Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **3**, September **2019**

Emergency Event Topic Clustering Based on Keyword Co-occurrence

Lyuchao Liao[1], Chenwei Wu[1,*], Zhijie Chen[2], Liying Qu[3], Bingyu Huang[4] and Zhengyu Chen[4]

[1]Fujian Provincial Big Data Research Institute of Intelligent Transportation,
Fujian University of Technology, 350118, China
[2] Fujian Early Warning Information Release Center, 350001, China
[3] Water Resources & Hydropower Research Institute of Fujian, 350001, China
[4] Evecom Technology Co., Ltd., 350001, China
Email: 943204646@qq.com

Received August 2019; revised September 2019

Abstract. *Internet information is an important source of urban emergency events; however, how to quickly identify urban emergency thematic events from large-scale data remains a challenge. At present, the existing common methods mainly use LDA to conduct topic analysis, and then the topic number is furtherly judged by its perplexity; but it often leads to problems with too many topics and even no obvious discrimination between topics. Thus, based on the perplexity of topics, an emergency event topic clustering method was proposed to furtherly analyze the LDA topics information by the co-occurrence degree of topic keywords. This method is applied in the microblog data set collected from seven official traffic microblogs in Fuzhou, and the results show that five key topic categories which are worthy of attention are discovered, including weather information in Fuzhou, safety tips for citizens, traffic news, traffic congestion and alteration of routes and stations. Verifications show that these topics can be considered as the main types of information published by the official Weibo related to transportation in Fuzhou; so it could provide a feasible method for the automatic perception discovery of urban emergency events.*
Keywords: Emergency events; LDA analysis; Topic models; Topic clustering; Keyword co-occurrence

1. **Introduction.** With the rapid development of social media in recent years, in China, Weibo is regarded as an important medium for disseminating information on emergency events. More and more people are willing to share what they have seen and heard on the road, such as sharing traffic information and road accidents during their journeys, and many government departments and large transportation companies also use Weibo to publish traffic information for public. Now, most traffic-related organizations often publish traffic information via Weibo, which includes not only various information about road traffic events, weather events, etc., but also a lot of other emergency-related information, so understanding these topics can provide an important fusion processing mechanism for big data of emergency.

In order to find the topic of text data, some scholars in the early research used the Space Vector Model VSM (Vector Space Model)[1; 2] to model the text, and then use the clustering algorithm to group the texts which belonging to the same topic under the same cluster, such methods can distinguish between text categories, but they cannot effectively

exploit the underlying semantic information contained in the text[3]. The LSA(Latent Semantic Analysis) model[4],compared with the text-word mapping proposed by the vector space model, proposes the concept of "semantic dimension", which transforms the text-word mapping into a text-semantic-word mapping, though it can effectively represent text semantic information, but the model is computationally complex and cannot solve the polysemy problem effectively. Then, Pan proposed a Fast clustering algorithms for vector quantization and saved more than 99% in the number of multiplications, as well as considerable saving in the number of additions[5]. In 2003, Blei proposed the LDA(Latent Dirichlet Allocation)[6] topic model, which regards text as the probability distribution in the topic space, and the topic is regarded as the probability distribution in the dictionary space, so the model is also a three-layer Bayesian model based on text-topic-word.

At present, many researches on text topic mining are based on the LDA topic model[7]. For example, Yang et al.[8] used the LDA topic model to mine the most important topics in newspapers during a specific time period. Griffiths et al.[9] had used LDA topic model to find research topic trends by looking at abstracts of scientific papers. Sun et al.[10] summarized the articles related to transportation journals uses the LDA topic model to conduct topic analysis, which reveals the research prospects in the transportation field. In addition, there are a series of improved topic models based on LDA topic model, such as Author-LDA[11],Topic-Link LDA[12], Twitter-LDA[13], Behavior-LDA[14] and so on.

