Longitudinal Data Based Research on Web User Interests Drift Modeling

Huimin Li, Liying Fang, Pu Wang, Jianzhuo Yan
College of Electronic Information & Control Engineering
Beijing University of Technology,
Beijing, China
wangpu@bjut.edu.cn, bgdl_hm@emails.bjut.edu.cn, fangliying@bjut.edu.cn, yanjianzhuo@bjut.edu.cn

***Abstract*—User interests drift modeling is the fundamental and key technology of Personal Recommendation System (PRS), and is arising more and more attention by researcher from home and abroad. At present, web user interests drift model is in its original phase for the reason of many aspects, and there also exist some problems need to be studied continually and deeply. Considering of web user interests might change with time flying, huge of surfing data and many influence factors need to be concerned, because user interests changes constantly with time flying. A novel web user interest drift pattern modeling method is proposed. First, this paper takes user interests as time series via using hidden Markvo model, which may properly map the sequential characteristics of user interests. Second, improved GSP searching algorithm is used to find the frequent pattern from user interest sequence. Finally, forgetting mechanism is used to solve interests drift problem, and time window is used to store current user interest pattern, and the old interests will elapse with the new interests coming constantly. Plenty of online experiments were done to verify the reasonability of the user interests drift model. Via experiments, conclusion can be drawn that this method can work perfectly and efficiently, which may firmly grasp user interests drift rule in time. Therefore, personal recommendation system using these interests drift model can recommend the latest information for web user, only according to user's potential interests changing.**

Keywords-Data Mining; Hidden Markov Model; Time Series; Interests Drift Pattern

I. INTRODUCTION

With the rapid development of Internet technology and sciences, information gained from Internet is explosion; it is easy for us to get lost in the information ocean. Nowadays, most of the web information is provided for public user, it seldom considers the differences among individuals, for example, traditional searching engines always returns the same results for the same query provided by people with different backgrounds, therefore, user cannot access fast and efficiently to their interests contents. The appearance of personalized recommendation technology makes it possible for information services personalization; provides specific information for different user. User interests drift model can accurately build user interests changing based on the user-related browser data; it is the key technology of personalized information services. User Interests Modeling has been researched in many areas for personalized information service; it greatly enhances the value of the World Wide Web. Grabtree and Soltysiak use fixed time window to tackle the problem of user interests drift in [1]. Maloof and Michaski adapt forgetting mechanism to select sample, each sample has age, and this 'age' can update automatically with the time flowing away, the sample will be forget, when the age of sample is old enough, therefore, only unforgotten sample can be used to train user interests model. This method can meet the requirement of interests drift, as in [2]. Ding Y and Ling X use time sensitivity function to realize a novel demo system; it can overcome the problem of user interests drift as in [3]. Han et al. proposed a user interest model based on folksonomy and ontology, developed a probability tree model to map user tags onto ontology, and applied to personalize Web searching as in [4]. Kim and Chan proposed an algorithm to cluster terms extracted from web pages, generated an User Interests Hierarchy （UIH） to represent the user interests from common to specificity in [5]. UIH can be used for personalized searching to recommend interesting information at different granularities and abstraction levels, and the concept of each node of UIH was represented only by the phrases. White et al. studied the correlation of the implicit indicators and explicit interests and the predictive power of different background information sources about user interests as in [6].

In a long run, one's browse interest is always changing constantly, and it can appear via behavior of browser. It indicates one's browse interests, if someone stays longer time on one website than another. So users browse contents and behaviors have to be anglicized, only in this way, can we depict user interests drift precisely.

II. PATTERN AND DEFINITIONS

Combining network with data mining, a novel user interest drift model is proposed. The model can be shown as follows, first, an access series is obtained from user browse behavior, then a group of vectors is used to represent the web pages, and cluster them; the access web page series are integrated to user interests and are formed as the user access interests database; at last, user interests

drift pattern can be gained by analyzing the interests database.

A. Basic Definitions

Definition 1: access web page series. We consider user surfing behavior as a case, several cases based time sequence can be joined to an access web page series, in some way, and we can treat them as longitudinal data.

Definition 2: for the given access web page series, replace the series with the core content, merge the same contents, then, a piece of user interest series can be formed.

Definition 3: we define the most frequent interest series coming from the web page as user interests drift pattern.

B. Model Construction

First, user web pages were collected, clean the unrelated web contents, add time stamp attribution to the data, and then form them to source data. Second, we do some preprocess to web page contents, it include characteristics selection and extraction, weight value computation, and represent the web page series as Vector Space Model （VSM）. Third, cluster the web page contents, and the possible result is one page maybe belongs to different class at the same time, and one kind of classification may include several pages. Finally, HMM (Hidden Markvo Model) [7] was used to combine user access series with user interests, i.e., replace the web page series with the interests class which belong to. Series pattern discovery refer to mine the frequent subsequence from the user database of interests series, User Interest Drift Pattern Discovery (UIDPD) is named.

