Microblog oriented user interest analysis and implementation in personalized information service

Lifang Tang

School of Business Administration, Zhejiang Gongshang University, Hangzhou, Zhejiang, China, 310018

School of Media and Design, Hangzhou Dianzi University, Hangzhou, Zhejiang, China, 310018

Corresponding author's e-mail: tlf@hdu.edu.cn

Received 1 December 2013, www.cmnt.lv

Abstract

This paper analyzes and studies the user interest in microblog data and the methods for personalized recommendation. It designs a microblog oriented personalized modeling system and explains the overall structure from a macro point of view. The personalized modeling system for microblog users includes two major parts: interest extractor and personalized model generator. In interest extraction and classification, we use a combined classifier by naive Bayes and support vector machine, to filter the microblogs unrelated to users' interest. At the phase of personalized model generation, we propose the indicators of long-term and short-term interest by analysis on rules of topic distribution. Two different updating mechanisms are adopted to meet users' demand for update quality of long-term and short-term interest. The experiments implement our research on a specific information push services system of user blog. Then the relative evaluation indicators in information retrieval verify the correctness and feasibility of the improved algorithm.

Keywords: microblog, user interest, personalized system, cluster, combined classifier

1 Introduction

In the past decades of years, the information on internet increases rapidly and people entered information overload period from information deficiency. Then, the transformation for people to obtain information method follows the mode from traditional artificial search to search engine, then, to current recommendation system. To recommend the usable information to users, the most important step is to acquire users' interest effectively. The emergence of social network such as microblog offers us a newly huge data source to analyze users' interest which has become the research focus. Microblog not only contains lots of users' generating text resources but it also has social network features [1]. It allows one user to concern other users to establish a social network. Due to these advantages of microblog, they attract lots of researchers to analyze and study microblog resources from various aspects, including abundant file resources, social network resources, attempting to extract useful information. However, most of these researches focus on the structure of social network and the content analysis of microblog. Few people study on the systematic model for modeling on the subjects that are interested by users.

Abel, etc al [2] propose to use microblog as information source which is used to analyze real-time interest of users. The authors define two kinds of keywords: content indicative words and action indicative words. In this way, the real-time interest of users can be expressed as the pair of this word. ZHENG, etc al [3] propose that users' interest can be mined from multiple data resources such as Twitter, Facebook, Linkedin. Useful interest is obtained from each data source at first and users' interest is expressed by several keywords. Then, by the information fusing strategy, a relatively complete users' interest set is generated. Zhao, etc al [4] compare two different kinds of methods to use words bag model to construct users' interest. Then they find that the effect of users to construct users' interest words bag, according to their own published microblogs, is better than uses' interest words bag which is generated corresponding to fans' microblogs. Duan, etc al [5] use the labels in microblog to construct users' interest. They extract several keywords from microblog content to be taken as user's personalized label to describe each user interest. The authors use two unsupervised keywords extraction methods, that is, TF-IDF and TextRank. They also take all the microblogs published by each user as a large document and use standard model to find potentially interesting subjects of each user.

Based on the above improvements, this paper designs and realizes a microblog oriented user interest personalized modeling system. It generally designs the purpose, overall framework, overall framework, system flow and the division between modules and functions. In addition, it also unfolds personalized modeling system of microblog users from macro prospective, which mainly includes the interest extractor and personalized model generator. For the design of interest extractor, the text classification technology is adopted. SVM algorithm and Naive Bayes algorithm are used for classification. Then, the classification results are combined to improve the classification accuracy and overall performance. In generating process of personalized model, the distribution law of users' interest subjects is analyzed and measuring indicators between long-term and short-term interest are put forward. Meanwhile, when two different kinds of updating mechanisms are adopted to guarantee the quality of long-term interest, the real-time updating requirement of short-term interest can be satisfied. Finally, theoretical research achievements are applied to a concrete information-pushing service system and the overall testing scheme in the system is provided. In addition, the evaluation indictors in information retrieval are used to verify accuracy and feasibility of the proposed algorithm.

