Microblog-oriented hot topic discovery and trend analysis

Chengfang Tan[1,2*], Caiyin Wang[2], Lin Cui[1,2]

[1]*School of Information Engineering, Suzhou University, Suzhou 234000, Anhui, China*

[2]*Intelligent Information Processing Lab, Suzhou University, Suzhou 234000, Anhui, China*

Received 1 June 2014, www.cmnt.lv

Abstract

Microblog has become an important means for people to obtain the first message. This paper proposes a method for hot topic detection and topic trend analysis in the massive microblog short text data set. Firstly, considering the microblog special style and clustering efficiency, we make appropriate improvements on the classical single -pass clustering algorithm to enhance the quality of clustering, and then put forward the topic heat evaluation method and rank topics of each topic cluster based on this method. Finally, in each time window, we calculate the strength of topics depending on the topic strength calculation formula, and discovery the topic evolution trend. Experiments show that the proposed method can not only effectively detect hot topics, but also can clearly find out the topic evolution process.

Keywords: Text Clustering, Microblog Text, Topic Discovery, Trend Analysis

1 Introduction

With the rapid development of Web2.0 technology and mobile communication technology, microblog has become a huge load information platform with its powerful user group, which is one of the important ways for user to access to information and disseminate public opinion. Due to the large number of microblog texts, the complexity of information, short content and other characteristics, it is difficult for users to browse all of microblog information. In addition, as more and more users begin to use microblog to express their views and comments, it has become an important platform to reflect public opinion. So it is significant to get hot topics from the mass microblog information. On the one hand, it can help users quickly find out the focus on each field of the society, on the other hand, it can provide the public opinion guide for public opinion monitoring field.

For the topic discovery and trend analysis, researchers have done a lot of researches. At present, abroad mainly researches the related topic detection on Twitter and Facebook. Topic detection technology in china focuses on the research of network news topic detection. Strzalkowski [l] applied automatic abstraction method to topic tracking. The main idea is to replace the original report with news abstraction, and then by comparing the relevant degree between topic and abstraction to determine the report whether belongs to the topic. Dragon [2] and UMass [3] used new tracking report to feedback on the topic model and continued to track with adaptive learning model [4]. Young woose and Katia sycara [5] proposed a new detection method, which combined single-pass algorithm with K-means algorithm, but the test results for news corpus have excessive dependence on the input sequence. Because microblog text is short, language style is casual and so on, so the traditional method is directly applied to the microblog often lead to large amount of calculation and low detection rate. This needs to study the suitable analysis method on microblog hot topic detection and trend analysis.

Based on the above research, according to the characteristics of microblog text, this paper improves classic single-pass clustering algorithm and clusters microblog texts. And we put forward topic heat evaluation method and then use this method to rank the topic to discovery hot topics. In addition, by calculating the strength of topic in each time window, the process of topic trend change can be seen.

2 Related theoretical basis

2.1 TOPIC DETECTION AND TRACKING

Topic detection and tracking (abbreviated as TDT) is launched by the US Defense Advanced Research Projects Agency (abbreviated as DARPA),which is mainly used for news discovery, tracking new event, engaging in news report automatic identification, locking and finding breaking news topics, and other related tasks. The research target is to automatically identify new topics and continuously monitor known topic. TDT evaluation conference divides topic detection and tracking into five subtasks, there are reports segmentation, topic identification, the first reported detection, topic tracking and correlation detection [6]. The involved key technologies for topic detection and tracking are mainly include representation model, similarity calculation, the organization mode based on similarity.

* ***Corresponding author's*** e-mail: 874036730@qq.com

2.2 THE BASIC PRINCIPLE OF TEXT CLUSTERING

Clustering is an important part of topic detection, which can effectively reduce search space to accelerate search speed, and then improve retrieval precision. Text clustering is mainly based on the famous clustering assumption: the same kind of document has greater similarity, otherwise has smaller similarity. Fast and high quality text clustering can organize large amounts of information into several meaningful clusters, so as to improve the retrieval performance. Text clustering method usually converts document into vectors in a high dimensional space by using the vector space model, and then clusters these vectors. The output of text clustering is generally a division of document collection. Clustering algorithm is an unsupervised learning algorithm, which can be divided into four types: hierarchical clustering algorithm, partition clustering algorithm, grid clustering algorithm and density- based clustering algorithm.

