User Dynamic Tag Cloud Construction in Social Tagging: A Postprint

Authors: Xie Mengyao, Pan Xuwei

Date: 2017-11-08T00:00:00+00:00

Abstract

【Purpose】Tag clouds can be utilized for information retrieval, recommendation, and navigation. Since user tagging exhibits temporal characteristics, we propose a method for constructing user dynamic tag clouds based on temporal evolution to effectively reveal the dynamic changes in user interests.

【Method】By leveraging the forgetting and reinforcement characteristics of memory from psychology to construct dynamic weights for tags, we establish user dynamic tag clouds that reflect changes in user focus.

【Results】Compared with existing tag cloud algorithms, the proposed user dynamic tag cloud algorithm can effectively rank tags according to dynamically changing user interests, demonstrates significantly superior performance in predicting user interest tags, and achieves higher recommendation accuracy.

【Limitations】Since user interests do not change substantially within short time periods, the predictive effectiveness of the dynamic method is not very pronounced in short-term cycles, but becomes more significant over longer time periods.

【Conclusion】The user dynamic tag cloud based on temporal evolution can effectively capture users' current interest hotspots and improve the effectiveness of personalized retrieval and navigation.

Full Text

Constructing Dynamic Social Tag Cloud for User Interests

Xie Mengyao, Pan Xuwei

(School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract

[**Objective**] Social tags can be used for information retrieval, recommendation, and navigation. Since user tagging exhibits temporal characteristics, this paper proposes a method for constructing dynamic user tag clouds based on temporal evolution to effectively reveal changes in user interests. [**Methods**] We constructed dynamic weights for tags by leveraging the forgetting and strengthening characteristics of memory from psychology, thereby establishing dynamic user tag clouds that reflect shifts in user focus. [**Results**] Compared with existing tag cloud algorithms, the proposed dynamic tag cloud algorithm can effectively rank tags according to dynamically changing user interests, demonstrating significantly higher predictive performance for user interest tags and achieving greater recommendation accuracy. [**Limitations**] Because user interests do not change substantially within short time periods, the dynamic method' s predictive effect is not very significant over short cycles but becomes more pronounced over longer periods. [**Conclusions**] The temporal-evolution-based dynamic user tag cloud can effectively capture users' current interest hotspots and improve the effectiveness of personalized retrieval and navigation.

Keywords: Social Tagging; Tag; User Interests; Dynamic Tag Cloud

Introduction

In recent years, Web 2.0 has brought substantive changes to the Internet, transforming users from passive recipients of online information into active creators. As a core component of Web 2.0, social tagging allows mass users to unconstrainedly annotate resources of interest based on their own understanding, with all users' annotations being mutually visible [1]. Social tagging sites centered on mass user participation, represented by Flickr and Delicious, have gradually grown and developed, becoming a new channel for users to effectively access information resources.

Tags, as the carrier of social tagging systems, have become an important information organization tool in the Web 2.0 era. Social tags contain rich information, and tag clouds—collections of tags with visual weight—have effectively solved the visualization problem of tag information, helping users quickly obtain valuable information from large numbers of tags. Current research on tag clouds mainly focuses on sorting algorithms [2], personalized recommendation [3], and visual layout [4-5], which has strongly promoted the development of tag cloud applications and theory. As an emerging retrieval and recommendation technology, tag clouds can reflect the importance of different tags through visual attributes, guide user browsing, and attract user attention to specific fields or regions. Millen et al. [6] studied users' query and browsing habits in tag cloud environments and found that social tags are an important way to improve social navigation. Hassan-Montero et al. [5] calculated tag usefulness by defining indicators such as the degree to which tags describe resources and the number of resources covered, and improved browsing experience through clustering algorithms. Additionally, as a navigation interface for social information, tag clouds can provide personalized search recommendations for different users after visualizing tag attributes and content classification. Xia et al. [2] processed information through tag cloud construction based on the structure and content of the Wikidata knowledge base, ultimately achieving information retrieval and page ranking.

Existing tag clouds primarily calculate tag weights quantitatively based on cumulative annotation frequency, using different colors or font sizes for intuitive visualization to facilitate user retrieval and browsing. Tags used by users can reflect their interests to a certain extent. However, as time passes, users' interest preferences and focus points change, and existing tag clouds built on cumulative frequency cannot well reflect these changes. Therefore, how to construct dynamic tag clouds for users based on tags used at different times to reveal changes in user interests and focus has become an important issue in leveraging tag clouds to better support user information retrieval and navigation. To address this, this paper studies methods for constructing tag clouds that reflect dynamic changes in user interests from the perspective of temporal characteristics of user annotations.

