IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 8, 2019, accepted April 25, 2019, date of publication April 30, 2019, date of current version May 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914097

Research on Topic Detection and Tracking for Online News Texts

GUIXIAN XU[1], YUETING MENG[1], ZHAN CHEN[1], XIAOYU QIU[2], CHANGZHI WANG[3], AND HAISHEN YAO[1]

[1]College of Information Engineering, Minzu University of China, Beijing 100081, China
[2]Library, Shandong University of Traditional Chinese Medicine, Jinan 250355, China
[3]Baidu APP Technology Platform Department, Baidu Company, Beijing 100085, China

Corresponding author: Guixian Xu (guixian_xu@muc.edu.cn)

This work was supported in part by the Project of Humanities and Social Science, Ministry of Education of China, under Grant 18YJA740059 and China Scholarship Council, under Grant (2018) 10038.

ABSTRACT With the rapid development of the Internet, the amount of data has grown exponentially. On the one hand, the accumulation of big data provides the basic support for artificial intelligence. On the other hand, in the face of such huge data information, how to extract the knowledge of interest from it has become a matter of general concern. Topic tracking can help people to explore the process of topic development from the huge and complex network texts information. By effectively organizing large-scale news documents, a method for the evolution of news topics over time is proposed in this paper to realize the tracking and evolution of topics in the news text set. First, the LDA (latent Dirichlet allocation) model is used to extract topics from news texts and the Gibbs Sampling method is used to speculate parameters. The topic mining using the K-means method is compared to highlight the advantages of using LDA for topic discovery. Second, the improved single-pass algorithm is used to track news topics. The JS (Jensen–Shannon) divergence is used to measure the topic similarity, and the time decay function is introduced to improve the similarity between topics with the similar time. Finally, the strength of the news topic and the content change of the topic in different time windows are analyzed. The experiments show that the proposed method can effectively detect and track the topic and clearly reflect the trend of topic evolution.

INDEX TERMS Latent Dirichlet allocation, topic detection, topic tracking, artificial intelligence.

I. INTRODUCTION

With the rapid development and popularization of Internet technology, how to obtain useful information from massive data has become a common concerned problem. Text mining technology [1], [2] can extract effective, useful and valuable information from a large number of texts, and it has gradually become one of the key technologies to solve the problem of topic discovery. Effective analysis of massive information on the network has also become a key research content by researchers in the field of machine learning and data mining. In text mining technology, the traditional text representation method usually adopts the space vector model (VSM), which is a commonly used model in natural language processing. But this model does not take the relationship of the underlying semantics of the texts into account. The topic model is a modeling method for extracting implicit topics from massive texts. It is widely used to mine topics from documents.

The topic model is a statistical probability model, which is used to find the statistics of topics from a large amount of texts. It broadens the scope of text mining technology and makes texts more expressive. The topic model can mine the potential semantic information and discover the potential topic information in the texts, which is the biggest difference from the traditional model. Therefore, the topic model can identify multiple topics hidden in the texts, helping people understand the text information from a semantic perspective. As a statistical probability model, the topic model has been widely studied and applied in natural language processing, text mining [3], [4], face recognition [5], content topic evolution [6] and other aspects. With the development of the topic model, how to use it to detect and track news topics has become a hot issue in the field of machine learning and text mining [7].

The associate editor coordinating the review of this manuscript and approving it for publication was Iztok Humar.

In this paper, the following contents will be divided into four aspects. Firstly, the relevant background of the topic model is expounded. The contributions of relevant researchers on the LDA topic model and topic detection and tracking are elaborated. Secondly, the proposed method is introduced to track the news topic and achieve the evolution of news topic content and strength. Thirdly, the experiments are carried out and the experimental results are analyzed and discussed. Finally, the proposed method is summarized and the next research direction is introduced.

II. RELATED WORK

A. TOPIC DETECTION

A topic is a collection of events caused by a seed event. TDT (Topic Detection and Tracking) is a study of the organization and utilization of information flows based on topical events, which began in 1996. At that time, the US Defense Advanced Research Projects Agency (DARPA) proposed to develop a new technology to automatically discover topics from the news data stream. The staff conducted some basic researches to find relevant topics and then automatically judge the new and old events that occur. Since 1998, with the support of DARPA, the National Institute of Standards and Technology has held an international conference on topic detection and tracking every year and conducted corresponding evaluations of TDT. The evaluation corpus was provided by the Language Data Alliance, which used manually labeled topics as standard answers. With the gradual deepening and continuous development of research, the focus of previous TDT evaluations is different, and the corresponding evaluation tasks are also changed. The story segmentation task in TDT in 2004 was no longer evaluated. The topic tracking task, topic detection and related detection tasks were retained. The supervised adaptive topic tracking tasks and hierarchical topic detection tasks were added. In recent years, TDT evaluation has received more and more attention from people. Many well-known universities, companies and research institutions have participated in this evaluation. Domestic research in this area has been carried out relatively late. Since 2003, domestic researchers have conducted research on TDT related technologies. At present, many scholars apply topic detection and tracking on social networks, forums [8], and microblogs, and have achieved good results.

The topic detection work provides the "raw material" of the topic for topic tracking. The topic detection work is usually done using clustering and topic models. Zhen *et al.* [9] improved the K-Means algorithm for the discovery of hot topics, which used the density function to optimize the selection of the initial cluster center. This method effectively detected topics in news corpus. Zhang Song [10] used the minimum distance and the average degree of aggregation to select the initial cluster center point. The number of clusters obtained by the CURE algorithm was taken as the K value. Finally, the improved K-means algorithm was applied to the microblog topic discovery. Experiments showed that the algorithm improved the accuracy of clustering results. Che and Yang [11] used a variety of feature fusion methods to discover news topics. Firstly, the news title and the position of the paragraph were weighted, and the similarity calculation based on the vector space model was performed. Then, LDA was used to calculate text similarity. Finally, these three combinations were used to improve the discovery of news by the Single-Pass algorithm. Hoffman *et al.* [12] proposed Online LDA in 2010 to detect topics in large-scale texts. It used the variational Bayesian method to infer the parameters and adopted a random natural gradient to determine the degree of convergence of the objective function. Wang and Zhang [13] used the LDA model to solve the initial cluster center selection problem in the K-Means algorithm. They used LDA to select the most important M topics in the text collection. Then preliminary clustering was performed so as to find the cluster center. Liu *et al.* [14] proposed LDA-K-means algorithm for topic discovery about food safety on network. They selected the topic probability distribution of documents to represent each document, and then used K-Means to cluster these documents to get the topics. Liu *et al.* [15] combined LDA, Single-Pass and hierarchical clustering to perform topic discovery on Weibo.

