Journal of Network & Information Security 4:2 (2013) 145-153

Study on Hot Topic Detection on Microblog

Libin PENG*, Guoyong CAI

Guilin University of Electronic Technology, Guangxi Key Lab of Trusted Software, Guilin 541004, China

Abstract

As a new platform which could post and share messages at unprecedented speed, Microblog has produced a great quantity and rich variety of messages. In order to detect the hot topics among the vast messages and topics, this paper proposes an algorithm based on a comprehensive semantic and context similarity measurement. First the algorithm extracts keywords which appear multiple times on microblog, and synthesizes them with comments, then a compound weight could be formed. We sort the compound weight *W* of key words, and apply semantic similarity and context similarity to cluster keywords. The experiments on real world microblog data have shown the effectiveness and efficiency of the approach to hot topics detection.

Keywords: Microblog; Hot Topics; Clustering; Comprehensive Similarity

1. Introduction

Microblog is the shortened form of micro blog (MicroBlog). It is a type of platform which allows users to disseminate/obtain/share information based on relationship between users. Users can set up personal community through WEB, WAP and other client terminals, update the information with no more than 140 words of text, and achieve instantly information sharing. The earliest famous microblog site is twitter, which appeared in USA in 2006. Sina Microblog, a private beta version Chinese portal Sina.com was launched in August 2009, is the first microblog service portal website in China. Till October 2011, Chinese microblog users have reached 249.8 million totally. Microblog has developed rapidly although it has been a short time since it appeared. It has become an important channel for public to transfer news and events promptly due to the simple content and convenience operations. Finding out hot topics from these entangling messages is significantly important for security of network and information management, also for public opinion analysis.

- This work is supported by the NSFC (#61063039), Guangxi Key Lab of Trusted Software (#kx201202).
* Corresponding author.
Email addresses: ccsupeng@163.com (Libin.PENG)

July 15, 2013

Recently, scholars have begun the research to dig out the hot topics on microblog. For instance, Sharifi [1] proposed a method which uses a sentence to summarize hot topics on microblog, so that users could understand a hot topic fast and correctly. Base on this research results, Inouye [2] proposed a method using multiple sentences to summarize hot topics on microblog, overcoming the defect that a single sentence cannot bear sufficient topic of the amount of information. Weihua Luo and Manquan Yus[3] group topics firstly and then classify them into micro-class with divide and conquer thoughts, combining with a variety of strategies to optimize in the cluster.

However, those methods mentioned above mainly focus on semantic similarity of words and the occurring frequency of words only. The position of keywords and context similarity are not considered in most of the methods. For example, if the topic is “recently, the sales volume of intelligent mobile phone keeps increasing continually, among them, the sales volume of apple mobile phone is the highest ", according to the semantic similarity , " Intelligent mobile phone " and " Apple mobile phone " are similar, but " intelligent mobile phone " and " sales up " are not similar, so the results obtained probably may be { " intelligent mobile phone ", " Apple mobile phone " } and { " sales up " }, instead of { " intelligent mobile phone ", " Apple mobile phone ", " sales up " }, which are the ideal topics.

Aiming at solving this issue, the paper proposes an algorithm based on comprehensive semantic similarity and context similarity, in order to detect hot topics among massive microblog messages. The remaining of this paper is organized as follows: In section 2, we provide a brief review of the related work. The introduction of our method is proposed in section 3, drawing on semantic similarity and context similarity as the comprehensive similarity to detect hot topics. In section 4, some analysis of our experimental results is given. Finally we make a conclusion and discuss our plans for future work in section 5.

