Detecting Hot Topics in Sina Weibo Based on Opinion Leaders

Donghui Li, Yuqing Zhang, Xin Chen, Long Cao, Chuanfeng Zhou

College of Information Engineering China University of Geosciences, Beijing, China
lidonghui1209@163.com, yqzhang@cugb.edu.cn

Abstract - Sina Weibo, as one of the most popular and fast growing social network, has gradually become the field where hot topics appear, propagate, and outbreak. In order to discriminate and find out hot topics in micro-blog information, we conduct a series of studies on Sina Weibo, and one of our key findings is that opinion leaders play a very important role in the propagation of hot topics. A smart discriminant model is proposed in this paper to detect hot topics in time, which takes the structure information and propagation characteristics of Sina Weibo as well as the users' influence into consideration. Moreover, word co-occurrence graph is used to extract and display topics. This model has some excellent characters such as a low coupling degree between modules and a low requirement for the amount of data. By experimental verification, it can detect hot topics effectively.

Index Terms - Sina Weibo, hot topic, discriminant model, opinion leader, word co-occurrence graph

1. Introduction

Micro-blog is a platform based on users' relationship which we can share, communicate and get information with the limitation of 140 words. The earliest micro-blog is twitter originated in the United States in 2006. During the next few years, there is an explosive growth of all kinds of micro-blog products. In August 2009, Chinese portal website Sina launched the beta version of "Sina Weibo", and then micro-blog entered the view of Chinese netizen. Till the first half of 2013, registered users of Sina Weibo have reached a number of 536,000,000. Through micro-blog, users can continuously comment on a piece of information and forward it without the restriction of time and space, which makes it possible that the information could be forwarded by hundreds of thousands times within a very short period of time, and then evolves into a social event that can arouse public concern.

In Sina Weibo, more than 100,000,000 micro-blogs are released every day. Many users login Sina Weibo through mobile terminals, which makes messages transmit faster. Because of the 140 words-limit, messages of Sina Weibo are short texts, in addition, the micro-blog terms are usually not normative and include much network language, which makes it very difficult to analysis the semantic information in Chinese micro-blogs. Therefore, there are three major difficulties in the detection for hot topics in Sina Weibo, vast quantities of data, various propagation forms of topics and complexity of Chinese semantics and data sparseness in short texts.

In order to solve the problems, we conduct a serious of investigations into the propagation characteristics of some hot topics sponsored in Sina Weibo and find that propagation usually expands when a micro-blog is forwarded by an opinion leader who has a large number of fans, however, when the micro-blog spreads to a normal user who has few fans, propagation often terminates here. Each outbreak of hot topics initiated in Sina Weibo has the participation of opinion leaders. Without opinion leaders and mainstream media, a topic just spreads in small circles and has little probability to break out into a social event.

Recent researches on Sina Weibo mainly focus on three directions. One kind of research proposes some methods to measure users' influence. Ref. [1] proposed a method to model user influence by taking consideration the factors including the number of followers and friends as well as the scale of the network. Ref. [2] ranked users by the number of followers and User PR (based on PageRank algorithm) respectively. Ref. [3] selected celebrities of Sina Weibo as the research object, and presented an evaluation model on the impact of micro-bloggers based on three aspects including attention, propagation and participation. Another kind of research is aimed at the propagation model and structural features of Sina Weibo. Sai Ref. [4] proposed a trigonometric sum algorithm to detect a threshold to the number of fans. Ref. [5] used correlative research methods and did some systematically analysis on friends relation network of the opinion leaders in Sina Weibo. Ref. [6] analyzed in detail the temporal aspect of trends and trend-setters in Sina Weibo. There are also some researches talking about topic detection in micro-blog. Ref. [7] provided a method based on word co-occurrence graph to detect news topic of micro-blog. Ref. [8] proposed a novel bursty topic detection technique based on an improved method by calculating term weight which took user weight and the number of listeners, replies and collections into account.

This paper presents a discriminant algorithm based on opinion leaders to detect hot topics in another point of view, and three modules are established to realize the topic detection: data collecting, hot micro-blog discrimination and topic extraction. Based on the model we conduct a serious of experiments, and the experimental results demonstrate that the model can detect hot topics effectively.

The rest of the paper is organized as follows. Section2 introduces our detecting model in detail and Section3 presents the experimental results and evaluations. Finally we conclude in section4.

2. Topic Detection Model

As shown in Fig.1, the model consists of three phases, including data collecting, hot microblog discrimination and topic extraction.

