An Efficient Microblog Hot Topic Detection Algorithm Based on Two Stage Clustering

Yuexin Sun[1], Huifang Ma[1], Meihuizi Jia[1], and Wang Peiqing[2]

[1] College of Computer Science and Engineering,
Northwest Normal University Lanzhou Gansu, China
{sunyuexintc,jmhuizi0424}@163.com, mahuifang@yeah.net
[2] Anding District People's Armed Forces Department, Dingxi Gansu, China
wangwsnwnu@126.com

Abstract. Microblog has the characteristic of short length, complex structure and words deformation. In this paper, a two stage clustering algorithm based on probabilistic latent semantic analysis (pLSA) and K-means clustering (K-means) is proposed. Besides, this paper also presents the definition of popularity and mechanism of sorting the topics. Experiments show that our method can effectively cluster topics and be applied to microblog hot topic detection.

Keywords: Probabilistic Latent Semantic Analysis, Topic detection, Microblog, K-means.

1 Introduction

Nowadays, Micro-blogging is increasingly extending its role from a daily chatting tool into a critical platform of spreading real-time information during emergencies. It provides a short and convenient way for users to express and share their attitudes instantly. Since it enables users to acquire near real-time information, it has become a major source for producing and spreading hot events on the internet. However, acquisition of information regarding real-time hot events and news poses significant challenges due to the extensive irrelevant personal messages and unstructured short texts. Recent research of hot topic detection started to focused on the micro-blog [1-3]. Vector space model is always adopted as the way to represent microblog [4]. A great amount of efforts have been made for microblog hot topic detection in a semantic analyzed way. Some people extend short texts information by Knowledge Base, such as WordNet or Wikipedia [5-8]. In addition, topic model is also an effective way for microblog hot topic detection [9]. However, all of the above methods take advantage of external knowledge and they fail to consider some prior knowledge for microblogs.

In this paper, we formulate the task of microblog hot topic detection as a two-stage clustering algorithm based on probabilistic latent semantic analysis (pLSA)[10] and K-means[11] clustering (K-means). Furthermore, the definition of popularity(hotness) and mechanism of sorting of the topics are presented.

The basic outline of this paper is as follows: Section 2 presents details of our approach. The experiments and results are given in Section 3. Lastly, we conclude our paper in Section 4.

Z. Shi et al. (Eds.): IIP 2014, IFIP AICT 432, pp. 90–95, 2014.

2 The Proposed Method

In this section, we will present the details of our algorithm for building topic model and hot microblog acquisition.

2.1 Basic Idea

The conservative pre-process approach is taken to be the first step, which means that the microblog will be filtered and segmented first. Then the stopwords are eliminated and the traditional Chinese are translated into simplified Chinese. All terms occurring less than 20 times are removed.

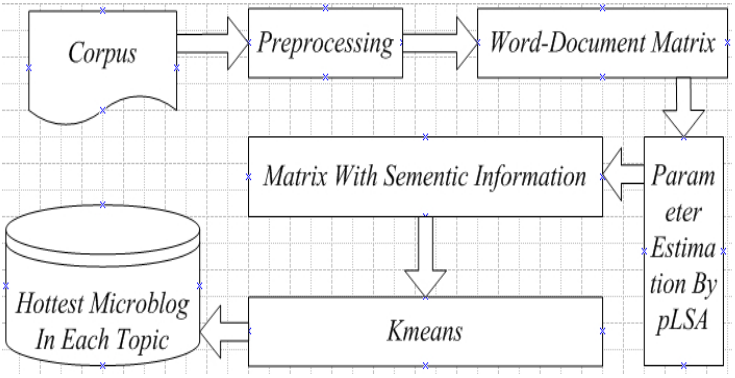
Corpus
Preprocessing
Word-Document Matrix
Matrix With Sementic Information
Param eter Estima tion By pLSA
Hottest Microblog In Each Topic
Kmeans

Fig. 1. Algorithm process

Specifically, pLSA model is first applied on the traditional term-document matrix to estimate two matrix denoting the distribution of topic under certain document and distribution of term under given topic, respectively. Then we multiply these two matrixes to obtain the word-document matrix with latent semantics, which becomes the input for the next stage clustering algorithm. We sort the probability of a document under each topic, the biggest one will be regarded as the centre for K-means clustering. Figure 1 shows the entire algorithm follow:

2.2 Our Approach

2.2.1 Pre-process of Microblogs

a). For microblog which is directed at another user, as denoted by the use of the @username token, we believe that the content of this field has little to do with real content. Therefore we just simply delete this field.

b). For microblog that contains "RT @" field, these microblogs are produced by someone else but the user copies, or forwards, them in order to spread it in his network. This is not treated as the real content. So we just keep the real content of this microblog.

c). For microblog that contains “# certain topic (user) #" microblog, the field can be considered as the noise, we just delete it.

d). We take advantage of stemmer for segmentation and remove the stop words at the same time. The stop list is based on 1028 stop words provided by sina.

