Topic Detection of Chinese News Based on Word Entropy

Bo Zhu[1, a], Min Hou[2,b] and Yuyin He[3,c]

[1,2]Broadcast Media Language Branch, National Language Resources Monitoring and Research Center,Communication University of China, China

[3]College of Foreign Languages, China

[a]zhubo812@ cuc.edu.cn, [b]houmin@cuc.edu.cn, [c]yuyinhe@huaa.edu.cn

Keywords: word entropy; topic detection; topic word co-occurrence net; modularity measure

Abstract. We propose a method of automatic news topic detection in large-scale data. First, topic words are detected based on their word entropy. Then, the topic word co-occurrence net is constructed via the semantic relationships of topic words represented by their orders in which they appear within the original text. Finally, implied communities are detected in the topic word co-occurrence net through modularity measures. Each implied community is regarded as a news topic. Experimental results show that this method can be used to effectively identify the key topic of each news report, with the presence of topic content in human-readable form.

Introduction

Topic detection task refers to discovering topics contained in text data without priori knowledge. In order to achieve this purpose, the text data relevant to the same topic should be aggregated and the content of the topic can be represented by a group of words. Thus topic detection task is similar to the clustering process of text data. Therefore, the previous studies of topic detection mainly focus on selection and optimization of clustering methods. For example, in early TDT evaluation the clustering algorithms often used are Kmeans[1] and Single-Pass[2], and then researchers make some improvement on the basis of the clustering algorithm[3,4,5]. The drawback of clustering algorithms is high complexity of time and space, and the quality of clustering centers will directly affect accuracy of the following topic detection.

Besides clustering methods, topic models have been widely applied in topic detection tasks. A representative topic model is LDA (Latent Dirichlet Allocation) model[6] proposed by Blei, et al. This model, which is also a topic model, is a probability generative model constructed based on PLSI model[7]. The model can process large-scale data, for its parameters will not increase linearly with the growth of the collected texts. However, whether a topic model or a clustering method, they are both based on bag-of-words model, thus damaging semantic relationships of Chinese lexicon. Jianhong Shi et al[8] use frequent item set algorithm, the FP-Growth, to uncover the collocation relationships of words, so as to obtain the semantic relationships among words.

Frequent item refers to repeated word combinations or high-frequency combinations of multiple words co-occurred in the texts. However, FP-Growth algorithm can only find relevant factors but is incapable of discovering their ordinal relation. Besides, some other researchers extract topic words and then construct a topic word co-occurrence map to discover topics[9,10]. These methods, with the same drawback of FP-Growth algorithm, fail to unveil sequenced relationships in the word map, which goes against representation of Chinese topics. In order to fill this research gap, this study makes use of word entropy to extract topic words and constructs a topic word co-occurrence net with sequenced orders among words, and finally achieves topic detection task through modularity measures.

The rest of the paper is organized as follows. Section 2 describes related work regarding word entropy and modularity measure. Section 3 details the proposed topic detection method based on word entropy. The experiments and result analysis come in Section 4 and Section 5 is conclusion.

Related Work

Word Entropy

The entropy of word indicates complexity of the word context. That is to say, the more types of words the word co-occurs with, the larger entropy of the word is. Huang et al[11] segment Chinese words by calculating the uncertainty of character strings on the left and right of a word string by means of entropy. In Chinese word segmentation, the larger value of the entropy of character strings on the left and right of a word string, the more uncertainty. This means the larger probability for the character string to become a word.