3 Microblog hot topic discovery

3.1 MICROBLOG TEXT REPRESENTATION

Text representation refers to extracting feature words from microblog text and then the feature set is formed, each feature has a corresponding feature value. The commonly text representation model are: Boolean model, vector space model (VSM) and the probability model, etc [7]. This paper uses vector space model to represent microblog text, which is expressed as:

$$M = (w_1, w_2, ..., w_n), \tag{1}$$

where w_i represents the weight of the *ith* feature item of microblog, the weight is calculated by the famous TF-IDF formula, which is given by:

$$w_{ij} = \frac{tf_{ij} * \log_2(\mathrm{N}/\mathrm{n}_i + 0.01)}{\sqrt{\sum_{j=1}^{n}\left[tf_{ij} * \log_2(\mathrm{N}/\mathrm{n}_i + 0.01)\right]^2}}, \tag{2}$$

where w_{ij} represents the weight of the *ith* word in microblog text j, and tf_{ij} represents the *ith* word frequency in microblog text j, N represents the total number of microblog texts, n_i represents the number of microblog text which contains the *ith* word. The denominator is the normalization factor.

3.2 SIMILARITY CALCULATION

Similarity calculation is based on the similarity of microblog text vector space model. Usually we use *sim (x, y)* to indicate the similarity between the text x and the text y. The value of *sim (x, y)* is proportional to the similarity degree of x and y. Similarity is generally a value between 0 and 1, that is *0<=sim (x, y) <=1*. In this paper, the greater similarity between two microblog texts, the more relevance they discuss. This paper uses Cosine distance to compute the similarity between two vector models. The distance is defined as follows:

$$sim(x, y) = \cos(\mathrm{x}, \mathrm{y}) \frac{\sum_{i=1}^{n} w_{xi} * w_{yi}}{\sqrt{\sum_{i=1}^{n} w_{xi}^2 * \sum_{i=1}^{n} w_{yi}^2}}, \tag{3}$$

where n is the dimension of vector space, w_{xi} is i-dimensional component of vector x, w_{yi} is i-dimensional component of vector y.

3.3 TOPIC IDENTIFICATION BASED ON SINGLE-PASS ALGORITHM

The classic single-pass algorithm is a kind of topic identification algorithm which is the most commonly used in topic detection. It has characteristics of fast calculation speed, simplicity calculation and so on. It adopts an incremental method, which does not establish its topic clusters at the beginning, but later according to the text input information to determine whether to establish topic clusters. The basic idea [8] is set a parameter named threshold in advance, when enter a text, the system will calculate the similarity between the input text and identified topic clusters, if the similarity is greater than the threshold, it indicates the input text has related content in microblog text and then incorporates into the related topic clusters, otherwise, the text will be used as a seed topic to create a new topic**.**

Single-pass algorithm sequentially processes text according to input sequence. It can determine the text belongs to which cluster when first read the text, but this also brings important disadvantage: text input sequence will have great impact on the results. From the theory, when the data source and the parameters are determined, the clustering results should be determined and should not vary.

In order to reduce the impact of input sequence on clustering results, by referencing Ref [9], this paper improves the single-pass clustering algorithm. In fact, the initial input text is more close to topic core and the more capable that the same topics will be clustered together. First, input data is sorted from high to low in accordance with the hit rate of high frequency words, which can be calculated accompanying with calculating TF-IDF value, so the calculation process does not increase the computation complexity of the algorithm. This makes the texts that can more represent the topic core are at the forefront of input sequence.

Based on the above analysis, this paper makes some improvement on single-pass clustering algorithm. Those reached texts according to the time sequence are divided by a certain time granularity. Microblog texts set which reach in the time window *t* can be expressed as:

$$D = (\mathrm{m}_1, \mathrm{m}_2, ..., \mathrm{m}_\mathrm{m}),$$

where m_i is one of texts. Improved single-pass clustering algorithm is described as follows.

Input: microblog texts sorted in accordance with the hit rate of high frequency words from high to low.

Output: each sub topic clusters T_1, T_2, …, T_k.

Step 1. Presets a clustering threshold k .

Step 2. Sequentially reads a microblog text m .

Step 3.Calculates the similarity between microblog text m and existed topic cluster T_i , that is:

$$sim(\mathrm{m},\mathrm{T}_i)=\frac{sim(\mathrm{m},t_1)+\cdots+sim(\mathrm{m},t_n)}{n},$$

where $t_1,\cdots,t_n$ are microblog texts of topic cluster T_i .

Step 4. Finds out the most similar sub topic cluster T, that is $\mathrm{sim}(m,T)=\max\mathrm{sim}(m,T_i)$.

Step 5. If $\mathrm{sim}(m,T)>k$, then m will be joined in sub topic cluster T, otherwise, takes the microblog text m as a seed topic to create a new sub topic T_x .

Step 6. If there is a new microblog text – go to Step 1.