B. TOPIC TRACKING

Topic tracking is based on the definition of a topic with one or more stories and then uses this topic to find all relevant news stories in the stories stream. There are two main research directions in traditional topic tracking research. They are based on rules and statistics. The key technology of rule-based research is to analyze the association and inheritance relationship between text content and use the related domain knowledge to summarize the related news texts. The statistical method is mainly based on the probability distribution of text features, using mathematical statistics to determine the relevance of text to topic models. In the topic tracking research, researchers have tried a variety of methods. Allan *et al.* [16] used the Rocchio algorithm to perform topic tracking. By adding weight to specific features in the topic model, the topic model was continuously updated during the tracking of the topic. Ren *et al.* [17] used K-Modes clustering to conduct adaptive topic tracking research. First, the texts were clustered. The topic model was represented by the center of each cluster, and the named entity vector was used as the cluster center. The cluster center was iterated continuously during the tracking process until the topic cluster center was stable. Bai *et al.* [18] proposed a method of tracking user relationship, which first mapped the blog to the feature space, and then used the improved K-Means algorithm to perform binary clustering to track related topics.

In recent years, with the development and popularity of topic models, the topic model LDA has also been frequently used in topic detection and tracking tasks. Zhang *et al.* [19] used the LDA topic model for topic tracking technology research. The topics of the texts were extracted by using the LDA model, and the relevant topics were tracked by

calculating the correlation. Griffiths and Steyvers [20] conducted topic detection and tracking on PANS text sets. Li *et al.* [21] studied the subject evolution of the Journal of Information Science. The OLDA model proposed by Alsumait *et al.* [22] constantly adjusted the topics that had been discovered by tracking the topic and used the evolution matrix to preserve the changes of the topics. Pei *et al.* [23] proposed a variable online LDA (VOLDA) based on the OLDA model. The model took into account the phenomenon of alternating old and new topics that existed in the OLDA model. They designed the dynamic weight calculation method and parameter optimization method. Finally, the reliability of the method was verified in the forum data. Gong *et al.* [24] proposed an adaptive topic tracking approach, which was based on an improved Single-Pass clustering algorithm with sliding time window. In the topic tracking process, a sliding time window strategy was used to guarantee the system accuracy and reduce the number of missed following stories. Finally, the experimental results showed that the approach achieved satisfying results.

III. PROPOSED METHOD

At present, topic tracking based on topic models mostly utilizes LDA model or LDA extension model. LDA is a document topic generation model, also known as a three-layer Bayesian probability model. It has three level structures which include words, topics, and documents. Based on pLSA, it introduces a K-dimensional implicit random variable obeying the Dirichlet distribution to represent the subject probability distribution of the document.

In this paper, the topic tracking method based on discrete time interval is proposed, and the time decay function is introduced to calculate the similarity between the documents. At the same time, the influence of the input order on the tracking results is also solved. The topic tracking is carried out by using the improved Single-Pass clustering method.

A. TOPIC DETECTION

1) LDA TOPIC MODEL

LDA (Latent Dirichlet Allocation) is a probabilistic model for modeling text data, which can model the topic information of text data. The LDA topic model can realize the dimensionality reduction representation of text in the semantic space, and it models the text with the probability of vocabulary, which alleviates the problem of data sparsity to some extent [25]. Therefore, in this paper, the LDA topic model is used to extract the topic of the text.

The LDA topic model is a hierarchical Bayesian model. In this model, it assumes that words appearing in the text are independent and irrelevant, and it considers that each document is composed of several implicit topics. And these hidden topics are made up of some specific words in the text. LDA is a typical probability model, determined by the parameters (α, β). α represents the relative strength between the implicit topics in the text set, and β represents the probability distribution of the implicit topic itself. The process of generating the LDA model is shown in FIGURE 1.

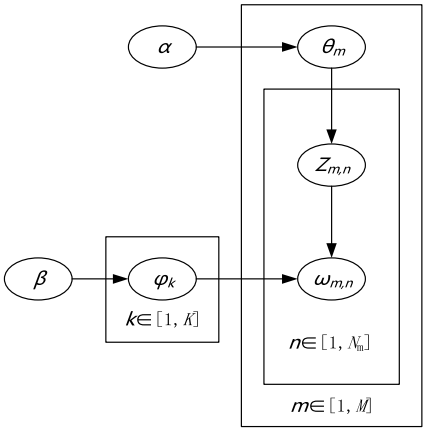


FIGURE 1. LDA diagram model representation.

TABLE 1. LDA model diagram parameters explanation.

Parameter symbol	Explanation
α	Priori parameters of documents - topics
β	Priori parameters of topics - words
M	Number of documents
K	The total number of potential topics in the document
N_m	Number of feature words in document m
$z_{m,n}$	The topic of the nth word in document m
θ_m	Polynomial distribution of the topic in document m
φ_k	Polynomial distribution of the word in topic k
$w_{m,n}$	The n^{th} word in document m

In FIGURE 1, α and β are generally determined empirically, $z_{m,n}$, θ_m, φ_k are determined by the LDA model. Each parameter explanation of FIGURE 1 is shown in TABLE 1.

α and β are the Dirichlet prior distributions of the LDA model. α represents priori parameter of the topic distribution over the entire document set. β represents priori parameter of the word distribution across all topics. Based on α, the distribution θ of topics in the document is generated. A topic Z is selected from the distribution θ in document. Based on β, distribution φ of feature word in the topic Z is generated. A term ω is obtained from the distribution φ of the topic Z. This process is repeated. The distributions of M documents are generated. The LDA generation process is shown in Algorithm 1.

Since Dirichlet is a conjugate prior distribution of polynomial distribution functions, the calculation of the model can be simplified. The model prior parameters α and β are usually empirically calculated as $\alpha = 50/K, \beta = 0.01$. In some cases, empirical Bayesian estimation of these two parameters can also be performed using the corpus.

Algorithm 1 LDA Generating Model

1: **for all** topics $k \in [1, K]$ **do**
2: sample mixture components $\varphi_k \sim Dir(\beta)$
3: **end for**
4: **for all** documents m $\in [1, M]$ **do**
5: sample mixture proportion $\theta_m \sim Dir(\alpha)$
6: sample document length $N_m \sim Poiss(\epsilon)$
7: **for all** words $n \in [1, N_m]$ **do**
8: sample topic index $z_{m,n} \sim Mult(\theta_m)$
9: sample item for word $w_{m,n} \sim Mult(\varphi_{z_{m,n}})$
10: **end for**
11: **end for**

The model parameter estimation of LDA is to obtain an estimate of the parameters according to a given optimization objective function. Generally, methods such as Variational Bayesian Inference (VB), Expectation Propagation (EP), and Collapsed Gibbs Sampling can be used. Each parameter inference method has different advantages, so choosing a suitable parameter inference method is a comprehensive consideration of indicators such as efficiency, complexity, and accuracy. In general, because the Gibbs Sampling method is simple to implement, it has become a widely used parameter estimation method in the topic model.

