Research on Multilingual News Topic Evolution Based on Topic Model

Shiqi Wang

School of Information, Yangzhou University, Yangzhou, China

Abstract: With the development of network technology and the popularization of online media, the number of online news has exploded. However, various news reports often include Chinese, English and other languages, and there is also a lot of redundant information. This actual phenomenon has influenced the research on the evolution of news topics. In order to solve this problem, this paper has constructed a multilingual news topic evolution method. Firstly, a text extraction method based on word clustering information entropy is proposed for new topics; for subtopic clustering, the problems of small clusters generated by single-pass algorithm and topic divergence of traditional LDA model are optimized and improved; When generating tags, based on TF-IDF, the words are given different weights according to their positions and entity elements, so as to select topic words. Finally, this article uses the relevant news of "the Belt and Road" as a corpus to carry out relevant experiments. The results show that the method proposed in this paper can achieve good results and has the feasibility of providing an effective way for network public opinion analysis.

Keywords*:* information entropy; LDA; topic discovery; topic evolution

1. Introduction

With the innovation and development of new technologies, people's production and lifestyles have begun to change. The media industry adapts to the development of the times, uses modern computer and network technology, and expands a new method of communication on the basis of traditional newspapers, radio, and television media, that is, Online media, also known as the fourth media. The emergence of online media has solved the problems of slow transmission of traditional media information, small scope of communication, and lack of effectiveness, which has brought great convenience to information dissemination. At the same time, new media such as self-media have been derived and the media industry has entered a new era. With the continuous maturity and popularization of the Internet and mobile Internet, the channels of information dissemination are diversified, information is presented in real-timedissemination, and the Internet has become an important carrier of information dissemination. On the one hand, online news has become the main source for the public to obtain information, and the public can obtain the latest information dynamics anytime, anywhere. And the access to information is no longer constrained by traditional technologies, and the liberalization of information is realized. But at the same time, due to the communication characteristics of the new media era, everyone is both a communicator and an audience, which reduces the credibility of the information. How to obtain real and effective information from the massive information has become a new problem. In addition, when mass information is disseminated via the Internet, it is very easy to generate online public opinion due to the impact of lowering the threshold for information release and the individual cognition of the communicator.

In view of the current problem of Internet information explosion and the need for public opinion control, a high-speed and convenient way to extract and process massive news has become an important research task in the field of natural language processing. At present, natural language processing technology is becoming more and more mature. The use of news extraction technology to extract new topic articles can effectively reduce the number of news and improve the efficiency of information management and control. In addition, the use of subtopic discovery technology to analyze the subsequent evolution of the article can provide researchers with insight into the direction of public sentiment and take timely countermeasures.

The essence of extracting articles on new topics is the topic discovery task. Topic discovery is a subtask in Topic Detection and Tracking (TDT) task [1]. In 1996, topic detection and tracking evaluation started. In the following year, many research scholars conducted research on TDT, including research on TDT related technologies and TDT task evaluation specifications. Papka [2] attempts to combine the advantages of different clustering algorithms for online topic detection. In order to improve the defects of the statistical model, Yang et al. [3] classified the prior reports of events, and then selected the best reports for analysis, which significantly improved the performance of new event detection. In order to identify words with higher discrimination among different events under the same topic, Kumaran et al. [4] introduced named entity recognition technology in natural language processing to express reports in three vector spaces, indicating that named entities can greatly enhance the degree of distinction among events. In practical applications, the corpus often contains multiple languages, Larkey, Leek, etc. have adopted different strategies for solving cross-

language problems [5,6]. The 2004 TDT evaluation ended, but the text topic is still a hot research. Sekiguchi et al. [7] conducted topic discovery research on blog content. Cselle et al. [8] conducted topic detection and tracking on the content in the email. Wang Wei et al. [9] proposed a multi-center model for hot topic discovery research in order to eliminate the influence of topic errors caused by different content focus under topic news. There are two main methods of research on current topic discovery: one is to use statistical probability information for analysis, and the other is to use features in linguistics for analysis. With the rise of topic models, LDA (Latent Dirichlet Allocation) [10] topic models have been widely used in topic discovery and evolution.