2. Dynamic Tag Cloud Construction

Constructing dynamic user tag clouds requires full consideration of the impact of temporal information on users' tag usage. Based on the forgetting characteristics and memory strengthening phenomena in psychology, we view the entire user annotation process as a combination of forgetting and repeated learning processes. That is, the importance of tags from more distant annotation times is lower, while repeatedly appearing tags have strengthened importance. Through this process, we dynamically calculate the weights of different tags to construct dynamic tag clouds and improve user browsing experience.

2.1 Dynamic Tag Weight Update User interests continuously change over time, and this change is a forgetting phenomenon. According to the forgetting characteristics and memory strengthening phenomena in psychology [7], the following basic features exist: (1) Interests closer to the current moment have higher weights, and interest weights gradually decline over time; (2) When the same interest repeatedly appears, there is a reinforcement process that merges with the original interest to form new user interests. Therefore, each interest undergoes both forgetting and repeated learning processes. Tags are expressions of users' own attitudes and interests, so we can utilize the forgetting and strengthening characteristics of interest memory to update the dynamic weights of user tags to reflect the impact of time on tag weights, thereby supporting dynamic tag cloud construction.

Using weight to measure users' interest level in tag t_k, the tag weight ktw undergoes decay and strengthening processes, forming a multi-stage decay process as shown in [Figure 1: see original paper]. In [Figure 1: see original paper], during a certain time period (e.g., from dn_0 to dn_1), the tag weight decays over time. When users continuously annotate in social tagging systems, the same interest periodically reappears (e.g., tag t_k reappears at times dn_1 and dn_2), and ktw is strengthened and rises again. Such repeated activities divide the entire user annotation process into multiple sub-stages, each being a new forgetting process. Therefore, based on similar forgetting curve formulas constructed by Yu et al. [7] and Yin et al. [8], we improve and propose a formula for calculating tag dynamic weights. The calculation of tag t_k' s dynamic weight ktw involves three main components: weight calculation at specific time points, forgetting decay, and memory strengthening.

(1) Tag Weight Calculation at Specific Time Points

The weight tw of tag t_k at a specific time point T is calculated using the TF (term frequency) method, i.e., the proportion of the number of times tag t_k is used to the total number of tag uses at that time point (e.g., a specific day). The calculation method is shown in formula (1):

$$tw = \frac{f_k}{\sum_{s=1}^{m} f_s}$$

where m is the total number of tags at that time point, and f_s is the frequency of tag t_s.

(2) Tag Weight Forgetting Decay

If tag t_k does not reappear, its weight ktw decays over time, which can be calculated using an exponential forgetting function. The quantitative function of the forgetting process is defined as formula (2):

$$ktw_d = ktw_{n-1} \times 2^{-\frac{d-d_{n-1}}{hlu}}$$

where ktw_d is the decayed tag weight, ktw_{n-1} is the weight when tag t_k appeared for the $(n-1)$th time (i.e., the initial value of the previous forgetting stage), hlu is the half-life of user u which varies with the user' s knowledge acquisition behavior cycle, and $d - d_{n-1}$ represents the time difference since the last appearance of tag t_k.

(3) Tag Weight Memory Strengthening

As shown in [Figure 1: see original paper], at three time points dn_1, dn_2, and dn_3, tag t_k repeatedly appears. The value of ktw'_d is composed of the remaining weight from the previous stage' s decay and the weight increase from the new annotation activity of the same tag t_k. Formula (3) calculates the initial interest degree for each forgetting stage:

$$ktw'_d = ktw_{n-1} \times 2^{-\frac{d_n-d_{n-1}}{hlu}} + tw_n$$

where ktw'_d is the initial tag weight at time point d_n, d_n represents the time point when tag t_k appears for the nth time, so $d_n - d_{n-1}$ is the time difference between two consecutive appearances of tag t_k; tw_n is the weight when tag t_k appears for the nth time, whose calculation method is given by formula (1) and represents the weight increase brought by annotation activity at time point d_n. The first term represents the remaining weight from the previous stage' s decay to time point d_n.

2.2 Dynamic Tag Cloud Construction Algorithm Based on the above dynamic tag weight update mechanism, we establish the following dynamic tag cloud construction algorithm.

Input: User u' s annotation history records (including annotation time, resources, and tags used)
Output: User u' s dynamic tag cloud

Algorithm Description:
Use formula (1) to calculate the initial tag interest weights from user annotations, obtaining each tag' s weight at different time points (typically counted by day).
Sort tags in chronological order and determine whether tag t_k reappears. If not, proceed to step to update the tag weight; if it does, proceed to step to update the tag weight.
Calculate the decayed tag weight using formula (2).
Calculate the strengthened tag weight using formula (3), which combines the decayed value from the previous stage and the weight increase from the new annotation activity.
Synthesize each tag' s weight and perform normalization to obtain user u' s dynamic tag cloud.

