Journal of Computer and Communications, 2014, 2, 59-66
Published Online December 2014 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2014.214006

Scientific Research

Research of Collaborative Filtering Recommendation Algorithm for Short Text

Chunxu Chao[1], Shouning Qu[2*], Tao Du[1]

[1]School of Information Science and Engineering, University of Jinan, Jinan, China
[2]The Center of Information Network, University of Jinan, Jinan, China
Email: chaochunxu@126.com, [*]qsn@ujn.edu.cn, dt@ujn.edu.cn

Received 20 October 2014; revised 15 November 2014; accepted 18 December 2014

Abstract

Short text, based on the platform of web2.0, gained rapid development in a relatively short time. Recommendation system analyzing user's interest by short texts becomes more and more important. Collaborative filtering is one of the most promising recommendation technologies. However, the existing collaborative filtering methods don't consider the drifting of user's interest. This often leads to a big difference between the result of recommendation and user's real demands. In this paper, according to the traditional collaborative filtering algorithm, a new personalized recommendation algorithm is proposed. It traced user's interest by using Ebbinghaus Forgetting Curve. Some experiments have been done. The results demonstrated that the new algorithm could indeed make a contribution to getting rid of user's overdue interests and discovering their real-time interests for more accurate recommendation.

Keywords

Short Text, Personalized Recommendation, Time Weight Function

1. Introduction

Recent years, like Facebook, Twitter, short texts are very popular in the social field all over the world. One of the most prominent short texts is micro-blog in China. Depending on the advantage of brief, real-time in information sharing, spreading and acquisition, weibo gains sharp development and begins to influence people's lives and their way of thinking. In July 2014, 34th China Internet network development state statistic report [1] given by CNNIC pointed out that, up to June 30, 2014, the scale of Chinese weibo users has reached 275 million. Micro-blog has already become one of the social networks used to broaden one's reach and realize social interac-

[*]Corresponding author.

How to cite this paper: Chao, C.X., Qu, S.N. and Du, T. (2014) Research of Collaborative Filtering Recommendation Algorithm for Short Text. *Journal of Computer and Communications*, **2**, 59-66. http://dx.doi.org/10.4236/jcc.2014.214006

tion, especially an important tool of acquiring latest information. Users play a role as information consumers; at the same time, they are data producers, too.

With the influx of large quantities of users, weibo surged in a short time. People have lost in the ocean of microblog information already. In the fast-pace today, how to acquire the most accurate information needed by users in the shortest time, has become a hot issue nowadays.

At present, there are two main recognized way to solve the problem of information overload: information retrieval and information filtering technology. Represented by Google, Yahoo, information retrieval technology has indeed achieved great success. However, it draws on the requirement that users must be able to accurately describe their personal needs. Once users cannot describe their demands well, information search quality of it cannot be guaranteed, which often leads to search results undesirable. Information filtering technology can solve this problem very well. As an important application of information filtering, recommendation system has become an indispensable part of individualized information service form among the new generation of Web applications. Collaborative filtering algorithm (CFA) is the most efficient recommendation algorithm at present. CFA analyzes user's interest and finds others who have the same interest with him and then integrates these similar users' evaluation with some information and forms recommendations for him. It is quite precise on locating users' interest. It can also filter some concepts complex and indescribable, which is obvious superior to other algorithms. However, CFA can't make a distinction between real-time interest and overdue interest well, which results in an unsatisfactory precision. This paper gives a new algorithm, time weight algorithm (TWA), which can tell user's real-time interest well and improve the precision of recommendation.

The rest of paper is organized as follows: Section 2 presents the research status home and aboard. Section 3 gives the preliminary concepts, regarding forgetting curve and the details of TWA. Section 4 analyzes the experiments results. Section 5 concludes and gives the pointer to the future work.