2 Personalized model design of microblog platform

2.1 OVERALL FRAMEWORK

Microblog oriented personalized model is based on personalized information of microblog users. It is used to apply personalized model technology to each microblog user, to establish a personalized model reflecting user's interest orientation and interest changing tendency. The purpose of microblog personalized modeling system [6] contains:

1) It has function to collect microblog users' information automatically. Because the homogeneity of different microblog network technology is strong and the webpage structure is similar, the data can be guaranteed extreme stability and unity when obtaining microblog data. Meanwhile, the completeness and accuracy of microblog personalized modeling data can also be ensured.

2) It can analyze user information and extract the interest type and changing tendency from user's information. The published articles and users' own classification are usually the key reflections of user interest types, while different texts published time also reflects users' interest changes tendency.

3) It provides effective and convenient model access interface for personalized application system of blog.

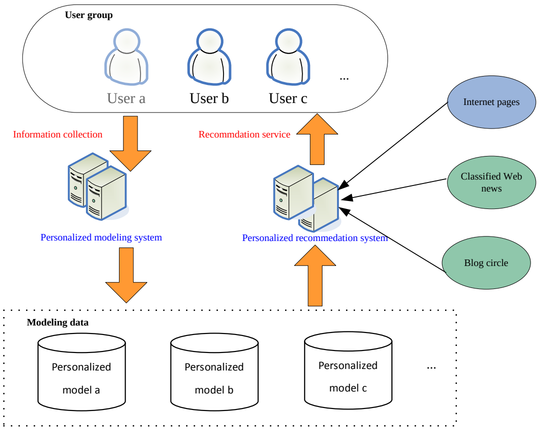

FIGURE 1 Overall architecture of system

For the overall application framework of the system, the personalized service system of blog can be divided into personalized modeling sub-system and personalized recommendation service sub-system of blog users. The function of personalized modeling sub-system aims to grasp blog webpage, analyze, mine blog users' personalized information and establish personalized interest model of blog users according to these information, including users' long-term interest model and short-term interest model. The function of personalized recommendation sub-system is based on users' personalized interest model to recommend blog users' interest resources actively, such as news information, blog site, advertisement, product, etc. Figure 1 shows the overall framework of microblog oriented personalized service system

Therefore, the user model is the main knowledge source of personalized recommendation generated by recommendation service sub-system. The accuracy of user interest and user authenticity reflected by users modeling results directly determines the personalized recommendation service quality, so user model is the core of the whole system.

2.2 MODULE PARTITION

The personalized modeling system of blog user divides the system into three sub systems according to the overall function target. They are blog information collector, interest extractor and personalized model generator, as is shown in Figure 2. Blog information collector is used to capture blog webpage from blog website and analyze the obtained blog information including blog username, the first page URL of users' blog, the name of blog user-defined classification, self-defined classification URL of blog users, blog article information, etc. They are all saved in the database.

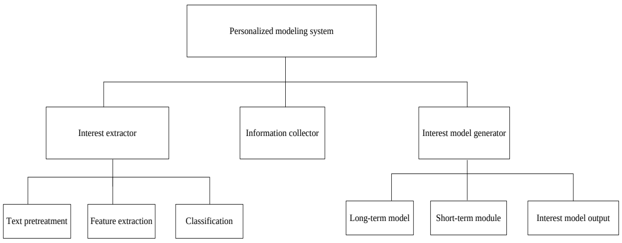

FIGURE 2 System modules of personalized microblog model

Interest extractor can calculate interest category of article and the probability belonging to corresponding interest category based on input articles. For the classification of blog users' articles, the output information of article categorization reflects users' interest type and interest degree.

Personalized model generator comprehensively adopts users' articles and published time of articles. It calculates the user's interest orientation and interest change tendency and output blog user's interest model with a general format including long-term interest model and short-term interest model. The long-term interest model is obtained by weight of all blog articles after text classification, while short-term interest model is obtained by weight of recent articles after text classification.