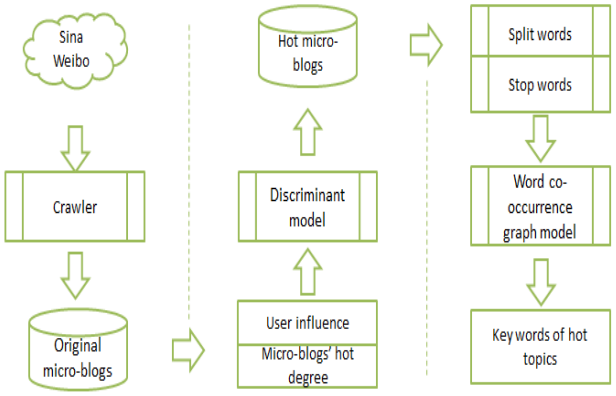

Fig. 1 General framework of our model

First, we make a list of opinion leaders and crawl their information from Sina Weibo at a fixed interval. In the second phase, a discriminant algorithm is devised to measure each micro-blog's popularity degree and screen out the hot micro-blogs. At last, we use the word co-occurrence graph to extract and display topics.

A. *data collecting*

In order to realize our idea, we collected 32 opinion leaders' information and their micro-blogs in Sina Weibo from April 1st 2013 to July 30th 2013. For each opinion leader, we crawled his user ID, nickname, fan number, follow number and blog number. Table I is the opinion leaders' list.

As illustrated in Table II, for each micro-blog, we crawled its information including message id, text, release time, like number, forward number, forward information, comment number and comment information every 30minutes. Several examples of forward information are shown in Table III. "forID" denotes the ID of a piece of forward, "forUid" and "forNick" represent the ID and the nick name of the user who forwards the micro-blog, "forAt" is the user who is mentioned in the piece of forward. Comment information shares the same format with forward information.

TABLE I Opinion leaders' list

OrderID	Nickname	User ID	Fan Number	Follow Number	Micro-blog Number
1	媒体人张刚	1035925393	9563	1316	3121
2	雷颐	1045529987	165224	919	18051
3	李承鹏	1189591617	7217225	609	3933
4	沈宏非	1191258655	2759569	612	3962
5	张颐武	1194868525	4963114	597	3697
6	方舟子	1195403385	4821922	28	11684
…	…	…	…	…	…
30	元芳视角	2688477682	61715	2463	1666
31	杂谈五味	2692988203	154491	1784	15810
32	光头王凯	1650305567	1678420	1005	11811
	Average		2388315	1071	10138

TABLE II A micro-blog's information list

User ID=1187986757 Message ID= 3569808928937298								
	text	Time	Catch Time	like	for Num	for Info	com Num	com Info
ΔT_1	救孩子同样要紧	2013-04-22 08:57	04-22 09:23	25	481		90	
ΔT_2			04-22 09:46	37	602		113	
ΔT_3			04-22 10:30	46	742		130	
…			…	…	…		…	
ΔT_n			04-23 08:55	73	1086		173	

TABLE III Forward information of a micro-blog

	forId	forUid	forNick	forAt
forInfo1	3569815124211619	1625676993	溜溜妹	LeoWong
forInfo2	3569815119373242	1806790127	Darlinglulu	NULL
forInfo3	3569815111262932	2186338954	Jasper	小小亮
…	…	…	…	…

B. *Discriminant model*

User influence in Sina Weibo can be reflected in three aspects: active degree which represents the users' frequency of releasing micro-blogs, dissemination degree which is related to the numbers of forwardings and comments of users' micro-blogs, and coverage degree which depends on the numbers of active fans of users. Based on the three aspects, user influence PR_u is defined in the following way.

$$PR_u = \frac{\sum_{j=1}^{n}(c_j + r_j)}{n} * \frac{N_u}{t} * fansNum * e \quad (1)$$

Where $\frac{\sum_{j=1}^{n}(c_j + r_j)}{n}$ denotes the dissemination degree; N_u / t represents the active degree; $fansNum$ is the coverage degree; e is standardization constant.

There are three main parameters for each micro-blog during each Δt : number of comments $C_{\Delta t}$, number of

forwardings $R_{\Delta t}$, number of likes $D_{\Delta t}$. D means that somebody has looked at the micro-blog and shows interest in it, so we treat $D_{\Delta t}$ as that ordinary users comment on the micro-blog with a word "like".

