A novel method of user interest drift detection engaging in individual background factors

Chonghuan Xu*

College of business Administration Zhejianggongshang University, Hangzhou, China

Center for Studies of Modern Business, Zhejiang Gongshang University, Hangzhou, China

Received 1 March 2014, www.tsi.lv

Abstract

Personalized service tends to be an emerging challenge in the field of interest mining on e-commerce platform, the issues of which include how to integrate the user's individual background factor, to hiddenly attain portal user interest behaviour, and to mine interest drift pattern. According to user interest drift problem of personalized service in network, this paper explains the user interest through an integration of individual background factor, user behaviour and interest. Meanwhile, it recommends the fuzzy logic thought to explain its impact factor weights comprehensively in order to reflect the level of the user interest on theme. And, it establishes the Hidden semi-Markov Model via user browsing path to detect whether the interest is drifted or not. Finally, the method is proved to be accurate through the experiment analysis.

Keywords: user interest, HSMM, background factors, drift detection

1 Introduction

The quality of personalized service depends on the accuracy of user interest mastered by the system. However, the user's interest will change ranging from time to other surrounding factors. In other words, the user interest drift occurred. The key issue of personalized service is how to identify the changes of user's interest accurately and timely in order to provide the interested service content.

Global scholars have made some research for user interest drift. Grabtree and Soltysiak [1] used the time window approach to solve the problem of user interest drift, whereas it only used one sample to train the user model as the recent visiting. Wu *et al.*[2] proposed a Semi-supervised classification algorithm for data streams of user interesting with concept drifts and Unlabeled data(SUN). Maloof and Michalski [3] used a forgetting mechanism to attenuate the sample. Koychev and Schwab[4] presented a method for dealing with drifting interests by introducing the notion of gradual forgetting. Ahmed *et al.*[5] proposed a novel framework to introduce a very useful measure, called frequency affinity, among the items in a HUP and the concept of interesting HUP with a strong frequency affinity for the fast discovery of more applicable knowledge. Nicoletti *et al.* [6] presented a novel method for topic detection of user interesting from online informal conversations.

This paper will continue to study the drift of the user's interest, not only from the perspective of way to proceed, but also from the perspective of the individual user background. And it would build Hidden Semi-Markov Model to detect whether the user interest drift or not, in light of the user interest combination of background, user behaviour and content.

2 User interest description and mapping

2.1 BASIC DEFINITIONS

Definition 1. The user interest content set *UIC* is the collection of interest content after classification of visited resource about all users in the website: $UIC=\{P_1,...,P_l\}\cup\{L_1,...,L_m\}\cup\{T_1,...,T_n\}=\{UIC_1,UIC_2,...,UIC_M\}$.

Where *P* is a web site component channels; *L* is a hyperlink content; *T* is a tag page; *UIC* is the classification of interest content used by the concept layered approach [7], and it has a corresponding interest concept set: $\Sigma=\{\sigma_x\,|\,1\le x\le Z\}$, $\exists UIC|\rightarrow\sigma_x$, σ_x is a characteristic concept of interest content, $|\rightarrow$ means the mapping relationship from interest content to characteristic concept.

Definition 2. The user background factors set *UBE* is a collection of various background factors existed in individual user *u*, mainly containing Region, Gender, Age, Marriage, Education and Income. It is defined as a user background set: *UBE* = {*Region*, *Gender*, *Age*, *Marriage*, *Education*, *Income*}.

Definition 3. The User interest behaviour set *UIB* is a collection of all possible behaviour operations when *u* visits the *UIC* on the page of website. In this paper, the behaviour data is divided into several types as follows

* ***Corresponding author*** e-mail: talentxch@gmail.com

when users are browsing the web: marking behaviour, such as increasing the bookmark (Book), saving the page (Save), etc.; operational behaviour, such as dragging the scroll bar (Scroll), visiting the page of time (Times), etc.; link behaviour, that is, whether click a hyperlink when you are browsing a page (Click). A set of interest behaviours is defined as follows: UIB = {*Book*, *Save*, *Scroll*, *Times*, *Click*}.