3.2 Visualization of User Dynamic Tag Cloud

To intuitively reflect the effect of dynamic tag clouds, we selected visualization results of typical users' tag clouds for comparison. We used both the existing cumulative annotation frequency method and the proposed dynamic tag cloud construction method to establish visual tag clouds, with font size distinguishing tag weights. [Figure 2: see original paper] shows the visualization results of tag clouds constructed by the two methods for an active Delicious user (UserID: 12116) at the end of the 6th month (i.e., the annotation deadline), displaying the top 50 popular tags, where larger font size indicates higher tag weight. This user performed 2,710 annotations within 6 months, covering 995 resources and using 424 tags, with the most frequently used tag appearing 447 times and the least frequent appearing only once.

The experimental data comes from two representative social tagging systems: Last.fm and Delicious. The Delicious data was collected by the DAIM research group at Peking University, covering over 185,000 users' social annotation data from the Delicious website between January and June 2009, available at: http://www.datatang.com/data/42989. The Last.fm data was collected by the Information Retrieval Group at the Autonomous University of Madrid, covering music annotation data from 1,892 users, available at: http://grouplens.org/datasets/hetrec-2011. The basic statistics of the experimental data are shown in .

shows sample annotation data. We selected users with high activity (i.e., number of annotated resources) for experimental study. The selected annotation data includes user, resource, tag, and annotation time information. If multiple tags are used to annotate one resource, multiple records are formed.

3.3 Evaluation of User Dynamic Tag Cloud Navigation Effect

To further verify whether the proposed dynamic tag cloud better reflects changes in user interests and provides better information retrieval and navigation, we conducted quantitative comparative experiments. The evaluation is based on the assumption that if a user is currently interested in a certain tag, they will continue to use it in the future. Therefore, we constructed evaluation indicators based on whether the top N tags with large weights at a certain time point are used again by the user in a future period, to characterize how well the tag cloud captures user interests. We first define two basic evaluation indicators as shown in formulas (4) and (5):

$$Acc = \frac{\sum_{i=1}^{N} f_i}{N}$$

$$Rec = \frac{\sum_{i=1}^{N} f_i}{\sum_{j=1}^{O} f_j}$$

where f_i is the frequency of the i-th tag among the top N tags used in a future period; N_n is the total number of times the top N tags are used in a future period; O_n is the total number of times all tags are used by the user in that future period. As can be seen from the definitions, Acc represents the average frequency of the top N tags used in a future period, while Rec represents the ratio of the total frequency of the top N tags to the frequency of all tags used by the user during that period. These two basic indicators characterize the usage of the top N tags in a future period from different perspectives. We therefore define a new comprehensive evaluation indicator combining both, as shown in formula (6):

$$AR = Acc \times Rec$$

Taking the Delicious experimental dataset as an example, we selected 15 users with relatively complete annotation histories (i.e., those with basically 6 months of continuous annotation activity) as subjects. The 6-month dataset was divided into 36 periods in chronological order, with each period being 5 days. For $k = 1, 2, \ldots, 36$, we took the tag weights at the end of the k-th period, ranked the top 10, 20, and 30 tags by weight, and calculated the usage of these tags in the next period of 5 days to compute the AR evaluation indicator value. For example, when $k = 2$, we calculated the top 10, 20, and 30 tags by weight up to day 10, then used the user' s annotation data from the next 5 days (i.e., the 3rd period from day 11 to day 15) to calculate the evaluation indicator values. Similarly, for Last.fm, we took 15 active users as research subjects. Since this dataset spans a large time period, we divided the 6-year dataset into 12 periods of 6 months each, with the same evaluation method as for Delicious.

To test the difference between the dynamic tag cloud constructed based on annotation temporal information and the static tag cloud constructed based on cumulative annotation frequency in reflecting user interest, we calculated the ratio of their comprehensive evaluation indicator AR values after computing both, as shown in formula (7):

$$Ratio = \frac{AR_D}{AR_S}$$

where AR_D and AR_S are the comprehensive evaluation indicator values calculated by the dynamic tag cloud method and static tag cloud method, respectively.

[Figure 3: see original paper] shows the comparison results for 15 selected users using both methods. It can be seen that for the top 10, 20, and 30 tags by observed weight, the dynamic method shows varying degrees of improvement over the static method. For the Delicious dataset, the improvement ranges from 4% to 11%, while for the Last.fm dataset, the improvement is more significant, ranging from 18% to 82%.