Weibo data contains a lot of valuable information. For example, Chen et al.[15] mined the social internal-body value model of passengers through Sina Weibo of China southern airlines. This study aims to apply the LDA topic modeling to the seven traffic-related official and enterprise Weibo data sets in Fuzhou. At the same time, based on the number of topics determined by perplexity, further categorize the results of the Weibo topic modeling by the topic keywords co-occurrence, and the classification result can be regarded as a summary of the information published by the official Weibo.

The rest of this paper is divided into the following structure. The section 2 mentions related work. The section 3 introduces the research methods used in this paper. In the section 4, the experimental process is described. In the section 5, we analyzed the results. And in the section 6, we discussed and summarized the results.

2. **Related Work.** In recent years, the development of self-media has made the network information more fragmentary. Short text data such as Weibo and Twitter are full of the whole network space, which contains a large amount of valuable information, and people are paying more and more attention to the thematic analysis of these short text data[16]. At present, many scholars use LDA topic model for text topic mining. For example, Wang et al.[17] used LDA topic model to conduct topic modeling of Chinese Sina Weibo data, so as to identify and describe health-related topics discussed by Chinese people on Sina Weibo. Hidayatullah et al.[18] used the LDA topic model to create thematic models of traffic-related information contained in Twitter messages released by Indonesia's traffic management center. Some scholars also use the LDA topic model to analyze the Twitter data set shared by the official Twitter account of BMKG on Java island and dig out several valuable topics[19]. However, when using LDA topic model to mine text data, it is a difficult problem to effectively determine the number of topics. Based on this, many scholars have carried out relevant researches, for example, Griffiths et al.[9] proposed to use Bayesian model to determine the optimal number of topics, but this method relies on Gibbs sampling process and has high computational complexity. Arun et al.[20] found that KL divergence can be used to measure topic similarity. When the number of topics is close to the optimal value, KL divergence is smaller, whereas it is large.

The work in this paper is to determine the number of topics through the concept of perplexity degree proposed by Blei et al. Perplexity is a common method to determine the number of topics in LDA topic model. As the number of topics increases, the perplexity value will show a decreasing trend. The number of topics can be determined by selecting a model with smaller perplexity value. However, the number of topics selected by perplexity degree usually tends to be large, which leads to the problems of large similarity and low identification between the extracted topics. In view of this, this paper proposes a method to classify similar topics in LDA topics mining results, starting from topic similarity, on the basis of initially determining the number of topics through perplexity degree.

3. **Model and Method.** In this paper, the perplexity value is used to judge the number of topics, and the LDA topic model is used to extract Weibo topics. Then we analyze the similarity of the extracted topics and classify the similar topics into one category.

3.1. **Latent Dirichlet Allocation.** LDA topic is a probabilistic generation model for mining potential topics in text, which was first proposed by Blei et al. (2003). It is a three-layer Bayesian model, which is divided into document layer, topic layer and keyword layer[21]. Each layer is controlled by corresponding random variables or parameter. The model holds that each document is formed by a series of topics mixed according to a certain probability, and each topic is formed by a series of words mixed according to a certain probability.

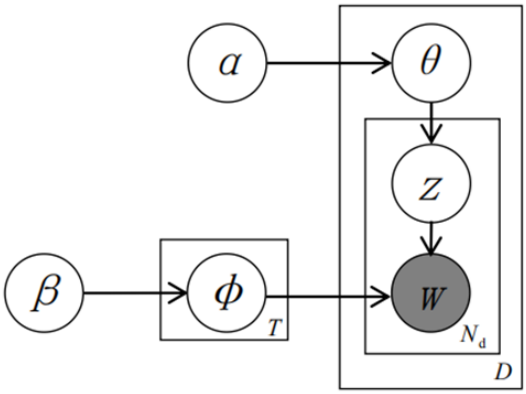


FIGURE 1. LDA topic model schematic

Where D represents the total number of documents in the database, T represents the number of topics, θ represents the topic distribution of the document and φ represents the word distribution of the topic. α and β are the Dirichlet Hyperparameter of θ and φ, respectively. Z represents the topic, and w represents the key word that belongs to the topic Z.