III. HMM BASED USER INTERESTS DRIFT MODEL

User interest series can be described as follows:

Q={(*s1_class,s1_time,s1_interest*),(*s2_class,s2_time,s2_interest*),…,(*si_class,si_time,si_interest*),…,(*sm_class,sm_time,sm_interest*)}. Where *si_class* represents the class of the page, *si_time* represents the time of page collected, and *si_interest* represents the interest degree of page.

As one page has several themes, that is to say, one page may simultaneously belong to different class; therefore it is convenient to denote the transformation from web page series to user interests via using Hidden Markvo Model (HMM). Suppose there are *N* interests classes after clustering, each class includes several pages, and a page can belong to different interests class, page belong to interests class via different probability distribution value. The user pages are taken as observed data, and the status of each page cannot be seen directly, it can perceive the feature at random process. Therefore, HMM could be used to gain the user interests series.

A. Hidden Markvo Model

HMM is made up of two random variable series, one is hidden Markvo chain $q_t \geq 0$; the other is observable random variable series $o_t \geq 0$. Fig. 1 shows the HMM.

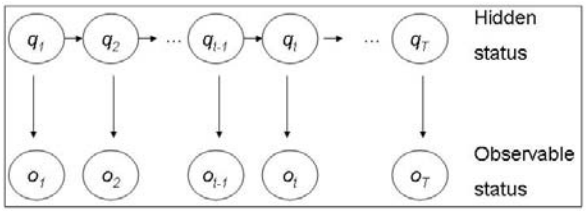

Figure 1. Hidden Markov Model

Generally speaking, HMM is made up of several elements, and it can be denoted as $\lambda = (N, M, A, B, \pi)$, where, parameter N represents the status number of Markvo chain, and the hidden status $S = \{S_1, S_2, ..., S_t, ..., S_N\}$, and it denotes Markvo chain of time t as q_t , obviously, $q_t \in \{S_1, S_2, ..., S_t, ..., S_N\}$. Parameter *M* represents the number of possible observable value, $V = \{V_1, V_2, ..., V_t, ..., V_M\}$, the observable value of time t is o_t , therefore, $o_t \in \{V_1, V_2, ..., V_t, ..., V_M\}$. Parameter A represents the matrix of status transformation probability, $A = (a_{ij})N \times N$, where a_{ij} represents the probability from status i $(0 \leq i \leq N)$ to status j $(0 \leq j \leq N)$. $a_{ij} = q_{t+1} = p(q_{t+1} = S_i \mid q_t = S_{i-1})$, s.t. $1 \leq i, j \leq N$. Parameter B represents the matrix of observable value probability, $B = (b_j(k))N \times M$, where $b_j(k)$ represents the probability that emergence $k(1 \leq k \leq M)$ observable value in status $j(1 \leq j \leq N)$. $b_j(k) = p(o_t = V_k \mid q_t = S_j)$ $1 \leq j \leq N, 1 \leq k \leq M$. Where o_t represents the observe value of time *t*. Parameter π represents the initial status probability vector, $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$, where $\pi_i (1 \leq i \leq N)$ denotes the probability of selecting status *i*. $\pi_i = p(q_1 = S_i)$, s.t. $1 \leq i \leq N$.

Three parameters are necessary for HMM, which are status changing probability matrix *A*, status outputting probability *B* and initial status probability distribution π , i.e., the problem is how to select $\lambda = (A, B, \pi)$, and make the observation series probability $P(O \mid \lambda)$ maximum, so observation series is used as training samples, and *A*, *B*, π is used as unknown parameters, that is to say, parameters evaluation is the key for modeling. Baum Welch algorithm is a kind of repeat processing of estimation, which can generate the optimum parameters of model through iteration. Evaluation of 3 parameters is as fig. 2:

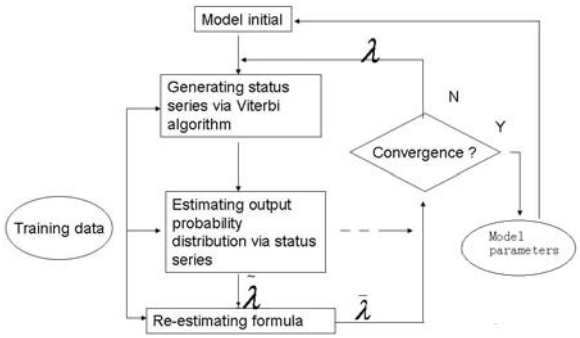

Figure 2. HMM parameters estimation method.

