2014 11th Web Information System and Application Conference

Topic trend prediction based on wavelet transformation

Mingyue Fang, Yuzhong Chen*, Peng Gao, Shuiyuan Zhao, Songpan Zheng
College of Mathematics and Computer Science
Fujian Key Laboratory of Network Computing and Intelligent Information Processing
Fuzhou University
Fuzhou, Fujian, China
yzchen@fzu.edu.cn

***Abstract*—The research of topic trend prediction can be a good reference for maximizing the propagation effects of network advertisements as well as guiding and controlling the network consensus. This paper proposes PTEP(the Prediction of Topic Energy Peak) method to model the life cycle of a topic and predicts the time when a hot topic will outbreak. Firstly, taking the number and the authority of followers and the interest of users to a topic into consideration, we design a topic-related user authority (TRUA) algorithm to measure the authority of users. Secondly, we calculate the energy value considering both the tweets and users authority related to the topic. Thirdly, we measure the fluctuation of the energy value based on wavelet transformation. Finally, we present rules to predict topic trend. Experimental results show that our method can effectively predict the peak of a topic in advance with a low omission rate.**

Keywords—microblog; topic trend prediction; user authority; aging theory; wavelet analysis

I INTRODUCTION

With the development of microblog, the explosive growth of microblog information makes microblogging platform an important public opinion field nowadays. However, it's a challenging task to understand and predict the trend of hot topics which netizens concern from the vast amount of microblog information. The hotness trend prediction research of topics has broad application prospects and high research value.

This paper proposes PTEP(the Prediction of Topic Energy Peak) method to model the life cycle of a topic and accents on predicting the time when a hot topic outbreaks, which is essential to public opinion monitoring. The main work and contributions of this paper are as follows:

1) The importance of a user within a topic is related to the number and the authority of its followers, as well as the level of their interest in the topic. Taking the factors above into account, we propose an algorithm, TRUA (Topic-Related User Authority), to measure the authority of users.

2) We develop a topic life cycle model to calculate the *energy value* of a topic based on aging theory, considering tweets and users authority related to the topic.

3) After calculating the *energy value* of a topic, we measure the *fluctuation* of the *energy value* based on wavelet transformation.

4) We present rules to evaluate the stage of a topic and predict the *peak* of a topic.

The remainder of this paper is organized as follows: Section II summarizes the related studies. Section III analyzes the problem. Section IV explains the proposed method. Section V presents the experimental results that validate the computational efficiency of our model and algorithm. Finally, Section VI concludes our work and describes some future works in this area.

II RELATED WORK

EDT(Emerging trend detection) is aiming to identify topics which are new or becoming popular from the text information and predict their trend [1]. EDT is different from TDT(Topic Detection and Tracking)[2], which focuses on detecting topics from data stream. ETD is more concerned about trend prediction of hot topics.

The research of hot topic trend prediction has achieved certain results. Overseas research in this area is mainly based on twitter dataset. L.Chen et al.[3] predicts presidential election through researching on the discussion of twitter users. S. Asur et al.[4]forecasts the movie box office performance by analyzing the movie reviews from a large number of users in a certain period of time. M. Cataldi et al.[5]develops a life cycle model of keywords for topic detection and tracking based on aging theory. The model takes the tweets containing the keywords as nutrition and calculates the energy value of a keyword by considering the weight of the keyword in tweets and the authority of its source. J. H. Lau et al. [6] proposes a novel approach based on LDA (Latent Dirichlet Allocation) [7,8] for on-line trend analysis. On each update, the model calculates the evolution of topics to detect newly emerged topics in the document collection. Domestic researches in the field mainly focus on BBS topic or web topic which is characterized by click ration. D. Y. Ou et al. [9] analyzes and forecasts the attention of social networking groups based on the statistical model. Y. Tian [10] establishes a future trend prediction model through regression analysis of microblog event trend samples.

Though the existing works have achieved certain results, there are still some limitations.

1) Most of the existing researches focus on content information of tweets while neglecting the social network relationship between users related to the topic.

2) Some works which consider the relationship of users, i.e. [5], ignore that the user authority is different within different topics .

