2012 4th International Conference on Intelligent Human-Machine Systems and Cybernetics

A Research of Hot Topic Detection through Microblogging

Feifei Peng Xu Qian
School of Mechanical Electronic & Information Engineering
China University of Mining and Technology, Beijing
Beijing, P.R.China
feifei.gaoren@gmail.com

Gaoren Li
Game Career Department
TieXue Science and Technology Limited Company
Beijing, P.R.China
ligaoren@gmail.com

Abstract—**Micro-blog is one of the internet applications, and it is newly developed online communication platform to achieve real-time information sharing. Micro-blog becomes quite popular. We analysis the researched approach of detecting hot topic, combine with a variety of related parameters of hot topic, and implement the detection of hot topic through Micro-blog. This paper presents a hot topic detection and extraction algorithm based on users' preference. The experimental results show that the approach can effectively detect hot topic in the Micro-blog, and can effectively remove the false hot topic which have ambiguous information.**

Keywords- user preferences; mash-up; pearson correlation coefficient; complete negative correlation; weight factor

I. INTRODUCTION

With the popularity of the network and internet technology advances, internet has become the platform of the expression of interest of social strata, emotional catharsis and thought collision. It is also the platform of expressing public opinion, discussing public affairs and participating in economic social and political life. Internet users become the main characters and publish, acquire, evaluate and propagate more than a number of one billion messages every moment. At the same time there is more and more spam in network [9].

Micro-blog with an open internet application is a platform of sharing, propagating and acquiring information based on user's relationship. Users can construct individual communities using WEB, WAP and other kinds of client components in Micro-blog. The length of text information in Micro-blog is limited in 140 words. And the text information owns instant sharing. Twitter is one of the earliest and most famous forums for Micro-blog in online social networks. According to the relevant public data, it owns 75million registered users until January 2010. Sina.com Micro-blog appears in August 2009 in China, and it became the first portal website to provide Micro-blog service [1]. For example, from "Miyagi Prefecture in Japan 7.4 earthquake" to "nuclear radiation" from the "nuclear radiation" to "salt crisis", and so on, more and more public topics were derived from the Micro-blog. In a word, it is essential to detect these topics from Micro-blog.

The remainder of this paper is organized as follows: in section2, we introduce previous works. We define the key terms and introduce the related ideas and adopted algorithms in this section. Section3 describes our novel method for hot topic detection. In section4, we demonstrate the superiority of our approach with comparative empirical results. Finally, in section5, we present our conclusions and some future research directions.

II. PREVIOUS WORKS

There are many related research for information propagation through networks. We provide the based concept----topic and hot topic, with pointers to more detailed survey works, and give some details around recent work.

A. Definition

- **Topic:** in 1999, a topic for some reason refers to the condition caused in a specific time, place, and may be accompanied by some of the inevitable result of an event [5]. However, a topic is defined as a seminal event or activity, along with all directly related events and activities [2]. Thus, we can infer that a topic consists of events and activities, both of which are defined in greater detail in. A TDT event is defined as something that happens at a specific time and place, along with all the necessary preconditions and unavoidable consequences. Such an event might be a car accident, a meeting, or a court hearing. A TDT activity is a connected series of events with a common focus or purpose that happens in specific places during a given time period [3]. For example, a TDT activity may be a scientific research, a natural and man-made disaster, an election campaign, an investigation, or a disaster relief effort and so on.
- **Hot topic:** a "hot topic" is defined as a topic that appears frequently over a period of time in [4]. The factor of "how often topics appear" indicates that topics are hot topics, but it is not the only condition. Everything has its life style----origin, development, recession and death [6], hot topic is no exception and it has its life time. So, time is a significant condition [11]. We can get the basic characteristics of hot topic: at first, simple. It is the basic requirement of hot topic. Hot topic is the topic of higher frequency in the huge amounts of information. Second, real-time. Hot topic is extracted from the current posts of

978-0-7695-4721-3/12 $26.00 © 2012 IEEE
DOI 10.1109/IHMSC.2012.53

IEEE computer society

network users. It must be real-time because we should ensure the "hot" of hot topic. Hot topic has its life cycle, so we must detect the topic in a specific period, and ensure hot topic's real time. Third, falsity. There are many important factors, such as the ambiguity of time and event information. Hence, there are a lot of false hot topics.