In the LDA model, the joint probability distribution of potential variables in an article [26] is shown in formula (1):

$$\begin{aligned} &P\left(\underset{w_m}{\longrightarrow}, \underset{z_m}{\longrightarrow}, \underset{\theta_m}{\longrightarrow}, \Phi\right) \\ &= \prod_{n=1}^{N_m} P\left(w_{m,n} | \underset{\varphi_{m,n}}{\longrightarrow}\right) P\left(z_{m,n} | \underset{\theta_m}{\longrightarrow}\right) \\ &\cdot P\left(\underset{\theta_m}{\longrightarrow} | \underset{\alpha}{\longrightarrow}\right) \cdot P\left(\Phi | \underset{\beta}{\longrightarrow}\right) \end{aligned} \tag{1}$$

In actual calculation, we use $P(z_i|z_{\neg i}, w)$ to simulate $P\left(\underset{w_m}{\longrightarrow}, \underset{z_m}{\longrightarrow}, \underset{\theta_m}{\longrightarrow}, \Phi\right)$. The formula is shown in formula (2):

$$P(z_i \mid z_{\neg i}, w) \propto \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t} \cdot \frac{n_{m,\neg i}^{(t)} + \alpha_k}{[\sum_{k=1}^{K} n_{m,\neg i}^{(t)} + \alpha_k] - 1} \tag{2}$$

In formula (2), we suppose $w_i = t$, z_i represents the subject corresponding to the i^{th} word, $\neg i$ indicates that the i^{th} item is removed, $n_{\cdot,\neg i}^{(\cdot)}$ indicates the number of items excluding the i^{th} term.

The Gibbs Sampling method is a special case of the Markov chain Monte Carlo method. The Gibbs algorithm is simpler when the joint distribution dimension is higher. In the actual processing, it is not necessary to integrate the parameter matrices θ and φ corresponding to the parameters Θ and Φ. Because they can be interpreted as correlation statistics among observable variables $w_{m,n}$ and related topics $z_{m,n}$ and Markov chain state variables.

The algorithm for Gibbs Sampling is given below:

Algorithm 2 Gibbs Sampling Algorithm

1: initialization
2: set $n_m^{(k)}, n_m, n_k^{(t)}, n_k$ to 0
3: **for all** documents $m \in [1, M]$ **do**
4: **for all** words $n \in [1, N_m]$ **do**
5: sample topic index $z_{m,n} = k \sim Mult(1/K)$
6: increment document-topic count: $n_m^{(k)} + 1$
7: increment document-topic sum: $n_m +$
8: increment topic-term count: $n_k^{(t)} + 1$
9: increment topic-term sum: $n_k +$
10: **end for**
11: **end for**
12: Gibbs Sampling Process
13: **while** not finished**do**
14: **for all** documents $m \in [1, M]$ **do**
15: **for all** words $n \in [1, N_m]$ **do**
16: decrement counts and sums: $n_m^{(k)} - 1, n_m - 1, n_k^{(t)} - 1, n_k - 1$
17: sampling topic index $k \sim P(z_i|z_{\neg i}, w)$
18: increment counts and sums: $n_m^{(k)} + 1, n_m + 1, n_k^{(t)} + 1, n_k + 1$
19: **end for**
20: **end for**
21: Check Convergence
22: **if** converged and L sampling iterations done since last read out **then**
23: read out parameter set $\emptyset$ and θ
24: **end if**
25: **end while**

The common method used in topic discovery is to treat topic discovery as clustering of events. The K-means clustering algorithm was used for topic discovery research. In this paper, the k-means algorithm is used as the baseline method to mine the topic of text. By comparing the topic mining effect of K-Means method and LDA model, the advantages of LDA for topic recognition are highlighted.

2) DETERMINATION OF THE NUMBER OF TOPICS

Perplexity is a measure of probability model or probability distribution prediction that can be used to evaluate a model. In this paper, the concept of Perplexity is used as a reference for determining the number of topics in the text set. The best number of topics is determined by selecting the model with the least perplexity. Smaller perplexity means that the model has a better predictive effect on new text. At the same time, the smaller the perplexity is, the more general the ability of the model is. The perplexity formula [27] is as follows:

$$perplexity\,(D) = exp\left(-\frac{\sum_{d=1}^{M} log\ p(w_d)}{\sum_{d=1}^{M} N_d}\right) \tag{3}$$

D represents the test set in the corpus, and the test set has a total of M documents. N_d represents the number of words

in each document d, w_d represents the word in document d, $p(w_d)$ is the probability of the word w_d in the document.

B. TOPIC TRACKING

Topic tracking is to organize news texts reporting the same topic in chronological order. Single-Pass is an incremental clustering algorithm for processing streaming data, often used for topic tracking. For the topic stream that arrives, one topic is processed in the order of input. According to the similarity between the current topic and the existing topic categories, the topic data is divided into one existing topic class with high similarity or create a new topic category to realize incremental and dynamic clustering of data. As the data flow increases, the Single-Pass algorithm will generate more and more topic categories, which will increase the time consumption of topic category judgment each round. And the topic tracking results are very sensitive to the order in which the data is entered.

In this paper, the topic tracking research based on topic model is conducted. Firstly, the LDA model is used to extract the topic information from the news texts of different time windows. Then the improved Single-Pass algorithm is used for topic clustering, and the JS divergence is used to measure the similarity between topics. The time decay function is utilized to improve the similarity between topics with close time. Finally, the changes in the strength and content of news topics in different time windows are analyzed.

1) TOPIC TRACKING ALGORITHM

When calculating the similarity of news topics for topic tracking, it is believed that the closer the two topics are, the more similar they should be. Therefore, in the topic similarity calculation, time is introduced as a parameter. The larger the time span is, the smaller the similarity possibility is. So the topic similarity of news topics for topic tracking is related to the time-based decay process. Therefore, the following time decay function is introduced to calculate the influence of time span on topic similarity. Its formula is as follows:

$$H(\Delta t) = e^{(-\Delta t/L)^k * \log(2)} \tag{4}$$

In formula (4), Δt is the time difference between the two topics, and parameters L and k are used to control the decay rate. L is the half-life time, that is, the correlation between the two topics with the gap L is attenuated to half. k controls the steepness of the curve. FIGURE 2 shows several decay trends at different L and k values. L is set 10 or 15. k is set 1, 1.8, 2 or 3. According to the observation experience, when $L = 15$, $k = 1.8$, the function is selected as the subsequent experiment.