3) The research of when a topic will burst remains to be studied. Most of the prediction rules are lack of the prediction of *energy peak*. And the precision of topic trend prediction needs to be improved

978-1-4799-5727-9/14 $31.00 © 2014 IEEE
DOI 10.1109/WISA.2014.37

III PROBLEM ANALYSIS

When a topic appears, people may be interested in it. As time goes by, people would be used to it and pay little attention to it. Thus, a topic can be considered as a life form that goes through the aging stages of ***infancy***, ***growth***, ***Stability***, ***decay***, **and** ***death*** [11,12]. Some existing works develop a life cycle of topic based on aging theory, taking the suitable resources as nutrition, such as number of tweets. And then *energy value* is defined to depict which stage a topic is in.

The stages of a topic are shown in Fig.1.

1) ***Infancy***: the topic has just appeared in this stage, where the related tweets and users are limited, the *energy value* is low, and the *energy value* has no obvious ***fluctuation***.

2) ***Growth*****:** with the spread of the topic, its *energy value* grows rapidly, and the *fluctuation* of *energy value* is obvious. Specially, the stage of "***growth***" contains a stage of "***burst***". It starts from the point '***a***' in Fig.1, where the *energy value* exceeds a threshold and the *fluctuation* is very obvious. The stage of "***burst***" signals that the *energy value* is reaching its *peak*.

3) ***Stability*****:** with the diffusion of the topic, the *energy value* reaches the maximum, and remains high for a period. That is, the *energy value* is the highest and has no obvious *fluctuation* in this stage**.**

4) ***Decay*****:** as time goes by, people pay less and less attention to the topic. In this stage, the *energy value* decays off.

5) ***Death*****:** in this stage, the topic is out of date, and the *energy value* is low.

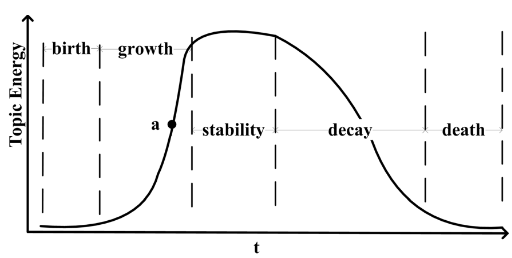


Fig.1. Life cycle of a topic

This paper aims to:

1) Quantify the *energy value* of a topic in an appropriate form to identify the life stage of the topic;

2) Predicting the time when a hot topic outbreaks, that is, identify the ***forecast point a*** in Fig.1, namely when the topic would grow explosively and hit the peak after a period of time.

IV PROPOSED METHOD

The proposed method PTEP includes four steps: measures the *authority* of users; calculates the *energy* and the energy *fluctuation* of a topic; presents rules for predicting topic trend.

A User Authority

The source of the tweets related to a topic is crucial to the spread of the topic. Topics on microblog are derived from a set of users. The higher the user authority is, the greater effect will be exerted on the topic. Therefore, measuring the user authority is the key point to quantify the importance of a topic.

In order to calculate user authority, we define a user's social network $G(V, E)$, where V is the set of users and E indicating the relationships between users. Given two users $u_i, u_j \in V$, the edge $e_{ij} = (u_i, u_j) \in E$ exists only if u_i is a follower of u_j.

It is generally known that influential users turn to have a lot of fans. The authority of a user is also related to the degree of authority of its followers. In addition, user's authority is different in different topics. Thus, we proposed a TRUA (Topic-Related User Authority) algorithm based on PageRank[13] to measure user authority on topics.

M. Cataldi et al. [5] Calculate user authority considering the number and the authority of followers based on PageRank. Hence, given a user $u_i \in V$, its authority is computed as follow:

$$Inf(u_i) = d \cdot \sum_{u_j \in fan(u_i)} \frac{Inf(u_j)}{\left|following(u_j)\right|} + (1-d) \qquad (1)$$

where $fan(u_i)$ represents a set of users who follow u_i, $following(u_j)$ represents a set of users that u_j follows, $d \in (0,1)$ is a dumping factor which is generally set to 0.85.

The user authority iteration process based on PageRank may be regarded as a random walk process: it randomly selects a node and transfers its authority value to one of its concern users with a certain probability in every step of the migration process. In formula (1), the transfer probability of each user is the same, that is:

$$\Psi_{j,i} = \frac{1}{\left|following(u_j)\right|} \qquad (2)$$

where $u_j \in fan(u_i)$.

However, the authority of the user is different within different topics. Thus, the transfer probability should be distinguished according to different concerned users. As shown in Fig.2, $U_{t,tp}$ is the set of users who talk about topic *tp* in time slot *t*, and $\overline{U_{t,tp}}$ is the set of users who are not involved in topic *tp* in time slot *t*.