Definition 4. Let the accessing process of user u among the session of time fragment T as a accessing sequential transaction tr, defined as a tuple: {$tr.u$, ($tr.content_1$, $tr.time_1$, $tr.background_1$, $tr.behaviour_1$),..., ($tr.content_p$, $tr.time_p$, $tr.background_p$, $tr.behaviour_p$)}. In which, $tr.u \in U$ denoted accessing user; Four tuples ($tr.content$, $tr.time$, $tr.background$, $tr.behaviour_1$) express as the per accessing operation of user, $tr.content \in UIC$ denotes the detail of interest in content objection, $tr.time$($tr.time_p$-$tr.time_1 \leq T$) denotes accessing timestamp; $tr.background \in UBE$ expresses as the specific background factors of the user; $tr.behaviour \in UIB$ expresses as the interests of specific behaviours of users. Therefore, consisting all accessing transaction tr to accessing transaction set of user in visiting the website by sequential session times: $TR_u = \{tr_i | 1 \leq i \leq |TR_u|\}$, $|TR_u|$ as total number of sessions of the user.

2.2 BACKGROUND

There are various differences among different backgrounds users, and different levels of interest in commodity. In light of the difference of their ages, occupations, backgrounds, interests, they focus on different emphasis on the information systems, and often merely focus on a subset of resources in specific areas.Internet users' interest properties are mainly determined by external factors and internal factors. External determinants include: cultural factors, social factors and family factors, while internal determinants include: life-cycle stages, occupational factors, income, lifestyle, personality factors, self-concept and psychological factors, etc., both these various factors of which will be integrated and have an influence on network behaviour of the users. In this paper, using the geographic, gender, age, marital status, educational background and income which are key impacts on the users' interest as indexes, combining with user behaviour and characteristics of its interest to obfuscate the content of user and to get their degree of interest value.

This paper introduces the idea of fuzzy logic to describe the joint mapping based on the factor weight in backgrounds and behaviour of interest.

Let the user's individual background factors be expressed as B_u=(u. $background$). FB_B= $Relation$(B_u, $UIC \cup UIB$) represents the fuzzy relationship between B_u and $UIC \cup UIB$ on the domain of $B_u \times (UIC \cup UIB)$, where u describes the process of interaction the user to access the background of the individual behaviour of interest. The definition of $W_B(content_k) \in [0,1]$ is a normalization that reflects the individual background FB_B weight.

Let FB_{TR} = Relation (TR_u, $UIC \cup UBE \cup UIB$) said $TR_u \times (UIC \cup UBE \cup UIB)$ domain u_{TR} and $UIC \cup UBE \cup UIB$ relationship between the fuzzy, interactive access to the process u described the behaviour of interest characteristics and evaluate the impact of the definition of W_{TR} (contentk) $\in$ [0,1] for the normalized reflected in the behaviour that FB_{TR} interest in weight.

Let user u interested in browsing the contents of the purchase process of change expressed as navigation path sequences: Su = {$tr_i.seq$} ($tr_i \in TR_u$, $| S_u | = | u_{TR} |$), $tr.seq$ each order record requests, the interest in the content of tr_i. $contentk$ where the ranks of the position. That FB_L = Relation (S_u, $UIC \cup UBE \cup UIB$), said $S_u \times (UIC \cup UBE \cup UIB)$ domain S_u and $UIC \cup UBE \cup UIB$ fuzzy relationship between describing the process of u interested in interactive access to content and degree of concern , the definition of W_L (contentk) $\in$ [0,1] is the normalization of interest that reflect the content of FB_L weight.

Thus, each combination of background, interests behaviour and interest Description in weight of user u can be expressed as W ($content_k$):
$W(content_k)=\theta_1 W_B(content_k)+\theta_2 W_{TR}(content_k)+ \theta_3 W_L(content_k)$ Where $\theta_1+\theta_2+\theta_3=1$ (θ_1, θ_2, $\theta_3 \in [0,1]$).

3 User interest drift mechanism based on HSMM

3.1 HIDDEN SEMI-MARKOV MODEL

A Hidden semi-Markov Model (HSMM) is an extension of HMM by allowing the underlying process to be a semi-Markov chain with a variable duration or sojourn time for each state. Therefore, in addition to the notation defined for the HMM, the duration d of a given state is explicitly defined for the HSMM. State duration is a random variable and assumes an integer value in the set D = {1, 2, . . . , D}. The important difference between HMM and HSMM is that one observation per state is assumed in HMM while in HSMM each state can emit a sequence of observations. The number of observations produced while in state i is determined by the length of time spent in state i, i.e., the duration d.

A parameter of the HSMM [8], can be expressed as a six-tuple: $\lambda = \{N, M, \pi, A, D, B\}$ where N indicates the number of states; M is the number of observations; π={π_i}; A={a_{ij}}; B={$b_j(k)$}; P={$p_i(d)$}. o_t represents observation of the t vector, which includes the first t requests objects and r_t between r_t and r_{t-1} with the time interval τ_t, that is o_t=(r_t,τ_t). Representatives from the first a one to one observation vector b sequence, represents the observation vector sequence. The length Tst represents t time state. ε_t represents the current state of the output will be the number of observations, $1 \leq t \leq T$.