2. Related Work

As a new thing, weibo filtering has not caused widely concern relatively. Western scholar Ernesto [2] combined with the effectiveness of micro-blog, found and ranked weibo topics, according this to recommend to users. Sriram [3] firstly divided weibo into several parts, such as news, transaction, private information and so on, then studied with different classes and achieved good results. Golbeck [4] presented some problems and challenges on weibo filtering. He demonstrated that some problems have been still existed among weibo filtering nowadays. Hannon [5] recommended similar users to specific user by using CFA based on content. Research of weibo on aboard has made some achievements, but these are only for western texts, most of them do not apply for Chinese texts. Domestic studies are still in its infancy. Wang Lin [6] proposed a filtering method faced with weibo, which is effective on noise discrimination and content similarity detection. Although the method could effectively purify micro-blog data, it needed constantly to gain new rules and characteristics to face the change of variety and feature about noise weibo. Shen Jing [7] in virtue of non-structured DM platform designed an efficient distributed text filtering algorithm, which acquired a good filtering result, but low efficiency. In addition, Shao Jianshuang [8] constructed a text filtering model based on concept lattice and gave its usage. For tracing and capturing the changing of users' interest, Xing Chunxiao [9] proposed data weight based on time-window and item-similar. They used linear function because they believe that the changing of user's interest follows the law of linear forgotten. Zhang [10] used exponential function as time function to solve the decline of recommendation quality with the changing of users' interest. These researches all made contribution to weibo filtering, but not very satisfactory.

Under this background, this paper proposed TWA based on Ebbinghaus Forgetting Curve [11] to further optimize CFA and improve the precision of recommendation.

3. The Design of TWA Based on Forgetting Curve

3.1. Forgetting Curve

German psychologist H. Ebbinghaus studied carefully and systematically about the phenomena of memory loss and made a forgetting curve using the testing results from the experiments about featureless syllables and letters. This is the famous Ebbinghaus Forgetting Curve, shown in **Figure 1**.

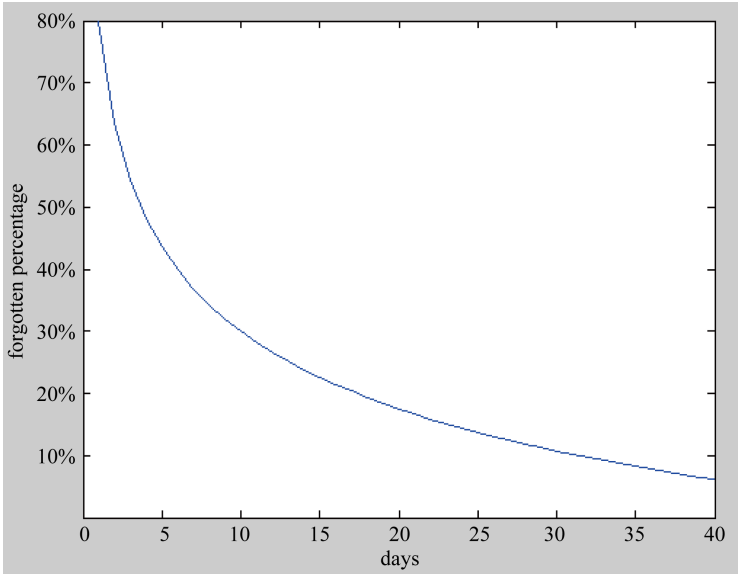

Figure 1. Ebbinghaus forgetting curve.

Among **Figure 1**, vertical coordinates of the curve representative the memorial quantity of a learner, while *x*-coordinates show the time after learning. As is shown by **Figure 1**, man forgets things not a simple process of gradual oblivion, but presents such a trend that oblivion in a short period of time after memorization was relatively quick and after a long interval oblivious speed slowed.

Weibo behavior of a man is a reaction of his psychology. So changing of user's interest on publishing, transmitting and commenting a weibo also follows this forgetting law. As the shape of forgetting curve much matches exponential function, we use exponential function to simulate user's interest changing over time. Yu Hong [12] took the advantage of ZGrapher [13] to fit Ebbinghaus Forgetting Curve and acquired a mathematical expression:

$$Y = 31.8 \cdot X^{-0.125} \tag{1}$$

where, X represents days after learning, Y is forgotten percentage.

3.2. TWA Based on Forgetting Curve

According to man's oblivious nature, we divide man's interest into real-time interest which includes long-term interest and recent interest, and overdue interest. Then use TWA based on Forgetting Curve to better explore user's real-time interest and get rid of his overdue interest to improve the precision of personalized recommendation. TWA formula is as follows:

$$\omega_t(k) = m \cdot \left(\frac{t_{\text{current}} - t_n}{t_n - t_{n-1}} \right)^{-0.125} + (1-m) \tag{2}$$

where, t_{current} represents current time, t_n is the publish time of target weibo and t_{n-1} is the publish time of the weibo before target weibo under the same class k. m, weight factor, values between 0 and 1.

We analyze the rationality of formula (2) from three aspects.