According to the text topic model schematic, the joint probability distribution of all explicit and implicit variables in the given hyperparameter case could be further derived [22]:

$$p(g_m, z_m, \theta_m, \Phi|\alpha, \beta) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(g_{m,n}|\varphi_{z_{m,n}}) \cdot p(z_{m,n}|\theta_m)}_{Words} \cdot p(\theta_m|\alpha)}^{Weibo-like\ messages} \cdot \underbrace{p(\Phi|\beta)}_{Topics} \tag{1}$$

In LDA, the process of generating documents is modeled as a three step process.

(1)For the document d, a distribution over topics Z is sampled from a Dirichlet distribution θ.

(2)For word in the document d, a single topic is chosen according to the distribution.

(3)Each word is sampled from a multinomial distribution φ over words specific to the sampled topic.

3.2. **Perplexity.** The perplexity value is a common index to evaluate the performance of the language model. It is considered that the language model which gives the test set a higher probability value is better. Generally speaking, the value of perplexity will show a decreasing trend with the increase of the number of topics. The formula as followed:

$$perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{2}$$

A total of M documents in the D form corpus. N_d represents the number of words in each document d , w_d represents the words in the document d , and $p(w_d)$ is the probability of the words w_d in the document.

3.3. **Similarity Evaluation for Topic Clusterting.** In this paper, for a certain topic set, the two topics where the largest co-occurrence degrees of keywords appear in the original data are similar topics. To be specific, we further classify the topics through the degree of co-occurrence of keywords on the basis of judging the number of topics by perplexity.

Step 1: Obtain the topic keywords through the LDA topic model.

Step 2: Combine the i^{th} topic with other topics into pairs and form the pairs set $S_i(i \in n)$, where n represent the number of topics. $S_{ij}(j \in n-1)$ is the element of S_i which denotes the pair combined by i^{th} topic and j^{th} topic, and the number of topic pairs in S_i is n-1. Then combined the keywords from i^{th} topic and j^{th} topic pairwise. Assumed there are λ keywords in one topic, which means the number of keywords pairs of S_{ij} will be λ^2.

Step 3: Calculate v and τ. v denotes how many Weibo text has contained the certain keyword pairs in set s_{ij}, while τ represents how many keywords pairs in set S_{ij} appeared in Weibo data.

Step 4: Compare the τ of each topic pairs in S_{ij}, if one topic pairs in S_{ij} has the maximum τ, these two topics in the topic pairs will be regarded as similar topics; if several topic pairs share the same maximum τ, then compare the v of each topic pairs and chose the pairs of topics with the maximum v to be similar topics; if some topic pairs share the same maximum τ and maximum v, all of these pairs of topics are treated as similar topics.

Step 5: Repeat the above steps until the most similar topic of each topic are found, and then categorize a combination of topics pairs that contain the same topic.

In summary, the proposed methods can reclassify the similar topics that extracted from LDA topic model more precisely.

4. **Experiments and Results.** In this paper, seven official Weibo posts related to traffic in Fuzhou were crawled through the Internet crawler. The time span was from 2011 to 2019, and the number of Weibo posts crawled totaled 22,444.

As Weibo text is short, in order to improve the modeling effect, literatures[23; 24] mentioned the method of developing the content of the short text, that is, integrating multiple twitters of the same author. In this paper, the original content and forwarded content of the blogger are joined in this study to expand the content of Weibo. In addition, Weibo contains a large amount of unstructured information, such as URL, tag symbol

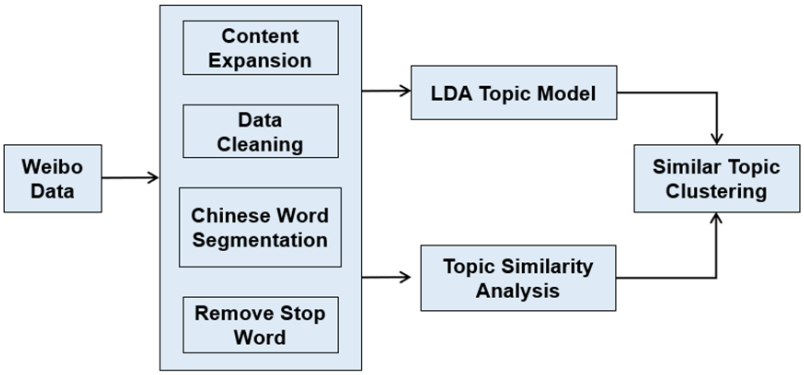


FIGURE 2. System Flow Chart for Emergency Event Topic Clustering

and emoji and so on, Referring to related work[25], these noise data are eliminated by regular expression.