B. User Interests Sequence Construction Via HMM

User surfing behavior on Internet can be called a random process $\{X_n\}$, suppose current user interests is *i*, and has browsed $i_0,...,i_{n-1}$, therefore, the probability of next interests *j* can be denoted as

$$p_{ij}(k) = p(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1},...,X_{n-k+1} = i_{n-k+1}). \quad (1)$$

In HMM model [8], *q* denotes the status of hidden Markvo chain, let's suppose initial status q_1, all user page set $E = \{e_i \mid i = 1,...,M\}$; for random knots q_i and q_j in access set *C*. Transformation probability $p(q_i \to q_j)$ can be represented as

$$p(q_i \to q_j) \approx num(q_i \to q_j) / num(q_i). \quad (2)$$

Where $num(q_i \to q_j)$ represents the time that they appear and follow in the same time, $num(q_i)$ represents the number in access series. In view of HMM decode description, i.e., user access series $O = O_1, O_2,..., O_T$ and model $\lambda = (A, B, \pi)$, it use Viterbi algorithm to select the optimum user interests series $Q^* = q_1^*, q_2^*,..., q_T^*$. This algorithm can degrade global optimum result to different local phase optimum results.

C. Interests Pattern Discovery Using Improved GSP Algorithm

As we all know, user interests sequence can be gained by HMM analyzing from user surfing records, then we would find the user interests drift pattern via using improved GSP algorithm. GSP algorithm is the extended version of Apriori algorithm [9], it use time window technique and constraint mechanism, it can give the optimum frequent sequence quickly. GSP algorithm includes two steps, one is maximum frequent sequences generation, and the other is rules generation. The improved GSP algorithm can be described in three steps: step 1 is the process of frequent series generation, step 2 is the maximum frequent sequences is generated (detail is as follows), step 3 is the rules generation.

Improved GSP Algorithm can be described as follows:

Step 1: scan time series database, gain series pattern L_1 with the length of 1, and is regarded as the initial seed set.

Step 2: according to seed set L_i with the length of *i*, by connecting and pruning operation, generate candidate series pattern C_{i+1} with the length of *i*+1.

Step 2.1: connection phase: if we cut down the first item of pattern s_1, which has the same effect with deleting the last item of pattern s_2, so s_1 can be connected with s_2, that is to say, the last item of s_2 should be added to the end of s_1.

Step 2.2: pruning phase: any sub sequence should be deleted in some candidate series, if it is not a series pattern.

Step 3: duplicate step 2, processing will stop until no more new series pattern and new series candidate series pattern generation.

Step 4: scan series database, calculate the supporting degree of each candidate series pattern, generate series pattern L_{i+1} with the length of *i*+1, and consider L_{i+1} as the new seed set.

Step 5: generate the maximum frequent sequences from frequent series.

Step 6: rule generation. For each frequent sequence L, generate every non-empty subset. Then, for each subset, if *support(L)/support(s)>min_conf*, then *output s->1-s*. Consequently, the interest pattern can be gained.

The step 5 can be described in detail as follows:

Input: frequence series L_k;

Output: the maximum frequent sequences L_{max};

Step a): delete the subsequence of the maximum series L.

L_{temp} initial;

For (i=0,i<L_{max}, i=i+1) do

Add *L* to L_{max},

For (j=1,j<*Result count*, j=j+1) do

If (Result is not subsequence of *L*)

Add non-subsequence of L_k to L_{temp};

Step b): delete the subsuqence in L_{temp}.

If (L_{temp} is not empty)

L=the maximum sequence item in L_{temp}.

Result =L_{temp}

Duplicate step a) and step b).

D. Forgetting Mechanism

As far as we concerned above, the user interests sequence $Q^* = q_1^*, q_2^*,..., q_T^*$ was gotten. Furthermore, user interests always change with the time elapsed, therefore, the user interests drift model must suit to this changing trend. and forgetting mechanism can solve this problem[10], here, time window constrain was leaded into experiments, which cover over several interests points in user interests pattern, $Q^* = q_1^*, q_2^*,..., q_T^*$. In other words, the user interests sequence can be looked as a queue with the length of 3~5 points. Figure 3 show the forgetting mechanism principal.