As an improved topic model, topic evolution model assumes that the topic changes with time. The model can identify topics from time series textand track the dynamic evolution of topics. According to the different ways of introducing time, the current research on topic evolution models is mainly divided into three types: first, modeling and then discrete analysis methods. This type of model is derived from the LDA model, that is, without considering the time factor, the LDA model is used on the entire corpus to obtain all the topics, and then the documents and topics are dispersed to the corresponding time points according to the time information of the documents. Although this type of model is simple and easy to implement, it assumes that all documents in the text set are interchangeable, thus failing to make full use of time information, resulting in a post-discrete topic evolution model that is more complicated than other models under the same modeling conditions [11], and the experimental results do not well highlight some short-term hot topics. The second is the method of introducing time variables. Common topic evolution models that introduce time variables are the topic over time model (TOTM) and continuous time dynamic topic model (CDTM). The former combines time, documents, and words as a model Parameters, using the Beta probability distribution model to model the change in topic heat within a given time range. But it only shows the change in topic heat, ignoring the change in topic content, and cannot reflect the evolution relationship of the topic [12]. The latter uses Brownian motion to simulate the evolution of the topic distribution in time, but it has certain requirements on the data itself, which makes the model's generalization ability poor [13]. The third is the method of first discrete and then modeling. The classic first discrete methods are dynamic topic model (DTM), online LDA (online LDA, OLDA) model and sequential LDA (sequential LDA, SLDA) model. These three models first aggregate the data set by time window, and the model parameters under each time window are dependent on the state of the previous time window, and then the model is learned. However, DTM has the problem of granular selection, and OLDA has the problem of topic association and topic detection. None of them can reflect the evolution of the topic content.

In terms of selecting the number of topics, there are mainly the following five methods: One is the experience-based method, but this method relies on subjective judgment and requires constant debugging. The second is a method based on perplexity, proposed by Blei [10] in the original paper of LDA. The third is a standard method based on Bayesian statistics. Griffiths et al [11] proposed the method of using log marginal likelihood function to determine the number of topics. The fourth is a non-parametric method. Teh et al [14] proposed the Hierarchical Dirichlet Processes (HDP). The topic itself is generated from data. The fifthis based on similarity among topics. Guan Peng et al. [15] proposed to use topic variance to measure the degree of deviation between the topic and its mean value, which is used to measure the overall stability of the topic.

This paper constructs a news topic evolution method for the problems of multilingual news and a large amount of redundant information caused by the explosive growth of the number of online news. Firstly, a new topic article extraction method based on word clustering information entropy is proposed, and then sub-topic discovery is performed on the extracted articles to build a topic evolution model. In the end, this paper carried out relevant empirical experiments using the news related to “the Belt and Road” as a multilingual corpus.

2. Theoretical Basis

2.1. Theme Model

As a new statistical method, the topic model analyzes the distribution of words in unstructured text through a probability distribution function to find the topics contained therein, and then uses the obtained topics for subsequent data mining and analysis (such as classification, clustering and evolutionary relationship analysis, etc.). LDA is currently a more popular and mature text topic mining model, which is essentially a three-layers Bayesian model including subject, document and topic, which is completely based on Bayesian inference mechanism and has good knowledge interpretation ability. At the same time, a large number of scholars have proposed a series of LDA-based improved models in combination with different application scenarios, such as the improvement of the model by combining information of authors, topics and links other than documents [16]. These improved topic models are not only widely used in text information analysis (such as Weibo's popular blogger mining [17] and sentiment mining [18], etc.), and has also been applied in the analysis of social networks, images, source code, and biological information.

The LDA model first introduces the probability distribution and Bayesian prior into topic analysis, and secondly uses the prior parameters to estimate the two parameters of the probability distribution of "document-topic" and "topic-feature word" through iterative calculation. The prior distribution of the LDA topic model satisfies the Dirichlet distribution, as shown in equation (1).

$$Dir(\vec{p}\mid\vec{\alpha})=\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_{k-1}} \quad (1)$$

The sample points of the LDA theme model are obtained by Gibbs sampling. Gibbs sampling is mainly based on Markov Chain Monte Carlo for generating high-dimensional distribution random numbers. Gibbs sampling can be understood as, in an n-dimensional distribution, a point is randomly selected as the initial point. Assuming that n-1 dimension is known, another dimension can be obtained by calculating the sampling, and a new calculation can be obtained after a round of calculation. Repeatedly calculating and sampling to get new sample points until the sequence converges. In order to obtain approximately independent sampling, sampling can be set every Literations during the sampling phase.

