1, \text{time}, e_1, \text{interest}), \dots, (e_i, \text{cluster}, e_i, \text{time}, e_i, \text{interest}), \dots, (e_m, \text{cluster}, e_m, \text{time}, e_m, \text{interest}) \rangle$, 其中, $e_i \in E, 1 \leq i \leq m, e_i, \text{cluster}$ 表示页面 e_i 所属的类别, e_i, time 表示页面 e_i 采集到的时间, $e_i, \text{interest}$ 表示页面 e_i 的兴趣度。

由定义 1 可知,对于给定的一个用户访问页面序列,如果将这些序列中的每个页面替换成它相应的类别表示,那么替换后的这个序列就是用户兴趣序列。

由于页面文档具有多主题性,即一个页面属于多个聚类类别,因此适合用隐马尔可夫模型来解决从访问页面序列到访问兴趣类别序列的转换问题。

设经过页面聚类后有 N 个聚类类别,每个类别里面包含有很多页面。一个页面可以属于多个类别,类别与所包含的页面之间由一组概率分布值来描述。将访问页面序列视为可以观察到的观察值序列,各个页面所属于的类别不能直接看到状态,而是通过一个随机过程去感知状态的存在及其特性。类别与类别之间的转移以及每次选取的类别被隐藏起来了,不能直接观察到。我们通过应用隐马尔可夫模型来获得当前观察到的用户访问页面序列的类别序列。

3.1 隐马尔可夫模型的基本概念

隐马尔可夫模型(如图 2 所示)由两个随机变量序列组成:一个是观察不到的马尔可夫链, $\{q_t \geq 0\}$;另一个是可以观察到的随机序列 $\{O_t \geq 0\}$ 。 $\{q_t \geq 0\}$ 称为马尔可夫链, $\{O_t \geq 0\}$ 称为其观察链。

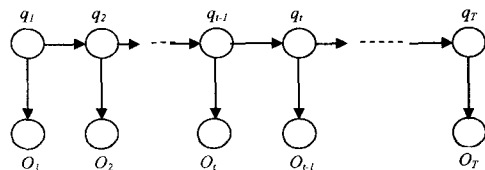


图2 隐马尔可夫模型

HMM 一个清晰的极好描述由 Rabiner 给出^[6]。HMM 由以下几种元素组成:

(1) N : 模型中 Markov 链的状态数目。记 N 个隐藏的状态集合 $S = \{S_1, S_2, \dots, S_N\}$, 并记 t 时刻 Markov 链所处的状态为 q_t , 显然 $q_t \in \{S_1, S_2, \dots, S_N\}$ 。

(2) M : 每个状态对应的可能出现的观察值数目。记 M 个观察值观察符号集合 $V = \{V_1, V_2, \dots, V_M\}$, t 时刻的观察值为 O_t , 显然 $O_t \in \{V_1, V_2, \dots, V_M\}$ 。

(3) A : 状态转移概率矩阵, $A = (a_{ij}) N \times N$, 其中 a_{ij} 表示从状态 $i (1 \leq i \leq N)$ 到状态 $j (1 \leq j \leq N)$ 的转移概率。

$$a_{ij} = q_{t+1} = p(q_{t+1} = S_j | q_t = S_i) \quad 1 \leq i, j \leq N$$

(4) B : 观察值概率矩阵, $B = (b_j(k)) N \times M$, 其中 $b_j(k)$ 表示状态 $j (1 \leq j \leq N)$ 出现第 $k (1 \leq k \leq M)$ 个观察值的概率。

$$b_j(k) = p(O_t = V_k | q_t = S_j) \quad 1 \leq j \leq N, 1 \leq k \leq M$$

式中, O_t 表示在 t 时刻的观察值。

(5) π : 初始状态概率矢量, $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$, 其中 $\pi_i (1 \leq i \leq N)$ 表示开始时候选取状态 i 的概率。

$$\pi_i = p(q_1 = S_i) \quad 1 \leq i \leq N$$

这样, 一个 HMM 可以记为 $\lambda = (N, M, A, B, \pi)$ 。

3.2 隐马尔可夫模型的构造

用户对 Web 的访问可以看作是一个随机过程 $\{X_n\}$, 如果当前浏览的兴趣为 i , 接下来浏览兴趣 j , 它的条件概率为 $P_{ij} = P(i|j)$, 这就是一阶 Markov 链。例如假定用户当前浏览的兴趣是 i , 并且已经浏览了兴趣 $i_{n-1}, \dots, i_0$, 接下来浏览的兴趣 j 的条件概率 $P_{ij}(k)$ 只依赖于兴趣类 $i, i_{n-1}, \dots, i_{n-k+1}$, 即:

$$P_{ij}(k) = P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_{n-k+1} = i_{n-k+1})$$