The changing analysis of the topic content can be measured by JS divergence. The probability topic distribution obtained by LDA topic model is a mapping in vector space. Calculating the topics similarity can be obtained by calculating the relative entropy between the topic distributions, that is, KL [27] (Kullback-Leibler Divergence) distance. It measures the difference between two probability distributions P(x) and Q(x)

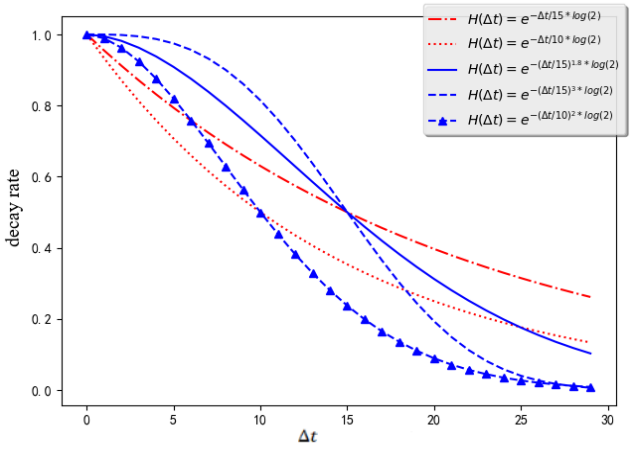


FIGURE 2. Decay function chart.

in the same space. The following uses $D(P||Q)$ to represent the KL distance, and the formula is as follows:

$$D(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \tag{5}$$

KL is asymmetrical, meaning that $D(P||Q)$ is not necessarily equal to $D(Q||P)$. And it does not satisfy the triangle inequality.

Therefore, in practical applications, JS divergence based on KL distance is used, and JS divergence is also an index to measure the similarity of two probability distributions, solving the asymmetry problem. The formula is as follows [28]:

$$JS(P||Q) = \frac{1}{2} KL(P||\frac{P+Q}{2}) + \frac{1}{2} KL(Q||\frac{P+Q}{2}) \tag{6}$$

The value of JS divergence is between 0 and 1. If the JS divergence of the two topics is smaller, it means the topic similarity is larger, the correlation between the two topics is greater, and the difference is smaller.

In this paper, the topic tracking algorithm is proposed and improved. Firstly, the time decay factor is introduced into the judgment of topic similarity to increase the similarity between topics with similar time. The topic with too long time distance does not need to perform similarity calculation. This strategy will effectively increase the processing speed in the large-scale topic tracking process without destroying the tracking effect. Secondly, the topic stream is processed in batches according to the time window, that is, the topics in the same time window are regarded as a batch, and the similarity results of each topic and the topic classes are recorded. When a batch of topic data is processed, the topic categories are divided and the topic centers are updated. The purpose is to solve the impact of the input order on the tracking results, and to select the most similar new topic as the topic category center.

The improved topic tracking algorithm is given below.

In the above algorithm, t refers to the time window. There are n time windows. And m represents the total number of news in the time window t. The topic modeling work is

Algorithm 3 Improved Topic Tracking Algorithm

Input: A total of M original news corpora
Output: Topic tracking results
1: **for all** news $m \in [1, K]$ **do**
2: segment, remove stop words
3: feature extraction
4: get a collection of text D_t based on the time window $t \in [1, T]$
5: **end for**
6: **for al** l times $t \in [1, T]$ **do**
7: get topic distribution by Algorithm 2 Gibbs Sampling Algorithm
8: **end for**
9: get all topics set S of D by time window
10: *category* $\leftarrow S_1$
11: **for all** $S_s \in [2, N]$ **do**
12: batch topic according to time window
13: **for** *cate* in *category* **do**
14: $\text{sim} = JS(S_s, cate)$
15: sort_ sim [cate] = sim $*H(\Delta t)$
16: **end for**
17: similarity ordering, get the maximum similarity value *Max*
18: **if** $Max > THRESHOLD$ **then**
19: Record all potential category of$S_s$and reverse the order of similarity
20: **end if**
21: **if** a topic processing of a time window is completed **then**
22: Assign all potential categories and choose the most similar one for the new topic center
23: **end if**
24: **end for**
25: **return** *category*

carried out according to different time windows, and then relevant experimental research on topic tracking is performed according to the time window.

2) TOPIC TRACKING RESULT ANALYSIS

The result analysis of the topic tracking is an analysis of the changes in the content and strength of similar topics in different time windows.

a: ANALYSIS OF TOPIC STRENGTH

The topic strength can be measured by the topic probability distribution of the news texts. In the Gibbs sampling calculation formula, $\theta^d_{z=j}$ can be obtained. The average strength of the topic is used to measure the change of the topic strength, where $\theta^d_{z=j}$ represents the probability of the topic j corresponding to the document d. At the same time, if the number of news related to the topic is taken into account, in the time window t, the topic strength$Hot(j)$ of the topic is obtained by the following formula:

$$Hot\,(j) = \mu \cdot \left(\frac{1}{M} \sum_{d=1}^{M} \theta^d_{z=j} \right) + \sigma \cdot \frac{n_j}{M} \tag{7}$$

In the above formula, M represents the total number of news in the time window t. n_j represents the number of news related to the topic j. μ and σ are the adjustment coefficients, and in this experiment, both are set 0.5.

b: ANALYSIS OF TOPIC CONTENT

Regarding the analysis of the topic content, the size of the topic content is measured according to the JS divergence between similar topics. It is also possible to intuitively analyze the topic content changes according to the topic feature vocabulary.

IV. EXPERIMENT AND ANALYSIS

A. TOPIC DETECTION EVALUTION INDEX

In the topic identification process, the evaluation method in TDT2003 [29] is usually adopted. This method is to weigh and evaluate depending on the loss rate and false positive rate of the relevant reports in the topic detection and tracking results. Its calculation formula is as follows:

$$(C_{det})_{norm} = \frac{C_{miss} \cdot P_{miss} \cdot P_{target} + C_{fa} \cdot P_{fa} \cdot P_{non-target}}{min\left(C_{miss} \cdot P_{target}, C_{fa} \cdot P_{non-target}\right)} \tag{8}$$

In the above formula, C_{det} represents the error recognition cost of the tracking system. The smaller the value is, the better the system tracking performance is. C_{miss} indicates the cost of tracking loss. C_{fa} indicates the cost of tracking the error report. P_{target} and $P_{non-target}$ are a priori probabilities of whether a report is related to a tracking topic. P_{miss} is the loss rate of the tracking report; P_{fa} is the false positive rate of the tracking report. In general, C_{miss}, C_{fa}, $P_{non-target}$, P_{target} are set to 1, 0.1, 0.98, and 0.02.

F_1 value are also used as evaluation indicators of the experiment. The specific formula is as follows:

The precision P, the recall rate R and the F_1 value are calculated as follows:

$$P = \frac{TP}{FP + TP} \tag{9}$$

$$R = \frac{TP}{FN + TP} \tag{10}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{11}$$

TP means the number of correctly identified news in the classification results of positive class. FP means the number of misidentified messages in the classification results of positive class. FN refers to the number of misidentified messages in the classification results of negative class. The F1 value is the harmonic mean of the precision and recall rate.

TABLE 2. Topic extraction display.