There are two kinds of users who are followed by u_j. One belongs to $U_{t,tp}$, such as u_i, u_2, u_3 in Fig.2 while the other belongs to $\overline{U_{t,tp}}$, such as u_4. Obviously, the transfer probability from u_j to users in $U_{t,tp}$ should be larger than that from u_j to users in $\overline{U_{t,tp}}$. Meanwhile the more common concerned users u_i and u_j follow, the more similar between them. And they are more likely to focus on the same topic.

Therefore, we define TDR (Topic dependence relationship) to denote the related degree between two users. TDR is defined as follows:

$$TDR_{j,i} = \frac{|UC_{j,i} \cap U_{t,tp}| + \xi \cdot |UC_{j,i} \cap \overline{U_{t,tp}}| + 1}{|UC_{j,i}|} \quad (3)$$

where $UC_{j,i} = following(u_j) \cap following(u_i)$.

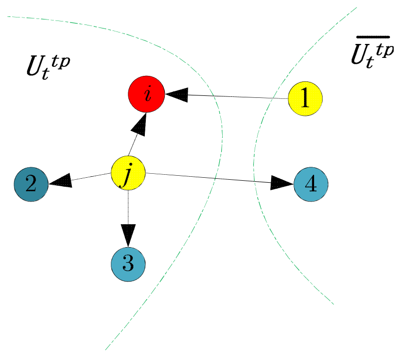


Fig.2. User relationships

Hence, the transfer probability is modified as followings.

$$\Psi_{j,i} = \frac{TDR_{j,i}}{\sum_{k \in following(u_j)} TDR_{j,k}} \quad (4)$$

Accordingly, we modify the user authority updating formula as:

$$Inf(u_i) = d \cdot \sum_{u_j \in fan(u_i)} \Psi_{j,i} \cdot Inf(u_j) + (1-d) \quad (5)$$

The initial authority of $u_i \in U$ is defined as total TDR value with his fans.

$$Inf^{(0)}(u_i) = \sum_{u_j \in fan(u_i)} TR_{i,j} \quad (6)$$

At each step, the algorithm recomputes the authority values as:

$$Inf^{(k)}(u_i) = d \cdot (\sum_{u_j \in fan(u_i)} \Psi_{j,i} \cdot Inf^{k-1}(u_j)) + (1-d) \quad (7)$$

The process terminates when a convergence condition is satisfied.

B Topic energy calculation method

Based on the aging theory, we define the topic energy by considering the number of tweets related to a topic and user authority in this section. It helps to quantify topic's activity in each time slot. According to aging theory, the *energy value* is transformed from the *accumulative support* of the topic in current time slot. And the attenuation of the *accumulative support* is the core process of aging theory, which is a weighted sum of the history *accumulative support* and the *nutrition* of current intake. Before introducing our aging scheme, we first explain how to calculate the nutrition value of the topic in each time slot.

The *nutrition* value of topic *tp* obtained in the *t*-th slot depends on the number of related tweets and the topic influence. The ***topic influence*** in time slot *t* is defined as the total authority of users who discuss the topic in formula (8):

$$TPInf_{t,tp} = \sum_{tw_i \in TW_{t,tp}} Inf(user(tw_i)) \quad (8)$$

where, $TW_{t,tp}$ represents the set of tweets related to the topic *tp* in *t*-th time slot, and $Inf(user(tw_i))$ represents the authority of the owner of $tw_i \in TW_{t,tp}$, which is calculated by TRUA.

Then, the nutrition value of topic tp in slot t is defined as:

$$Nutr_{t,tp} = |TW_{t,tp}| \cdot TPInf_{t,tp} \quad (9)$$

After giving the definition of *nutrition*, we use the recursive decay(RD) as our aging scheme. It employs an exponential smoothing approach to model a topic's energy decay. Since we apply this aging scheme to calculating the *accumulative support*, the *accumulative support* value of topic *tp* obtained from time slot 1 to *t* can be represented by the follow expression:

$$S_t = \alpha \cdot [\beta S_{t-1} + (1-\beta) Nutr_{t,tp}] \quad (10)$$

where, $S_0 = 0$, $0 \le \alpha \le 1$, $0 \le \beta \le 1$, parameter α is the support transfer factor, which decides the influence of *nutrition* on the life of a topic and parameter β, called the support decay factor, governs the pace of aging [12].