Generally, topic words will repeatedly appear in news reports. However, it is hard to distinct the topic words from other high-frequency words that are unrelated to the topic merely through word frequency computation. Compared with other non-topic high-frequency words, the topic words in news reports appear in the title, lead and text repeatedly, and describe the same or similar content. Collocations of the topic words in the context with punctuation centered are relatively stable from the perspective of entropy. Therefore, the word entropy can be used to evaluate the possibility of a word as a topic word. Larger word entropy means smaller semantic discrimination, whereas smaller word entropy indicates larger semantic discrimination of a word. The content of core topic in a text unit generally extends through the full text, but for a punctuation sentence, the words that constitute the core topic lack the capability of semantic distinction. Therefore, after the basic stop words are filtered out, the word of larger entropy than threshold value can be taken as the topic. Eq.1 is the entropy calculation formula of the ith word in a text:

$$H\left(w_i\right)=-\sum_{i=1}^{N} p\left(n_{w_i},N\right)\log p\left(n_{w_i},N\right) \tag{1}$$

Wherein, N represents the total number of words in a punctuation sentence which contains the word w_i; and n_i represents the frequency of the word w_i in the punctuation sentence. Punctuation sentence is the unit for calculating the entropy of a word in a text set. All the punctuation sentences in a news report constitute a text set for calculating the word entropy of the news report.

Modularity Measure

Modularity measure[12] is often used for measuring community stability in the network. The topic words extracted in the text are used to construct the topic word co-occurrence net, from which the topics are detected via the modularity measure. The text adopts the FastUnfolding algorithm proposed by Blondel et al to realize the value of modularity.

The basic steps are as follows:

(1) Initialize node data, and classify every node into different communities.

(2)Traverse every node, and calculate the modularity gain of each node according to Eq.2. If the maximum gain is larger than 0, classify it into the adjacent community; otherwise, keep it in the original community, till the community which the node belongs to does not change.

(3) Construct a new map, and the points in the new map represent the different communities generated in the last stage. Weight of sides is the sum of the weights of all the sides between every two nodes in two communities. Repeat the second step until the maximum value of modularity is obtained.

The aforesaid three steps can be divided into two stages. The first stage comprises Step (1) and (2) to determine the community of each node. The second process contains Step (3) for constructing the new map and repeating the first stage until the value of modularity does not change.

The specific calculation formula of modularity measure is as follows:

$$\Delta Q=\left[\frac{\sum in+k_{i,in}}{2m}-\left(\frac{\sum tot+k_i}{2m}\right)^2\right]-\left[\frac{\sum in}{2m}-\left(\frac{\sum tot}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right] \tag{2}$$

Topic Detection Method Based on Word Entropy

In the topic detection method based on word entropy, firstly, words of the text are segmented and the basic stop words are filtered out in the preprocessing stage. Then entropy of the word is used for obtaining topic word candidate set to construct the topic word co-occurrence net. Constructing the topic word co-occurrence net requires that the multiple topic words in the same punctuation sentence form the Bigram combination to maintain the semantic relationships of the topic words. Finally, after modularity measures applied in the topic word co-occurrence net, every implied community in the topic word co-occurrence net becomes a topic. The basic procedure of this method is shown in Fig.1.