Category	Method	Top 20 words
Tourism	LDA	Tourism, Visitor, Beijing, Golden week, China, Activities, City, Work, Culture, Transportation, Shanghai, Passenger, Reception, Scenic spot, Area, Increase, Holiday, Department, Telephone, Country
	Baseline	Tourism, Golden week, Visitor, China, Reception, Increase, Passenger, Scenic spot, Beijing, Travel agency, Holiday, City, Country, Attractions, Route, Activity, Work, Shanghai, Leisure, Department
Car	LDA	Automobile, Market, Company, Enterprise, China, Brand, Model, Technology, Sale, Development, Price, Product, Engine, Car, Independent, Design, System, Power, Production, Consumer
	Baseline	Automobile, Market, Model, Brand, China, Engine, Car, Sale, Price, Product, Company, Sale, Technology, Power, Enterprise, System, Increase, Listing, Design, Independent
Education	LDA	Candidate, Professional, Volunteer, Admission, Student, Recruit, School, Exam, University, College, Education, University, Report, Undergraduate, Program, Teacher, English, College entrance examination, Study, Result
	Baseline	Candidate, Examinations, Specialty, Student, Recruit, School, Volunteer, Time, University, Admission, College entrance examination, Study, Education, Problem, Enroll, University, Accounting, Teacher, English, Plan
Culture	LDA	China, Life, History, Problem, Social, World, Culture, Children, Boxer, Works, Relationship, Crime, Research, Art, Time, Spirit, Work, Performance, Literature, Sir
	Baseline	China, Beijing, Work, Life, Culture, Time, Children, Works, Company, World, Activity, Japan, Sir, Problem, Country, Social, History, Art, Development, Shanghai
Military	LDA	China, US, Japan, Army, Training, Military, Missile, System, Manoeuvre, Weapons, Operation, Air force, Equipment, Liberation army, Navy, Russia, Data, Capability, US, Iran
	Baseline	US, System, Missile, Training, China, Manoeuvre, Army, Military, Operation, Air army, Weapon, Equipment, Technology, Russia, Flight, Iran, Company, Navy, Manufacture, Capability

B. TOPIC RECOGNITION COMPARSION EXPERIMENT

The method of topic tracking proposed in this paper is based on the LDA model to identify the topic of news text. In order to verify the advantages of the proposed method in topic tracking, it is first necessary to verify the advantages of the LDA topic model used in the topic representation work. By comparing with the K-means method, the advantages of topic extraction using the LDA model are highlighted.

The corpus used in the topic identification experiment comes from the labeled news published by Sogou Lab, which has 3,000 news reports in ten categories: automobile, culture, education, finance, health, IT, military, recruitment, sports, and tourism. Each category contains 300 articles. In the experiment, the NLPIR word segmentation tool of the Chinese Academy of Sciences is used to cut the words. At the same time, in order to filter the meaningless vocabulary, the stop word list of Harbin Institute of Technology was expanded including a total of 1379 stop words.

From the labeled news corpus, the five categories of tourism, automobile, education, culture, and military are selected for topic extraction experiments. The LDA method and the K-Means clustering method based on TF-IDF weighting are used for text topic mining comparison experiments. K-Means is the Baseline method. In the LDA experiment, the model prior parameters α and β are usually empirically calculated as $\alpha = 50/K$, $\beta = 0.01$. TABLE 2 shows the top 20 words of the potential topic.

As can be seen from TABLE 2, the results of the extraction of potential topics by K-means clustering and LDA topic model are strongly related to the topics that have been tagged. However, from TABLE 2, it is difficult to judge the quality of the topic extraction from the combination of literal words. From the ten tagged news categories, each time five categories are randomly selected, and five topic extraction experiments are performed using K-means clustering and LDA topic model. K-Means is as the Baseline method. The selected five data sets are: D_1 = {tourism, car, education, culture, military}, D_2 = {education, finance, culture, sports, IT}, D_3 = {health, culture, car, military, recruitment}, D_4 = {IT, culture, military, finance, sports}, D_5 = {car, IT, travel, military, health}. The process of topic recognition is a kind of clustering. The number of the clusters is set 5 in the two kinds of experiments. After the topics are extracted, the original texts are classified depending on the identified topics. The classification performance evaluation is performed according to the label tag of the texts. The experimental results are shown in TABLE 3.

From TABLE 3, precision, recall, F_1 and C_{det} of the two topic recognition methods are compared, as is shown in FIGURE 3, FIGURE 4, FIGURE 5, FIGURE 6.

As can be seen from FIGURE 3,4,5,6, for the various indicators such as precision, recall, F_1, and error recognition cost, the LDA topic model is almost better than the Baseline method. It can be known that LDA has obvious advantages in discovering potential topics and subject classifications in the corpus. It is because LDA has the strong generalization ability. So the LDA model is applied to topic tracking experiments.

C. TOPIC TRACKING EXPERIMENT

In the experiments, the collected network news data is utilized for the analysis of topic tracking. The news corpus used in the

TABLE 3. News classification experiment resules.

Data set	method	precision	recall	F_1	P_{miss}	P_{fa}	C_{det}
D_1	Baseline	0.8590	0.6867	0.7632	0.3133	0.0783	0.1422
	LDA	0.8324	0.8060	0.8190	0.1940	0.0485	0.0881
D_2	Baseline	0.6685	0.6013	0.6332	0.3987	0.0997	0.1811
	LDA	0.7537	0.7060	0.7291	0.2940	0.0735	0.1335
D_3	Baseline	0.7648	0.6313	0.6917	0.3687	0.0922	0.1674
	LDA	0.8282	0.8053	0.8166	0.1947	0.0487	0.0884
D_4	Baseline	0.6678	0.5567	0.6072	0.4433	0.1108	0.2013
	LDA	0.8221	0.7713	0.7959	0.2287	0.0572	0.1039
D_5	Baseline	0.8138	0.6893	0.7464	0.3107	0.0777	0.1411
	LDA	0.8337	0.8133	0.8234	0.1867	0.0467	0.0848

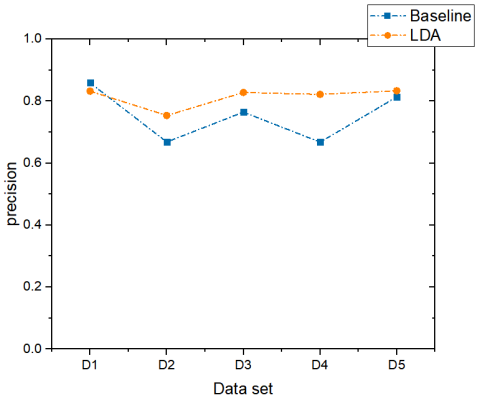


FIGURE 3. Precision of topic classification.

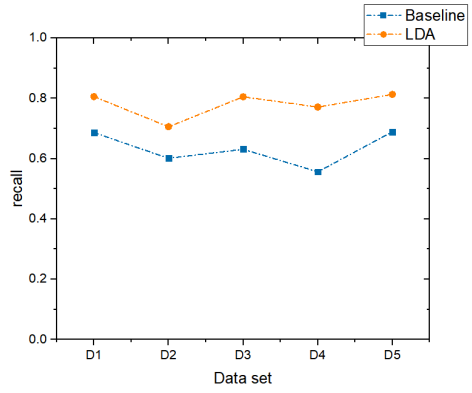


FIGURE 4. Recall of topic classification.