2. Related Works

A topic usually refers to the condition caused in a specific time, place, and may be accompanied by some of the inevitable result of an event in literature [4]. A “hot topic” is defined as a topic that appears frequently over a period of time. James Allan, RonPapka and vieto: Lavrenk were the first to do research in the field of TDT (Topic Detection and Tracking) [5]. In recent years, experts at home and abroad have undertaken an in-depth study for TDT. We could divide the strategy of hot topics detection into the following classes:

(1) Clustering model. This model mainly uses the method of clustering to dig out hot events. For instance, in 1998, Allan[5] used a single-pass clustering algorithm which combining a novel threshold model, realized an online news monitoring system; In 2010, Lu Rong[6] used a method which combines double K-mean clustering and hierarchical clustering, and then integrate it with implicit subject, and he detected the hot topics successfully. In 2011, Fengjing Yin and Weidong Xiao [7] used an improvement Single-pass algorithm, combining the concept of“generation”, and they realized a Network-oriented topic, named ICTC.

(2) Extended topic model. For microblog messages, a topic model is established directly based on traditional topic model LDA (Latent Dirichlet Allocation).One can use the basic model and

extract topic directly. For example, in 2007, Blei [8] established a new topic model-CTM (Correlated Topic Model), which models the correlation between topics through normal distribution; In 2008, Asuncion [9] proposed an improvement LDA based on distributed algorithm and HDP (Hierarchical Dirichlet Process) topic model; In 2010, Remage [10] constructed a L-LDA, which can personalize users' information by characterization of users in Twitter.

3. Hot Topic Detection and Extraction

Compared with the topics on traditional internet, the data of Chinese microblog is a series of short text essentially, which keeps increasing constantly. At the same time, some special symbols and formats showing the communication and interaction between users will appear in the text. For example, we use the format of "@ somebody" to represent "communication with" a user, the format of "#topic#" to show the discussion about a particular subject, and we also use "RT" to represent forwarding messages. The system of hot topics on microblog mainly cover the steps of data preprocessing,word segmentation,word clustering,text combining and hot topic evaluation. The main framework of detecting hot topic in this paper is shown in figure 1.

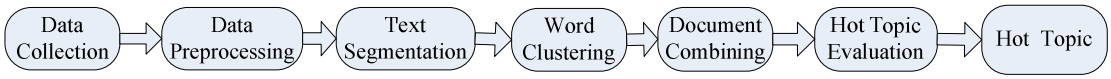

Fig. 1 Flow Chart of Hot Topic Detection

3.1. Vector Space Model

Microblog document d is given by

$$d = (t_1, w_1, t_2, w_2, ...t_n, w_n) \tag{1}$$

Where $t_1, t_2, t_3 ... t_n$ is the keyword of d, $w_1, w_2, w_3 ... w_n$ is the weight of keywords. The paper improves the method of weight calculation from traditional TF-IDF, which obtains the weight of keywords according to the comments of microblog and the position of the keywords. $F(t) = TF(t) * N_1 * N_2 (TF(t)$ is the number of times the keyword t appears in the main post, N_1 is the number of the comments which contained keyword t, N_2 is the position of keyword, N_2 is shown as 2 if keyword t appears in the main post, otherwise it is shown as 1 in other conditions.) The formula of weight w of the keyword is given as follows:

$$w_t = \frac{F(t)\log(\frac{N}{DF(t)} + 0.01)}{\sqrt{\sum_{t \in d} F^2(t) + \log^2(\frac{N}{DF(t)} + 0.01)}} \tag{2}$$

Where, w_t is the weight of keyword t, $F(t)$ is the frequency of keyword t, $DF(t)$ is the number of keyword t contained in the text, N is the number of the keywords.

Topic T is consist of related documents, given by

$$T = \{d_1, d_2, d_3, ...d_n\} \tag{3}$$

The core vector of topic T is topic T', given by

$$T' = (t_1, w_1, t_2, w_2, ...t_{n,} w_n) \tag{4}$$

Where $t_1, t_2, t_3 ... t_n$ is the pivotal keyword of T, $w_1, w_2, w_3 ... w_n$ is the weight of keywords .