B. *Related Approach*

In the research of burst event, some related posts have strong synchronicity. Time is playing a constructive role in the affairs of the judgment of emergencies [10]. In [7], time-based language model considers the correlation between time and events. There are two main drawbacks. There is the identification process with no time-related events. Xiaoyan Li test result selected manually. In [8], Fernando puts forward some characteristics of time and related reference index. Propose an approach of query accuracy of prediction based on time, this approach estimates probability of a particular time in the query, and then, propose four time-related features based on the probability.

Much previous research investigating the flow of information through networks has been based upon the analogy between the spread of disease and the spread of information in networks. Topic is time-sensitive and can be discussed drastically in a short time, then, concerned extent rapids decay. During the whole propagation process there is once significant peak in the propagation rate curve. Besides, users with similar interesting always discuss the same topics. We will detect hot topics in this community.

III. Hot Topic Detection and Extraction

This paper is concerned only with the propagation characteristics of a sudden event topic. The content of such topic is relatively simple, and it can reflect the propagation characteristics of hot topic in social network. In order to solve above problems mentioned in the previous section, this paper researches a hot topic detection and extraction algorithm----HTDEA. Next, this paper will introduce detailed algorithm used in the parameters and methods, and the implementation of the algorithm process.

A. *Pretreatment*

At first, collect metadata from Micro-blog. In order to facilitate fast clustering, we aggregate network information resources based on mash-up and sort information according to weight factor. Users' post views and users' post reply quantity in the Micro-blog can reflect the concerned extent. And the words in concerned posts most likely to be hot topics. Besides, we also consider the temporal characteristics of hot topics and their life cycle. Weight factor is calculated as:

$$\lambda(post) = x * Published\_Number(post) + y * Forwaded\_Number(post) \tag{1}$$

Where,

- $\lambda(post)$ is weight factor of post.
- *Published_Number(post)* is the number of users' published post.
- *Forwaded_Number(post)* is the number of users' forwaded post.
- x and y is variable coefficient, and the sum of x and y is 1.

B. *Clustering*

Hot topic is described in this paper as:

$$\begin{aligned} HT(t) = (&HW_1, HW_2, \ldots HW_i; \\ &TF_1, TF_2, \ldots TF_j; \\ &U_1, U_2, \ldots U_k) \end{aligned} \tag{2}$$

Where,

- HW_i is on the behalf of hot words, that is, words with higher frequency and relate with hot topic.
- TF_j is the frequency of words.
- U_k is the users of owning hot words.

Clustering is the key technology of hot topic detection, and correlation (user preferences) is an important factor in the results of clustering. Correlation is measured by distance. This paper uses Pearson Correlation Coefficient to measure related distance. A user preference is calculated as:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{N}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{N}\right)\left(\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{N}\right)}} \tag{3}$$

Where,

- r measures an extent of linear correlation between variable X_i and variable Y_i. It will return a number in [-1~1].
- When r=1, it means that two users have the same topic.
- When r=0, it means two users are independent.
- When r=-1, it means complete negative correlation.
- n is topics' number.
- N is the number of topics.

The weight of hot words is calculated as:

$$W_i = \frac{TF_i(t,p)\log(\frac{N}{DF(t)} + 0.01)}{\sqrt{\sum_k TF_i^2(t,p)\log^2(\frac{N}{DF(t)} + 0.01)}} \tag{4}$$

Where,

- W_i is the weight of hot words i.

- $TF_i(t, p)$ is the topic t frequency in post p.
- N is the whole of post number.
- $DF(t)$ is number of post containing t.

Then, the average of r is calculated as:

$$AVG(r) = \frac{\sum_{i=1}^{n} w_i r_i}{\sum_{i=1}^{n} w_i} \qquad (5)$$

C. Algorithm Design

There are many types of text and data, such as structured, semi-structured and unstructured text and data. In order to make for preprocessing these text and data, this paper uses mash-up technology based on context classify, and aggregates types of text and data. Then, we extract words and sort these words with the weight of post *λ(post)*. In the clustering process, we calculate the related extent *r* and words frequency $TF_i(t, p)$. And then calculate the average of *r* through the formula of weighted average. We can measure the concerned extent with the average of *r*. Finally, we use bubble sort to sort the extracted words, and take the first ten as a set of hot topic.

