Model Bloggers' Interests Based on Forgetting Mechanism

Yuan Cheng[1], Guang Qiu[2], Jiajun Bu*[2], Kangmiao Liu[2], Ye Han[3], Can Wang[2], Chun Chen[2]
[1]College of Software Technology [2,3]College of Computer Science
Zhejiang University
Hangzhou 310027, China
*Corresponding Author, +86 571 87952148
[1]star_926@msn.com [2]{qiuguang, bjj, lkm, wcan, chenc}@zju.edu.cn [3]hanye.zju@gmail.com

ABSTRACT
Blogs have been expanded at an incredible speed in recent years. Plentiful personal information makes blogs a popular way mining user profiles. In this paper, we propose a novel bloggers' interests modeling approach based on forgetting mechanism. A new forgetting function is introduced to track interest drift. Based on that, the Short Term Interest Models (STIM) and Long Term Interest Models (LTIM) are constructed to describe bloggers' short-term and long-term interests. The experiments show that both models can identify bloggers' preferences well respectively.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, Retrieval models.*

General Terms
Algorithms, Design, Experimentation, Human Factors.

Keywords
Blog, short term interests, long term interests, interest forgetting

1. INTRODUCTION
Blog, considered as an online diary publishing platform, has been in favor with numerous web users. People are more willing to share feelings and interests with friends through this interactive web media. Flourish of blog provides researchers with an ideal channel to access people's personal information especially their interests which contain credible value for business use.

Four Characteristics make blogs unique from other web applications [1]. (1) Highly personal and rapidly evolving content. (2) All posts lists in chronological order. (3) Highly related to users' interests. (4) Linking bloggers form a community. So far, most researches focused only on one of these features. [1] proposed a combined classifier (Naïve Bayes, SVM and Rocchio) to identify bloggers' interests from single posts. [2] analyzed how blog communities can be discovered through interconnected blogs as a form of social hypertext. However, few covered multiple blog features. [3] discussed the feasibility to detect interests using textual, temporal, and interactive features but it did not give out a practical modeling method.

In this paper, we model bloggers' interests based on forgetting mechanism. Our method combines textual and temporal features, which not only well identifies users' interests from posts but also threads them in chronological order using a forgetting function. In addition, our approach follows two general ideas of psychologists: (1) Similar to memory, human interests wane as time goes by (2) Forgetting speed slows down gradually and the accumulated interests become more stable. Based on the two principles above, we construct two interest models for different purposes. STIM: bloggers' recent interests with weak stability; LTIM: accumulated interest status after a long period with strong stability.

WWW 2008, April 21-25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. MAIN APPROACH

2.1 Model Representation
In our approach, interest models are represented by vectors V= $\{(c_1, w_1, t_j), (c_2, w_2, t_j),...,(c_m, w_m, t_j)\}$. We predefine m interest categories $\{c_1, c_2, c_,...,c_m\}$ and each category c_i is associated with a weight w_i according to its cohesion with users' interests. The term t_j denotes the timestamp when the model is constructed.

Post vectors are the foundation of our blogger modeling work. Each category c_i in the vector is assigned with a weight w_i which denotes the probability this post belongs to c_i. The probability is given through the process of text classification (we use Naive Bayes classification model in our work).

2.2 Interest Forgetting Function
As an important part of our approach, the forgetting function is implemented to simulate the attenuation of users' interests as shown in the following equation:

$$F(t) = e^{-\frac{\ln 2 \times (t - est)}{hl}} \qquad (1)$$

where forgetting coefficient $F(t)$ means which percent the original interests have declined to; t means the current date; *est* means the date when the original model was established; *hl* denotes the half-life (in days) controlling speed of forgetting. The larger *hl* is the slower interests fall. When t-*est*=*hl*, $F(t)$ falls to 1/2.

2.3 Short Term Interest Modeling
To construct a STIM on a specific day, we firstly maintain a time window of size S (the number of days) over the most recent posts that are used for modeling. To determine the window size, we define four constants: *MinDayLimit, MaxDayLimit, MinArticleNum* and *MaxArticleNum.* Then we adopt the following three strategies to determine the size of window. (1) If *PAN* (posted articles number) >= *MaxArticleNum* within the past *MinDayLimit* days, *S*=*MinDayLimit.* (2) If *PAN* < *MaxArticleNum* within the past *MinDayLimit* days but *PAN* >= *MinArticleNum* within the past *MaxDayLimit* days, *S*=*MaxDayLimit.* (3) If *PAN* < *MinArticleNum* within the past *MaxDayLimit* days, this instance can not be considered as a valid timestamp to construct a STIM as there are too few articles within the time window to capture the interests of users.

Suppose we get N posts from the time window of size S by T_s, we convert these post vectors from V= $\{(c_1, w_1, t_j), (c_2, w_2, t_j),...,(c_m, w_m, t_j)\}$ to $\{(c_1, w_1, f_j), (c_2, w_2, f_j),...,(c_m, w_m, f_j)\}$. The forgetting

coefficient f_j is calculated through function (1) where $cur=T_s$ and $est=t_j$. We set hl to be a small value leading to fast forgetting. Weights of categories c_i in the STIM are calculated as follows:

$$W_i^s = \sum_{j=1}^{n} w_i(j) f_j \qquad (2)$$

Where W_i^s means the weight of c_i in the STIM; $w_i(j)$ means the weight of c_i in the vector of the j[th] article. At last, we normalized W_i^s to the interval [0, 10] using the following method:

$$W_i^{s\prime} = \frac{W_i^s}{\sum_{c_i \in C} W_i^s} \times 10 \qquad (3)$$

The final format of STIM: $V=\{(c_1,W_1^{s\prime},T_s)(c_2,W_2^{s\prime},T_s)...(c_n,W_m^{s\prime},T_s)\}$