1) User has involved frequently on the theme k in the past, but recently does not focus on it [14]. This shows that user has been very interested in the theme k in a period of time in the past, but now he is not interested in the theme. In the algorithm, there are many weibo under k and time intervals between them are very short. For they published in the past, all of them have long time intervals with the current time, that is to say, denominator $t_n - t_{n-1}$ very small, but numerator $t_{\text{current}} - t_n$ very big, so $\frac{t_{\text{current}} - t_n}{t_n - t_{n-1}}$ very big and $\omega_t(k)$ very small. This situation describes user's interest rightly.

2) User has involved frequently on the theme *k* in the past and recently. This shows that user has been

interested in k and k is user's long-term interest. In the algorithm, time intervals between weibo under theme k are relatively long,so denominator $t_n - t_{n-1}$ very big, $\omega_t(k)$ very big.This situation describes user's interest rightly.

3) User has involved frequently on the theme k recently. This shows that theme k is user's recent interest. In the algorithm, there are many weibo under k, and time intervals between each other are very short, but they all published in the recent, have short time intervals with the current time, that is to say, $\frac{t_{\text{current}} - t_n}{t_n - t_{n-1}}$ would be smaller than the first situation, so $\omega_t(k)$ would be bigger than the $\omega_t(k)$ in the first situation. This situation describes user's interest rightly, too.

3.3. Recommendation Algorithm and Process Description

The algorithm of improved ITC [15] is the improvement of TF-IDF. It includes two parameters that the Information of the term in a category $\left(\text{DDEC}(t_i)\right)$ and the weight of position distribution $\left(w_{\text{pos}_{t_i}}\right)$. It has already been demonstrated that it is very useful and efficient on short text classification. In our algorithm, it is used to acquire the weight of each item preliminarily. The process description of our new algorithm is given as follows:

Algorithm: compute $\omega_t(k)$ of user u's interests under the theme k

Data: k, t_{current}, t_n, t_{n-1}, threshold, $n \geq 1$;

Result: $\omega_t(k)$, sum;

Foreach weibo that user published

  Firstly pre-processing, then usingimproved ITC to acquire the weight of every item, then making use of cosine feature to get u's preliminary kinds of interests;

  Foreach weibo under the same theme

    Let $t_0 = 0$, t_1 is the first weibo in the theme;

    Using formula (2), compute $\omega_t(k)$. For each weibo whose $\omega_t(k) >$ threshold, we mark its value as 1, otherwise, mark its value as -1;

    n++;

  end

  put all weibo under a theme together and plus their values as sum. If sum > 0, we mark this theme as user's real-time interest, otherwise, mark the theme as overdue interest;

end

The flow chart of this algorithm is given in **Figure 2**.

4. Experimental Evaluation

4.1. Data Set

Data set in this paper is grasped from the open platform [16] provided by Sina Micro-blog. It mainly includes 7113 weibo which was published by 15 users in recent 3 months. Most of weibo include weibo id, user id, user name, screen name, re-tweeting id, content, weibo url, resource, picture url, audio url, video url, geographic coordinate, re-tweeting number, comment number, the number that who like it, publishing time.

For those weibo that user re-tweets or comments, we regard the previous weibo and the content of user's comment as its real content. Then we begin to pre-process. Pre-processing covers eliminating the stop list and function words, such as "haha", "too", "also" and so on. We treat weibo whose number of words after pre-process less than 5 as pointless weibo and wipe out it. In addition, this experiment is only for Chinese content. If weibo is completely foreign language, we will wipe it out, too. Thus, after pre-process, our data set still contains 4981 weibo. Then, we select 14 randomly from 15 users and regard their 4707 weibo which were tweeted