3.2 USER INTEREST BEHAVIOUR

For the user's interest drift, this will create two hidden semi-Markov models. One is used to describe the stable interest and behaviour profile of one or a group of users, the other HSMM is used to outline the behaviours of transferred interests. To get classified in accordance with the path sequence of the user access behaviour, and make sure the observation value set corresponding to every state according to the training data (Unchanged behaviour of user interest) of unchanged user interest behaviour. Take the path sequence to access the web page of users as the basis of the classification of the drift behaviour patterns. Make sure the interest behaviour patterns have the similar path sequence into the same category.

This paper selected the following two observations to describe the user's browsing behaviour:

1) the path sequence of the user access to Web browsing;

2) the time interval between the two adjacent pages. As the user's browsing behaviour is usually from one page to another, therefore, let us assume the user's browsing actions are consistent with the characteristics of Markov chain, and can describe the chain from a state perspective. The set of all states is expressed as $S = \{S_1, S_2, ..., S_N\}$, the corresponding set of observations expressed as $V = \{v_1, v_2, ..., v_M\}$, discrete integer seconds time interval, set as $I = \{1,2, ...\}$. For the typical user's browsing behaviour of a class, the number of its navigation path link is another random variable, which can be considered state of the output in a given number of observations , the set of which is represented as $\{1, ..., D\}$. The sequence of user browsing path is expressed as user browsing Web content objects and user time interval between *rt*-1 and *rt* from one page to another page. *O* is a model of two-dimensional sequence of observations. $B = \{b_i (v, q)\}$ is the output probability matrix model, where as a given state $i \in S$, $b_i (v, q)$ matches the state of the user in a page $r_t = v \in V$ and a page with the previous time interval To $\tau_t = q \in I$ probability. $P = \{p_i(d)\}$ represents the output under a given state *i* the number of observations for the d $\in \{1, ..., D\}$ of probability, and meets $\Sigma d\ p_i(d) = 1$, that is, *P* is the HSMM model the status of the residence time probability matrix. $\pi = \{\pi_i\}$ represents the initial state probability vector, where πi represents the initial state $i \in S$ the probability. $A = \{a_{ij}\}$, represents the state transition probability matrix, where a_{ij} ,is transfered from state $i \in S$ to $j \in S$ the probability. The user's behaviour is an important record of interest which is defined as: $U_{interest} = \{user, timestamp, content, background, behaviour\}$.

3.3 USER INTEREST DRIFT DETECTION

First, the user browsing behaviour data from collection system are used as a sequence of observations, after pretreatment to form a training sequence to train the model. After the model parameters are determined, the model can be used for drift detection. After a pre-measured data is required by observations, by calling HSMM algorithm module, it can calculate the average of the number of contingent probabilities. Then, user interest module in the same sentence will be to get the same interest in the value of user behaviour. If the value of the user's interest in the normal range, the user data will be added to the training data set used in the background update HSMM model parameters, and enter the service queue; Otherwise, the user will be considered to be interested in drift, and to other modules (interested in change processing module) for processing.

Drift detection implementation is shown in Figure 1.

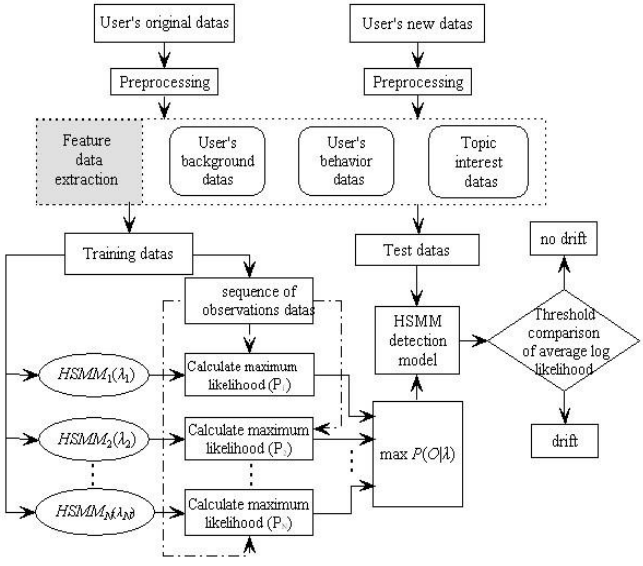

FIGURE 1 Interest drift detection

In this paper, the training data set of all sequences, the average number of contingent probabilities of the mean *lkh* normal behaviour as a reference point. An observation sequence can be pre-defined length threshold *T*0, when the user's browsing path l to T_0, the user can calculate the average number of contingent on the probability of $lkh^{(l)}$. By comparing $lkh^{(l)}$ and *lkh*, you can get the user relative to the model deviation. The smaller the absolute value of difference is, the lower the deviation is, and the smaller the degree of interest drifts is.