3.2. The Calculation of Similarity

The traditional methods are based on the semantic similarity of the two words to judge the similarity of the keywords. That is to say, the similarity of two words is fixed, which means that the related words are much more similar, but the unrelated words are not similar. This paper proposed a method of comprehensive similarity based on semantic similarity and context similarity. The formula of comprehensive similarity is shown as follows:

$$S = aSim_1 + bSim_2 \tag{5}$$

Where, as the weighted coefficient, a, b reflect the contribution of different similarity to the overall similarity. Sim_1 represents the semantic similarity; Sim_2 represents the similarity of context.

This paper refers to literature [14] for the calculation of semantic similarity, given by

$$sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} sim_j(S_1, S_2) \tag{6}$$

where, $sim_j(S_1, S_2)$（j＝1,2,3,4）is the similarity of the four different sememes S_1, S_2 correspond to. β_j is an adjustable parameter. $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1, \beta_1 > \beta_2 > \beta_3 > \beta_4$ The significance of this parameter method is that the similarity of main sememes can restrict minor sememes and maintains the contribution status that the similarity of main sememes contributes to the overall similarity.

The context similarity is given by as follow:

$$p(w_1 \mid w_2) = \frac{F(w_1 w_2)}{F(w_2)} \tag{7}$$

That is to say, the context similarity equals the ratio of the number of w_1, w_2 appeared in message at the same time and the number of w_2.

3.3. Clustering

After obtaining data of microblog through the crawler program, we must eliminate noise data and filter out the irrelevant information of the text as far as possible during the procedure of microblog preprocessing. After pretreatment, we are going to perform word segmentation for the text. There are many kinds of tools to treat with word segmentation. The ICTCLAS system of the Chinese Academy of Sciences used in this paper is a tool that often being used in Chinese text processing, which has a good effect on word segmentation and could support for named entity recognition, location recognization and organization recognization. Then, we sort the weight w of keywords and cluster the keywords based on comprehensive similarity. The algorithm is shown in figure 2.

```
Algorithm: An algorithm of detecting hot topics
  INPUT：Keywords list T, Threshold D
  OUTPUT：Clustering result
            Count=1
      Initialization the first keyword to cluster, Cluster(1) ⟵ {t1}
       Repeat
      Count++
          Input the next keyword, compare the similarity
          if (sim(ti,Cluster(i))<D)
                    new cluster cm, cm ⟵ {tj}
              else   Cluster(i)=Cluster(i)+{tj}
          end
      until (Count >n)
```

Fig.2 Description of the Proposed Algorithm

3.4. The Calculation of Hot-degree

We can detect the topics through the above method, but the quantity is enormous. Users could only know the topics content on microblog, and they can not get the topic which been mention and discussed frequently. In this paper, we proposed a method which combining the quantity of comments, the fans and the times of being forwarded comprehensively. And the formula is as following:

$$H_i = \log(fl_i + 1) + \sqrt{f_i} + c_i \tag{8}$$

H_i refers to the hot degree of topic i, fl_i refers to the fans of microblog users of topic i, and f_i refers to the amount which topic i being forwarded, and c_i refers to the comments for topic i.We could obtain the hot degree of each topic through the above algorithm, and then get the hot topics after ranking them.

4. Experiments

4.1. Data source

The experimental data are captured from the largest microblog site named Sina microblog. Through API provided by Sina, about 8 million original microblog data from August 13th 2012 to August 29th are captured. After preprocessing data, we reject some messy independent data. Then after doing segmentation and combining microblog document, finally we get about 3 million microblog data. We artificially mark 10 pieces of hot topics such as "the 2012 London Olympic Games", "the voice of China", "Price war", "Diaoyu Island ". In equation (5), we set a as 0.4 and b as 0.6 to reflect the influence of the parameters on the result. Table 1 is the results of hot topics, and the first column is fragments of keywords, the second one is corresponding events.

Figure 3 shows the relationship between the number of combined microblog documents and the time in the hot topic. The horizontal represents the amount of the hot topic on microblog of the day, the ordinate represents the number of combined microblog documents of the hot topic.