一般隐马尔可夫模型用一个五元组 $\lambda = (N, M, A, B, \pi)$ 表示, N 是用户所有可能的兴趣类的集合, M 是用户访问的所有页面的集合, A 是兴趣类到兴趣类的转移概率, B 是在某一状态下用户访问页面隶属于该状态的概率, π 是初始状态的选择概率。

(1) 用户访问页面经过聚类后发现的类别主题为 HMM 的状态节点 q , 给定一个虚拟的初始状态 q_1 ;

(2) 所有的用户访问页面集 $E = \{e_i | i = 1, \dots, M\}$;

(3) q_j 节点包含 E 的一个子集 $E' = \{e_1^j, e_2^j, \dots, e_n^j\}$ 。这个子集就是用户访问页面进行聚类后属于兴趣类 q_j 的页面集合;

(4) 两个节点 q_i, q_j 之间存在着一个转移概率 $P(q_i \rightarrow q_j)$; 在访问类别序列集 C 中计算任意两个节点 q_i, q_j 的概率:

$$P(q_i \rightarrow q_j) \approx \frac{\text{count}(q_i \rightarrow q_j)}{\text{count}(q_i)} \quad (1)$$

式中, $\text{count}(q_i \rightarrow q_j)$ 表示在访问类别序列集 C 中 q_i, q_j 同时出现且 q_j 紧随 q_i 的次数; $\text{count}(q_i)$ 为 C 中含有 q_i 的次数。

(5) 在一个兴趣类节点 q_j 上, 属于该节点的页面集 E' 与这个节点之间存在一个概率分布 $P(e_1^j | q_j), P(e_2^j | q_j), \dots, P(e_n^j | q_j)$, 即为 HMM 中状态节点的观测概率。 $P(e_n^j | q_j)$ 的计算方法如下: 用式(2)计算该节点上所有页面与该兴趣类 q_j 的聚类中心 c_j 之间的相似度, 再经过归一化后的取值便是 $P(e_n^j | q_j)$ 的值。

$$\text{Sim}(e_i^j, c_j) = \cos\theta = \frac{\sum_{k=1}^n e_{ik}^j \times c_{jk}}{\sqrt{(\sum_{k=1}^n (e_{ik}^j)^2)(\sum_{k=1}^n (c_{jk})^2)}} \quad (2)$$

通过上述数据准备过程以及相应的 HMM 定义, 就可以建立一个 HMM 模型。

在此, 我们借鉴隐马尔可夫模型中的解码问题的描述, 即已知用户访问序列 $O = O_1, O_2, \dots, O_T$ 和模型 $\lambda = (A, B, \pi)$, 运用 Viterbi 算法选择相应的在某种意义上最佳(能最好解释观察序列)的用户兴趣序列 $Q^* = q_1^*, q_2^*, \dots, q_T^*$ 。

Viterbi 算法用于搜索能够生成观察序列的最大概率的状态序列, 该算法能够找到最佳解, 其思想精髓在于将全局最佳解的计算过程分解为阶段最佳解的计算。我们所需要的最佳用户兴趣类别序列 Q^* 就可用 Viterbi 算法求出。 Q^* 具有访问类别序列的最大概率。

4 基于序列模式挖掘的用户兴趣迁移模式发现

分析 Web 用户访问历史记录, 识别出用户访问兴趣序列后, 就可以对它进行挖掘, 发掘出隐藏在数据背后的规律和模式。本文中的序列数据都带有时间特征, 挖掘的目的是发现序列中事件之间的模式。为了发现用户多次页面访问中用户最感兴趣的访问规则和模式, 引入了序列模式挖掘方法。

序列模式挖掘在关联模型中增加了时间属性, 把数据之间的关联性与时间联系起来, 寻找的是事务之间在时间上的先后次序关系。应用序列模式挖掘, 再根据数据随时间变化的趋势, 发现某一时间段内数据的相关模式后, 可以利用这些模式预测将来可能出现的情况。

通过对序列模式挖掘相关算法的特性和适用范围进行理论上的分析和比较, 本文将 GSP 算法^[7]应用到用户兴趣序列的挖掘中。本文结合 Web 页面内容来分析用户的访问行为模式, 将页面序列集转换为页面所属的类别序列集, 从整个类别序列集中找到频繁序列。挖掘出的这些频繁序列反映的是用户访问兴趣类别序列随时间推移发生了哪些变化, 用户访问类别序列之间有什么变化规律, 这些体现变化和规律的访问模式就是用户的兴趣迁移模式。由于篇幅有限, 具体内容见文献[8]。