3.4 TOPIC HEAT EVALUATION

By clustering microblog texts, we can get a lot of microblog text clusters, where each cluster is a topic, so the clustering result can be seen as a topic set. By referencing Ref [10], this paper introduces its BBS heat evaluation thoughts to microblog, considering the number of released topic, topic attention degree, the number of topic comment, the number of forwarding number and other factors to establish hot topic evaluation model for topic heat score, then find out hot topics in microblog. Since each factor has different dimension, which need to normalize the value of each factor.

Assuming the microblog text set has n topics, the influence of the number of released topic t can be calculated as follows:

$$R(t)=\frac{m_t}{\sqrt{\sum_{i=1}^{n}m_i^2}}, \tag{4}$$

where m_t represents the number of texts related to topic *t*, and m_i represents the number of texts related to the *ith* topic.

Attention degree of topic *t* can be calculated as follows:

$$\mathrm{F}(t)=\frac{\sum_{j=1}^{m_t}f_{t_j}}{\sqrt{\sum_{i=1}^{n}(\sum_{j=1}^{m_t}f_{t_j})^2}}, \tag{5}$$

where f_{t_j} represents the attention degree of a microblog text, there is $f_{t_j}=N_u$, N_u is the current number of concerns for this microblog.

Impact factor of comment number about topics t can be calculated as follows:

$$\mathrm{C}(t)=\frac{\sum_{j=1}^{m_t}c_j}{\sqrt{\sum_{i=1}^{n}(\sum_{j=1}^{m_t}c_j)^2}}, \tag{6}$$

where c_j represents the comment number of the *jth* microblog text about topic *t*.

Impact factor of forwarding number about topics *t* can be calculated as follows:

$$\mathrm{S}(t)=\frac{\sum_{j=1}^{m_t}s_j}{\sqrt{\sum_{i=1}^{n}(\sum_{j=1}^{m_t}s_j)^2}}, \tag{7}$$

where s_j represents the forwarding number of the *jth* microblog text about topic *t*.

Based on the above, the topic heat evaluation formula eventually obtained as follows:

$$\mathrm{H}=\mathrm{R}(t)+\mathrm{F}(t)+\mathrm{C}(t)+\mathrm{S}(t). \tag{8}$$

Through the above topic heat evaluation formula, we can calculate each topic heat obtained after clustering, and then sort these topics, select the first n topics within the specified time, that are hot topics.

4 Topic trend analysis

The whole process of topic is consisting of the following four stages: emergence, development, decline and demise, this is called the life-cycle model [11]. Topic trend analysis expresses the change process of topic with the time change. Microblog topics are related to the reality events. With the passage of time, events will also change. In this paper, trend change is defined as follows: the change of topic strength in the continuous time window.

Microblog real-time is relatively high, so the number of topics will change with the time change. Based on the actual requirements of topic trend analysis, those reached texts according to the time sequence are divided by a certain time granularity. Topic development trend can be considered from two aspects: topic growth ratio between adjacent time windows and the change of topic heat, which is also named topic strength.

Set the strength of topic *T* at time *t* is s_t, it is defined as follows：

$$s_t=(H_t-H_{t-1})+\frac{n}{N}-\alpha\bullet\frac{t}{t+1}, \tag{9}$$

where H_t represents the heat value of topic *T* at time *t*, H_{t-1} represents the heat value of topic *T* at time *t-1*, *n* represents the number of increased topics about topic *T* from *t-1* to *t*, N represents the total number of increased topics from *t-1* to *t*, α is decay rate. With the passage of time, the topic will be gradually weakened, so the decay rate multiplies *t/t +1*, when *t* is greater, the *t+1* will also more and more great.

5 Experiments and Analysis

5.1 EXPERIMENTAL TOOLS

The experimental tools are segmentation software ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System), statistical analysis tools SPSS and Excel, Java integration tools Eclipse 3.7.3 and JDK 1.6.0, 7 6 and the numerical matrix tool Matlab7.6.0.

5.2 EVALUATION INDEX

In the TDT test standard, people usually use false detection rate, missing detection rate and cost function to evaluate the quality of topic discovery. Missing detection rate is the ratio between the reported number that system did not detect for topic (defined as B) and the reported number that should be detected (defined as A+B), where A is the reported number that the system correctly identified for the topic. False detection rate is the ratio between the reported number that the system false detection (defined as C) and the reported number that does not belong to the topic (defined as C+D), D is the reported number not belong to the topic that the system correctly detected. The above indicators can be formulated as follows:

$$P_M = \frac{B}{A+B} , \tag{10}$$

$$P_F = \frac{C}{C+D} . \tag{11}$$

Cost function considers comprehensively the cost of false detection and missing detection, which is expressed as follows:

$$C = C_M P_M P + C_F P_F (1-P) , \tag{12}$$

where C_M represents the cost factor of missing detection, C_F represents the cost factor of false detection, P_M represents the missing detection rate, P_F represents the false detection rate, P is priori probability. According to TDT evaluation standard, set $C_M = 1$, $C_F = 0.1$, $P = 0.02$.