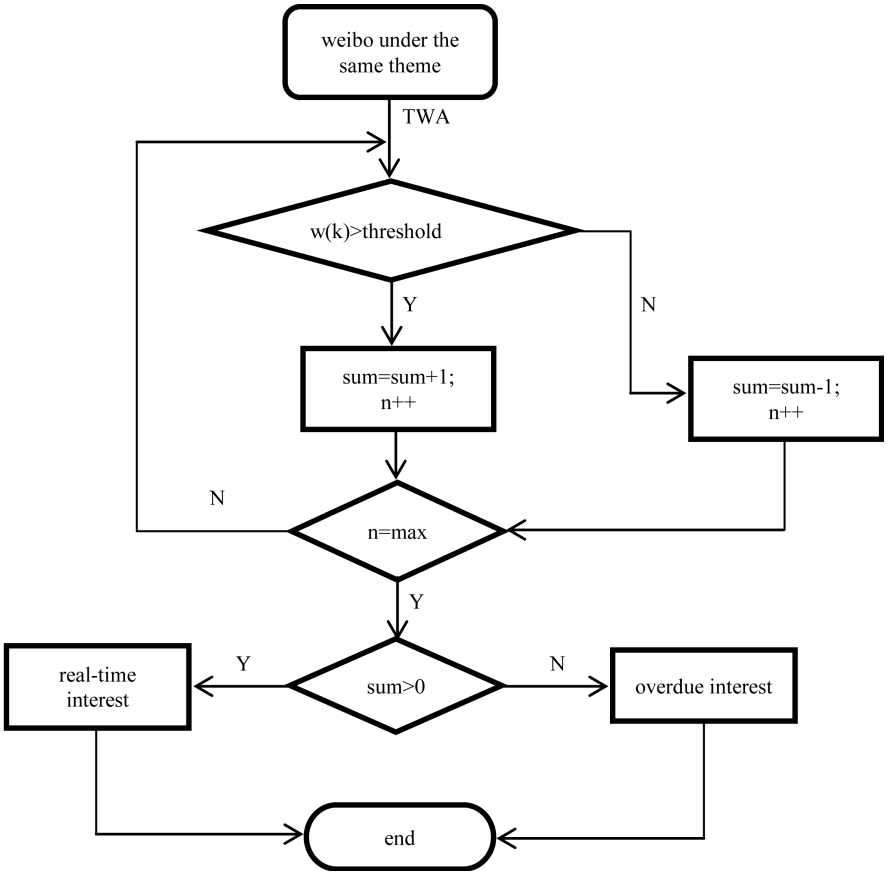

Figure 2. Flow chart of TWA.

by them in recent 3 months as training set. 274 weibo published by another one comes into being test set. The classification of training set and test set is shown in **Table 1**.

We try to manually annotate weibo of training set to tell user's real-time interest and overdue interest. After training and calculating, we reach when m take 0.4, the result simulated by formula (2) is the most similar to the result we mark. What's more, we depend on the statistics and set threshold. Then, we bring $m = 0.4$ into formula (2) and test on the test set.

4.2. Evaluation Criterion

According to TWA based on forgetting curve, we compare the weight of weibo with threshold. If the weight is greater than threshold, we set the result of weibo as 1. Otherwise set its result as −1. Then put all weibo under a theme k together and plus their values as sum, if sum > 0, we mark this theme as user's real-time interest, if not, mark the theme as overdue interest.

In this paper, we take Precision, Recall and MAPE as evaluation criterion.

Precision is the radio of the number of related documents which were retrieved and the number of all documents which were retrieved. It measures the precision of a recommendation system. Recall is the radio of the number of related documents which were retrieved and the number of all related documents. It measures the comprehensive radio of a recommendation system. Their formulas are shown as follows:

$$\text{Precision} = \frac{\left|\{\text{Relevant}\} \cap \{\text{Retrieved}\}\right|}{\left|\{\text{Retrieved}\}\right|} \tag{3}$$

$$\text{Recall} = \frac{\left|\{\text{Relevant}\} \cap \{\text{Retrieved}\}\right|}{\left|\{\text{Relevant}\}\right|} \tag{4}$$

where, $\{\text{Retrieved}\}$ is the set of documents which were retrieved, $\{\text{Relevant}\}$ is the set of documents which were related with request.

MAPE (Mean Absolute Percentage Error) measures the precision of algorithm by computing the mean absolute percentage error between predicted value and true value. The smaller the value of MAPE, the smaller the gap of predicted value and true value, which means predicting much closer to the true choice of user, the higher precision of recommendation. We make prediction score set of user $\{r_1, r_2, \cdots, r_n\}$ and the true score set $\{p_1, p_2, \cdots, p_n\}$, then the formula of MAPE is:

$$\text{MAPE} = \left| \frac{\sum_{i=1}^{n} (p_i - r_i)}{n} \right| \tag{5}$$

4.3. Experimental Result

Experiment one show the traditional difference of Precision and Recall between TWA and improved ITC. We set two classifications on this experiment. The result is shown in **Table 2** and **Table 3**.

For a better intuitive effect, we give histograms of **Table 2** and **Table 3** in **Figure 3** and **Figure 4**.

Table 1. The classification of training set and test set.