In this study, python Jieba word segmentation tool was used to word segmentation. As the data of Weibo was derived from official and enterprise accounts related to traffic in Fuzhou, part of urban areas and road names in Fuzhou were imported into the word segmentation database to improve the precision of word segmentation.

Baidu stop words list, Harbin Institute of Technology stop words list and Sichuan machine laboratory stop words list were combined into one stop words list to improve the removal effect of stop words.

4.1. **Topic Number Determination with Perplexity.** Generally speaking, the smaller the perplexity value is, the smaller the uncertainty of the document and the better the generalization of the model will be[26]. As shown in figure 3, with the increase of the topics numbers, the perplexity value shows a decreasing trend. And we set the number of topics to 65 because the perplexity value is the smallest at this point.

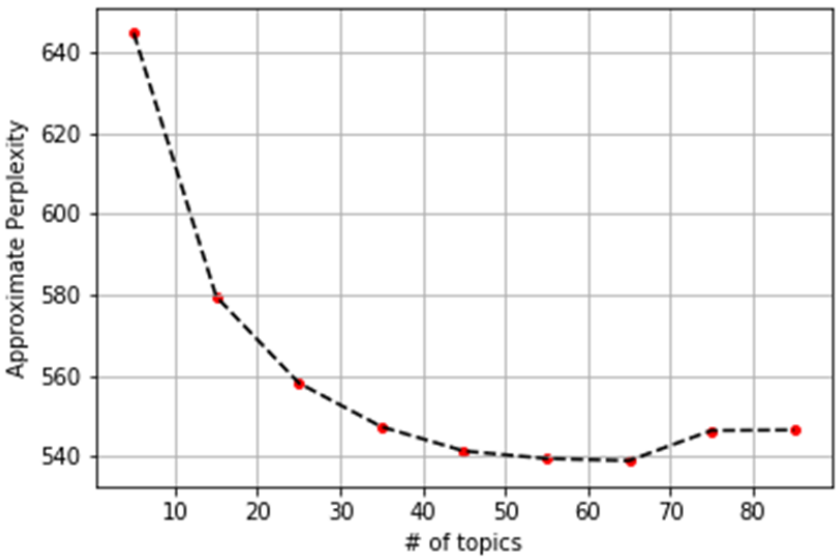


FIGURE 3. Perplexity value varies with the number of topics

4.2. **LDA Topic Modelling.** In this paper, the topic modeling of Weibo is realized by Latent Dirichlet Allocation (LDA). In addition, Latent Dirichlet Allocation in the Sklearn library in the python programming language was be used. And set the parameter α to 50/topic, set β to 0.01, set the number of topics to 65.

Table 1. Part of the topic and its keywords

Topic	Keywords
Topic No.32	行车 车速 距离 途经 减速 降低 控制 车距 交警部门 注意事项 发动机 前车 转向 路网 驾驶
Topic No.35	发生 汽车 交通事故 超车 刹车 减速慢行 引发 变道 隐患 操作 碰撞 盲区 动作 行驶 驾驶
Topic No.59	中心 发布 全市 超过 登陆 暴雨 信号 防范 气象台 局部 阵雨 气象 风力 乡镇 大雨
Topic No.61	气温 市区 天气 地区 高温 下降 模式 冷空气 感觉 降温 改善 降雨 阴天 不用 空气