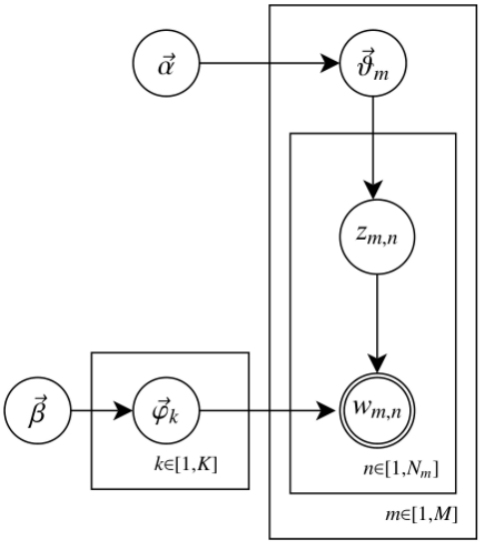

Figure 1. LDA diagram model

LDA is a very important model in the field of text processing and analysis. It can perform topic discovery on articles and calculate the similarity of different articles. The graph model of LDA is shown in Figure 1. Among them, M is the number of documents, K is the number of topics, N_m is the total number of words in the m-th article. $\vec{\alpha}$ and $\vec{\beta}$ satisfy the Dirichlet prior distribution, which determines $\overrightarrow{\theta_m}$ and $\overrightarrow{\varphi_k}$, $\vec{\alpha}$ and $\overrightarrow{\theta_m}$ are K dimensions, $\vec{\beta}$ and $\overrightarrow{\varphi_k}$ are |V| dimensions, |V| is the number of words in the dictionary. $\overrightarrow{\theta_m}$ is the subject distribution probability vector of document m. From $\overrightarrow{\theta_m}$ randomly determine a $z_{m,n}$ to constitute the document-topic distribution, $z_{m,n}$ represents the implicit classification or topic $z_{m,n} \in [1, K]$ corresponding to the n-th word in the m-th article. $\overrightarrow{\varphi_k}$ is the word polynomial distribution vector of topic k, and the value of k in $\overrightarrow{\varphi_k}$ is determined according to the value of $z_{m,n}$. $\overrightarrow{\varphi_k}$ determines the words in the dictionary to constitute $w_{m,n}$, and obtain the topic-word distribution

The goal of the LDA model is to estimate the parameters $\overrightarrow{\varphi_k}$ and $\overrightarrow{\theta_m}$ in the model. $\overrightarrow{\varphi_k}$ describes the relationship between words and topics, $\overrightarrow{\theta_m}$ describes the relationship between documents and topics. The idea of the LDA algorithm is:

First, the joint probability distribution formula is given, as shown in equation (2). Among them, $\vec{\alpha}$ and $\vec{\beta}$ are known a priori Dirichlet distribution parameters, which can generally be determined according to experience or according to the basic idea of Bayesian school. All component values are 1. In the case of known parameters, find $\vec{z}$ and $\vec{w}$, $\vec{w}$ represents the lexical distribution in the corpus, which is the observed known data, and $\vec{z}$ represents the topic distribution in the corpus, which is an implicit variable.

$$p(\vec{\omega},\vec{z}\mid\vec{\alpha},\vec{\beta})=p(\vec{\omega}\mid\vec{z},\vec{\beta})p(\vec{z}\mid\vec{\alpha}) \quad (2)$$

Then the conditional probability formula is derived from the joint distribution according to the Dirichlet distribution formula, which is used for Gibbs sampling. $\overrightarrow{n_z}$ in formula (3) refers to all words in the z-th topic. $\overrightarrow{n_m}$ in formula (4) refers to a word in a topic in the m-th article, $\overrightarrow{n_m}$ is a vector of K elements, each element represents a topic

$$p(\vec{\omega}\mid\vec{z},\vec{\beta})=\prod_{z=1}^{K}\frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})},\vec{n}_z=\{n_z^{(t)}\}_{t=1}^{V} \quad (3)$$

$$p(\vec{z}\mid\vec{\alpha})=\prod_{m=1}^{M}\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})},\vec{n}_m=\{n_m^{(k)}\}_{k=1}^{K} \quad (4)$$

The sample points satisfying the joint distribution are obtained through Gibbs sampling, so as to calculate the topic to which each word belongs, and calculate the formula (5). The implicit subscript of i is (m, n), which represents the n-th word in the m-th article. $z_i = k$ means that the topic of a certain word is classified as k. For known documents and words, randomly select a topic for the words in each document, and then perform Gibbs sampling according to the conditional probability formula until the samples converge.

Finally, after the topic is determined, the distribution parameters θ and φ are derived by formulas (6) and (7).

$$p(z_i=k\mid\vec{z}_{-i},\vec{\omega})=\frac{p(\vec{\omega},\vec{z})}{\vec{\omega},\vec{z}_{-i}}\alpha\frac{n_{k,7i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,-i}^{(t)}+\beta_t}\left(n_{m,-i}^{(k)}+\alpha_k\right) \quad (5)$$

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_m^{(k)}+\alpha_k} \quad (6)$$

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t} \quad (7)$$

2.2. Clustering Algorithm

The current clustering techniques for text topic analysis are mainly partitioning clustering methods and hierarchical clustering methods [19]. But for the task of topic discovery, the number of articles is constantly changing, and new articles are often added. In view of this situation, the above method often requires a global recalculation. Therefore, on the basis of partitioning clustering and hierarchical clustering, some scholars have proposed an incremental algorithm [20] to perform incremental calculations on newly added articles, such as the Single-Pass clustering algorithm [21].