In order to measure the popularity degree of micro-blogs more accurately, we conduct two judgements to each comment of a micro-blog: Whether the user who makes this comment is an opinion leader? Whether the user mentions other opinion leaders? Thus we can divide C into four new parameters: CU(ordinary users comment on this micro-blog), CV(opinion leaders comment on the micro-blog), CUA(ordinary users comment on the micro-blog and mention other opinion leaders at the same time), CVA (opinion leaders comment on the micro-blog and mention other opinion leaders at the same time). In the same way, we can get four parameters of R: RU, RV, RUA, RVA. Thus, $C_{\Delta t_i}$ and $R_{\Delta t_i}$ can be modified as the following form.

$$C_{\Delta t_i} = \sum_q CU_q + \sum_j CV_j * PR_j + \sum_m CUA_m * PR_m * P_1 + \sum_n CVA_n * PR_n * \Pr_n * P_2 + \sum_k D_k \quad (2)$$

$$R_{\Delta t_i} = \sum_q RU_q + \sum_j RV_j * PR_j + \sum_m RUA_m * PR_m * P_1 + \sum_n RVA_n * PR_n * \Pr_n * P_2 \quad (3)$$

Where PR_n denotes the influence degree of the user who makes the comment or forwarding; $\Pr_n$ represents the influence degree of the user who is mentioned in this comment or forwarding. P_1 and P_2 are standardization constants.

In Eq. (2) and Eq. (3), the numbers of comments and forwardings are weighed and calculated based on the influence degree of opinion leaders instead of simply calculate their numbers.

The hot degree of a micro-blog at Δt_i is defined as follows:

$$H_{\Delta t_i} = C_{\Delta t_i} + R_{\Delta t_i} \quad (4)$$

Since that influence degrees of different users' vary from each other, so do the popularity degrees of different micro-blogs. In view of this situation, an exclusive threshold value K for each user in our opinion leaders' list is defined as follows.

$$K = nH^j_{\Delta t_i} / \sum_{j=1}^{n} H^j_{\Delta t_i} \quad (5)$$

Where $H^j_{\Delta t_i}$ denotes the popularity degree of micro-blog j at Δt_i; $\frac{1}{n}\sum_{j=1}^{n} H^j_{\Delta t_i}$ means the average popularity degree at Δt_i of micro-blogs released by the same user during the past few weeks.

As shown in Fig.2, when K is greater than an appointed threshold at Δt_i, the micro-blog is considered to be popular and sent into hot micro-blog library.

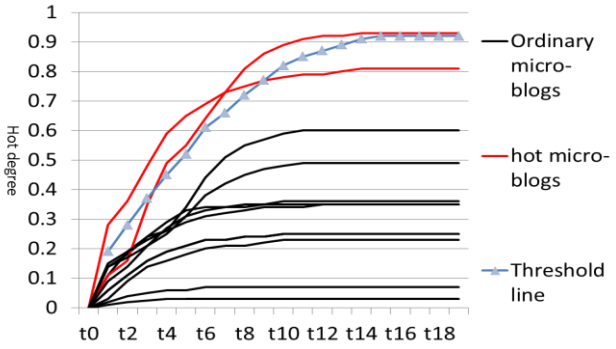

Fig. 2 Discrimination principle of our discriminant

C. *Topic extraction*

In this paper, we choose word co-occurrence graph [7] as our model to extract topics in micro-blog information, and topics are displayed by several key words.

3. Experimental Results and Analysis

In this section, we first present the experimental data used in our test bed and then introduce our experimental method, finally we evaluate the experimental results.

A. *Experimental data*

The experimental data are crawled from 32 opinion leaders in Sina Weibo since May 14th to May 22th with a total of 2376 pieces of micro-blogs. We manually find out four hot topics in these 2376 micro-blogs:

Topic1: A headmaster in Hainan Province raped girls.
Topic2: China Food and Drug Administration emphasizes the issue of food safety.
Topic3: The incident of NONGFU Spring.
Topic4: The incident of TienanLiu in National Energy Bureau.

B. *Experimental method*

In this step, we separate the micro-blogs into three groups and respectively extract topics from these three data sets by using the word co-occurrence graph algorithm.

Data set 1: All the micro-blogs obtained by the crawler without being filtered by the discriminant model.
Data set 2: The micro-blogs selected by the discriminant model which have higher popularity degrees.
Data set 3: The micro-blogs rejected by the discriminant model which have lower popularity degrees.

Table IIII shows the micro-blog numbers of each data set with different threshold values of K.

C. *Experimental results and analysis*

By using the control variate method, we note that different threshold values achieve different performances and the discriminant model achieves best when K=1.0 and the threshold value of co-occurrence degree ranges from 0.3 to 0.5. Table V presents the topics extracted in data set 1, data set

2, and data set 3. Compare the experimental results in these three datasets, we can see that the four topics listed above can be found out in the micro-blogs which are expected to be popular by our discriminant model, however, key words extracted in data set 1 are mostly insignificant words and almost no topic can be found in this data set. In addition, the key words in data set 3 whose micro-blogs are rejected by our model are mostly about entertainment or related to some other inconsequential affairs.