	Number of user	Number of weibo
Data set	15	4981
Training set	14	4707
Test set	1	274

Table 2. The precision and recall of improved ITC.

	Economy	Tourism	Health	Education	IT	Car
Economy	39	1	2	0	1	1
Tourism	0	43	2	0	1	2
Health	1	0	50	1	0	1
Education	2	3	3	14	0	0
IT	2	2	1	1	11	0
Car	0	2	2	0	0	29
Recall	0.830	0.811	0.769	0.778	0.714	0.829
Precision	0.886	0.843	0.833	0.875	0.769	0.879

Table 3. The precision and recall of TWA.

	Economy	Tourism	Health	Education	IT	Car
Economy	35	1	2	0	1	1
Tourism	0	42	2	0	1	2
Health	1	0	50	1	0	1
Education	2	3	3	12	0	0
IT	2	2	1	1	11	0
Car	0	2	2	0	0	23
Recall	0.814	0.808	0.769	0.75	0.714	0.793
Precision	0.875	0.840	0.833	0.857	0.769	0.852

Figure 3 and **Figure 4** show clearly that Precision and Recall of interest classification of user u totally goes down by using TWA. It's because that TWA aims at separating user's real-time interest from his overdue interest and abandons those documents of overdue interest. It leads to the decrease of numerator $\left|\{\text{Retrieved}\}\cap\{\text{Relevant}\}\right|$ of formula (3), (4), but denominator $\{\text{Retrieved}\}$, $\{\text{Relevant}\}$ of formula (3), (4) remains unchanged. Both of them make contribution to the decline of Precision and Recall. These two traditional parameters going down, is the inevitable result of optimization on the basis of the original one. But on the other hand, it is more certain the necessity and importance to optimize the original algorithm.

Experiment two shows the difference between improved ITC and TWA at capturing user's real-time insterest. For MAPE, we set $r_i = 1$ if i is the designated kind of interest by user u. Otherwise, $r_i = 0$. Then use algorithm to estimate user's interest and compare with r_i. If the kind of interest corresponds to $r_i = 1$, we set $p_i = 1$, otherwise $p_i = -1$. The result of MAPE is in **Table 4**.

From **Table 4**, we can see that compared with traditional algorithm, using TWA makes the result of MAPE decline 16.6%. It also demonstrates that TWA based on collaborative filtering algorithm is obviously prefer to the traditional algorithm. TWA indeed promotes the quality of recommendation. It can be more precise to capture the change of user's real-time interest.

5. Conclusion and Future Work

Traditional collaborative filtering algorithms haven't consider sufficiently about the change of user's interest.

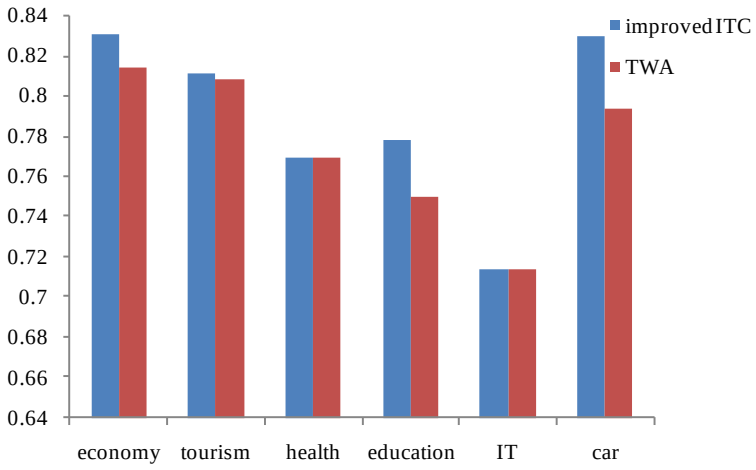

Figure 3. The recall of two different algorithms.