3 Improvement of interest extraction classification algorithm

The principle idea is that many classifiers anticipate in prediction and some strategies including one-vote negation system, majority voting system and the obedience of minority, and so on, will determine the final results. It can also adopt Bayesian vote and assign the weight based on classification performance of each basic classifier and vote according to obtained weight [7, 8]. Considering the two classifiers in this paper, with different eigenvalues, this paper adopts linear integration. The linear integration means that each classifier output multiplies corresponding weight and accumulates results as final classification result.

We use m_i to denote a piece of microblog. m_i is divided into two class "related to interest topic" and "unrelated to interest topic", marked as class c_1 and c_2. For naive Bayes classifier, the probability of m_i belonging to c_1 is $P_b(c_1 | m_i)$, the probability belonging to c_2 is $P_b(c_2 | m_i)$. For SVM classifier, we set its class to m_i as G_k, $G_k = c_1$ or $G_k = c_2$. So we get a combined classifier by the linear integration of two classifiers, whose probability to separate m_i into class c_i is

$$P(c_i | m_i) = \alpha + w_1 \cdot P_b(c_1 | m_i) + w_2 \cdot \begin{cases} 0 \; if \; G_k \neq c_i \\ 1 \; if \; G_k = c_i \end{cases} \tag{1}$$

w_1 and w_2 denote the weight of naive Bayes and SVM classifier. Let $X = (1, P_b, P_{SVM})$, $\hat{\beta} = (\alpha, w_1, w_2)^T$, the above equation is equivalent to

$$P(c_i | m_i) = X\hat{\beta} \tag{2}$$

Give n group of observation value

$$X = \begin{pmatrix} 1 & P_b^1 & P_{SVM}^1 \\ 1 & P_b^2 & P_{SVM}^2 \\ \dots & \dots & \dots \\ 1 & P_b^n & P_{SVM}^n \end{pmatrix}$$

and $Y = (y_1, y_2, ..., y_n)^T$. We get the least squares estimation as $\hat{\beta} = (X^T X)^{-1} X^T y$.

4 Personalized model generator

After the interest extraction of microblogs, the results are the weights on each interest class of blog users. These personalized interest information are used to generate the personalized interest model of users. They can be simply divided into two classes: long-term and short-term interest [9]. The long-term is stable relatively, while the short-term has big fluctuations. Therefore we the personalized model generator adopts long-term interest model and short-term interest model respectively to maintain the interest of users. The former reflects users' interest for a long time, and it is computed by text categorization of all the papers published by blog users; the short-term interest records the interest within recent concern by users, and it is computed by text categorization of recent papers published by blog users. Figure 3 depicts the whole working process of personalized generator, from which we can clearly see the data relationship between data direction in model generator and interest extractor.

To well capture the law of user interest and describe them, this paper adopts tracking window [10] and the browsing content in it for expressions. Since the amount of browsing records of users is very large and the price for update is also huge, it is infeasible in actual implementation. Thus, no matter long-term interest or short-term interest, we need to define the tracking windows: we set n documents as tracking windows, generated by the nearest browse and recorded as $\beta_1, \beta_2, ..., \beta_n$. The documents are ordered as the sequence of time from front to back. The document is newer when it is closer to the right of window. With the increase of browsed documents the windows with fixed length will move to right along.