4.2. Experimental Result

Choose evaluation standards are as follows:

$$\text{Accuracy: } P = \frac{A}{A+B}$$

$$\text{Recall rate: } R = \frac{A}{A+C}$$

$$\text{F-measure: } F = \frac{2PR}{P+R}$$

Where, A is the number of related documents detected, B is the number of irrelevant documents detected, C is the number of related documents not detected. To compare the similarity of the algorithm (S_HTDEA) in this paper with classic single-pass algorithm [7], we found S_HTDEA algorithm has higher accuracy and recall rate than single-pass. As shown in figure 4.

Table.1 The Results of Hot Topic

Time	Hot Topics	Corresponding Event
2012/8/13	Summer/Olympics/closed	The 30th Annual Summer Olympics closed in London.
2012/8/14	Jingdong/Suning/Gome /price war	Qiangdong Liu, the Jingdong Mall CEO, announced to launch a price war to Suning and Gome.
2012/8/17	China/voice/identity /questioned	The voice of China's players Yong Huang and Hongyu Zhou's identities have been questioned.
2012/8/19	Diaoyu Island/across the country/boycott/Japanese goods	After Diaoyu Island incident, people across the country resolutely boycott of Japanese goods.
2012/8/20	Chinese/farmer/invent /Ultraman/ sliced noodle/robot	The speed of the Ultraman sliced noodles robot invented by Chinese farmer Cui Run quan can reach 150 per minute.
2012/8/21	Bo Gu Kailai /Zhang Xiaojun/final judgment	Bo Gu Kailai and Zhang Xiaojun's case got the final judgment.
2012/8/23	happy/Tanabata Festival	Happy Tanabata festival.
2012/8/24	Harbin/bridge/collapse/ heavy-duty	Harbin Yangmingtan Bridge was collapsed, The last paragraph of the main bridge was collapsed by a heavy-duty freight car.
2012/8/27	Hero/movie/Batman/Spider-man	Two superhero movies, "Batman 3" and"extraordinary Spider-Man" ,are on the domestic market.
2012/8/29	Summer/Paralympic Games/open	2012 Summer Paralympic Games opens.