4 Performance analysis

The testing data is extracted from a shopping site, and this paper obtains the background factors and the most interesting of the three themes according to the user's registration information and historical data. Background factors mainly contain region, age, marital status, income, and education; while theme is divided into clothing, ornament, beauty, digital, home, motherhood, food, sports and entertainment. The user's background factors and the initial interest theme form are shown in Table 1, where the theme is arranged by the interest weight. In other words, interest topic 1 > interest topic 2 > interest

topic 3, and the empty means the interest does not exist. This paper allows user access to 150 pages, and then regards user interest category as the training set. Through user awareness, human factors analysis, the user browsing the page follows the regular pattern, in a total of nine categories, the main user interest categories of which are "Digital", "food".

We adopt the fuzzy logical thinking to describe user's interest weight jointly by providing a fuzzy processing of the user's background factors, browsing behaviour and the subject content of interest. These specific methods are defined as definition 6 to 8.The interest weight reflects the level of the user interest in some subject as well as tests on the user interest drift if the weight changes. For the online consumers, the paper analyses the nine interest subjects to judge the interest drift. We use the 1st to 15th user visit, and each visit contains 100 samples. Table 2 takes the user1 as an example, the first line of each concept is the observation sequence of change process of user interest, where the number in the table presents the probability of this characteristic and empty presents no appearance. Among the 15 visits of user1, the original interest subjects (clothing, accessories and entertainment) change into (maternal, infant and clothing).

TABLE 1 User's background factors and the initial theme interest form (part)

Number	Gender	Region	Marriage	Income	Age	Education	Interest topic 1	Interest topic 2	Interest topic 3
1	female	eastern	No	8000	27	undergraduate	clothing	digital	amusement
2	male	northeast	Yes	6000	26	master	digital	sports	
3	female	eastern	Yes	7000	27	undergraduate	ornament	hairdressing	food
4	female	eastern	No	5500	33	undergraduate	maternal-infant	household	clothing
5	male	south	No	14000	31	doctor	hairdressing	amusement	
6	female	south	Yes	4500	38	undergraduate	digital	household	sports
7	female	eastern	No	5000	40	undergraduate	household	maternal-infant	sports
8	female	northeast	Yes	14000	46	undergraduate	clothing	hairdressing	food
9	male	southwest	No	14000	31	doctor	clothing	sports	amusement
10	male	eastern	Yes	8500	42	senior	digital	sports	amusement
11	male	south	Yes	15000	35	doctor	clothing	household	
12	male	eastern	No	7500	39	master	clothing	digital	household
13	male	eastern	No	23000	30	doctor	digital	food	sports
14	female	northeast	Yes	9000	28	undergraduate	ornament	clothing	hairdressing
15	female	north	No	15000	25	master	digital	ornament	sports
16	female	eastern	Yes	6500	46	senior	clothing	hairdressing	
17	male	north	No	6000	42	senior	clothing	hairdressing	amusement
18	male	eastern	Yes	4500	29	senior	digital	amusement	food
19	female	northeast	No	6500	27	undergraduate	clothing	ornament	sports
20	female	southwest	No	15000	26	doctor	amusement	clothing	food
21	male	eastern	Yes	4500	38	senior	household	digital	
22	female	south	No	8000	29	master	clothing	amusement	food
23	female	north	No	4000	27	undergraduate	amusement	ornament	amusement
24	female	eastern	Yes	3000	30	senior	sports	clothing	food
25	male	eastern	Yes	12000	34	master	amusement	digital	household
26	female	eastern	No	8000	34	undergraduate	clothing	household	food
27	male	southwest	Yes	15000	35	doctor	digital	maternal-infant	sports
28	female	southwest	No	6000	30	undergraduate	clothing	amusement	
29	male	north	No	13000	45	master	digital	household	sports
30	female	eastern	Yes	5000	36	senior	hairdressing	clothing	maternal-infant
31	male	northeast	Yes	8000	30	undergraduate	clothing	digital	ornament