This paper researches a hot topic detection and extraction algorithm----HTDEA. The HTDEA is to detect and extract topics from large amount of data sets. The algorithm description of HTDEA is shown in figure 1. It details description of realizing process of the algorithm.

```
Algorithm: An algorithm of detecting hot topic
Function:  HT_Dectection (int p)
Begin
Initialization parameters;
Hot words←hotwords_extraction(P);
Printout Hot words[k];
Printout words frenquency[i];
Sort_hotwords(Hot words);
Printout Sort_hotwords(Hot words);
While ∃ i∈ n do
r'  ← r(i);
w' ← w(i);
AVG'(r') ← AVG(r);
Printout AVG'(r');
While (cluster (i), cluster (j)>number) do
                cluster (j) = cluster (j) + cluster (i);
    j++;
    printout result(hot topic);
End
```

Figure 1. Discription of algorithm

In order to better understand the execution of the algorithm steps, as shown in figure 2, the flow chart of algorithm is given.

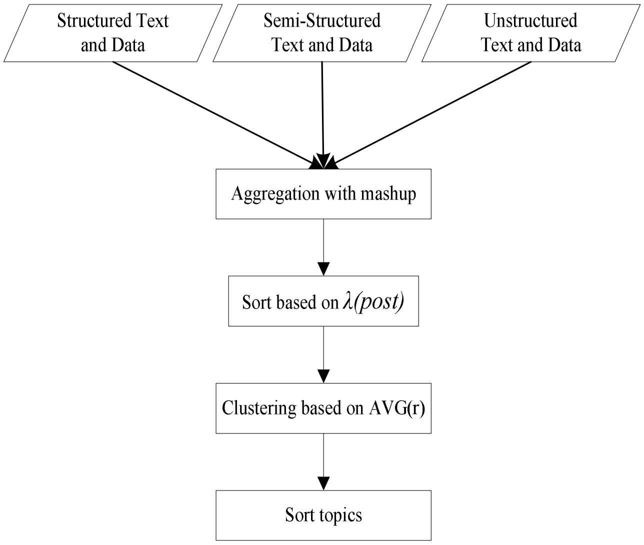
Structured Text and Data
Semi-Structured Text and Data
Unstructured Text and Data
Aggregation with mashup
Sort based on λ(post)
Clustering based on AVG(r)
Sort topics

Figure 2. Flow chart of algorithm

IV. EMPIRICAL VERIFICATION

A. Experimental Environment and Data Source

Use the server of laboratory and programming language----python.

Different approaches are available to solve the same problem, but the quality of an approach would affect the approach as well as the efficiency of the procedure. Randomly select fifty users' posts as a test corpus through sina.com Micro-blog within a month, from April 15th, 2011 to May 14th, 2011. We can get 25,087 messages published in every week. Then, retrieve each user's posts. At last we collect a total of 101,032 posts from fifty users. As shown in figure 3.

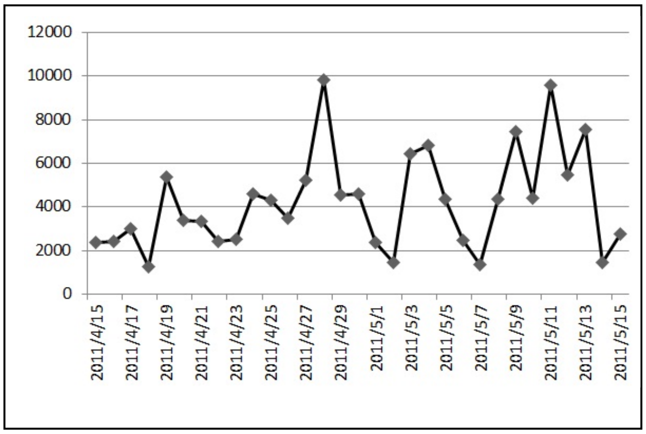

Figure 3. Number of postings in a month

B. Evaluation Standards Experimental Result

Choose evaluation standards are as follows:

Accuracy: $Accuracy = \frac{num_i}{num_i + num_j}$

Recall rate: $recall = \frac{num_i}{num_i + num_k}$

F-measure: $F - measure = \frac{2 \times Accuracy \times recall}{Accuracy + recall}$

Where, num_i is the number of hot topics extracted and they can reflect the theme of the posts and the interest of the user; num_j is the number of hot topics extracted and they cannot reflect the theme of the posts and the interest of the user; num_k is the number of hot topics cannot extracted but these hot topics can reflect the theme of the posts and the interest of the user.