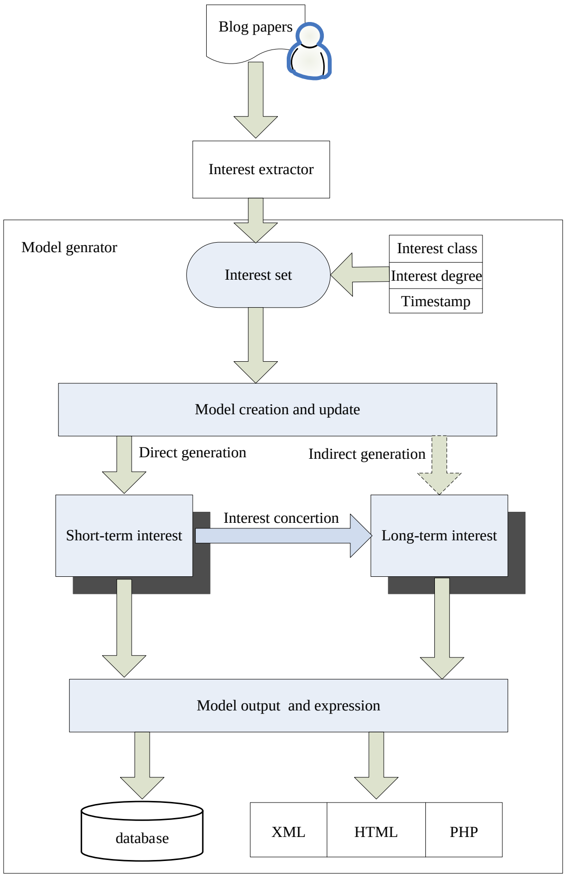

FIGURE 3 Flowchart of personalized model generator

The development of long-term interest needs long visit and it is stable relatively after establishment. So we need longer tracking window and filtering the documents before analysis. The documents with discrete distribution and regular rules can represent the long-term interest of user. The short-term interest is short-lived and it represents the burst characteristics whose distribution is centralized and erratic. To capture two classes of interest distribution rules, we first adopt the improved clustering algorithm for the requirement of user interest analysis. A threshold α is set to denote the same cluster when the similarity of two documents is bigger than this threshold. The centroid of cluster is computed as:

$$c = \frac{1}{n} \sum_{i=1}^{n} \frac{c_1}{\| c_i \|^2} . \tag{3}$$

If the set of document cluster for browsing behaviors is S, the document β_1 which is far from the time in distance is taken as the centroid c_{O1} of the first cluster, $S = \{c_{O1}\}$. The procedures in detail:

Input: Browsing records of tracking windows
Output: Clusters set S

Step 1: Select c_i ($1<i<n$) by time order. For each $c_{Oej} \in S$, compute the similarity $s_{ij} = sim(c_i, c_{Oej})$. Then the maximum is obtained as $s_{\max} = \max(s_{ij})$;
Step 2: When $s_{\max} > \alpha$, separate c_i into c_{Oej} and recalculate the centroid of c_{Oej};
Step 3: When $s_{\max} < \alpha$, we add c_i to the new cluster $c_{Oe(m+1)}$;

Then all the documents of tracking window are process by above procedures until new document is generated at the tail by window sliding. The new document is categorized to existing or new clusters. The oldest document is removed and the centroid of its cluster will be recalculated.

The webpage cluster reflecting short-term interest of user is recent and continuous, which is directly related to the visiting time of webpages. So we introduce the idea of interest freshness [11], combined with visiting time factors, to capture the behavior features. The interest freshness is normalized value of the mean value of visiting time for effective clusters, recorded as δ.

The mean value of visiting time is calculated as:

$$M(t) = \frac{1}{n}\sum_{i=1}^{n} t_i \,. \tag{4}$$

If there are x effective clusters, the mean value of visiting time of these clusters will construct a vector $V = \{M_1(t), M_2(t), ..., M_x(t)\}$. Their freshness can be calculated as:

$$V_f = \frac{M(t)}{\|V\|^2} = \frac{1}{n\|V\|^2}\sum_{i=1}^{n} t_i \,. \tag{5}$$

The reason to update user's interest in such way is that the long-term is usually stable and continuous and a few browsing can not cause influence to it. Therefore, the setting of update period under this condition can avoid unnecessary update to save the resources. The update model can be finished based on expression of user interest or semantic extension process; for short-term behaviors, they may be influenced by fewer new visit, so we need real-time process to avoid ineffective recommendations. We assume that the new document β belongs to the cluster $\beta_{OE(m+1)}$ after clustering. To determine whether it is a short-term interest we get:

$$\theta = \frac{V_f}{V_d} \,. \tag{6}$$

When θ is bigger than the threshold, we perform incremental update algorithm based on term weight to the short-term interest, which is described as follows:

Input: New document β
Output: Updated short-term vectors

Step 1: Express document β with improved vector space and scan all the terms f_i and weights w_i;
Step 2: If there is not any short-term interest model of f_i, we add its feature and weight $< f_i, w_i >$ to it directly;
Step 3: If f_i exists, we will merge the weights of original weight w_0 after attenuation by time attenuation function $\tau = e^{-\chi(t-t_0)}$;

χ is the attenuation factor of feature with time going and it is decided by original and new weight of f_i. If $w_i > w_0$, the attenuation is slow; otherwise, the attenuation is fast. The idea lies in that when the topic attention degree of short-term interest is increasing continuously, the feature is believed to has slow attenuation; otherwise, when the attention degree is low, it is believed to deviate the attention on this topic and we should accelerate its attenuation.