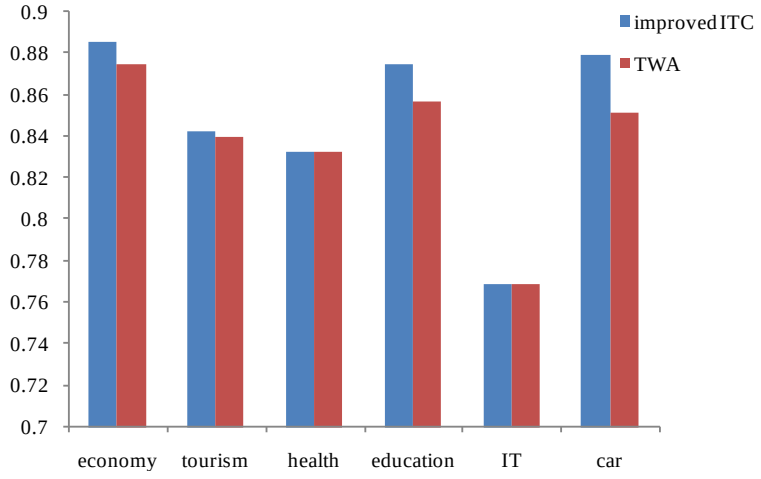

Figure 4. The precision of two different algorithms.

Table 4. MAPE of using two algorithms.

Algorithm	MAPE
Improved ITC	0.333
TWA	0.167

This leads to a big difference between the result of recommendation and user's real demands. Under this context, we propose the TWA based on collaborative filtering algorithm. As the experimental result suggests that TWA is obviously prefer to other traditional algorithms on the precision and it can promotes the quality of recommendation in a large extent. It can do user more effective personalized recommendation indeed. However, for the limitation of Sina Micro-blog open platform, our privilege is so low that we can only test for 15 users one time, which inevitably leads to the experimental subjects slightly single. We expect that Sina Mirco-blog open platform could open more user privilege in future. Thus we can trace and test more users in real-time. It can not only further improve the precision of personalized recommendation, but also be the highlight of our next work.

References

[1] (2014) China Internet Network Development State Statistic Report. China Internet Network Information Center, Beijing, 27.

[2] Diaz-Aviles, E., Drumond, L. and Gantner, Z. (2012) What Is Happening Right Now…That Interests Me? Online Topic Discovery and Recommendation in Twitter. *Proceedings of the* 21*th ACM International Conference on Information and Knowledge Management*, 1592-1596.

[3] Sriram, B., Fuhry, D. and Demir, E. (2010) Short Text Classification in Twitter to Improve Information Filtering. *Proceedings of the* 33*rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 841-842.

[4] Globeck, J. (2012) The Twitter Mute Button: A Web Filtering Challenge. *Proceedings of the* 30*th International Conference on Human Factors in Computing Systems*, 2755-2758.

[5] Hannon, J., Bennett, M. and Smyth, B. (2010) Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches. *Proceedings of the* 4*th ACM Conference on Recommendation Systems*, **9**, 199-206.

[6] Wang, L., Feng, S. and Xu, W.L. (2012) A Filtering Approach for Spam Discrimination and Content Similarity Double Detection for Microblog Text Stream. *Computer Applications and Software*, **2**, 25-29.

[7] Shen, J. and Jiang, Q. (2011) A Distributed Short Text Filtering Algorithm. *Journal of Sichuan Ordnance*, **32**, 151-153.

[8] Shao, J.S., Li, G.Y. and Zhang, J. (2011) Design of Text Filtering Model Based on Concept Lattice. *Computer Engineering and Design*, **32**, 1047-1050.

[9] Xing, C.X., Gao, F.R. and Zhan, S.N. (2007) A Collaborative Filtering Recommendation Algorithm with User Interest Change. *Journal of Computer Research and Development*, **44**, 296-301.http://dx.doi.org/10.1360/crad20070216

[10] Zhang, Y.C. and Liu, Y.Z. (2010) A Collaborative Filtering Algorithm Based on Time Period Partition. *The Proceeding of* 3*rd International Symposium on Intelligent Information Technology and Security Informatics*, 777-780.

[11] Forgetting Curve. Wikipedia. http://zh.wikipedia.org/wiki/%E9%81%97%E5%BF%98%E6%9B%B2%E7%BA%BF

[12] Yu, H. and Li, Z.Y. (2010) A Collaborative Filtering Recommendation Algorithm Based on Forgetting Curve. *Journal of Nanjing University* (*Natural Science*), **46**, 522-523.

[13] Palam Software. ZGrapher-Grahping Calculator Software.

[14] Wang, G.X. (2013) User Interest Analysis and Personalized Information Recommendation Based on Microblog. Shanghai Jiao Tong University, Shanghai.

[15] Li, L.L. and Qu, S.N. (2013) Short Text Classification Based on Improved ITC. *Journal of Computer and Communication*, **1**, 22-27. http://dx.doi.org/10.4236/jcc.2013.14004

[16] Sina Micro-Blog Open Platform. http://open.weibo.com/