C. Experimental Result

We use *λ(post)* sorting these posts at first. And then we use a method of hot topic detection based on user preferences. At last, only take the topic of the top ten. The preliminary result is shown in Table I.

TABLE I. EXPERIMENTAL RESULT

NO	Topic	Topic Frequency	AVG(*r*)
1	rent rises	1155	0.96
2	drunk	759	0.85
3	Palace Stolen	578	0.73
4	The Fast and The Furious5	466	0.72
5	Tsinghua Anniversary	379	0.69
6	Wenchuan Earthquake	353	0.58
7	Prince William	231	0.55
8	NBA	199	0.49
9	Bin Laden	176	0.45
10	Cheongsam	115	0.39

According to the hot topics got from the proposed algorithm---HTDEA comparison with hot words offered from some related sites, we could know that hot topics in Table I were of higher heat, and had higher representative.

TABLE II. PERFORMANCE COMPARISON

Comparisons	Algorithms		
	HTDEA	Time	*Word Frequency*
accuracy	82.6%	75.2%	74.8%
recall rate	76.9%	66.3%	64.7%
F-measure	79.6%	70.5%	69.4%

Besides, in allusion to collect the total of 101,032 posts from fifty users, this paper compares HTDEA with other traditional detection and extraction algorithms: detection and extraction algorithm based on word frequency and detection and extraction algorithm based on time distribution. We extract hot topics from above obtained posts. We find that the accuracy of HTDEA is higher than other two algorithms', and then preliminary results are shown in Table II.

The preliminary results show that the HTDEA could more effectively extract a hot topic from micro-blog and more effectively remove the false hot topics. The extraction efficiency of HTDEA is higher, thus it could ensure a soft real-time sharing of information.

V. CONCLUSION

Micro-blog is a chart platform for network users. More and more users use it for communicating with other network users. Micro-blog is popular with people's life, study and work and so on. In this paper, we introduce the definition of topic and hot topic at first. The next, we analyze related parameters of hot topic detection, such as time, word frequency, user preferences etc. And then we propose an approach of hot topic detection based on weight factor and user preferences. The experimental result shows that this method can effectively detect hot topics.

ACKNOWLEDGMENT

China's postdoctoral scientific fund projects (NO: 2011M500035);

Higher school the special research fund for the doctoral program special (NO: 20110023110002)

REFERENCES

[1] Micro-blog, http://baike.baidu.com.

[2] The 2004 Topic Detection and Tracking (TDT '04) Task Definition and Evaluation Plan, http://www.nist.gov/speech/tests/tdt/, 2004.

[3] TDT 2004: Annotation Manual Version 1.2, http://www.nist.gov/speech/tests/tdt/, Aug. 2004.

[4] K. K. Bun and M. Ishizuka, Topic Extraction from News Archive Using TF*PDF Algorithm, Proc. Third Int'l Conf. Web Information Systems Eng. (WISE '02), pp. 73-82, 2002.

[5] Allan J, Carbonell J, Doddington G, et al. Topic Detection and Tracking Pilot Study: Final Report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop [C]. Lansdowne, Virginia. February 1998.

[6] C. C. Chen, Y. T. Chen, Y. Sun, and M. C. Chen, Life Cycle Modeling of News Events Using Aging Theory, In Proceedings of the fourteenth European Conference on Machine Learning, pp. 47-59, 2003.

[7] Xiaoyan Li and W. Bruce Croft, Time-Based Language Models, Proceedings of the twelfth international conference on Information and knowledge management, 2003.

[8] Fernando Diaz, Rosie Jones, Using Temporal Profiles of Queries for Precision Prediction, ACM 2004 Article, 2004.

[9] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgrlio Almeida, Detecting Spammers on Twitter, CEAS 2010 - Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, July 13-14, 2010.

[10] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, WWW2010, April 26-30, 2010.

[11] Dazhen Lin, Shaozi Li, Donglin Cao, Blog Emergent Event Detection Based on Temporal Distribution, Computer Engineering and Sciemce, pp. 145-148, 2010.