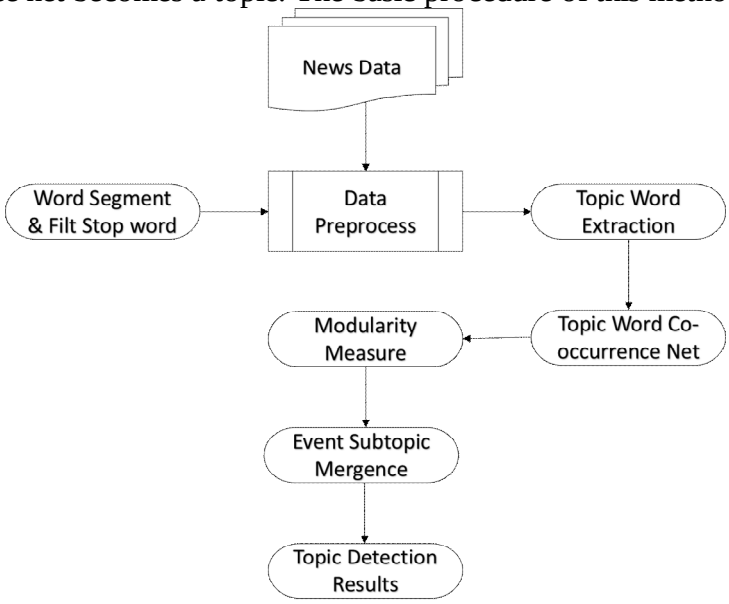

Fig.1 Flow diagram of topic detection

Extraction of Topic Words

Obtaining the topic of a document means to detect its main content, which will generally repeat several times in the full document. A document can be regarded as a set constituted by multiple punctuation sentences. The content of the topic will appear repeatedly in multiple punctuation sentences, thus the semantic distinction degree of topic words in the punctuation sentences is relatively low, while the entropy is high.

Table 1 example of entropy of candidate topic words

NO.	candidate topic words	entropy
1	Guinea	1.6277
2	Virus	1.540761
3	Ebola	1.53025
4	Death	1.276342
5	Epidemic situation	1.226117
6	March	1.22211
7	People	1.123355
8	Outburst	0.845367
9	Case of illness	0.793766
10	59	0.75715

Construction of Topic Word Co-occurrence Net

After the topic words candidate set is obtained, the multiple topic words in the same punctuation sentence are constructed as the Bigram combination according to the order in which they appear in the original text, and then count them. For instance, Example 1 and Example 2 are punctuation sentences of a same news report. Every topic word is taken as a node of the network, and the Bigram itself records the direction of the side between two nodes, and the number records the weight of the sides.

Example 1 Guinea before no find have been/done Ebola virus

Example 2 Ebola virus generally by blood and other body fluid diffuse

Table 1 includes three topic words which are “Ebola”, “virus” and “Guinea” in Example 1 and Example 2. And the Bigram combination is shown in Table 2.

Table 2 bigram combination constituted by topic words

Bigram combination	frequency
Guinea Ebola	1
Ebola virus	2

The bigram combinations that reach threshold value are selected to construct the topic word co-occurrence net, whose result is shown in Fig.2.

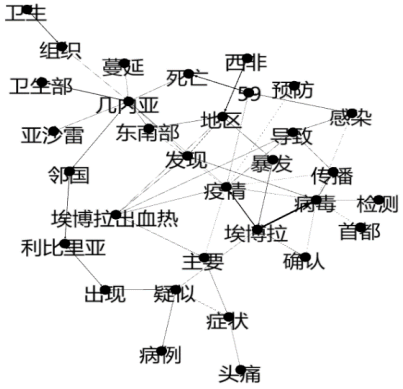

Fig.2 drawing of topic word co-occurrence net

The bigram combination can present ordinal collocations of the topic and filter them by means of entropy of the words. Any words that compose the topic words combination with the weighted entropy below the threshold value will be deleted.

Topic Detection

The topic word co-occurrence net contains multiple topics in the original data. In order to detect the independent topics from the topic word co-occurrence net, we regard the topic word co-occurrence net as a network and the implied communities contained in the network as topics. Therefore, topic detection can be seen as an issue of community discovering. This study adopts modularity measures to acquire the implied communities within the network.