5 实验分析

我们的实验数据集是使用 Winpcap 在局域网中捕获用户一个月以来所发送的数据包, 然后提取出其中有效的 URL, 接着依据这些 URL 使用网页爬虫程序取得用户浏览网页。通过对捕捉到的浏览页面进行 URL 分析汇总后得到 712 个浏览页面, 分析后取了其中的 302 个有用页面进行挖掘, 分别编号为 1, 2, ..., 302。

考虑到文档的多主题特性, 本文将基于后缀树的文本聚类方法应用于页面文本聚类中。通过聚类发现这 302 个页面分布于 8 个类, 分别是艺术、计算机、经济、环境、医药、军事、体育、交通。

对采集到的用户历史访问页面建立隐马尔可夫模型, 初始状态的选择是随机的。模型中有 8 个状态值, 也就是经过兴趣聚类出来的 8 个兴趣概念。构建了隐马尔可夫模型后, 利用隐马尔可夫模型中的经典 Viterbi 算法将用户在这一个月内的访问序列转变为用户兴趣序列, 结果如表 1 所列。访问 ID 表示用户访问时间的标识符, 如“1”表示用户在实验中访问 Web 的第一天。

表 1 用户兴趣序列挖掘的结果

访问 ID	用户访问序列	用户兴趣类别序列
1	1, 2, 3, 4, 5, ...	计算机, 军事, 经济, 环境
2	17, 18, 19, 20, ...	体育, 军事, 经济, 环境
3	31, 32, 33, 34, ...	军事, 经济, 环境, 计算机
4	46, 47, 48, 49, ...	艺术, 军事, 经济, 环境
5	61, 62, 63, 64, ...	艺术, 计算机
6	77, 78, 79, 80, ...	军事, 艺术
7	91, 92, 93, 94, ...	艺术, 经济, 计算机
.....		

通过 HMM 挖掘的过程, 我们得到了用户兴趣序列。将用户在这一个月内的兴趣序列组成兴趣序列数据库, 从中挖掘出用户的兴趣迁移序列。

(下转第 219 页)

[8] Gori M, Pucol A. A random-walk based scoring algorithm with application to recommender systems for large-scale ecommerce [C]// Proceedings of WEBKDD'06, Philadelphia, USA, 2006: 127-146

[9] Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites[J]. Machine Learning, 1997, 27: 313

[10] Zhang Y-C, et al. Recommendation model based on opinion dif-

fusion [J]. Europhys. Lett., 2008, 80: 68003

[11] Zhou T, et al. Bipartite network projection and personal recommendation [J]. Phys. Rev. E, 2007, 76: 046115

[12] Zhang Y-C, Blattner M, Yu Y-K. Heat Conduction Process on Community Networks as a Recommendation Model [J]. Phys. Rev. Lett., 2007, 99: 154301

[13] Zhou T, et al. Accurate and diverse recommendations via eliminating redundant correlations[J]. New J. Phys., 2009, 11: 123008

(上接第 177 页)

用户访问序列经过 HMM 处理后的用户兴趣类别序列作为序列模式挖掘的数据源, 存在数据库中, 形成序列数据库, 共包含 30 个序列。GSP 算法运行时的最小支持度可根据实际情况设置, 支持度的大小会直接影响到算法执行的时间和挖掘出的模式。例如, 如果模式级别定义得比较高, 那么在预处理阶段就有可能清洗掉很多低级类别的用户兴趣浏览记录。此时如果将最小支持度定义得比较大, 就很难发现长序列模式。算法运行时所需数据直接从序列数据库中获取。将最小支持度分别设置为 3, 4, 5 时, 算法执行结果如表 2 所列。

表 2 在不同的支持度阈值下 GSP 算法执行的结果

最小支持度=3		最小支持度=4		最小支持度=5	
用户兴趣 迁移模式	模式 支持度	用户兴趣 迁移模式	模式 支持度	用户兴趣 迁移模式	模式 支持度
军事→经济→ 环境→计算机	4	军事→经济 →环境	7	军事→经济→ 环境	7
		经济→军事→ 计算机	4		
		经济→环境→ 计算机	4		