TABLE 2 User interest sequence and chances of its weight

Interest	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
clothing	**3.44**	**3.43**	**2.65**	**2.21**	**2.32**	**3.11**	**2.76**	**1.94**	**2.02**	**1.88**	**2.36**	**2.58**	**2.61**	**2.29**	**2.39**
ornament	**2.31**	**2.62**	**1.98**	**0.96**	0.83	0.72	0.2								
hairdressing				0.86	**1.32**	**1.56**	**1.63**	0.94	0.32	0.18					
digital															
household									0.2	0.3	0.4	0.4	0.3	0.4	0.4
maternal-infant							0.03	0.04	0.05	**0.98**	**1.32**	**2.63**	**2.91**	**3.67**	**5.41**
food	0.36	0.36	0.52		**1.33**			**1.89**	**1.75**		**1.98**		0.43		0.96
sports					0.53	**0.92**	**1.31**	**1.62**	**1.18**	**1.53**	1.21	**1.67**	**2.28**	**3.02**	**3.52**
amusement	**1.89**	**2.64**	**2.53**	**1.62**	0.98	0.67									

Define the accuracy of shift algorithm: Detection accuracy = correct identification of a particular interest in the number of users / number of a particular user interest. The accuracy of detection is used to measure the fit degree of the adjustment of the user's interest and the actual change in the interest by the algorithm. The higher accuracy indicates that the more algorithms meet the change of user's actual interest. There are three respective algorithm accuracy comparison (HSMM model, sliding windows, and progressive forgetting), and thus the results are shown in Figure 2.

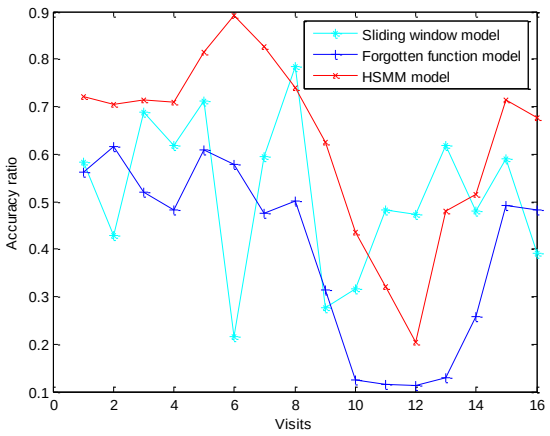

FIGURE 2 Comparison of drift detection methods

5 Conclusions

User interest extraction and user drift detection have great significance on interest mining applications. This paper explains the joint mapping based on the factor weight in interest background and interest behaviour, taking into account of the user's explicit and implicit interests as well as adopts HSMM model to detect whether or not the user's interest is drifted. These methods can express the user's personal interest comprehensively, and improve the accuracy and predictability on the basis of personal interest mining. There requires, however, a further discussion on how to improve the situation of interest drift after the detection.

Acknowledgments

This work was supported in part by Ministry of Education, Humanities and Social Sciences project (Grant No. 13YJCZH216), Natural Science Foundation of Zhejiang Province (Grant No. LQ12G01007),Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20093326120004), Zhejiang Science and Technology Plan Project (No. 2010C33016, 2012R10041-09), and the Key Technology Innovation Team Building Program of Zhejiang Province (No. 2010R50041).

References

[1] Crabtree Barry, Soltysiak Stuart J 1998 Identifying and tracking changing interests *International Journal of Digital Libraries* Springer Verlag 2 38-53
[2] Xindong Wu, Peipei Li, Xuegang Hu 2012 Learning from concept drifting data streams with unlabeled data *Neurocomputing* **92**(1) 145-55
[3] Marcus A Maloof, Ryszard S Michalski 2000 Selecting Examples for Partial Memory Learning *Machine Learning*, **41**(1) 27-52
[4] Ivan Koychev, Ingo Schwab 2000 Adaptation to drifting user's interests *Proceedings of ECML2000 Workshop: Machine Learning in New information Age 2000*
[5] Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Ho-Jin Choi 2011 A framework for mining interesting high utility patterns with a strong frequency affinity *Inform Sciences* **181**(21) 4878-94
[6] Matias Nicoletti, Silvia Schiaffino, Daniela Godoy 2013 Mining interests for user profiling in electronic conversations *Expert Syst Appl* **40**(2) 638-45
[7] Kim Hyoung-rae 2005 Learning implicit user interest hierarchy for Web personalization: Florida Institute of Technology
[8] Yu Shun-Zheng 2010 Hidden semi-Markov models *Artif Intell* **174**(2) 215-43

Authors

Chonghuan Xu, born on November 14, 1983, Hangzhou, China

Current position, grades: lecturer, Zhejiang Gongshang University.
University studies: B.S. and M.S. degrees in Computer and Information Engineering from Zhejiang Gongshang University, Hangzhou.
Scientific interest: Operations research, electronic commerce, data mining.
Publications: 15.