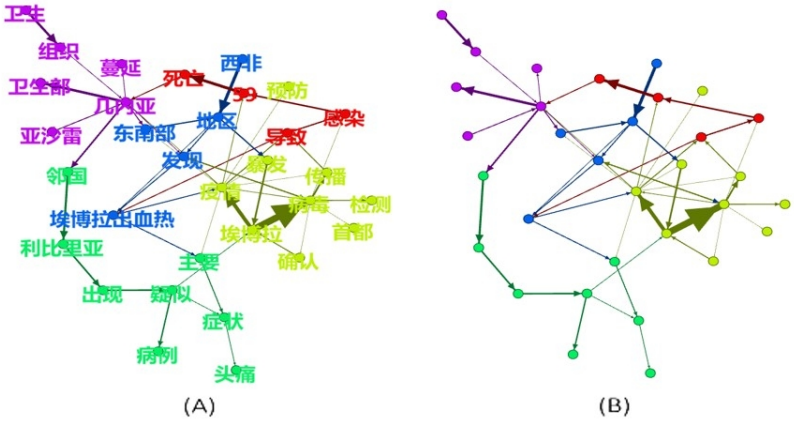

Fig.3 results of modularity measures

As shown in Fig.3, each node is a topic word, and every two nodes are connected by a directional side which represents the order in which the two nodes on the two ends of the side appear in the flow of speech. The thickness of the sides represents the degree of strength. The thicker the side, the higher the frequency of co-occurrence of the nodes on the two ends. Fig.3 shows the relevant topics of an event which contains the core topic and other sub-topics. Every topic in the topic word co-occurrence

net should be mapped to the original text to construct the text-topic matrix in processing the data set that contains multiple news events. In the end the different texts with the same core topic are correlated.

Sub-topic merging

Generally, a news report is corresponding to one core topic and multiple sub-topics, and a topic is corresponding to multiple news reports, as shown in Fig.4. The core topic and sub-topics of a text are extracted according to the method introduced in 3.3. A topic word co-occurrence net contains multiple topics of multiple texts. Judging the relationships of the topics in the topic word co-occurrence net mainly depends on whether the multiple topics belong to the same text set. If the multiple topics belong to the same text set, the topic of the highest side weight is the core topic, while other topics are sub-topics. For example, in Fig.4, Topic 2 exists in documents d_1, d_2 and d_4, while Topic 3 exists in documents d_2 and d_4. Thus Topic 2 and Topic 3 belong to the same text set, and if the side weight of Topic 2 is larger than that of Topic 3, Topic 2 is the core topic.

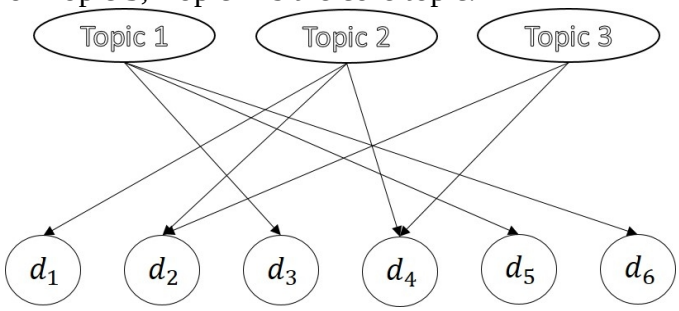

Fig.4 mapping of topic-document

Generally, every core topic has its own independent text set, but the sub-topics that are subject to different core topics might be the same or similar. In addition to detecting the core topic of a text, the topic detection method based on word entropy can also detect multiple sub-topics. Topic merging is not only capable of detecting the core topic in the data, but also capable of obtaining the sub-topics subordinated to the core topic as well.

Experiment and result analysis

Data Description

The data used in the news topic analysis is totally 16,447 news reports in 2014 from five major portal news websites, namely, Sina, Sohu, Netease, Tencent and PhoenixNet. There are 4,086 reports related to “Ebola epidemic situation”, 11,908 reports related to “Malaysia Airline incident”, and 1,389 reports related to “occupying central district of HongKong”. Every text is preprocessed, including sentence divisions, word segmentations via the CUC word segmentation software, stop words filtering and then topic words extraction.

Experimental Procedure

(1) Words of the text are segmented and the basic stop words are filtered out in the preprocessing stage.

(2) The entropy of each word in a news text is calculated and the average word entropy is taken as the threshold value for topic word extraction.

(3) The topic word co-occurrence net is constructed based on topic phrase bigrams.

(4) Modularity measures are applied in the topic word co-occurrence net for topic recognition.

(5) All the topics of a news event are merged via topic-document mapping.