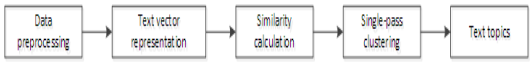

Figure 2. Single-Pass is used for topic discovery process

Table 1. Single-Pass clustering algorithm

Input: text feature vector, similarity threshold Output: multiple clusters
Steps: (1) Use the first document as a seed, create a new topic cluster, and set a similarity threshold □. (2) For subsequent documents, the similarity calculation is performed with the existing topic clusters, and the maximum similarity value between the document and each topic cluster is taken; (3) If the document's similarity value is greater than the threshold □, the document is divided into corresponding topics; if the document's similarity value is less than the threshold □, indicating that the document does not belong to an existing topic cluster, and then a new topic cluster is created and categorize the documents into this cluster. (4) Repeat steps (2) and (3) for new documents until no new documents are added and the clustering ends.

The flowchart of Single-Pass clustering for text topic discovery is shown in Figure 2. The Single-Pass clustering algorithm does not need to set the number of clusters in advance, and mainly depends on setting the similarity threshold to determine whether a document is divided into a certain cluster or create a new topic cluster. The similarity distance can be the distance from the document to the center of the cluster, or the average distance from the document to all samples in the cluster. The specific process of Single-Pass clustering algorithm is shown in Table 1.

3. Evolutionary Analysis Framework of News Topics

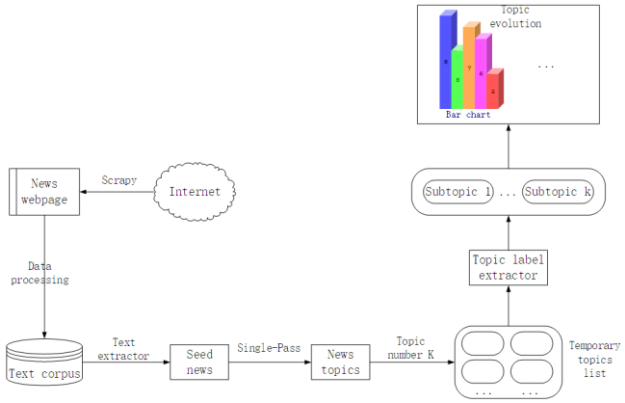

Figure 3. Overall structure of the frame

This paper builds a framework for topic discovery and topic evolution analysis of multilingual news texts based on the topic model. The overall process is shown in Figure 3. First, the news text content obtained through the web crawler technology, and then clean the text data to remove the impurity information in the text. Afterwards, for the processed news information, the news is first extracted in multiple texts, and the extraction target is new topic news in the text. Then use the extracted news text as seed news, and cluster topics according to the seed news. Next, discover the subtopics of the news under the same topic, and extract the topic tags for each subtopic. Finally, according to the change of the number of articles in each subtopic, the evolution trend of the corresponding subtopic is analyzed to determine whether the subtopic is in the ascending, descending, or stable phase, so as to achieve the goal of providing support for the public opinion control of the relevant departments.

3.1. News Extraction based On Word Clustering Information Entropy

In view of the current status of mass news and the effectiveness of public opinion control, this paper extracts new topic news from a large number of texts, which aims to reduce the number of text analysis and improve the efficiency of public opinion control. Since old topic news has been disseminated and fermented for some time, it lacks certain research value. New topic articles can often represent the occurrence of new events. Conduct subtopic research and evolution analysis on them to quickly understand the propagation and evolution effects of new events. Many scholars have studied news extraction, but most of them extract keywords and sentences in a single text. In order to extract articles in multiple texts, this paper proposes a text extraction method based on information entropy of word clustering, aimed at extracting articles in multiple texts. The process is shown in Figure 4.

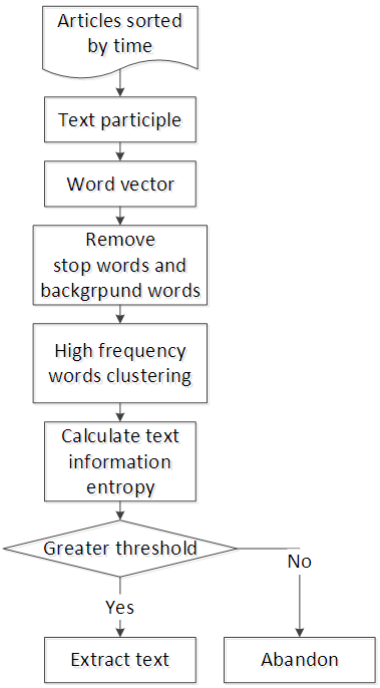

Figure 4. Text extraction process

Information entropy is used to reflect the uncertainty of text content in natural language processing. This paper uses the method of calculating information entropy to quantitatively analyze the changes of news texts, so as to extract articles with large changes in information. When the information entropy surges, it means that the degree of confusion in the system increases, the text information changes greatly, and new content or topics often appear.