2.2.2 pLSA Parameter Estimation and K-means Clustering

We first use TF*IDF[4] to represent the importance of each word, and the word-document matrix is constructed. pLSA is then used to estimate the parameters: distribution of topic under document d_i $P(z_k|d_i)$and distribution of term under topic z_k $P(w_j|z_k)$.

pLSA uses a generative latent model to perform a probabilistic mixture decomposition. This results in a more principled approach with a solid foundation in statistical inference. More precisely, pLSA makes use of a temperature controlled version of the Expectation Maximization algorithm for model fitting, which has shown excellent performance in practice. The word-document matrix with semantic information is then obtained by multiplication of these two matrixes. This matrix also becomes the input for K-means clustering.

K-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This algorithm is sensitive to the predefined centre. In our algorithm the biggest probability of a document under each topic serves as the centre for K-means clustering.

We also define the heat[12] of microblog:

$$HT_i = N\_com_i + \sqrt{N\_rel_i} + \log\left(fan_i + 1\right) \tag{1}$$

HT_i represents the heat of the microblog *i*, N_com_i is the comment number of microblog *i*. N_rel_i is the forward number of *i*. fan_i is the fans number of the microblog. We therefore can sort the hottest microblog in each topic by its heat.

3 Experiments

3.1 Data Corpus

Two thousand two hundred and twenty one microblog entries were collected in two phases. Both the training and test data of microblog entries are manually collected from sina microblog sites during 2011-11-01 to 2011-11-03 and 2013-03-10, respectively. The training dataset is composed of 37 topics.

Two baseline methods K-means and LSA+Kmeans are adopted to verify effectiveness of our model. As for the parameters of our approach, to give these algorithms some advantage, we set the number of clusters equal to the real number of microblog clusters.

3.2 The Experimental Results and Analysis

The experiments include three parts: 1) Overall evaluation of our model by comparing with that of 2 other algorithm; 2) Clustering results of our method; 3) Hot topic Visualization.

3.2.1 Overall Evaluation

We run our algorithm on the training data. We have run 5 times for each evaluation, and the average performance scores are reported. Four clusters with largest number of documents are selected. To evaluate the performance of our algorithm, we use precision[13] as our measures. Table 1 and Table 2 show the results.

From the experimental results, we can see that our method is superior to the former two algorithms in terms of precision. This is because the traditional VSM method simply considers the term occurrence information while ignores the semantic information, therefore it has difficulty in represent high dimensional and sparse vector. LSA performs decomposition of matrix with some negative numbers. All these characters are not in conformity with the microblog' feature.

Table 1. Performance of Precision on training datasets of our method

Topic	Correct cluster	Incorrect cluster	Precision
1	25	12	0.68
2	22	13	0.63
3	33	6	0.85
4	26	10	0.72

Table 2. Performance of Precision on training datasets of 3 methods

Methods	Precision
Kmeans	0.55
LSA+Kmeans	0.65
Our method	0.72

3.2.2 Clustering Results

Figure 2 demonstrates the clustering results on the test data. In the figure, each colour stands for one topic. Since the coordinate system is two-dimensional, we select each microblogs' two largest vector dimensions to show the cluster results. We can see that the clustering result is basically in conformity with clustering standard.

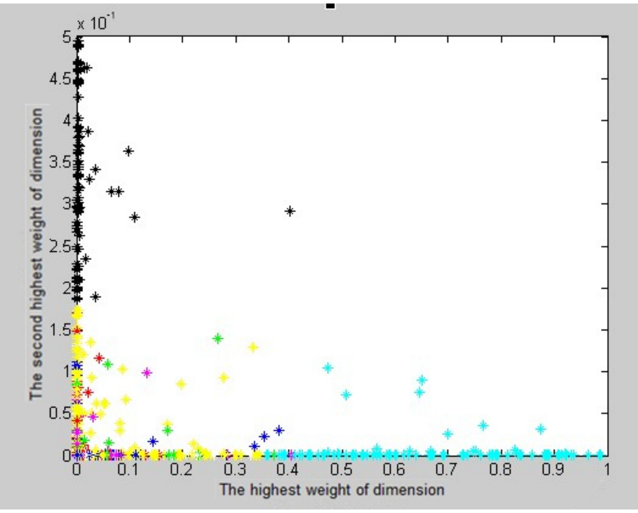

Fig. 2. Cluster result

3.2.3 Hot Topic Visualization

We choose the hot topic with highest HT value from five clusters with largest number of microblogs. Table 3 lists the five hottest topics on the test data, while table 4 lists terms with highest probabilities for each topic. It is clear that our method is able to discover meaningful topics with much better readability and the experimental results are consistent to our intuition.