Evaluation

The indicators such as precision and recall are used to evaluate topic recognition results. Their calculation are Eq.3 and Eq.4 as follows:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

Wherein, TP represents the number of correctly recognized texts, FP represents the number of incorrectly recognized texts and FN represents the number of texts which have not been recognized. F value is used to show the experiment effect by combining evaluation results of the indicators of both precision and recall. The calculation of F value is Eq.5 as follows:

$$F - measure = \frac{2PR}{P + R} \tag{5}$$

Result of Topic Detection

The entropy of words in every news report is calculated according to Eq.1. Because the lengths of every news reports are various, the threshold value of the word entropy should be adjusted in line with each news report.

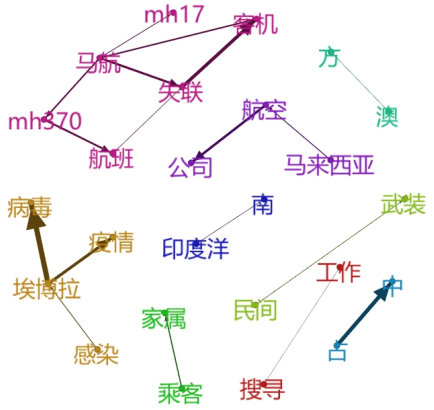

Fig.5 topic word co-occurrence net composed by Top 20 topic words

The topic word co-occurrence net is identified according to the number of communities obtained in Fig.5. Every side in the figure is directional, and the direction denotes the order in which the nodes on the two ends of the side appear in the flow of speech. The thickness of the sides represents the degree of strength. The thicker the side, the higher the frequency of co-occurrence of the nodes on the two ends. Fig.5 above is divided into 9 independent topics, but parts of the topics are sub-topics subordinated to the core topic. Every topic in Fig.5 is mapped to a group of texts and the relation of the topics are detected through the texts reflected by each topic.

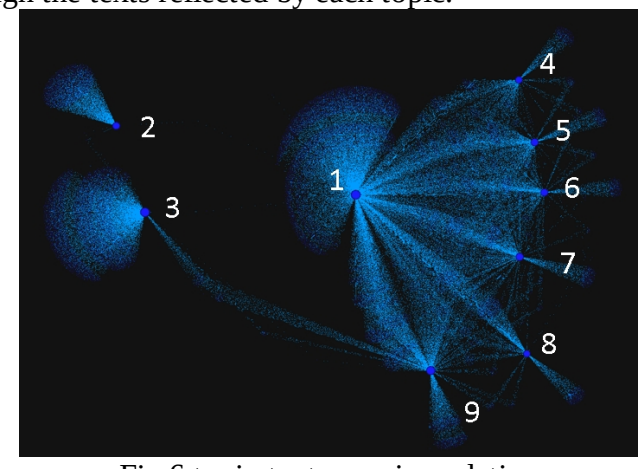

Fig.6 topic-text mapping relation

As shown in Figure 6, the nodes represented by the numbers in the figure are topics, and other splashes in light color stand for the texts related to the specific topics. Some texts only contain one of the 9 topics in Fig. 5, and some other texts contain multiple topics. For example, the Topics 4 to 9 and Topic 1 in the figure are mapped to the same texts. That is to say, these texts contain multiple topics which include one core topic and several sub-topics. For the texts containing multiple topics, its core topic is determined by judging the weight of each topic in the topic word co-occurrence net. The topic of the

highest weight is the core topic and other topics are sub-topics of this topic. The result of topic merging is shown in the following Fig.7.

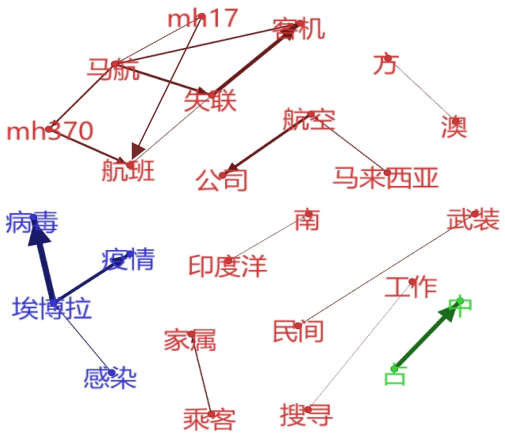

Fig.7 example of topic detection result

The table below is the statistical result of topic detection. The table shows the influence of the topic word co-occurrence net constituted by the topic words of different number on the accuracy rate and recall rate of topic detection.