4.3. **Topic similarity evaluation.** As shown in Table 1, there are many traffic safety related words in Topic 32 and Topic 35, such as "车速(vehicle speed)", "车距(vehicle speed)", "刹车(brake)", "减速(deceleration)", etc.; The key words in topic 59 and 61 are mostly related to weather, such as "暴雨(rainstorm)", "气象(meteorology)", "气温(temperature)", "天气(weather)" and so on. When determining the number of LDA topics by the perplexity value, the number of predicted topics tends to be too large, resulting in a large number of similar topics as shown in table 1. In order to classify topics like LDA topic extraction results into one category, this paper proposes a classification method based on the degree of keyword co-occurrence. The specific steps are as follows:

Firstly, we obtained 65 topics determined by perplexity and combined any two topics into pairs. As a result, there were 65 topic sets and 64 topic pairs in every certain topic set, and each topic has 15 keywords.

Secondly, for topic A and topic B in any topic pairs, we selected any keyword from topic A and topic B respectively and combined them pairwise. (Fig.4) Hence each topic pair has 225 keywords pairs.

Thirdly, the number of keyword pairs τ of a certain topic pair that appeared in Weibo data were calculated. Those two topics in the topic pair with the maximum τ were the most similar topics. For several topic pairs shared the same maximum τ, we calculated the occurrence number v of the 255 keyword pairs respectively, and regarded the topic pairs with the maximum v as similar topics. Those topic pairs that had the same maximum τ and v were categorized into one category that contain similar topics.

Finally, all the topics were reclassified into several categories. Topics in each category were more similar then those classified by the LDA topic model.

Two keywords from different topics are connected by a link which represent the combination of these two keywords. The numbers above the links indicate the number of co-occurrence of corresponding keyword pairs on Weibo.

5. **Discussion.** The determination of the number of text topics has always been a difficult problem in the field of text data mining, especially for short text data such as Weibo, which is difficult to determine the number of topics because of its short length and small number of words[27]. To determine the number of topics by perplexity value is a common

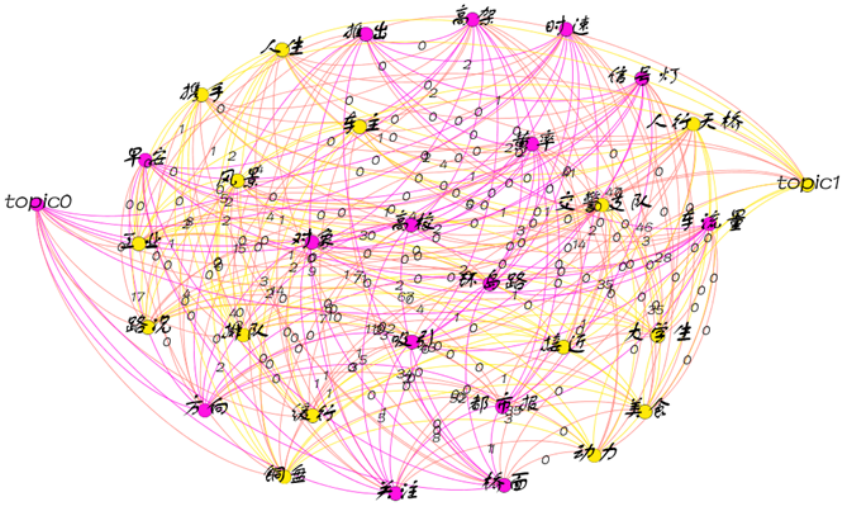


FIGURE 4. Topic and keywords relationship charts

method for text topic mining. In this paper, the number of topics is preliminarily judged by the perplexity value, and then the similar topics extracted from the LDA topic model are classified according to the degree of keyword co-occurrence.

In this paper, the classification of similar topics in Weibo text topic mining results can be summarized into two steps. The first step is to determine the number of topic based on the perplexity value, after that, the topic mining of Weibo is carried out by the LDA topic model, and the keywords of each topic will be obtained. The second step is to combine every two topic into one group, then to combine the keywords between different topics in each topic group, and get a series of keyword combinationsAand then to judge the similarity of different topic by the degree of co-occurrence of keyword combination under each topic group in Weibo data.