Before calculating the information entropy, first determine the random variable describing the text. The minimum granularity of the description text is words, so this paper starts from the granularity of words, clusters the selected high-frequency words, and then uses the clustered phrases as a random variable to calculate the information entropy. That is to say, after clustering high-frequency vocabulary, the frequency of occurrence of all vocabularies in each category is used to calculate the information entropy. The calculation formula of text information entropy is shown in equations (8) and (9). Among them, $p(x_k)$ represents the frequency of occurrence of all words in category k, $label_k$ represents the k-th category after word clustering, and $count(w)$ is the number of occurrences of a word in category k in the article, $count(words)$ represents the total number of words in the article.

$$p(x_k) = \frac{\sum_{w \in label_k} count(w)}{count(words)} \tag{8}$$

$$H(X) = -\sum p(x_k)\log(p(x_k)) \quad (k = 1,2,...,20) \tag{9}$$

3.2. Text Similarity Clustering Based On Single-Pass

The traditional Single-Pass algorithm matches new text with existing clusters. If it matches a certain cluster successfully, the new text is divided into this cluster. Otherwise, create a new class cluster. However, the single-pass clustering algorithm is prone to generate many small clusters, and need to consider merging this small cluster in the future. In order to eliminate the shortcomings of the traditional single-pass, the output of the single-pass algorithm in this paper is only one cluster, and there is no problem of generating small clusters. The similarity is calculated using cosine similarity. The modified clustering algorithm flow is shown in Table 2.

Table 2. Modified Single-Pass clustering algorithm

Input: text feature vector, similarity threshold Output: a cluster
Steps: (1) Take the first article as the first cluster and set it as the CITIC of that cluster. (2) For subsequent articles, the similarity calculation is performed with the existing topic clusters, and the maximum similarity value between the articles and each topic cluster is taken; (3) If the similarity value of the document is greater than the threshold, the document is divided into this cluster, and the center of the cluster is updated by communication; if the similarity of the document is less than the threshold, the document does not belong to the existing topic cluster。 (4) Repeat steps (2) and (3) for new documents until no new documents are added and the clustering ends.

3.3. Subtopic Discovery Based On LDA Topic Model

When LDA topic model is used for topic discovery, the determination of topic data has a great influence on the final result. And the topic evolution is mainly to discover the subtopics of the same topic, and the number of subtopics of different topics also changes dynamically. In order to obtain better results, the number of subtopics often needs to be dynamically adjusted according to the topic.

In this paper, the method of selecting the number of topics based on the Bayesian model is used to calculate the probability $P(W|K)$ of the model given the text data through formula (10), so that the maximum probability K value is the number of topics. Among them, $\Gamma()$ is the Gamma function, V is the vocabulary in the entire text data set, and C_{wj} is the number of times the word w belongs to the topic j.

$$P\ （W \mid K）= (\frac{\Gamma(V\beta)}{\Gamma(\beta)^V})^K \prod_{j=1}^{K} \frac{\prod_V \Gamma\ （C_{\omega j} + \beta)}{\Gamma(\sum_{\omega} C_{\omega j} + V\beta)} \quad (10)$$

This paper is based on the results of single-pass clustering. LDA is used for topic modeling. Since news reports in specific fields are often highly correlated, it is necessary to optimize the vocabulary with small subtopic discrimination in the text. Therefore, before the topic modeling, the use of word vectors to cluster similar words reduces the input dimension of the model and improves the problem of subtopic divergence under the same topic.

3.4. Topic Evolution Trend

The topic evolution proposed in this paper includes two aspects. One is to analyze the evolution trend of the topic according to the change of the number of articles in the topic; the other is to calculate whether there is similarity or evolution relationship between adjacent topics. The first goal can be determined by dynamically counting the number of articles involved in the topic. The second goal can be determined by calculating the KL distance between topics. The smaller the KL distance, the greater the correlation between the two topics, and there may be an evolutionary relationship. Therefore, a distance threshold needs to be set. If the KL distance between two topics is less than the threshold, it is considered that there is a similarity or evolution relationship between the two topics. When the KL distance approaches zero, the two subtopics can be merged. Since the KL distance is not symmetrical, formula (11) is used in this paper to calculate the distance between the two topics.

$$d_{KL}(p,q) = \frac{1}{2}(D(p \| q) + D(q \| p)) \quad (11)$$

3.5. Topic Tag Generation

In view of topic tag generation, this paper uses commonly used extraction methods. Based on TF-IDF, the keywords in each subtopic are extracted, and the ten keywords are selected as the topic tags of each subtopic.

In order to avoid topic words with repeated meanings in the extracted topic tags, synonymous words are merged, and a representative word is selected as the topic word in the merged phrase group. When selecting topic words in phrases, different weights are given to different position elements of vocabulary in the text, which are used to measure the importance of vocabulary in different positions. Give a greater weight to the words and named entities that appear in the title and in the first and last paragraphs, and keep the original values of other words unchanged. According to the calculation result of the vocabulary in the phrase, the vocabulary with the largest value is selected as the subject word.