And we adopt a sigmoid function as our energy function, which is defined as follows:

$$E(S) = \begin{cases} \frac{S}{1+10 \cdot S}, & S > 0 \\ 0, & otherwise. \end{cases} \quad (11)$$

The energy function transforms the unlimited range $[0, \infty)$ of the *accumulative support* S_t into a limited *energy value* within $[0,1)$. With a limited range, we can give a meaning to each segment of the range and interpret the status of an event by its *energy value*.

The energy function needs to satisfy the following constraints:

$$E(\min\{S_1, S_1, ..., S_T\}) \ge s_1 \quad (12)$$

$$E(\max\{S_1, S_1, ..., S_T\}) \ge s_2 \quad (13)$$

where $1 \ge s_2 \ge s_1 \ge 0$, s_1 represents the lowest energy threshold value of a topic in its life cycle, and s_2 represents the lowest energy threshold value that the *energy peak* of a topic must exceed it. These two threshold values guarantee that the topic *energy value* is large enough in the ***stability*** period of a topic's life cycle, and the sum of *energy value* of each time slot in a topic's life cycle is much larger than that outside the cycle. Simultaneously, we can obtain those two parameters through converting the problem of the calculation of them to a constraint satisfaction problem(CSP) problem [14]. We will discuss it in the experimental section.

C Topic trend prediction

Section III has introduced that there are two important indicators to analyze the stage of a topic in the life cycle: 1) the *energy value* of a topic; 2) the *fluctuation* of the *energy value* of a topic in a short period of time.. Part B has introduced the method to calculate the *energy value* of a topic. And we will introduce a wavelet analysis method to quantify the extent of *fluctuation* in a sequence of *energy value*s.

1) Wavelet transformation

Wavelet analysis is used to accurately characterize the frequency characteristics of the signal change in the timeline. It is a localized analysis of the signal at the time (space) frequency. The signal ultimately achieves time segmentation in area of high frequency and frequency segmentation in the area of low frequency through multi-scale thinning on it gradually. Wavelet transformation is the core of wavelet analysis, The basic idea of it is to use a wavelet family to represent or approximate a signal. Wavelet family is a set of linearly independent wavelets, which are generated by scaling and translating a chosen mother wavelet . Since we take a time period of the energy curve as a finite-length discrete signal, we use orthogonal Haar wavelet, used in discrete wavelet transformation (DWT) commonly, as the mother wavelet.

Given the mother wavelet, we can deal with a fragment of the energy curve by multiscale wavelet decomposition. Firstly, the energy sequence in a time period from *t-w* to *t* is set to a signal, i.e. $Signal_t = \{E_k \mid k = t-w+1+1\cdot r, t-w+1+2\cdot r, ..., t\}$, where, *w* represents the number of review time slots and *r* is the signal sampling rate. We take an *energy value* every *r* time slots in the review period. Through the multi-scale decomposition, a signal can be represented by the following approximate expression:

$$Signal_t = \sum_{j=1}^{N_j}\sum_{i} C_{j,i}\cdot\psi_j^i(t) \tag{14}$$

where $\psi_j^i(t)$ is the *i*-th wavelet in scale *j*, N_j is the total number of layers scaling, valued $\log_2\frac{w}{r}$, and $C_{j,i}$ is the wavelet coefficient. Then, the signal energy in the *j*-layer wavelet decomposition is:

$$SE_j = \sum_i C_{j,i}^2 \tag{15}$$

Thus, the whole energy of the wavelet is:

$$SE_{all} = \sum_{j=1}^{N_j} SE_j \tag{16}$$

We use wavelet energy of the energy curve to quantify the *fluctuation* of the energy. We define *TEF*(Topic Energy Fluctuation) on topic tp in time slot t as:

$$TEF(t,w) = SE_{all}^t \tag{17}$$

The larger the *TEF* value is, the more obvious the *fluctuation* is.

2) Prediction rules

After defining the topic *energy* and *TEF*, we combine these two indicators and present a topic trend prediction rule displayed in Algorithm 1. It can judge in which stage of the life cycle a topic is, as well as predict when a topic will burst in advance.