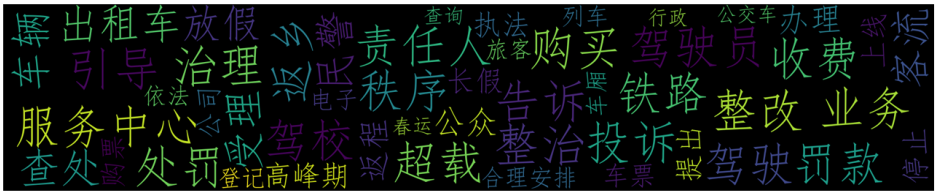
We get 65 topics by perplexity value in the process of Weibo topic mining, but many of these topics are similar, they either share the same keywords or their keywords have similar semantics. Therefore, through the degree of keyword co-occurrence, this paper divides these 65 topics into 5 categories, they are about Fuzhou weather information, citizen travel safety tips, road news, traffic congestion and bus line site change. Some of the key words under each category are shown in Figure 5.

Category 1 contains the keywords related to weather, such as “暴雨(rainstorm)”, “阵雨(shower)”, “阴天(overcast)”, etc., and appears keywords which reflect temperature change, such as “高温(high-temperature)”, “冷空气(cold air)”, “降温(hypothermia)”, etc. Moreover, there are “气象(meteorological)”, “天气(weather)” and other words, show that the main theme of the cluster associated with weather information. Weather is closely related to traffic conditions, and this cluster also reflects that most official traffic-related microblogs share weather information with citizens.

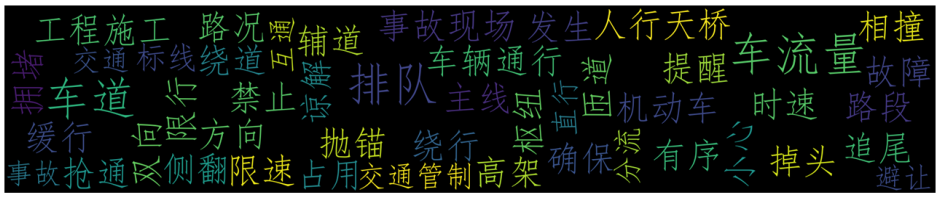
Category 2 is about citizen travel safety tips. People must keep a safe distance from the vehicle in front when driving, so the keywords of this cluster include “车距(vehicle-distance)”, “距离(distance)”, “前车(preceding vehicle)”, etc. Driving speed is also often closely related to vehicle safety, so there are the keywords such as “车速(vehicle speed)”, “减速(deceleration)”, “超速(overspeed)”, etc. And traffic accidents are often related to drivers’ violation driving, “超员(overstaffing)”, “闯红灯(red light runners)”, “超车(overtaking)” and other causes of traffic accidents can reflect the theme of bloggers’ safety tips to citizens.

(a) Category No.1: Weather Information Release

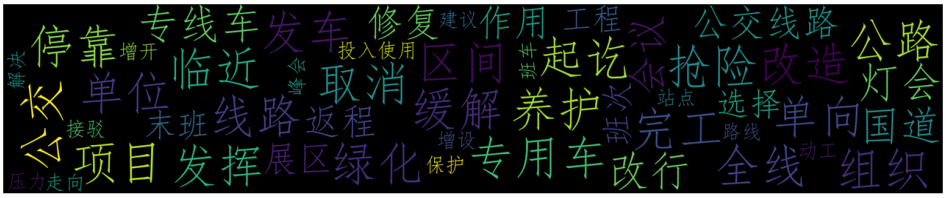
(b) Category No.2: Safe Travel Tips



(c) Category No.3: Road News Release



(d) Category No.4: Traffic Congestion



(e) Category No.5: Alteration of Routes and Stations

FIGURE 5. Part of keywords in similar categories

Category 3 is about road condition news. As can be seen from the keywords, most of the published news focuses on violation driving, traffic regulation, complaints and suggestions, and travel during holidays, etc. So this cluster includes “罚款(fine)”, “处罚(punishment)”, “依法(lawfully)” and other words that the punishment of violation driving. contains “整治(regulation)”, “治理(treatment)”, “交通管理(traffic management)” and other words to reflect the traffic management department renovation, also appears the keywords such as “投诉(complaint)”, “公众(public)”, “服务中心(service center)”, etc., which reflect the complaints and Suggestions of relevant departments, and also includes “春运(Spring Festival Transportation)”, “高峰期(peak)”, “返程(reverse)”, which indicate bloggers’ release of information related to holiday travel. This cluster contains the most topics, which also