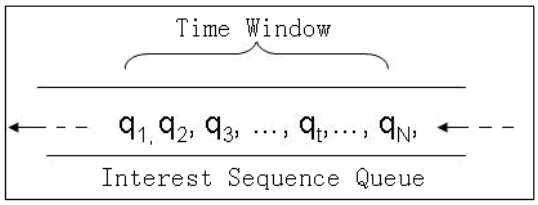

Figure 3. Principal of Forgetting Mechanism

In fig. 3, q_1 represents user interest in the past, and q_N represents the latest interest. The queue window size is set to 3~5, via experiments, we found, 4 is the optimum selection of window size, we know from trials, the size of the interests sequence impact the model precise. Every time, user interests sequence was selected from right direction, if there is a new interests value q_{N+1} coming, the oldest interests q_1 will be deleted automatically. Only in this forgetting mechanism way, can our interests drift model adapt the changing of status in user surfing behaviors.

IV. Experiments and Analysis

Plenty of experiments were done to verify this modeling method; the experimental data come from China Telecom of Beijing branch. Winpcap was used to catch data package during two month, and filter the useless contents out, then only effective URL leave behind. Webpage reptile was used to obtain web content according to URL, via analyzing and integration, 508 page contents is selected to this experiment. In view of the diversity of page themes, suffix tree method was used to cluster the pages, and get eight categories, including Art, Economy, Computer, Environment, Medicine, Military, Sports, Traffic. HMM model based on user access page was built, the initial status is stochastic, eight status of model corresponding to eight user interests. Then, the classical Viterbi algorithm was used to transform user access series to user interest sequence, table 1 shows part of the user interests sequence by HMM.

TABLE I. User Interests Pattern

Time	User surf sequence	User interests sequence list
1	1,2,3,4,5,…	Computer, Military, Art, Economy
2	13,14,15,…	Military, Art, Economy, Computer
3	23,24,25,…	Economy, Sports, Military, Environment
4	36,37,38,…	Art, Computer, Sports, Environment
…	…	…
57	440,441,…	Sports, Medicine, Computer, Traffic
58	461,462,…	Art, Military, Medicine, Traffic
59	480,481,…	Art, Military, Computer, Traffic
60	…,507,508	Computer, Military, Art, Medicine

For user interest sequences, HMM method was used to process them, and 60 interests items was gotten and used as source database. Then improved GSP algorithm was used to mine frequent interests pattern, here, the minimum support degree was used to indicate the precise and pattern. Long sequence is hard to be mined, if the minimum support degree is set too big. To be contrast, some slave sequence pattern will be ignored, if smaller support degree threshold value is used.

Our experiment hardware is Dell 760 with a dual core CPU and 4GB memories, and software Matlab 7.1 is used. We set the minimum support degree with 3 in improved GSP process, and obtain four frequent interests in 32ms, it is military-art-economy-computer at first part of the sequences, and in the second part, and it is art-military-computer-traffic. That is to say, the user interests has changed with the time elapsing.

V. Conclusion

Personality recommendation system can meet the need of rapid development of network and online service, which can learn user behaviors and interests via analyzing user surfing information, and realizes recommendation initiatively. The key technology in personality recommendation service needs correctly grasp user interests drift rule, and updates the content of recommendation in time. In view of this, a novel user interest drift modeling method is proposed based on HMM and forgetting mechanism. Via experiments, we found this modeling method can work perfectly and efficiently. Only in this way, tracking user interests and changing rule timely, can personality service individuals more perfectly.

Acknowledgment

This work is supported by the Outstanding Talented Person Cultivation foundation of Beijing (No. 2010D005015000001); the NSF of China (No. 61174109); the New Century National Hundred, Thousand and Ten Thousand Talent Project (2010). Here, sincere thanks are given to them.

References

[1] Grabtree, Soltysiak s. Identifying and Tracking Changing Interestss. International Journal of Digital Libraries, 1998, 2:38-53.

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

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

[4] Han, X.G., Shen, Z.Q. Folksonomy-based ontological user interests profile modeling and its application in personalized search. AMT 2010, LNCS 6335,34-46.

[5] Kim, H.R. and Chan, P.K. Learning implicit user interests hierarchy for context in personalization. Applied Intelligence. 2008, 28(2), 153-166.

[6] White, R.W., Bailey, P. Predicting user interestss from contextual information. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009, 363-370.

[7] WANG Shi, GAO Wen. Mining Interests Navigation Patterns Based on Hidden Markov Model. CHINESE J. COMPUTERS 2001,24(2),152-158.

[8] JIN wei, ZHANG ke-Jun. Research on Distributed Web Interests Transfer Pattern Mining. Computer Engineering. 2006, 32(24), 44-48.

[9] ZHANG Ke-Jun, YANG Bing-Ru. Distributed Web interests conversion pattern mining based on localization property.Systems Engineering and Electronics. 2008,30(10),1995-1999.

[10] WANG You-Wei, ZHANG Jian-Bin. A new user interests shift modelling algorithm for websites with hierarchical classification structure. System Engineering theroy and practice. 2008,95-102.