5 Experiments and results analysis

5.1 FILTERING EFFECT OF COMBINED CLASSIFIER

We capture 6000 microblog articlets from Sina microblog and these microblogs are from 100 different users. For each microblog, if this microblog is forwarding microblog, its content will be added to user's evaluation as the actual content of this microblog. We pretreat each microblog and word segmentation. Meanwhile, the microblog which is less than 5 words after segmentation will be deleted. If one microblog only contains foreign language, this microblog will be also deleted. After pretreatment, the rest of effective microblog will be 4580. Then, we randomly set 4000 of them as training set and another 580 microblog will be taken as test set. Then, these microblogs are artificially marked into two classifications. They are the microblog which can reflect user interest and cannot reflect user interest. After manual annotation, all effective microblogs and the microblog amount which respectively belongs to the first class and the second class in training set and testing set are shown as Table 1. Then, the methods mentioned above are respectively used. Naive Baysian classifier and SVM classifier are trained on training set and their composed weight of combined classifier can be learned. Then, these classifiers are used to be tested on test data.

TABLE 1 The amount of training set and test in each class of microblog

	Amount of microblogs	Amount of microblogs with users' interest	Amount of microblogs without users' interest
All effective microbe	4688	1765	2919
Training set	4001	1554	2465
Test set	680	220	461

Figure 4 is the test set curves on naive Bayesian classifier and combined classifier in training set. Since the output of SVM classifier is not probability, but -1 and 1, we only get one point. From this figure it can be seen that after features under social relation net classification are removed, classifier curve does not basically change. However, after removing features under any three classifications, the obtained classifier effect will decrease. This indicates the adopted features belonging to social relation net classification. For instance, the times of this microblog which are forwarded by friends and the times of this microblog which are commented by friends are not closely related to recognize whether a microblog can reflect the user interest. This can be also explained that whether a microblog is related to user interest or not, it may be forwarded by friends.

5.2 UPDATE PERFORMANCE OF INTEREST MODEL

In order to verify the feasibility of mixed user interest model and updating learning technology of model which have more superior performance, we compare the updating learning technology of several mainstream models and updating learning technology of mixed model in this paper. We adopt progressive forgetting method, Rocchio method and time window sliding method [12] to be compared with the improved method in this paper. We totally implement eight experiments and these results are in statistically treatment, to return the precision, recall and values of these four kinds of algorithms. The results are shown as Figures 5 and 6.