4. Empirical Analysis

In the verification of results in this paper, the PRF value is mainly used to verify the experimental results of text extraction and clustering. For topic discovery and topic evolution, comparison and display will be made through charts and line charts. For data in languages other than Chinese, it is verified after translation into Chinese.

4.1. Empirical Objects and Data Source Selection

The corpus used in this paper is related to "the Belt and Road" articles searched by search engines. The corpus contains the article's title, source, publication time, and text information. The corpus used for the experimental analysis of this article is 19782, which includes six languages: Chinese, English, German, Thai, and Japanese. 967 articles that included the "Lanzhou Investment and Trade Fair" event were labeled for topic evaluation and used to evaluate clustering algorithms. In addition, according to the weekly event summary released by the Belt and Road Initiative, new topic news in “the Belt and Road” related news from July 1, 2019 to August 23, 2019 is marked. The number of new topics is shown in Table 3. As shown. Table 4 lists some of the new topic events from July 1 to July 7, 2019.

Table 3. Number of new topics in the time window

News Date	Number of new topics
July 01-July 07	18
July 08-July 14	35
July 15-July 21	35
July 22-July 28	38
July 29-August 04	35
August 05-August 11	35
August 12-August 18	40
August 19-August 23	43

Table 4. Some new topic events from 2019.7.1 to 2019.7.7

Number	New topic events
1	Xi Jinping Holds Talks with Bulgarian President and Turkish President
2	Li Keqiang: Abolition of Foreign Stock Ratio Restrictions on Securities and Futures Life Insurance Next Year
3	Re-relaxation of Foreign Investment Access: China ’s Total Trade in Goods with Countries along “the Belt and Road” Exceeds US $ 500 Billion
4	Registration of Professional Visitors from Home and Abroad for the 2nd Import Expo Started

4.2. Analysis of Text Extraction Results

Table 5. Evaluation of text extraction results with different information entropy thresholds

Information entropy threshold	Language	P	R	F
0.6	Chinese	78.3	59.5	67.6
	Foreign language	58.8	67.2	62.7
0.7	Chinese	74.5	76.2	75.3
	Foreign language	70.1	68.9	69.4
0.8	Chinese	69.5	68.3	68.8
	Foreign language	73.2	54.4	62.4

In this section, the comparative experiments on different information entropy thresholds are carried out. The experimental results are shown in Table 5. Among

them, P represents the precision rate, R represents the recall rate, and F value represents the comprehensive index evaluation. P represents the proportion of real new topic articles in the extracted articles, and R represents the proportion of the extracted real new topic articles in all the new topic articles. The following table is an average consideration of all extraction results from July 1st to August 23rd. It can be seen that when the information entropy threshold is 0.7, the text extraction effect is optimal. After that, this paper sets the information entropy threshold to 0.7, and extracts the article with week as the time window.

4.3. Subtopic Discovery and Evolution

After the articlesare extracted, the Single-Pass clustering algorithm is used to cluster similar articles for each new topic article to form a topic cluster. The experimental analysis of topic discovery and evolution analysis in this section will take the "Lanzhou Investment and Trade Fair" as an example. Figure 5 shows the evaluation result of the clustering of the "Lanzhou Investment and Trade Fair" events through different similarity thresholds.

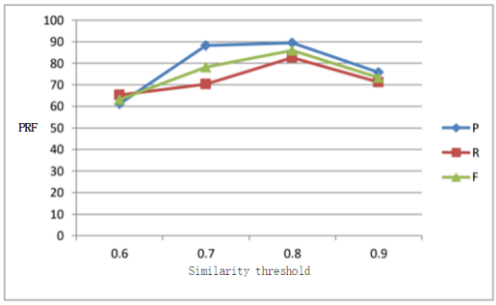

Figure 5. Evaluation graph of clustering results with different similarity thresholds

Table 6. Subtopic hashtag generation results

Topic name	Topic tags (ten)	Subtopic description (manual annotation)
Topic 1	Sasseur; Outlets; project; settle in; attract merchants; attract investment; social interaction; cinema; consumption; International	Sasseur Outlets Opened in Anning District
Topic 2	NetEase; joint; Scientific innovation; park; Sign a contract; innovation; development; knowledge; cloud computing; big data	Northwest's First NetEase Joint Innovation Center Unveiled in Lanzhou
Topic 3	Cpgroup; food; exposition; Chinese food industry; Mutangxiang; New Oriental; Gold emblem wine; Silk Road; commercial trade; Conference	Grand Opening of Longshang Food Expo
Topic 4	green; ecology; investor; favor; Jiayuguan; Zhangye; tourism; service; at home and abroad; hub	Green Ecological Industry is Favored
Topic 5	library; audience; collection; theater; Cultural and Creative; classics; literature; knowledge; Cultural communication; art	Gansu Provincial Library's Cultural and Creative Products on Display