Table 3 statistical result of topic detection

Number of bigrams	Accuracy rate	Recall rate	F-value
20	0.997617	0.770638	0.86956
25	0.997639	0.781568	0.876483
30	0.99735	0.801128	0.888534
35	0.99739	0.812128	0.895275
40	0.99744	0.829201	0.905573
45	0.96861	0.818989	0.887542
50	0.93634	0.796525	0.860775

As shown in Table 3, the method based on word entropy has favorable effects on accurate detection of the topics of the texts. Within a certain scope the number of bigrams in the topic word co-occurrence net is from 20 to 40. As the number of topic words increases, the accuracy rate and recall rate will grow with it; but when the number topic words is around 40 to 50, the accuracy rate and recall rate both begin to decrease. That means that, in the experiment, when the number of topic words in the topic word co-occurrence net is 40, the three different news events can be better distinguished. However, when the number of topic words is above 40, some topic words of worse distinguishing ability will enter into the topic word co-occurrence net with the increase of topic words, leading to a decrease of accuracy rate. On the whole, the recall rate is low. The reason for the low recall rate is mainly because some news reports are short and their content is just one aspect of the news events. The entropy of the core topic words is too low, thus the recall rate is reduced.

Contrastive Analysis of Topic Detection

The contrastive analysis of the results of topic detection mainly refers to the comparison between the result of the topic detection method introduced in this study and the topic detection result of LDA[14] model. LDA model uses a k-dimensional implied random variable complying with Dirichlet Distribution to represent the subject of the document and to simulate the generation process of the document. LDA model essentially is a three-layer Bayesian model which comprises three layers, namely, word, subject and document. Each document is represented as a text formed by multiple potential subjects, where each subject is a polynomial distribution in the fixed wordlist. All the

documents in the text set can share these subjects. The following Fig.8 is a representing graph of LDA model.

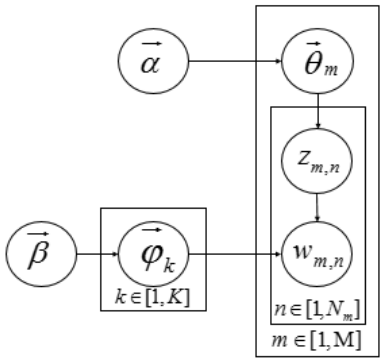

Fig.8 generation of LDA model

In the experiment the two kinds of news reports data used are the same. The parameter setting of the LDA model is as follows:

Topic Number = 3;

alpha=50/3;

beta=0.1;

iteration=200;

Wherein, Topic Number means the number of categories of the topics. Alpha and beta are two priori probabilities. Iteration refers to the times of iteration in the sampling process. LDA model represents the topics in the form of word clustering. Below a topic is represented with 20 topic words, and the statistical result of topic detection is as follows:

Table 4 statistical result of LDA model topic detection

Accuracy rate	Recall rate	F-value
0.79503	0.461283	0.583825

Comparing the statistical results of Table 3 and Table 4, the effect of topic detection based on word entropy has great improvement compared with LDA model. As for topic representation, the following table shows the topic representation result of LDA model.