In general, our discriminant model can filter out many micro-blogs which are inconsequential but have a bad infect on the extraction of topics. Without our discriminant model, it's hard to find out topics in data set 2 and data set 3.

TABLE IV Number of micro-blogs in each dataset

	K=0.8	K=1.0	K=1.2	K=1.4	K=1.6	K=1.8	K=2.0
Data set 2	695	546	424	360	322	276	245
Data set 3	1681	1835	1952	2016	2054	2100	2131
Data set1	2376						

TABLE V Topics extracted in data set 1, data set 2, and data set 3

		Dataset1	Dataset2	Dataset3
		Key words		
Co-occurrence >0.3	topic1	中国、爆料、红会社监委	造谣污蔑、刘铁男、国家能源局	中国、爆料、歌唱类节目
	topic2	爆料	幼女、强奸猥亵、嫖宿幼女罪、性侵	微博、国家能源局、人身攻击
	topic3	爆料	农夫山泉、翠宫饭店、央视	浙江工贸学院
	topic4	NULL	NULL	NULL
Co-occurrence >04	topic1	央视	造谣污蔑、刘铁男、国家能源局	爆料
	topic2	中国、卫三畏	幼女、嫖宿幼女罪、校长	微博
	topic3	中国、卫三畏	农夫山泉、央视、中国	中国、歌唱类节目
	topic4	NULL	食药监总局	NULL
Co-occurrence >0.5	topic1	飞马旅	造谣污蔑、刘铁男	爆料
	topic2	NULL	幼女、嫖宿幼女罪	微博
	topic3	NULL	农夫山泉、中国	中国、歌唱类节目
	topic4	NULL	食药监总局	NULL

4. Conclusion

In this paper, we present a holonomic model based on opinion leaders to detect hot topics in Sina Weibo. In the model, we take the structure information and propagation characteristics of Sina Weibo into account to solve the problems in the detection for micro-blog topics. Moreover, the word co-occurrence graph model is used to avoid the problem of clustering for Chinese short texts. The experimental results indicate that the model can discriminate and find out hot topics in Sina Weibo effectively. However, due to the lack of data and various propagation forms of topics in Sina Weibo, our model is far from consummate. Therefore, we will further research on Sina Weibo and focus on the applicability of our model.

Acknowledgment

This work was supported by "the Fundamental Research Funds for the Central Universities" of CUGB (2652013102, CUGB), Subject Supportive Project of CUGB (2013) and the Innovative Experiment Plan for College Students of China University of Geosciences, Beijing.

References

[1] Yanchao Zhang, Yun Liu, Hui Cheng, Fei Xiong, Changlun Zhang, "A Method of Measuring User Influence in MicroBlog", JCIT: Journal of Convergence Information Technology, vol. 6, no. 10, pp. 243 ~ 250, 2011

[2] Hong Liang, Gang Lu, Nanshan Xu, "Analyzing user influence of microblog," Advanced Computational Intelligence (ICACI), 2012 IEEE Fifth International Conference on , vol., no., pp.15,22, 18-20 Oct. 2012.

[3] Yuan Zhang, Yuqian Bai, "Research on the Influence of Micro-bloggers -- Take Sina Celebrity Micro-blog as an Example," Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference on , vol., no., pp.189,192, 22-24 Oct. 2012.

[4] Sai Zhang, Ke Xu, Haitao Li, "Measurement and Analysis of Information Propagation in Online Social Networks Like Microblog", Journal of Xi'an Jiaotong University,2013,(02):124-130.

[5] Jie Tian, Yongcheng Li, Yuliang Lu, Hao Guo, "A Research on Famous-User Network of SINA-Weibo," Instrumentation, Measurement, Computer, Communication and Control (IMCCC), 2012 Second International Conference on , vol., no., pp.57,62, 8-10 Dec. 2012.

[6] Yu, L.L., Asur, S., Huberman, B.A., "Artificial Inflation: The Real Story of Trends and Trend-Setters in Sina Weibo," Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom) , vol., no., pp.514,519, 3-5 Sept. 2012.

[7] Wenqing Zhao, Xiaoke Hou, "News topic recognition from chinese microblog base on word co-occurrence graph", CAAI Transactions on Intelligent Systems, 2012.07(5):444-449. DOI:10.3969/j.issn.1673-4785.201205045.

[8] Yanyan Du, Yanxiang He, Ye Tian, Qiang Chen, Lu Lin, "Microblog bursty topic detection based on user relationship," Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint International , vol.1, no., pp.260,263, 20-22 Aug. 2011.DOI: 10.1109/ITAIC.2011.6030199.