Table 7. Distance matrix between each subtopic

Distance	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	0	0.872	0.724	0.925	0.896
Topic 2	0.872	0	0.954	0.881	0.913
Topic 3	0.724	0.954	0	0.825	0.847
Topic 4	0.925	0.881	0.825	0	0.908
Topic 5	0.896	0.913	0.847	0.908	0

It can be seen from Figure 5 that the optimal similarity threshold for single-pass clustering of "Lanzhou Investment and Trade Fair" events is 0.8, which is the article under this topic after clustering in the time window. There are 488 articles in Chinese and 362 articles in foreign languages. Calculate the 850 articles according to formula (10), and the optimal number of subtopics is 5. The generated tags of each sub-topic are shown in Table 6, and the subtopic description part is manually labeled topics. The distance matrix between subtopics calculated according to formula (11) is shown in Table 7. According to the distance matrix, it is judged whether there is a similarity or evolution relationship between the two topics. The total number of reports of each subtopic in a week is shown in Figure 6, and the trend of the number of reports in a week is shown in Figure 7.

Figure 6 shows that Chinese reports pay more attention to topic 1, topic 2, topic 4, and topic 5. Foreign-language reports have the same reporting trend, but compared to Chinese reports, the number of reports on topic 1, topic 2 is significantly higher than the number of reports on other topics. From the trend graph of the number of topic news reports corresponding to Figure 7, for Chinese and foreign languages, topic 1 has a relatively high report volume at the beginning, and it is in a stable fluctuation state within a week. The total number of reports is relatively close. Topic 2 didn't have a high number of reports at the beginning. It didn't add many related reports until the fourth day, and then the number of reports stabilized. Topic 3 has the same overall reporting trend. However, the number of reports in foreign languages is significantly less than that in Chinese, indicating that foreign languages pay less attention to topic 3. The coverage of topic 4 is relatively high at the beginning, and the subsequent coverage keeps fluctuating around the initial coverage. For foreign languages, unlike the trend of Chinese reports, at the beginning there was a higher coverage of topic 4, and the subsequent coverage gradually decreased to a lower range of coverage. The coverage of topic 5 fluctuated smoothly within the initial coverage. However, the coverage of topic 5 in foreign languages is significantly less than that in Chinese.

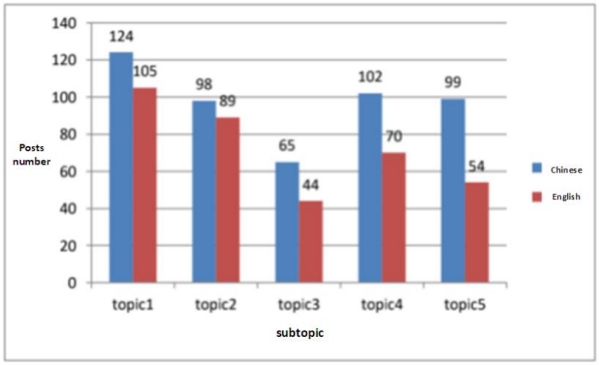

Figure 6. Comparison of the number of reports on various subtopics in Chinese and foreign languages

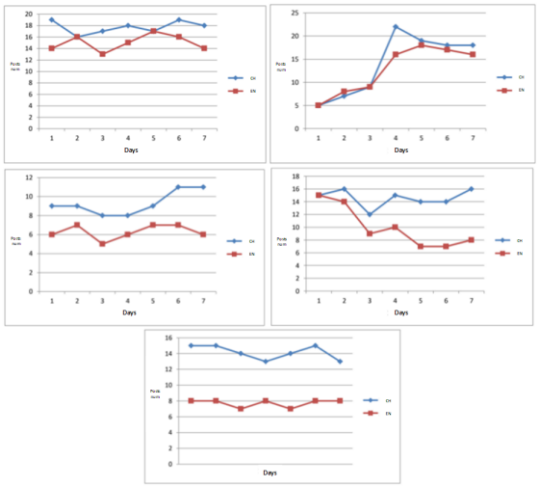

Figure 7. Trends in the number of news reports

(From left to right, from top to bottom are topic 1, topic 2, topic 3, topic 4 and topic 5)