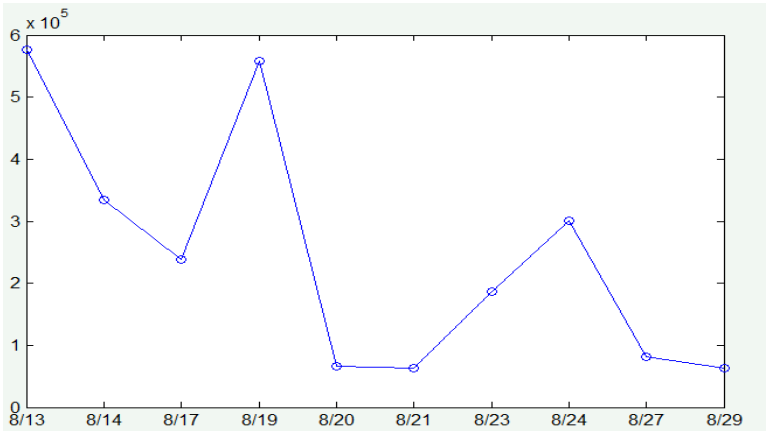

Fig.3 The Change Number of Documents

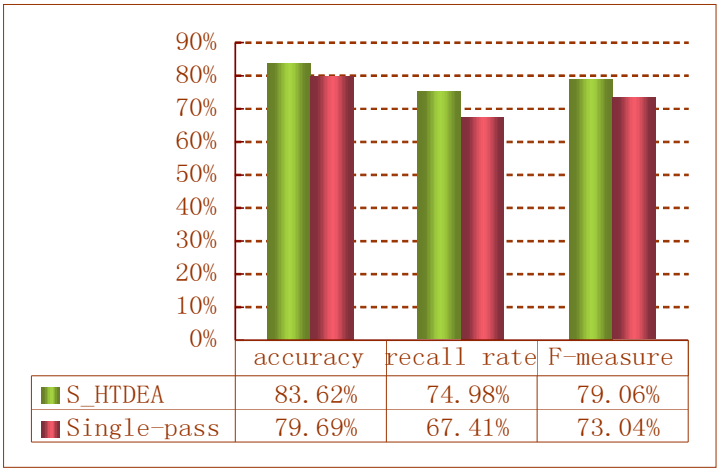

Fig.4 Performance Comparison

5. Conclusions

Microblog is mainly used to keep a record of the matter what is happening around the word and to strengthen communication among users.By detecting hot topics in microblog, we can acquire internet hotpots.In this paper, we present an algorithm based on comprehensive similarity. We synthesize semantic similarity and context similarity to cluster key words. Finally, hot topics list could be obtained. The experimental result has shown that this algorithm could detect hot topics accurately. In our future work,we will optimize the algorithm, to make it have better performance, on the basis of this, we will start to do research on the evolution of topics.

Acknowledgment

We extend our gratitude to the anonymous reviewers for their insightful comments.

References

[1] SHARIFI B M, HUTTON A, KALITA J K. Automatic microblog classification and summarization[C] Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg: The Association for Computational Linguistics, 2010: 685-688.

[2] INOUYE D.Multiple post microblog summarization [R]Colorado Springs,GA: University of Colorado at Colorado Springs, 2010.

[3] W. Luo, M.Yu, H. Xu. etc The Study of Topic Detection Based on Algorithm of Division and Multi-level Clustering with Multi-strategy Optimization[J]. Journal of Chinese Information processing 2006, 20(1):

29-36.

[4] K. K. Bun and M.Ishizuka, Topic Extraction from News Archive Using TF*PDF Algorithm[C]. In Proceedings of the 3rd Web Information Systems 2002, 73-81.

[5] James Allan, RonPaPka, VietorLavrenko.Online new event detection and tracking [C]. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.Melboume: ACM Press. 1998.37-45.

[6] R.Lu, L.Xiang, M.R.Liu. Extracting News Topics from Micro logs based on Hidden Topics Analysis and Text Clustering[A] // Proceedings of the sixth International Conference on Information Retrieval 2010[C]. 2010: 291-298.

[7] F.Yin,W. Xiao, B.Ge, F. Li. Incremental algorithm for clustering texts in internet-oriented topic detection[J]. Application Research of Computers, 2011.28(1): 54-57.

[8] BLEID M, LAFFERTY J D. A correlated topic model of science[J]. Annals of Applied Statistics, 2007,1(1): 17-35.

[9] ASUNCION A,SMYTH P,WELLING M. Asynchronous distributed learning of topic models [C]// NIPS 2008: Proceedings of the 22th Annual Conference on Neural Information Processing Systems. Atlanta: Curran Associates Inc, 2008: 81-88.

[10] RAMAGE D, UMAISS T, LIEBLING D J. Characterizing microblogs with topic models[C] //Proceedings of the Fourth International Conference on Weblogs and Social Media.Menlo Park: AAAI Press, 2010:130-137.

[11] Bejan, C. A. and Harabagiu, S. 2008. Using clustering methods for discovering event structures[C]. In Proceedings of the 23th National Conference on Artificial intelligence 2008(3): 1776-1777.

[12] Y. Hong, Y. Zhang, T. Liu. Topic Detection and Tracking Review[J]. Journal of Chinese information processing, 2007, 21(6):71-85.

[13] Y.Zeng,H.Xu.Research on Internet hot spot information detection [J]. Journal Communications, 2007, 28(12):141-146.

[14] Q. Liu, S. Li. Word Similarity Computing Based on How-net[C]//The 3rd Chinese Lexical Semantics Workshop. 2002: 59-76.