TDT detection usually needs to normalize cost function *C*, so the normalized cost calculation formula is as follows:

$$C_U = \frac{C}{\min(C_M P, C_F (1-P))} . \tag{13}$$

5.3 EXPERIMENTAL DATA

Since there is no universal Chinese microblog data test set, this paper obtains data by combining Sina Weibo API interface with web crawler, crawling the original microblog data between May 15, 2014 and May 22, 2014. After primary screening, we filter out too simple and meaningless microblog text, select the length of 4 characters or more microblog text, and obtain the number of forwarding, the number of comments, released time, the number of fans, the number of user attention, etc. By manual annotation the major topics in this period, there are five hot events:"520", "hammer phone", "graduation season" and the "Haibo Huang event" and "Cannes Film Festival".

Before the experiment, we process data set with word segmentation, removing stop words and POS filtering pretreatment, where word segmentation use ICTCLAS, removing stop word use Chinese stop vocabulary and English stop vocabulary, and POS filtering is based on the annotation of word segmentation system.

5.4 ANALYSIS OF EXPERIMENTAL RESULTS

Experiment 1. Select the clustering threshold

For different similarity thresholds, there may be different results, so this experiment will use different thresholds to compare missing rate, false rate, normalized overhead. The comparison results are shown in Figure 1.

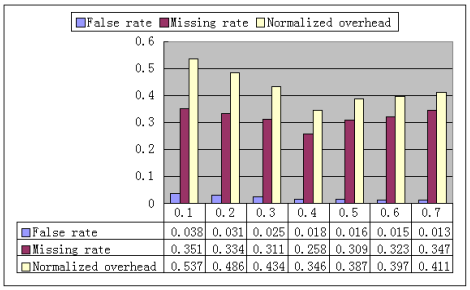


FIGURE 1 The comparison of results under different thresholds

Where the abscissa represents preset clustering threshold, the vertical axis represents the value of missing rate, false rate and C_U . From Figure 1, we can see that different thresholds have a certain impact on the value of missing rate, false rate and C_U . With the increase of the threshold, the three values gradually decreases, but when reaching a certain value, the increase of threshold also lead to increase of values. When threshold is 0.4, the value of missing rate, false rate and C_U is relatively small. Therefore, by comprehensively consideration, we determine the cluster threshold is 0.4 in this paper.

Experiment 2. Comparison of classical single-pass algorithm and improved single-pass algorithm

The experiment adopts missing detection price C_M *=1*, false detection price C_F *=0. 1*, the prior probability *P=0.02*. This paper selects 5 hot topics of experimental data, firstly uses the classical Single-Pass algorithm as clustering algorithm to detect the hot event, and then compares with improved Single-Pass algorithm on the missing rate, false rate and normalized overhead. The results are shown in table 1 and table 2.

TABLE 1 Classic single-pass clustering results

Experimental data set	Cannes Film Festival	Haibo Huang event	520	hammer phone	graduation season
Missing rate	0.387	0.301	0.358	0.335	0.357
False rate	0.035	0.021	0.027	0.029	0.033
The average missing rate	0.348				
The average false rate	0.029				
Normalized overhead	0.489				

TABLE 2 Improved single-pass clustering results

Experimental data set	Cannes Film Festival	Haibo Huang event	520	hammer phone	graduation season
Missing rate	0.287	0.226	0.271	0.238	0.265
False rate	0.021	0.015	0.019	0.018	0.018
The average missing rate	0.258				
The average false rate	0.018				
Normalized overhead	0.346				

From table 1 and table 2, we can see that the missing rate, false rate and normalized overhead obtained by classical single-pass algorithm is respectively 0.348, 0.029, 0.489, the improved single-pass algorithm is respectively 0.258, 0.018, and 0.346. Thus, the improved single-pass algorithm plays good effect on the topic discovery, enhances the quality of clustering.

Experiment 3. Topic trend analysis

Among these identified microblog topics, in order to get topic trend change process, this experiment selects "520" and "Haibo Huang event" as tracking topics, which are respectively named topic 1 and topic 2. This time window is set by the day, so the data set is divided into 8 time window. According to the formula (9), the strength of each topic can be calculated in 8 time window. The experimental result is shown in Figure 2.