4.4. Analysis of experimental results

Through the above experimental process, it is found that when the information entropy threshold is set to 0.7, the text extraction effect is best, the accuracy rate of the Chinese corpus is 74.5%, the recall rate is 76.2%, and the F value is 75.3%; the accuracy rate of the foreign language corpus is 70.1 %, the recall rate is 68.9%, and the F value is 69.4%. When the similarity threshold of Single-Pass clustering is set to 0.8, it has the best effect on text topic clustering, with an accuracy rate of 70.1%, a recall rate of 68.9%, and an F value of 85.9. In this article, the extracted article is the topic event of "Lanzhou Investment and Trade Fair", and its subtopic discovery and evolutionary analysis are carried out. Through experimental analysis, it is found that there are five subtopics under the topic of "Lanzhou Investment and Trade Fair". In this paper, topic label extraction and evolution trend display within one week are performed on the five subtopics, and the distance between each subtopic is calculated. According to the topic label, we can see that the subtopics under this topic mainly discuss the five aspects of business, technology, food, environmental protection and culture. By setting the distance threshold between topics to 0.5, the corresponding subtopic distance matrix is obtained. It can be found that the distance between each subtopic is higher than the set threshold, and there is no subtopic of similar subtopics or evolutionary relationships. On the one hand, it is verified that the optimal number of topics is obtained by calculating the topic number selection strategy in this paper.

5. Conclusion

Based on the current generation and dissemination of network news information, this paper constructs a multilingual news topic evolution method for multilingual news reports and a large number of redundant information problems in online media. By introducing information entropy in text extraction, the single-pass clustering algorithm for news media information is specifically improved to avoid the problem of small clusters, and at the same time, the word vector is used to optimize the topic divergence of traditional LDA models. Finally, this article makes a relevant empirical analysis based on the corpus of "the Belt and Road" news. The results show that the method proposed in this paper can achieve good results and has the feasibility of providing an effective way for network public opinion analysis. In the following work, we will do further exploration and research on long text vector representation and multi-field news.

References

[1] Zhang, X.Y.; Wang, T. Research of Technologies on Topic Detection and Tracking. Journal of Frontiers of Computer Science and Technology, **2009**, 003(004): 347-357.

[2] Ron, P. On-Line New Event Detection, Clustering, and Tracking. **1999**.

[3] Yang, Y.; Pierce, T, Carbonell, J.G. A Study of Retrospective and On-Line Event Detection// Sigir 98: International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, **1998**.

[4] Kumaran G, Allan J. Text classification and named entities for new event detection//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. **2004**: 297-304.

[5] Larkey, L.S.; Feng, F.; Connell M, et al. Language-specific models in multilingual topic tracking//Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. **2004**: 402-409.

[6] Leek, T.; Jin, H.; Sista, S.; et al. The BBN cross lingual topic detection and tracking system//Working Notes of the Third Topic Detection and Tracking Workshop. **2000**.

[7] Sekiguchi Y, Kawashima H, Okuda H, et al. Topic detection from blog documents using users' interests//7th International Conference on Mobile Data Management (MDM'06). IEEE, **2006**: 108-108.

[8] Cselle, G.; Albrecht, K.; Wattenhofer, R. BuzzTrack: topic detection and tracking in email//Proceedings of the 12th international conference on Intelligent user interfaces. **2007**: 190-197.

[9] Wang, W.; Yang, W.; Q, H.F. Network Hotspot Topic Detection Algorithm Based on Multi-center Model. **2009**.

[10] Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. Journal of machine Learning research, **2003**, 3(Jan): 993-1022.

[11] Griffiths, T.L.; Steyvers, M. Finding scientific topics. Proceedings of the National academy of Sciences, **2004**, 101 (suppl 1): 5228-5235.

[12] Hall, D.; Jurafsky, D.; Manning, C.D. Studying the history of ideas using topic models. Proceedings of the 2008

conference on empirical methods in natural language processing. **2008**: 363-371.
[13] Wang, X.; McCallum, A. Topics over time: a non-Markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. **2006**: 424-433.
[14] The, Y.W.; Jordan, M.I.; Beal, M.J.; et al. Sharing clusters among related groups: Hierarchical Dirichlet processes. Advances in neural information processing systems. **2005**: 1385-1392.
[15] Guan, P.; Wang, Y.F. Research on the Method of Determining the Optimal Topic Number of LDA Topic Models in the Analysis of Scientific and Technological Information. New Technology of Library and Information Service, **2016**, 9: 42-50.
[16] Xu, G.; Wang, H.F. The Development of Topic Models in Natural Language Processing. Chinese journal of computers, **2011** (08): 75-88.
[17] Jianshu Weng, Ee-Peng Lim, Jing Jiang, etc. Twitter Rank: finding topic-sensitive influential twitterers. 2010.
[18] Mei Q, Ling X, Wondra M, et al. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. Proceedings of the 16th International World Wide Web Conference (WWW '07). ACM, **2007**.
[19] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier, 2011.
[20] Zhang, Z.X.; Wu, Z.X.; Zhao, Q.; etc. A Summary of Technical Methods of Extracting Content Objects in Unstructured Text. **2008**.
[21] Tax, Y.D.; Qu, L.Y.; Huang, H.K. A New Topic Detection and Tracking Approach Combining Periodic Classification and Single-Pass Clustering. Journal of Beijing Jiaotong university, **2009**, 33(5): 85-89.