Table 3. Hot topic with highest HT value from five clusters

Topic	Microblog
1	Happy New Year 2013! (祝大家2013年新年快乐！欢度元旦节!)
2	Since many countries' policies are difficult to change, China's Internet should find reasons from itself. (中国互联网应从自身找原因，很多国家政策是难改变的，只能适应。）
3	In order to implement the value of its currency reduction goals, zero interest rate policy guides the domestic capital to high interest rates countries by the means of carry trade. (零利率政策主要借助套利交易，引导境内资本流向高利率国家，实现降低本币汇率的目标。)
4	Luo Yufeng is recently named the most shameless person in China by CNN, who has defeated Yang Erche Namu and sister Furong, following Xiao Shenyang who has achieved the China's most vulgar person. It seems that Americans appear to be fair on the issue of shame. (继小沈阳被美国国家有线广播电视台CNN评为中国最低俗的人之后，最近罗玉凤凤姐一路打败杨二车娜姆和芙蓉姐姐,被CNN评为中国最不要脸的人，看来美国人在判断不要脸这个问题上非常之公正。)
5	Yesterday, China's government would manage melamine by the real-name system. It's estimated that Mengniu is related to melamine. Anyway, we had better not drink milk. We should not believe the words that our management is good. (昨天国内高调报道中央要以实名制管理三聚氰胺。虽然不知道蒙牛到底出了什么事，但估计又是三聚氰胺惹的祸。反正牛奶是再也不能吃了。绝不能相信说管理得很好的话。)

Table 4. Terms with highest probabilities for each topic

Topic	Keywords				
1	New Year (新年)	Wish (愿望)	Promised (许下)	Adjust (调整)	Notice (公告)
2	Quantization (量化)	Lead (引导)	Carry out (实施)	Policy (政策)	domestic currency (本币)
3	World (世界)	Currency (货币)	Loose (宽松)	Risk (风险)	Capital (资本)
4	Entertainment (娱乐)	Luo Yufeng (凤姐)	Choose (评为)	Vulgar (低俗)	Shameless (不要脸)
5	Mengniu (蒙牛)	Find (发现)	Report (报导)	Milk (牛奶)	Confiscate (查收)

4 Conclusion

In this paper, we have described a new algorithm for hot topic detection on microblogs. While it is based on two stages clustering, the second stage Kmeans can fully take advantage of the prior information provided from the first stage pLSA. We also define the concept of topic hotness. In the future, we will take users' role into consideration.

Acknowledgement. Supported by the National Natural Science Foundation of China (61363058), the Scientific Research Fund for Colleges and Universities of Gansu Province (2013B-007,2013A-016), Youth Science and technology support program of Gansu Province (145RJZA232,145RJYA259), and Promotion Funds for Young Teachers in Northwest Normal University (NWNU-LKQN-12-14).

References

1. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, New York, USA, pp. 851–860 (April 2010)
2. Chen, J.F., Yu, J.J., Shen, Y.: Towards Topic Trend Prediction on a Topic Evolution Model with Social Connection. In: Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 153–157. IEEE Computer Society, Washington (2012)
3. Chen, Y., Xu, B., Hao, H., Zhou, S., Cao, J.: User-defined hot topic detection in microblogging. In: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, New York, USA, pp. 183–186 (August 2013)
4. Salton, G.: The SMART retrieval system—experiments in automatic document processing, Upper Saddle River, NJ, USA (1971)
5. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th International Conference on World Wide Web, New York, USA, pp. 377–386 (May 2006)
6. Yih, W.T., Meek, C.: Improving similarity measures for short segments of text. AAAI 7(7), 1489–1494 (2007)
7. Zhai, Y.D., Wang, K.P., Zhang, D.N., Hunag, L., Zhou, C.G.: An algorithm for semantic similarity of short text based on WordNet. Acta Electronica Sinica 40(3), 617–620 (2012)
8. Banerjee, S., Ramanathan, K.: Clustering short texts using wikipedia. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, pp. 787–788 (July 2007)
9. Ma, H.F., Wang, B.: Microblog Online Event Analysis Based on Incremental Topic Model. Computer Engineering 39(3), 191–196 (2013)
10. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1-2), 177–196 (2001)
11. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1(281-297), p. 14 (1967)
12. Sun, S.P.: Chinese microblog hot topic detection and tracking technology. Beijing Jiaotong University, Beijing (2011)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)