As seen from the visualization in [Figure 2: see original paper], the static tag cloud constructed using existing cumulative annotation frequency methods and the dynamic tag cloud constructed using the proposed temporal information approach yield different relative tag weights, thus producing different effects in guiding users' relative priority for information retrieval and navigation.

Furthermore, under the above Last.fm experimental data and the same evaluation metrics, we conducted comparative experiments between the dynamic method and other methods, selecting the cumulative frequency-based method (TF) and the tag time weight strategies proposed in literature [8], including TF time weight and TFIDF time weight. As shown in [Figure 4: see original paper], the dynamic method outperforms other methods with higher recommendation accuracy.

From the current experimental results, the improvement effect on the Delicious dataset is not very significant because interests typically do not change substantially within a short 6-month period. However, for the Last.fm dataset with a longer time cycle, the dynamic method shows significant improvement over the static method, indicating that the dynamic tag cloud constructed based on annotation temporal information can better reflect changes in user interests and better support information retrieval and navigation.

Conclusion

To enable tag clouds to better reflect users' current dynamic interests, we propose a method for constructing dynamic user tag clouds based on temporal evolution, considering the dynamic characteristics of user interests and the temporal features of social tagging. This method constructs dynamic tag weights based on the forgetting and strengthening characteristics of memory in psychology. Experimental results show that the visualization of dynamic tag clouds differs from static tag clouds built on cumulative annotation frequency. Compared with existing tag sorting algorithms, the dynamic method is superior to other algorithms and can characterize and capture users' current interests, which is beneficial for helping users better utilize tag clouds for information retrieval and navigation. Tags are only single words reflecting user interests, while user interests are often characterized by themes formed by multiple tags. Therefore, further mining of user interest themes based on the current dynamic tag cloud will be the focus of future work.

References

[1] Wei Jianliang, Zhu Qinghua. Advances in Personalized Information Recommendation Based on Social Tagging [J]. Journal of the China Society for Scientific and Technical Information, 2010, 29(4): 625-633.

[2] Xia Yu, Ge Jiaqi, Ma Xiu, et al. Research on Searching Algorithm Based on Wikidata and Tag Cloud [J]. Software Guide, 2016, 15(8): 42-46.

[3] Zeng Ziming, Zhang Zhen. A Personalized Recommendation Approach Based on Community Tag Cloud in Social Tagging System [J]. Journal of Intelligence, 2011, 30(10): 128-133.

[4] Bi Qiang, Zhou Shanshan, Ma Zhiqiang, et al. Study on Optimization Mechanism of Tag Cloud for Knowledge Relation [J]. New Technology of Library and Information Service, 2014(5): 33-40.

[5] Hassan-Montero Y, Herrero-Solana V. Improving Tag-Clouds as Visual Information Retrieval Interfaces [C]//Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies, Merída, Spain. 2006.

[6] Millen D R, Feinberg J. Using Social Tagging to Improve Social Navigation [OL]. [2016-09-30]. https://www.researchgate.net/publication/228904554_{{{Using}}{{social}}}{{{tagging}}

[7] Yu Hongtao, Cui Ruifei, Dong Qinqin. Micro-blog User Interest Model Based on Forgetting Curve [J]. Computer Engineering and Design, 2014, 35(10): 3367-

3372, 3379.

[8] Yin Guisheng, Cui Xiaohui, Ma Zhiqiang. Forgetting Curve-based Collaborative Filtering Recommendation Model [J]. Journal of Harbin Engineering University, 2012, 33(1): 85-90.

[9] Zhao Kaihui. Recommendation Method of Personalized Information Based on Socialized Tagging [J]. Information Science, 2015, 33(6): 39-42.

Author Contributions

Xie Mengyao: Method research and data experiments, manuscript drafting.
Pan Xuwei: Research idea proposal, experimental design, and manuscript revision.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data [1] is available in the online version of the journal at http://www.infotech.ac.cn; supporting data [2-4] are self-archived by the authors and available via E-mail: panxw@zstu.edu.cn.

[1] Xie Mengyao, Pan Xuwei. Experimental Data.zip. Selected user experimental data from Delicious and Last.fm.
[2] Xie Mengyao, Pan Xuwei. Static vs Dynamic Method Comparison.xlsx. Improvement effect of dynamic method over static method.
[3] Xie Mengyao, Pan Xuwei. Dynamic Method vs Other Time Weight Methods Comparison.xlsx. Improvement effect of dynamic method over other methods.
[4] Xie Mengyao, Pan Xuwei. Experimental Database File.sql. Database code.

Received: 2016-10-12
Revised: 2016-12-09

Note: Figure translations are in progress. See original paper for figures.