indicates that the main news released by traffic management departments is about road news.

Category 4 is about traffic congestion information sharing. The keywords such as "拥堵(congestion)", "限速(speed limit)", "排长队(queue length)", etc., indicate that most of the topics are related to road traffic conditions. There are many reasons for road congestion, such as traffic accidents, road construction, performance in key words, "追尾(rear collision)", "侧翻(side tumbling)", "抛锚(anchorage)", "工程施工(engineering construction)", and so on. For congested roads, bloggers also advise drivers to detour or drive carefully, which is also reflected in the key words "小心(careful)", "绕道(detour)", "缓行(walking slowly)".

Cluster 5 is a hint with regard to alteration of routes and stations. This cluster contains keywords concerning city construction (e.g. "工程(engineering)", "动工(begin construction)", "绿化(greening)", "养护(maintenance)") and urban activities (e.g. "会议(conference)", "峰会(summit meeting)", "灯会(lantern show)"), which leads to different degrees of traffic pressure. In order to facilitate the travel of citizens and relieve traffic pressure, traffic management department will change the transit routes and stations accordingly. Therefore, cluster 5 also contains this kind of keywords (e.g. "增设(additionally build)", "起讫(origin and destination)", "专线(special railway line)", "公交路线(bus route)", "改行(change route)", "接驳(transship)") as long as keywords that reflect the blogger's suggestions on travel routes of the public (e.g. "建议(suggestion)", "选择(choice)"). To sum up, this cluster is mainly concerned about the advice of bloggers on the travel routes after the change of transit routes and stations.

6. **Conclusion.** Determining the number of topics in a topic model by perplexity is currently a common method, and the value of perplexity can be used as criteria for evaluating the quality of the model[28]. However, the number of topics selected by perplexity is often too large, which will lead to greater similarity between different topics. In order to solve this problem, this paper distinguishes similar topics by the co-occurrence of topic keywords based on using perplexity to determine the number of topics. In this way, the topics with higher similarity can be effectively classified into one category.

In this experiment, the number of topics judged by the perplexity value is 65. After analysis, there are often great similarities among these topics. As shown in table 1 above, 65 topics are finally divided into five categories by using the method proposed in this paper. However, the deficiency of this article lies in that only statistical methods are used to distinguish the similarity of topics, and semantic information is not used to classify the topics, which makes the classification results not detailed enough. For example, category 3, "Road News", includes 25 sub-topics, and the content of these topics involves in illegal driving, traffic control and other aspects of the news. But the method proposed in this paper cannot further detailed classification of these topics, it will be our next research goal.

Acknowledgements. This work was supported in part by Projects of the National Science Foundation of China (41971340, 41471333, 61304199), project 2017A13025 of Science and Technology Development Center, Ministry of Education, project 2018Y3001 of Fujian Provincial Department of Science and Technology, projects of Fujian Provincial Department of Education (JA14209, JA15325, FBJG20180049). The Fujian Early Warning Information Release Center and the Water Resources & Hydropower Research Institute of Fujian are also acknowledged for system experimental supporting.

REFERENCES

[1] G Salton, A Wong, C-S Yang, A vector space model for automatic indexing, Communications of the ACM, vol. 18, no. 11, pp. 613-620, 1975.

[2] L Jing, M K Ng, J Z Huang, Knowledge-based vector space model for text clustering, Knowledge Information Systems, vol. 25, no. 1, pp. 35-55, 2010.