Table 5 LDA result of news topics extraction

NO.	Ebola epidemic situation	Malaysia Airline incident	occupying central district of HongKong
1	Ebola	Malaysia airline	Hong Kong
2	America	Plane	China
3	Virus	Malaysia	Problem
4	Ukraine	Airliner	Government
5	Epidemic situation	Lost contact	Country
6	Infection	Look for	Represent
7	Russia	Flight	Society
8	Staff	MH370	Relation
9	Country	Family member	Develop
10	MH17	Search and rescue	Action

As shown in the table above, on the whole each topic is distinguished, but a few of topic words are wrongly placed. For example, “MH17”, “armed forces” and “crash” appear in the topic of “Ebola epidemic situation”. These topic words should belong to the topic “Malaysia Airline incident”. These topics in Fig.7 are not wrongly classified into the subnet of other topics. In addition, the topic words

contained in the three topics all include “represent”, but the word itself has no ability of distinguishing different events. However, LDA model is incapable of excluding this kind of words that have no semantic distinction ability. These topic words can only be excluded through the stop words list defined by users. However, if the data to be processed is entirely unknown, the stop words list defined by users might cause loss of key topic words, and then cause inaccurate expression of the topic content. The topic detection method based on word entropy first uses the comprehensive word entropy, and then filters out the words that cannot be distinguished existed in multiple texts. Finally, compared with the topic representation of LDA model, topic representation based on topic word co-occurrence net is more readable and accurate in expressing meaning.

Conclusions

It is significant to extract topic content with complete semantic information from the massive text data for a better and quicker understanding of text data, and thus we propose a topic detection method based on word entropy so as to achieve this goal. This method detects news topic through steps of preprocessing, topic words extraction and topic word co-occurrence net construction. The results of the experiment prove that the method can effectively detect the topic content in the text data and judge the news events which the specific news reports belong to according to their core topic.

There are some limitations in our study. Although this study is capable of effectively detecting hot topics with concentrated data, the news topic could be left out when the number of news reports related is too small. In addition, the news report that contains multiple topics belonging to different news events might be falsely detected to some extent.

Acknowledgements

The research was supported by Project Funds of Humanities and Social Sciences from National Language Resources Monitoring & Research Center under grant YZYS15-04 and the Project Funds of Humanities and Social Sciences from Ministry of Education of China under grant 11YJA740030 and Beijing Municipal Project Funds of Humanities and Social Sciences under grant 13WYB012.

References

[1] Ron Papka. On-line new event detection, clustering, and tracking. PhD thesis, University of Massachusetts Amherst, 1999.

[2] James Allan, Ron Papka, and Victor Lavrenko. On-line new event detection and tracking. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 37–45. ACM, 1998.

[3] Yi Xiaolin, Zhao Xiao, Ke Nan, and Zhao Fengchao. An improved single-pass clustering algorithm internet-oriented network topic detection. In Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on, pages 560–564, June 2013.

[4] Yaohong Jin. A topic detection and tracking method combining NLP with suffix tree clustering. In Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on, volume 3, pages 227–230, March 2012.

[5] Chunshan Li, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, Shengchun Deng, and Xiaofei Xu. Clustering based topic events detection on text stream. In Intelligent Information and Database Systems, pages 42–52. Springer, 2014.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.

[7] Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57. ACM, 1999.

[8] Jianhong Shi, Xingshu Chen and Wenxian Wang. Topic Discovering of Chinese Weibo Based on Implied Subject Analysis [J]. Application Research of Computers,2014,03:700-704.

[9] Wenqing Zhao and Xiaoke Hou. Topic Dtection of Chinese Weibo News Based on Word Co-occurrence [J]. CAAI Transactions on Intelligent Systems,2012,(5).

[10] Zhongming Han, Hui Zhang, Meng Zhang, and Jinhui Huang. Fast Topic Discovering Method and Evaluation Research of Massive Short Texts [J]. Application Research of Computers,2015,(3).

[11] Jin Hu Huang and David Powers. Chinese word segmentation based on contextual entropy. In Proceedings of the 17th Asian Pacific conference on language, information and computation, pages 152–158, 2003.

[12] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008, 2008.

[13] G Doddington and J Fiscus. The 2002 topic detection and tracking (tdt2002) task definition and evaluation plan. Technical report, Technical Report, 2002.

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.