2.4 Long Term Interest Modeling

Different from STIM, the LTIM describes users' interest after a long period. Thus we adopt all their posts as modeling foundation. To simulate the interest accumulation process, whenever the blogger publishes a post, the system updates the original LTIM by integrating the new article's interest vector. The main formula for long term interest modeling is described as below:

$$V_{nl} = V_{ol} F_d \times \frac{N-1}{N} + V_p \times \frac{1}{N}, \qquad (4)$$

where V_{nl} denotes the updated LTIM; V_{ol} denotes the old LTIM established when last article was posted; V_p is the vector of new article. $(N\text{-}1)/N$ and $1/N$ represent the weight of $V_{ol}F_d$ and V_P. The initial V_{ol} is the vector of the first posted article of blogs. Here, half-life for LTIM is not a constant any more. As is described in the introduction section, bloggers' interests are inclined to be more stable as time goes by. Based on that, we use the following forgetting formula to calculate coefficient F_d:

$$F(t) = e^{-\frac{\ln 2 \times (t-est)}{hl_0 + d_{acc} \times s}} \qquad (5)$$

Where hl_0 represents an initial half-life value; a_{acc} denotes how many days the original LTIM has evolved. Constant s reflects the impact of d_{acc} on forgetting speed. By involving factor $d_{acc} \times s$, bloggers' interests fall more slowly than before. The final format of STIM is $V=\{(c_1,W_1^{l\prime},T_l)(c_2,W_2^{l\prime},T_l),...,(c_n,W_m^{l\prime},T_l)\}$

3. EXPERIMENTS AND RESULTS

We use NetEase Blog Directory (blog.163.com) to train our classifier. There are fifteen first-level categories in total: *Education*, *Sport*, *Game*, *Multimedia*, *Modernlife*, *Star*, *Tour*, *Entertainment*, *Digital*, *Healthcare*, *Workandfinacing*, *Technology*, *Society*, *Family* and *Art*. For each category, we collect 5000 articles as training sets. In the modeling phase, 100 popular blogs and 31409 contained posts are selected as test data.

Currently, there exists no recognized method to evaluate the performance of interest modeling approaches. In our experiments, we adopt the precision of interest prediction (*IPP*) to evaluate the relativity between our modeling results and users' interests. For a given interest model, we extract the first three dominant categories *{C_1, C_2, C_3} order by corresponding weights* as prediction candidates. Then we classify the succeeding N posts and find out the interest category C_p with highest weight. If C_p belongs to *{C_1, C_2, C_3}*, we think interests reflected by this article has been correctly predicted. Avg(*IPP*) below shows the average precision of *IPP*s for 100 selected blogs.

$$Avg(IPP) = \frac{\sum_{i=1}^{100} (n_i / N)}{100} \qquad (6)$$

Where n_i denotes the number of correctly predicated article in the i[th] blog. To evaluate the performance of our approach, we set up a comparison group based on a general modeling method without forgetting mechanism, in which LTIMs are simply constructed by summing all posts' vectors. We adopt both of them to construct LTIMs using the first 75% articles of bloggers. The rest 25% are kept for interest prediction. Here we set hl_0=10, s=0.5 for LTIM. Through experiments, Avg(*IPP*) of the general modeling is 78.3%; Avg(*IPP*) of our approach is 82.0%. We can see that our approach has evidently higher Avg(*IPP*) than that of the general modeling. It proves interest accumulation is not only a simple summing process but also involving forgetting influence. We also notice that the difference between two Avg(*IPP*)s is not that far as we expected. The reason is that some categories of bloggers' interests are so dominant that both approaches get the same candidate set.

In the second part of the experiments, we compare the predication characteristics of the STIM and LTIM. We select 10 representative bloggers who post articles regularly and construct two models on the date when their 200[th] article was posted. Both models are used to predict the interests of the succeeding N posts (N=10, 20…100). For short term interest model, we set hl=20, *MinDayLimit*=20, *MaxDayLimit*=30, *MinArticleNum*=10, *MaxArticleNum*=5. Figure 1 shows the predication results.

According to the figure, we find that STIM performs better when the predicated articles number N is relatively small. It manifests that STIM is more closed to users' recent interest status. However, with increase of the article number, LTIM's predication ability goes up gradually and finally transcends STIM's. The prediction traits of STIM and LTIM revealed from Figure 1 successfully validate our original intention to set up these two models.

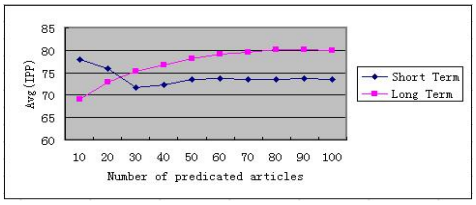

Figure 1. Avg (*IPP*) of STIM and LTIM for Different N

4. CONCLUSIONS AND FUTURE WORK

This paper proposes a comprehensive modeling approach to detect bloggers' profiles. Forgetting mechanism is used aiming at dynamically simulating interest drift. Both STIMs and LTIMs are proved to be good identifiers of bloggers' interests respectively. We are currently extending our work by integrating more features to identify bloggers' interests such as comments, blogger communities. In addition, we also try to propose a method to determine personalized half-life hl for every blogger which reflects people's distinct interest forgetting characteristics.

5. REFERENCES

[1] Xiaochuan Ni, et al. "Automatic Identification of Chinese Weblogger Interests Based on Text Classification". In *WI' 2006.*

[2] Alvin Chin, et al. "A Social Hypertext Model for Finding Community in Blogs". In *HT'06*.

[3] Chun-Yuan Teng, et al. "Detection of Bloggers' Interests: Using Textual, Temporal, and Interactive Features". In *WI'2006*