Algorithm 1 *TrendPrediction*(*t*)

1. $status \leftarrow null$
2. **if** $s_1 < energy[t] < s_2$ **then**
3.   **if** $!hasBurst = false \,\&\, h(t,w) < h_1$ **then**
4. $status \leftarrow infancy$
5.   **else if** $isEnergyUp(t,w) \,\&\, h(t,w) > h_2$ **then**
6. $status \leftarrow burst$
7.   **else if** $isEnergyUp(t,w) \,\&\, h_1 < h(t,w) < h_2$ **then**
8. $status \leftarrow growth$
9.   **end if** $isEnergyDwon(t,w) \,\&\, h(t,w) > h_1$ **then**
10. $status \leftarrow decay$
11.   **end if** $hasBurst \,\&\, h(t,w) < h_1$ **then**
12. $status \leftarrow death$
13.   **end if**
14. **else if** $energy[t] > s_2$ **then**
15. $status \leftarrow stability$
16. **end if**
17. **return** ***status***

Threshold parameters h_1 and h_2 are used in the rule, where, h_1 represents the maximum *TEF* value a topic own when we believe it of low *fluctuation*. It can help to determine whether the topic is in its ***infancy***, ***stability*** or ***death*** period or not. And h_2 represents the minimum TWE value when we believe a topic in its ***burst*** period, that means the number of the topic's related discussions increase rapidly in this period and the topic is reaching the *peak* of the discussion. Note that the ranges of topic *energy value* and topic energy *fluctuation* are both similar in ***infancy*** and ***death*** period, so line 3 and 11 in the rule need to consider whether the topic has experienced the *burst* period.

V EXPERIMENTS

A Experimental Dataset

Dataset used in the experiments is crawled from Sina Weibo, which spans from 2013/06/01 to 2014/03/15. After pre-processing, we choose 792,000 valid tweets and 335,862 users including the users who post the tweets in the dataset and their followers and their concerned users. We figure out 120 topics, divided into four categories: entertainment, politic, sports and accident/disaster. Number of Labeled events in every category is 30.

B Training parameters α and β

Section IV converts the problem of calculating parameters α and β into a CSP problem. In this section we use aima-java toolkit to solve the optimal parameter values. Just like [12], we train the parameters for different categories respectively and the results are listed in Table 1. Fig.3 shows four life cycle graphs of four corresponding topics: "Yang Mi and Hawick Lau wedding", "Two-child policy", "Evergrande F.C wins Afc champion" and "Deadly knife attack in Kunming".

The parameter α indicates the ability of *nutrition* value being converted to the *energy value*. The topic of political has less discussing audiences than the other three categories. In order to distinguish the level of activity at different periods of its life cycle more clearly, it needs higher conversion factor α. So, the political topic gains the largest α value of 0.0014121 in the training result.

The decay factor β governs the pace of aging. We can learn from Table 1 that *disaster*(*accident*) related topics have higher decay factor relatively which means slower decay speed. As Fig.3d) shows, the topic “Deadly knife attack in Kunming” is related to public life safety, which causes not only the majority of ordinary users of expressing their views, but also a large number of government agencies and public figures to track the course of the event. It makes the topic decay slowly and maintain in a highly active state in a long time.

Besides, the *entertainment* and *sports* related topics gain smaller β. These two types of topics will lead to a lively discussion from a lot of users before and after the occurrence of them. However, attention will soon fade. Both "Yang Mi and Hawick Lau wedding" and “Evergrande F.C wins Afc champion” shown in Fig.3a) and Fig.3c) have shorter period of time in the stage of ***stability*** and have relatively shorter survival period than “Deadly knife attack in Kunming”. With narrow user groups participating in discussion, most political topics surge and plummet and decay faster, like topic “Two-child policy” shown in Fig.3b) .

In summary, the method we proposed to depict the stages of the life cycle of a topic conforms to the actual situations. Besides distinguishing the categories of topics can reflect the actual situations more accurately.