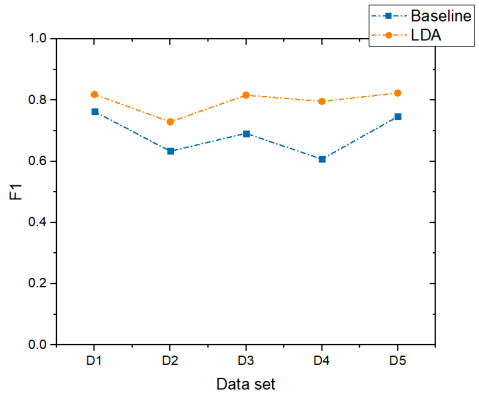


FIGURE 5. F_1 of topic classification.

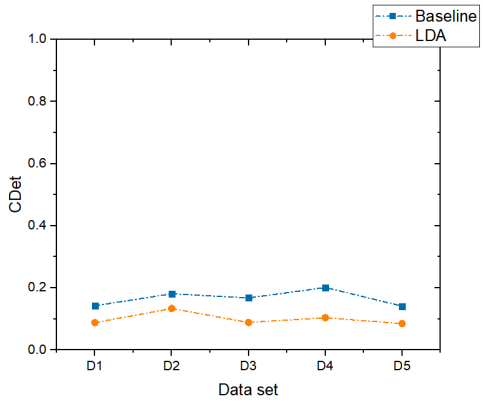


FIGURE 6. The error recognition cost of topic classification.

experiments comes from Sina News Network, Tencent News Network, Netease News Network, People's Network, Xinhua Net, China News Network. From the captured news corpus, the news of February, 2018 is selected as experimental data. The number of news is 123,569.

In the following experiments, the LDA model is adopted to extract news text topics and the best number of the topics per day is verified and determined by multiple experiments. Then the improved Single-pass algorithm is used for topic tracking. Finally, the strength and the content evolution of the topic tracking are analyzed.

1) DETERMINATION OF THE NUMBER OF TOPICS

The lower the perplexity is, the better the model is. The relation experiments of the perplexity and the topics number are conducted. The number of different topics can be set and then the corresponding perplexity of the model can be calculated. Through many experiments on the data for one day, the results of the perplexity are shown in FIGURE 7.

As can be seen from FIGURE 7, the model has the lowest level of the perplexity when $K = 150$. In subsequent

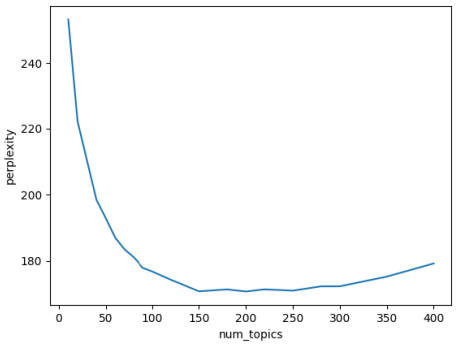


FIGURE 7. The corresponding perplexity of different topics.

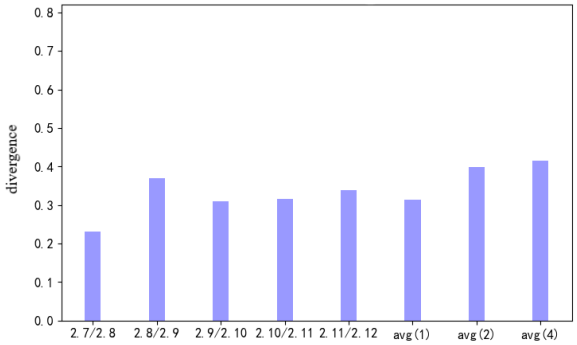


FIGURE 8. Analysis of the topic divergence of the Taiwan hualien earthquake.

experiments, the number of topics in the time window of one day is set to 150.

2) TOPIC TRACKING EXPERIMENT

In the experiments, the time window is set to one day, and the average number of web pages collected per day is about 5,000. From February 6th, 2018 to February 12th, 2018, seven days of news are selected for news topic tracking experiments. In the results of topic tracking, two representative topics are selected for analysis, namely ''Taiwan Hualien Earthquake'' and ''Stock Market Fluctuation''. The topic centers of the two topics in each time window are displayed in TABLE 4.

As can be seen from TABLE 4, the first news topic related to the Hualien earthquake in Taiwan began to appear on February 7th, 2018. The number of the news is 213. The Hualien earthquake occurred at 23:50 on February 6th, 2018 (Beijing time). Therefore, it's logical that the related topics started on February 7th. Judging from the topic of February 7th, among the first 15 topic vocabularies, there are words such as ''injury'', ''depth'' and ''aftershock'', which can reflect the characteristics of the report at the beginning of the earthquake. From February 8th, there have been words such as ''died'' and ''rescue'', which can reflect that the reports have changed from reporting the earthquake situation to reporting the rescue after the earthquake. Beginning on February10th, the emergence of the keywords such as "family", "remains" and "authority" indicated that the focus of the news topic turned to the aftermath of the government after the earthquake. The keywords such as ''tourist'' and ''mainland'' have been running through this topic event from the beginning to the end. At the same time, the keywords ''tourist'' and ''mainland'' locate in a relatively advanced position among the topic keywords. It can be seen that the news media of China mainland focuses on the safety of the mainland tourists traveling in Taiwan. From February 7th to February 12th, a topical report about the Hualien earthquake in Taiwan appeared in the continuous time window, which indicated that this incident was a major event that occurred during this tracking period.

The topic of the stock market is a daily topic. from february 6th, ''fall'', ''plunge'' and ''fluctuation'' keywords appeared in the forefront of the topic. The word "panic" appeared in the topic keywords on February 7th. It can be seen that the stock situation was quite bad. and the rank position of keyword "US" is in front of the "global" position among the topic keywords that day. In addition to the fact that the United states is the center of the global economy, this stock market volatility was also caused by the US stock market's first plunging and then spread to the world. Continued until February 9th, the keyword ''plunge'' also appeared in the forefront each day during the topic tracking process. the word ''fall'' has been continuing at the whole topic tracking process.

3) ANALYSIS OF TOPIC CONTENT CHANGES

The relevant content of the two topics has been analyzed on the above experiments. the following is an analysis of the changes on the contents of hualien earthquake reports based on js divergence (difference), as is shown in FIGURE 8.

In FIGURE 8, avg(n) represents the average of the difference degrees of the associated topics with a time window interval of n days. For example, avg(1) is the average of the first five histograms.

In the first five histograms with one day interval, it can be seen that the news reports of February 7th and 8th have the smallest difference, and the news reports of February 8th and 9th have the largest difference. But the difference does not exceed 0.4 and is less than avg(2) and avg(4). Such differential changes are in line with the actual development of news reports on disaster events. The content of the initial report is mostly about the disaster, so the difference of the topics is small at the beginning. With the development of time, the focus of news reports shifted to rescue information, and then it was biased towards reporting aftercare work. From the analysis of the top 15 keywords in TABLE 4, there are more than ten co-occurrence words in each time window. Although there are many co-occurrence words, it does not mean that there will be little difference between topics. In the actual calculation, it is also necessary to combine the contribution rate of the keyword to the topic.