本文在测试时把最小支持度设置为 3, 程序运行时间为 62ms, 发现的序列模式很多, 最长的序列模式长度为 4。下面以长度为 4 的序列模式(军事、经济、环境、计算机)为例解释该模式的意义: 该模式表示在所有用户访问兴趣序列中, 至少有 4 天的时间用户在 Web 上都有这样一个访问路线: 军事→经济→环境→计算机。可以看出用户对这 4 个类比较感兴趣。

序列模式挖掘出来的结果是从全局的角度挖掘出的用户兴趣类别, 反映了一个用户具有持久的兴趣倾向和用户访问 Web 的偏好。通过分析可知, 最长序列模式中包含的兴趣类别最能反映用户访问兴趣的稳定性。本系统中将挖掘出的最长序列模式中的兴趣类别作为用户兴趣模型中的稳定兴趣。在不同支持度阈值的设定下提取的用户稳定兴趣结果如表 3 所列。

表 3 用户稳定兴趣提取结果

最小支持度=3		最小支持度=4		最小支持度=5	
用户兴趣 迁移模式	用户稳 定兴趣	用户兴趣 迁移模式	用户稳 定兴趣	用户兴趣 迁移模式	用户稳 定兴趣
军事→ 经济→ 环境→ 计算机	军事、 经济、环境、 计算机	军事→经济 →环境	军事、 经济、 环境、 计算机	军事→ 经济→ 环境	军事、 经济、 环境

从表中可以看出, 当算法中最小支持度为 3 时用户稳定兴趣为军事、经济、环境、计算机; 算法中最小支持度为 4 时用户稳定兴趣为军事、经济、环境、计算机; 算法中最小支持度为

5 时用户稳定兴趣为军事、经济、环境。

为了评估本文中提出的兴趣迁移方法的性能, 定义算法的查全率和精确度:

查全率 = 正确识别出用户的兴趣数 / 专家推荐的用户兴趣总数;

精确度 = 系统正确标记的用户的兴趣类别 / 系统标记的用户兴趣类别总数。

考虑到专家的知识是主观的, 不能提供精确的定量分析, 我们邀请了 5 组专家来做同一个试验, 以期得到一个比较客观的结果。经过分析, 在本实验中, 专家推荐的用户兴趣总数最终定为 5, 即计算机类、体育类、环境类、经济类、军事类。

实验中, 当把支持度阈值设为 3 时, 提取出了 4 个稳定兴趣, 算法的查全率为 0.8; 支持度阈值设为 4 时, 提取出了 4 个稳定兴趣, 算法的查全率为 0.8; 支持度阈值设为 5 时, 提取出了 3 个稳定兴趣, 算法的查全率为 0.6。

结束语 个性化 Web 服务通过收集和分析用户信息来学习用户的兴趣和行为, 从而达到主动推荐的目的。个性化服务技术能充分提高站点的服务质量和访问效率, 从而吸引更多的访问者。实现个性化服务的关键就是正确分析 Web 用户浏览信息, 准确地描述用户的兴趣。只有准确地把握用户的浏览兴趣以及兴趣的变化情况, 才能将用户感兴趣的资源推荐给用户, 更好地描述用户兴趣。

参 考 文 献

[1] Land C, Soltysiak S. Identifying and Tracking Changing Interests[J]. International Journal of Digital Libraries, 1998, 2: 38-53

[2] Cxand W, Kubat M. Learning in the Presence of Concept Drift and Hidden Contexts[J]. Machine Learning, 1996, 23: 69-101

[3] Maloof M, Michalski S. Selecting Examples for Partial Memory Learning[J]. Machine Learning, 2000, 41: 27-52

[4] Koychev I, Schwab I. Adaptation to Drifting User's Interests [C]// Proceedings of ECML2000 Workshop: Machine Learning in New Information Age, Barcelona, Spain, 2000: 36-45

[5] Ding Y, Li X. Time weight collaborative filtering [C]// Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM). New York, NY, USA: ACM Press, 2005: 485-492

[6] Rabiner L R. A Tutorial on Hidden Markov Models and selected Applications in Speech Recognition [J]. Proceeding of the IEEE, 1989, 77: 257-286

[7] 刘立军, 崔杰, 梅红岩. GSP 与 PrefixSpan 算法的比较与分析 [J]. 辽宁工学院学报, 2006(05)

[8] 谭薇. 基于 Web 访问信息的用户兴趣迁移模式的研究[D]. 西安: 西安邮电学院, 2010

[9] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2008

[10] 刘河生, 高小榕, 杨福生. 隐马尔可夫模型的原理与实现[J]. 国外医学: 生物医学工程分册, 2002, 25(6): 253-259