C   *Topic trend prediction*

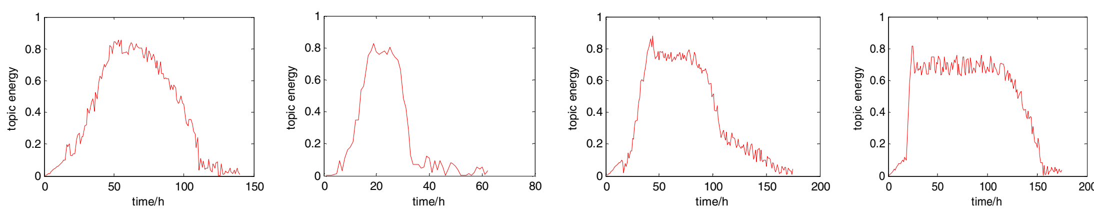
1)   Trend prediction analysis

According to prediction rule proposed in part C of section IV, we will use a specific example to explain how to analyze the trend of a topic. Wherein each parameter is set to: $\xi = 0.7$, $w = 4$, $r = 1$, $h_1 = 0.005$, $h_2 = 0.01$. We choose a movie “Personal Tailor”, which is directed by Feng Xiaogang and released on December 19, 2013 as an example in this section. Fig.4 shows the topic energy and *TEF* value changing with time.

As shown in Fig.4, a small increase of the topic *energy value* and the *TEF* value appears at slot 25 when some related publicity carried by the movie organizers cause small part of users to discuss the upcoming movie on the microblog, but both of them are small and the topic is in its ***infancy*** period. As time closer to the movie Showtimes, with the intensified publicity and the influence of internet water armies, the topic energy and topic *TEF* value increase distinctly, when the topic arrives its ***growth*** period. At the 57^{th} time slot, when the move releases, the topic *TEF* value reaches a maximum that means the topic reaches the ***burst*** point and the energy of topic will hit the peak in a short time. Then, the topic enters its ***stability*** since its topic *energy value* is greater than s_2 and its topic *TEF* value is smaller than h_1 after time slot 60. After maintaining the stage of ***stability*** of 10 time slots, the energy begins to decline, but the topic *TEF* value is greater than h_1, we update the topic status ***decay***. With the passage of time, the *energy value* gets smaller and *TEF* value tends to 0, which illustrates that the topic is dying. The experimental analysis above shows that the topic *TEF* value can react the trend well.

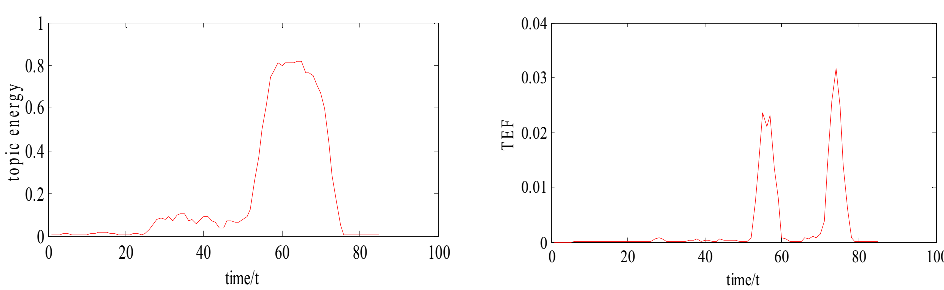
TABLE.1.   TRAINING RESULTS OF PARAMETERS α AND β

Category	Entertainment	Politic	Sports	Accident /Disaster
Energy threshold	s_1	0.002	s_2	0.85
$\overline{\alpha}$	0.0001016	0.0014121	0.0001052	0.0001121
$\overline{\beta}$	0.0035	0.0021	0.0064	0.0101



a) Y M & H L wedding   b) Two-child policy   c) E.F.C wins Afc champion   d) Deadly knife attack in Kunming

Fig.3.   Life curve of four topics



a) Topic energy   b) Topic energy fluctuation

Fig.4.   The energy and energy fluctuation of the film “Personal Tailor”

TABLE.2. PREDICTING RESULTS USING DIFFERENT APPROACHES

Methods	The average time between the forecast point and the energy peak(hour)	omission rate = the number of topics that the forecast point is after the energy peak / total number of topics(%)
PTEP	3.32	2.21
The method in [15]	5.53	10.63
Threshold θ =0.8*s2	4.93	12.37
Threshold θ =0.85*s2	4.56	19.33
Threshold θ =0.9*s2	3.26	26.76
Threshold θ =s2	2.33	37.33

2) Evaluation of predicting result

In order to verify the effectiveness of the proposed topic trend prediction rule, we must prove that the rule can accurately predict topic forecast point, since people are more interested in when topics will burst and be discussed widely. Thus, the public opinion monitoring centers could know in advance the topic forecast point and make the necessary measures before topic erupts.