4) TOPIC STRENGTH ANALYSIS

The LDA model can directly obtain the probability distribution of the topic-document. The strength of the topic in the time window from February 6th to 14th can be computed

TABLE 4. Topic tracking display.

Topic	Date	Topic content (top 15)	Number of news
Taiwan Hualien Earthquake	2018_02_07	Earthquake, Taiwan, Hualien, Mainland, Building, Staff, Sea area, Hotel, Injured, Compatriots, Tourists, Populace, Aftershock, Commander, Depth	213
	2018_02_08	Earthquake, Taiwan, Hualien, Mainland, Tourists, Building, Staff, Died, Rescue, Female, Compatriots, Injured, Aftershock, Disaster relief, Populace	133
	2018_02_09	Earthquake, Hualian, Building, Staff, Mainland, Tourists, Taiwan, Canada, Rescue, Beautiful, Died, Life, Female, Room, Live	74
	2018_02_10	Hualien, Earthquake, Taiwan, Staff, Mainland, Rescue, Died, Building, Tourists, Remains, Confirmation, Room, Gas, Family, Help	70
	2018_02_11	Hualien, Building, Earthquake, Staff, Died, Mainland, Tourists, Remains, Taiwan, Rescue, Demolition, Work, Family, Hotel, Beautiful	43
	2018_02_12	Taiwan, Hualien, Mainland, Earthquake, Building, Died, Staff, Tourists, Rescue, Cross-Strait, Demolition, Authority, Family, Compatriots, Taiwan Office	72
Stock Market Fluctuation	2018_02_06	Stock market, Index, US, Fall, Market, Decline, Plunge, Rise, Economy, Global, Dow, Monday, S&P, Investor, Fluctuate	163
	2018_02_07	Stock market, Index, Market, US, Fall, Fluctuate, Plunge, Global, Investor, Rise, Decline, Dow, Transaction, Panic, Vix	101
	2018_02_08	Index, Stock market, Market, Fall, Rise, Fluctuate, Investor, Risk, US, Decline, Stock, Futures, Plunge, Closing, Performance	75
	2018_02_09	Fall, Index, Stock market, Decline, Plunge, Hong Kong stock Market, Closing, Dow, Sector, S&P, SSE, Over, Rise, Stock	233
	2018_02_10	US, Index, Stock market, Fall, Market, This week, Gold, Data, Friday, Fluctuate, Decline, Global, Economy, Investor, Analysis	76
	2018_02_11	Market, Fall, US, Index, Stock market, Rise, Fluctuate, Price, Global, Stock, Gold, Investor, This week, Transaction, Decline	60
	2018_02_12	Fund, Index, Sector, Security, Gain, Stock, Start an undertaking, Decline, Fall, Rebound,	137
		Rise, Continue, SSE, Investor, Stock price	

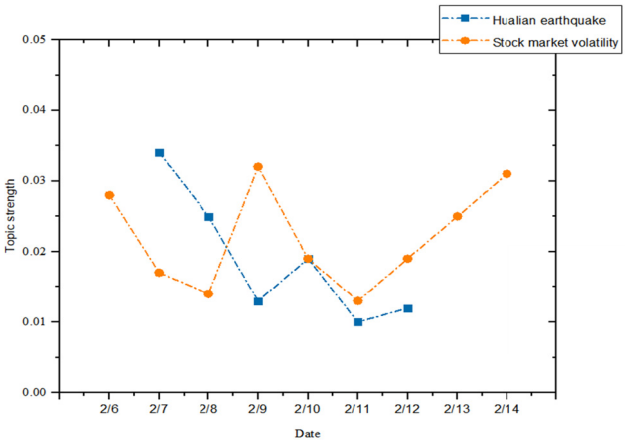


FIGURE 9. Topic strength change chart.

by means of formula (7). Then the topic strength with time changing can be analyzed. FIGURE 9 shows the topic strength analyses of the "Taiwan Hualien Earthquake", "the stock market shocks".

It can be seen from FIGURE 9 that after the earthquake incident in Hualien, Taiwan on February 7th, the strength of the topic reached its peak. Afterwards, although the strength of the topic declined, it still maintained a high level. The trend of the topic is consistent with the reporting rule of disaster events. As for the relevant reports on the stock market turmoil, it can be seen from the FIGURE 9 that with the first round of plunging, on February 6th, the topic strength gradually reached its first peak. With the second round of plunging, it came to the top strength on February 9th. Then the heat is declining, and from the 11th to the 14th, the topic heat is rising. The overall trend of topic strength is very "fluctuating". The analysis of the strength of the above topics is in line with the reality.

V. SUMMARY AND OUTLOOK

In this paper, the topic tracking research is conducted based on topic model. Firstly, the LDA model is used to extract the topic information from the news texts of different time windows. Then the improved Single-Pass algorithm is used for topic tracking, in which the time decay function and the JS divergence are used to measure the similarity between the topics. Finally, for the results of topic tracking, the content and strength of the topics are analyzed. In the experimental part, the topic discovery experiment is first carried out on the tagged corpus. It is found that the topics discovered by the LDA model are more reliable than the k-means clustering in topic recognition. For topic tracking, the perplexity degree

is used to determine the optimal number of the topics in the time window. The proposed topic tracking algorithm is used to trace the real data coming from some Chinese websites. The feasibility of the proposed topic tracking method is effectively proved by these experiments. However, the optimal number of topics needs to set the number of topics in advance. The number of topics per day is a dynamic range. How to adaptively determine the number of daily topics will be a research direction. At the same time, for the characteristics of the news, the position factor is not considered, which will be the direction of the next work.