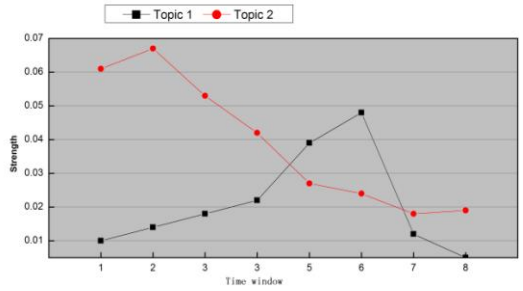


FIGURE 2 The process of topic trend

As can be seen from Figure 2, in the second time window, strength of topic 2 is significantly great. In the 6th time window, topic 1 begins to increase its strength, and becomes a hot topic, although strength of topic 2 is relatively reduced, but its strength is still great, this explains that topic 2 has very strong influence and attracts more attention. In the 7th time window, strength of topic 1 is suddenly reduced, when time goes to the 8th time window, topic 1 is almost disappeared, while strength of topic 2 is still continuing.

6 Conclusion

This paper improves the classic single-pass algorithm, before clustering, in accordance with the hit rate of high frequency words to sort microblog texts, which enhances the clustering quality. Based on the topic identification, topic heat evaluation model is created by the number of released topic, topic attention degree, the number of topic comment, the number of topic forwarding and other factors. By using this model, hot topics in microblog can be found. In addition, through the calculation of topic strength, the topic trend change is realized. The experimental results are consistent with the actual situation, which indicates our method is effective. The next step we will consider how to further explore the characteristic of the topic, and to better find out the association between topics.

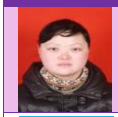
Acknowledgements

This work was supported by Key University Science Research Project of Anhui Province (No.KJ2014A250) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2014YKF41) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2013YKF19) and Software engineering projects (No.2013zytz074) and The training system and training platform construction based on engineering training and innovation ability of computer professional (No.2013cgtg032) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2011YKF10).

References

[1] Strzalkowski T, et al. (1999) Lightweight Topic Tracking System. *Proc.of the DARPA Workshop.*
[2] Yamron J, Knecht S, Van Mulbregt P. (2000) Dragon's tracking and detection systems for the TDT2000evaluation.*Proc. of Topic Detection and Tracking Workshop*, 75-80.
[3] Allan J, Lavrenko V, Frey D,et al. UMass at TDT. (2000) *Proc.of Topic Detection andvTracking Workshop*.USA, 109-115.
[4] Lam W, Mukhopadhyay S, Mostafa J,et al. Detection of shifts in user interests for personalized information filtering. (1996) *Proc.of the 19th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM. 317-325.
[5] Young-Woo seo, Katia Sycara. (2004) Text Clustering for Topic Detection. *Pittsburgh:Camegie Mellon University*, 5-6.
[6] TDThomepage.*http://www-tl.nist.gov/iad/894.01/tests/tdUindx.htm*.
[7] Zhang Xiao-Ming, Li Zhou-Jun, Chao Wen-Han. (2012)Research of automatic topic detection based on incremental clustering. *Journal of Software*, 23(6):1578-1587.
[8] Ron Papka, James Allan. (1998)On-line new event dectection using single pass clustering. *UM-CS-1998-021*, Amherst.
[9] XuePing Feng. (2013) Research on Topic Detection and Tracking Method of Microblog. *Huazhong University of Science & Technology*.
[10] KaiMei Lan. (2011) BBS Hot Topic Detection and Monitoring System. *Beijing Jiaotong University*.
[11] Ronald P. (2012) Facing scalability:Naming faces in an online social network. *Pattern Recognition*, 45(6):2335-2347.

Authors

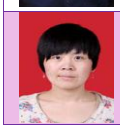


Chengfang Tan, 16. 02. 1981, China

Current position, grades: researcher at Intelligent Information Processing Lab, Suzhou university, China.
University studies: master degree in education technology from Nanjing Normal University, China in 2007.
Scientific interest: information retrieval, sentiment analysis and text mining.

Caiyin Wang, 16. 09. 1978, China

Current position, grades: researcher at Intelligent Information Processing Lab, Suzhou University, China.
University studies: master degree in computer science and technology from Hefei University of Technology, China in 2009.
Scientific interest: P2P and information retrieval.



Lin Cui, 19. 08. 1979, China

Current position, grades: researcher at Intelligent Information Processing Lab, Suzhou university, China.
University studies: master degree in computer science and technology from Hefei University of Technology, China in 2008.
Scientific interest: information retrieval and Semantic Web.