[3] LC Liao, XH Jiang, FM Zou, A spectral clustering method for big trajectory data mining with latent semantic correlation, Chinese Journal of Electronics, vol. 43, no. 5, pp. 956-964, 2015.

[4] S Deerwester, S T Dumais, G W Furnas, et al., Indexing by latent semantic analysis, Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[5] JS Pan, FR Mcinnes, MA Jack, Fast clustering algorithms for vector quantization, Pattern Recognition, vol. 29, no. 3, pp. 511-518, 1996.

[6] D M Blei, A Y Ng, M I Jordan, Latent dirichlet allocation, Journal of Machine Learning Research, 3(Jan), pp. 993-1022, 2003.

[7] Q Chen, L Yao, J Yang, Short text classification based on LDA topic model, International Conference on Audio, Language and Image Processing (ICALIP), pp. 749-753, 2016.

[8] T-I Yang, A Torget, R Mihalcea, Topic modeling on historical newspapers, Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 96-104, 2011.

[9] T L Griffiths and M Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences of the United States of America, 101(suppl 1), pp.5228-5235, 2004.

[10] L Sun and Y Yin, Discovering themes and trends in transportation research using topic modeling, Transportation Research Part C: Emerging Technologies, vol. 77, pp. 49-66, 2017.

[11] M Rosen-Zvi, T Griffiths, M Steyvers, et al., The author-topic model for authors and documents, Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 487-494, 2004.

[12] Y Liu, A Niculescu-Mizil, W Gryc, Topic-link LDA: joint models of topic and author community, Proceedings of the 26th Annual International Conference on Machine Learning, pp. 665-672, 2009.

[13] WX Zhao, J Jiang, J Weng, et al., Comparing twitter and traditional media using topic models, European Conference on Information Retrieval, pp. 338-349, 2011.

[14] M Qiu, F Zhu, J Jiang, It is not just what we say, but how we say them: Lda-based behavior-topic model, Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 794-802, 2013.

[15] S Chen, Y Huang, W Huang, Big data analytics on aviation social media: The case of china southern airlines on sina weibo, 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), pp. 152-155, 2016.

[16] W X Zhao, J Jiang, J He, et al., Topical keyphrase extraction from twitter, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 379-388, 2011.

[17] S Wang, M J Paul, M Dredze, Exploring health topics in Chinese social media: An analysis of Sina Weibo, Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[18] A F Hidayatullah, M R Ma'arif, Road traffic topic modeling on Twitter using latent dirichlet allocation, 2017 International Conference on Sustainable Information Engineering and Technology (SIET), pp. 47-52, 2017.

[19] A F Hidayatullah, S K Aditya, Karimah, S T Gardini, Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA), IOP Conference Series: Materials Science and Engineering, vol. 482, no. 1, pp. 012033, 2019.

[20] R Arun, V Suresh, C V Madhavan, et al., On finding the natural number of topics with latent dirichlet allocation: Some observations, Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 391-402, 2010.

[21] M Steyvers and T Griffiths, Probabilistic topic models, Handbook of Latent Semantic Analysis, vol. 427, no. 7, pp. 424-440, 2007.

[22] G Heinrich, Parameter estimation for text analysis, Technical report, 2005.

[23] J Weng, E-P Lim, J Jiang, et al., Twitterrank: finding topic-sensitive influential twitterers, Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261-270, 2010.

[24] L Hong and B D Davison, Empirical study of topic modeling in twitter, Proceedings of the First Workshop on Social Media Analytics, pp. 80-88, 2010.

[25] A Hidayatullah and M Ma'arif, Pre-processing tasks in indonesian twitter messages, Journal of Physics: Conference Series, vol.801, no.1, pp. 012072, 2017.

[26] F Sun and H Chen H, Feature extension for Chinese short text classification based on LDA and Word2vec, 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1189-1194, 2018.
[27] Z Wang and B Li, Topic derivation in weibo through both interactions and content, 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), pp. 932-935, 2018.
[28] Y Xu, H Xu, L Zhu, et al., Topic discovery for streaming short texts with CTM, 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-7, 2018.