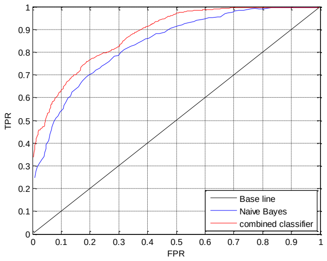

FIGURE 4 ROC curve of combined classifier on test data set

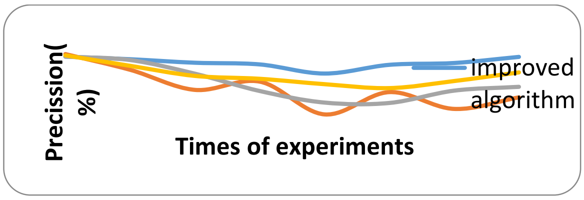

FIGURE 5 Precision statistics comparison of 4 interest models

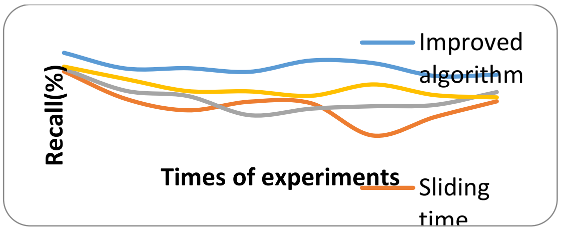

FIGURE 6 Recall statistics comparison of 4 interest models

Based on above results, we can summarize the updating learning technology features of these interest models. The proposed learning method of mixed users' interest model is better than the other threes on precision and recall. Thus, the improved method in this paper is superior to the other three methods on overall performance. Meanwhile, this paper has analyzed the respective features of long-term interest and short-term interest and adopts different updating learning methods according to its features. For short-term interest model it adopts time window sliding while long-term interest method adopts combination between Rocchio and learning method of progressive forgotten method, to effectively and accurately update the interest model.This experiment verifies the feasibility of updating interest model of mixed user and learning algorithm in this paper, and it shows better performance in comparison with other updating learning technologies.

6 Conclusion

For the microblog characteristics, this paper puts forward a method to analyze the users interest and modeling from microblog data. Then it applies mined users interest to recommend personalized information to users. Our job is mainly divided into three parts: At first, microblog oriented framework of personalized service system is put forward. Then, the noise filtering method in microblog data is improved and the microblog data which is unrelated to users' interest are filtered. The weight of users' interest features is determined by tracking subject distribution of viewed content in window. This effectively improves the accuracy of user interest model, and the updating mechanism of two kinds of interests. Experimental results show the improved algorithm can effectively mine users' long-term and short-term user's browsing behavior features in comparison to traditional keywords notation model. Meanwhile, user interest analysis and the information source content waiting to be recommended can use the same topic model for analysis, so it has better recommendation effect.

References

[1] Sun Q, Chen S 2013 New Media User Attention Index Model Simulation Analysis *Computer Simulation* **31**(1) 428-32

[2] Abel F, Gao Q, Houben G-J 2013 Twitter-based user modeling for news recommendations *Proceedings of International Joint Conference on Artificial Intelligence* 2962-6

[3] Zheng F, Miao D, Zhang Z 2012 News Topic Detection Approach on Chinese Microblog *Computer Science* **39**(1) 138-41

[4] Zhao X, Ma Z, Ding L 2012 A novel interesting recommendation system based on tags mining *Proceeding of International Conference on Information Management, Innovation Management and Industrial Engineering* 493-7

[5] Duan L, Guo W, Zhu X 2013 Constructing spatio-temporal topic model for microblog topic retrieving *Geomatics and Information Science of Wuhan University* **39**(2) 210-3,243

[6] Gao M, Jin C Q, Qian W N 2013 Real-Time and Personalized Recommendation on Microblogging Systems *Chinese Journal of Computers* **37**(4) 963-74

[7] Wen K, Xu S, Li R 2012 Survey of Microblog and Chinese Microblog Information Processing *Journal of Chinese Information Processing* **26**(8) 27-37

[8] Wang S,Wang Z, Zhang M 2012 Personalized Recommendation Algorithm on Microblogs *Journal of Frontiers of Computer Science & Technology* **6**(10) 895-902

[9] Ratkiewicz J, Conover M, Meiss M 2011 Truthy: Mapping the spread

of astro turf in microblog streams *Proceedings of International Conference Companion on World Wide Web* 249-52

[10] Zhang HP, Zhang RQ, Zhao YP 2013 Big data modeling and analysis of microblog ecosystem *International Journal of Automation and Computing*, **11**(2) 119-27

[11] Cetintas S, Si L A, Hans B K 2011 *IEEE Transactions on Learning Technologies* **4**(4) 292-300

[12] Ge H, He Y, Chen Q 2013 Micro Blogging Users Classification Method Based on the Time Slice *Journal of Chinese Computer Systems* **34**(11) 2441-5

Author

Lifang Tang, 1981.3, Yongzhou, Hunan, P.R. China

Current position, grades: lecturer of School of Media and Design, Hangzhou Dianzi University, China.
University studies: M.S. in 2006, currently is a PhD candidate of Zhejiang Gongshang University.
Scientific interest: research on internet user behaviour.
Publications: more than 5 papers.
Experience: 8 years teaching experience, 3 scientific research projects.