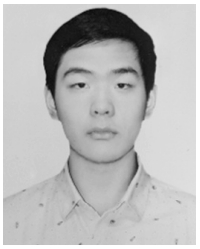
REFERENCES

[1] J. Yuan, D. Zhu, Y. Li, L. Li, and J. Huang, "Survey of text mining technology," *Appl. Res. Comput.*, vol. 23, no. 2, pp. 1–4, Mar. 2006.
[2] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," *J. Comput. Linguistics Lang. Technol.*, vol. 20, no. 1, pp. 19–62, May 2005.
[3] Q. Mei and C. X. Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Chicago, IL, USA, Aug. 2005, pp. 198–207.
[4] S. Xu, "Topic model on image classification and their applications in high spatial resolution remote sensing image," Ph.D. dissertation, School Electron. Inf. Elect. Eng., Shanghai Jiao Tong Univ., Shanghai, China, Aug. 2012.
[5] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 195–200, Jan. 2003.
[6] K. Cui, "The research and implementation of topic evolution based on LDA," Ph.D. dissertation, Dept. Comput. Sci. Technol., Nat. Univ. Defense Technol., Changsha, China, Mar. 2010.
[7] L. Qiu and J. Yu, "CLDA: An effective topic model for mining user interest preference under big data background," *Complexity*, vol. 2018, May 2018, Art. no. 2503816.
[8] Y. Chen, X. Cheng, and S. Yang, "Outburst topic detection for Web forums," *J. Chin. Inf. Process.*, vol. 24, no. 3, pp. 29–36, May 2010.
[9] Z. Lei, L. D. Wu, Lei Lei, and Y. Y. Huang, "Incremental K-means method base on initialisation of cluster centers and its application in news event detection," *J. China Soc. Sci. Tech. Inf.*, vol. 25, no. 3, pp. 289–295, Jul. 2006.
[10] Y. Zhang and A. Song, "Application of improved algorithm based on K-means in microblog topic discovery," *Comput. Syst. Appl.*, vol. 25, no. 10, pp. 308–311, Dec. 2016.
[11] L. Che and X. Yang, "News topic discovery model of multi feature fusion text clustering," *J. Nat. Univ. Defense Technol.*, vol. 39, no. 3, pp. 85–90, Jan. 2017.
[12] M. D. Hoffman, D. M. Blei, and F. R. Bach, "Online learning for latent Dirichlet allocation," in *Proc. Int. Conf. Neural Inf. Process. Syst.* Sydney, NSW, Australia: Curran Associates, Nov. 2010, pp. 856–864.
[13] C. Wang and J. X. Zhang, "Improved K-means algorithm based on latent Dirichlet allocation for text clustering," *J. Comput. Appl.*, vol. 34, no. 1, pp. 249–254, Jan. 2014.
[14] J. Liu, Y. Peng, L. Zhang, Y. Zhang, and J. Deng, "LDA-K-means algorithm of network food safety topic detection," *Eng. J. Wuhan Univ.*, vol. 50, no. 2, pp. 307–310, Apr. 2017.
[15] H. Liu, W. Li, and Y. Zhang, "Microblog topic detection based on LDA model and multi-level," *Clustering Comput. Technol. Develop.*, vol. 26, no. 6, pp. 25–30 and 36, Jun. 2016.
[16] J. Allan, V. Lavrenko, and M. E. Connell, "A month to topic detection and tracking in Hindi," *ACM Trans. Asian Language Inf. Process.*, vol. 2, no. 2, pp. 85–100, Jun. 2003.
[17] X. D. Ren, Y. K. Zhang, and X. F. Xue, "Adaptive topic tracking technique based on K-modes clustering," *Comput. Eng.*, vol. 35, no. 9, pp. 222–224, May 2009.
[18] W.- Y. Bai, C. Zhang, K.-F. Xu, and Z.-M. Zhang, "A self-adaptive microblog topic tracking method by user relationship," *Acta Electronica Sinica*, vol. 45, no. 6, pp. 1375–1381, Jun. 2017.
[19] X. Y. Zhang, T. Wang, and X. B. Liang, "Use of LDA model in topic tracking," *Comput. Sci.*, vol. 38, no. 10, pp. 136–139, Oct. 2011.
[20] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, Apr. 2004.
[21] X. Li, J. Zhang, and M. Yuan, "On topic evolution of a scientific journal based on LDA model," *J. Intell.*, vol. 33, no. 7, pp. 115–121, Jul. 2014.
[22] L. Alsumait, D. Barbará, and C. Domeniconi, "On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*, Pisa, Italy, Dec. 2008, pp. 3–12.
[23] K. Pei, Y. Chen, and J. Ma, "Variable online theme evolution model based on OLDA," *Inf. Sci.*, vol. 35, no. 5, pp. 63–68, Jun. 2017.
[24] Z. Gong, J. Zhe, L. Shoushan, T. Bin, N. Xinxin, and X. Yang, "An adaptive topic tracking approach based on single-pass clustering with sliding time window," in *Proc. Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, Harbin, China, Dec. 2011, pp. 1311–1314.
[25] J. Yu and L. Qiu, "ULW-DMM: An effective topic modeling method for microblog short text," *IEEE Access*, vol. 7, pp. 884–893, Jan. 2019.
[26] J. Ji, C. Liu, and Z. Sha, "Bayesian belief network model learning, inference and applications," *Comput. Eng. Appl.*, vol. 39, no. 5, pp. 24–27, May 2003.
[27] T. He and H. Wang, "Evaluating perplexity of Chinese sentences based on grammar & semantics analysis," *Appl. Res. Comput.*, vol. 34, no. 12, pp. 3538–3542 and 3546, Jan. 2018.
[28] Y. Zhang, G. Han, N. Lu, B. Jiang, and Y. Zhi, "The health assessment of railway vehicle door system based on Jensen–Shannon divergence," *Mach. Des. Manuf. Eng.*, vol. 46, no. 11, pp. 122–127, Dec. 2017.
[29] V. Giedraitis *et al.*, "Genome-wide TDT analysis in a localized population with a high prevalence of multiple sclerosis indicates the importance of a region on chromosome 14q," *Genes Immunity*, vol. 4, no. 8, pp. 559–563, Nov. 2003.

GUIXIAN XU was born in Changchun, Jilin, China, in 1974. She received the B.S. and M.S. degrees from the Changchun University of Technology, in 1998 and 2002, respectively, and the Ph.D. degree in computer software and theory from the Beijing Institute of Technology, in 2010.

Since 2002, she has been a Teacher with the Information Engineering College, Minzu University of China. She is also an Associate Professor. Her research interests are data mining and machine learning.

YUETING MENG was born in Shijiazhuang, Hebei, China, in 1996. She received the B.S. degree in computer science and technology from the Hebei University of Science and Technology, in 2018. She is currently pursuing the master's degree in software engineering with the Minzu University of China. Her research interests include artificial intelligence, natural language processing, and data mining.



ZHAN CHEN was born in Yangzhou, Jiangsu, China, in 1998. He is currently pursuing the bachelor's degree in computer science with the Minzu University of China. His research interests include data mining, math, and algorithm.

XIAOYU QIU was born in Jinan, Shandong, China, in1982. He received the M.S. degree in computer sciences from Shandong Normal University, in 2008. He is currently a Librarian with the Library of Shandong University of Traditional Chinese Medicine. His current research interests include different aspects of pattern recognition, artificial intelligence and distributed systems.

CHANGZHI WANG was born in Chuzhou, Anhui, China, in 1992. He received the B.S. degree in software engineering from South-Central Minzu University, in 2014, and the master's degree in software engineering from the Minzu University of China, in 2018. He works with the Baidu APP Technology Platform Department, Baidu Company. His research interests include data mining, natural language processing, and artificial intelligence.

HAISHEN YAO was born in Heze, Shandong, China, in 1989. He received the B.S. degree from the Shandong University of Science and Technology, in 2016. He is currently pursuing the master's degree in software engineering with the School of Information Engineering, Minzu University of China, Beijing, China. He is interested in scientific activities such as data mining and natural language processing.