We define the *forecast point* of a topic as the moment when the topic energy *fluctuation* maximizes in the *burst* period of its life cycle. Wherefore a topic will erupt in a short time after it reaches its own forecast point. We carry out the comparison among our method, the method proposed in [15] and fixed energy threshold methods. The comparison includes the value of the time between *forecast point* and *energy peak* and whether *forecast point* occurs after the *energy peak*.

Table 2 shows that the omission rate of the our method is 2.21%, and the time distance between the *forecast point* and the *energy peak* is 3.32h on average, that meets the time requirements of the public opinion monitoring center on anticipation of topic *energy peak* and makes our prediction method more meaningful when applied in micioblog data. Though the method proposed in [15] has the similar time distance value with ours, but its omission is much greater. Moreover, simply fixed energy threshold methods performs poorly that higher energy threshold values lead to shorter time distance but lager omission rate and lower energy threshold value leads to longer and meaningless time distance.

VI CONCLUSION

This paper develops a model of topic life cycle and accents on predicting the time when a hot topic will outbreak. Firstly, we measure the user authority by taking the number and the authority of its followers and users' interest to a topic into consideration. Then we calculate the *energy value* considering tweets and users authority related to the topic. Thirdly, we measure the *fluctuation* of the *energy value* based on wavelet transformation. Finally, we present rules to evaluate which stage a topic is in and predict the time when a hot topic outbreaks. Experimental results show that the method proposed can effectively predict the outburst of a topic in advance with a low omission rate.

There are still some improvements need to be researched. The future work will be focused on gaining more appropriate parameter α and β to the current detecting topic. And we will explore new wavelet function, which is more suitable for the characteristics of microblog in subsequent research work.

ACKNOWLEDGEMENTS

The authors would like to thank the support of the Technology Innovation Platform Project of Fujian Province under Grant No. 2009J10027, the Key Project of Fujian Education Committee under Grant No. JK2012003, the Program of National Natural Science Foundation of China under Grant No. 61103175 and 61370210, the Natural Science Foundation of Fujian Province under Grant No. 2013J01232.

REFERENCES

[1] A. Kontostathis, L. M. Galitsky, W. M. Pottenger, et al, "A survey of emerging trend detection in textual data mining," Survey of Text Mining. Springer New York, 2004, 185-224.

[2] J. Allan, J. G. Carbonell, G. Doddington, et al, "Topic detection and tracking pilot study final report". 1998.

[3] L. Chen, W. Wang, A. P. Sheth, "Are Twitter users equal in predicting elections? A study of user groups in predicting 2012 US Republican Presidential Primaries" Social Informatics. Springer Berlin Heidelberg, 2012. pp. 379-392.

[4] S. Asur, B. A. Huberman, G. Szabo, et al, "Trends in Social Media: Persistence and Decay" Technical report. HP Laboratories, 2011.

[5] M. Cataldi, L. Caro, C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM, 2010.

[6] J. H. Lau, N. Collier, T. Baldwin, "On-line Trend Analysis with Topic Models:#twitter Trends Detection Topic Model Online," in Proceedings of the 24th International Conference of on Computational Linguistics, 2012, pp. 1519-1534.

[7] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.

[8] T. L. Griffiths, M. Steyvers, "Finding scientific topics," Proceedings of the National academy of Sciences of the United States of America, vol. 101, pp. 5228-5235, 2004.

[9] D. Y. Ou, Y. H. Xiao and W. Wang, "Statistical Model—Based Analysis and Prediction of Collective Attentions in Social Networks," Journal of Computer Research and Development, vol. 70, pp. 378-384. 2010.

[10] Y. Tian, "On Trends Analysis and Prediction Based on Micro-Blogging Platforms," Wuhan University, 2012.

[11] C. C. Chen, Y. T. Chen, Y. Sun, et al, "Life cycle modeling of news events using aging theory," Machine Learning: ECML 2003. Springer Berlin Heidelberg, 2003, pp. 47-59.

[12] C. C. Chen, Y. T. Chen, M. C. Chen, "An aging theory for event life-cycle modeling," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 37, pp. 237-248, 2007.

[13] L. Page , S. Brin , R. Motwani, et al, "The PageRank citation ranking: Bringing order to the web,". 1999.

[14] V. Kumar, "Algorithms for constraint-satisfaction problems: A survey," AI magazine, vol. 13, pp.32, 1992.

[15] R. Lu, Y. Zhang, Q. Yang, "Predicting News Trend in the Social Network," Jouranl of Chinese Information Processing, vol. 26, pp.85-90 2013.