Scientometrics (2022) 127:3767–3792
https://doi.org/10.1007/s11192-022-04439-x

Evolution analysis of online topics based on 'word-topic' coupling network

Hengmin Zhu[1] · Li Qian[1] · Wang Qin[2] · Jing Wei[1] · Chao Shen[1]

Received: 2 January 2021 / Accepted: 7 June 2022 / Published online: 21 June 2022

Abstract

Analyzing topic evolution is an effective way to monitor the overview of topic spreading. Existing methods have focused either on the intensity evolution of topics along a timeline or the topic evolution path of technical literature. In this paper, we aim to study topic evolution from a micro perspective, which not only captures the topic timeline but also reveals the topic status and the directed evolutionary path among topics. Firstly, we construct a word network by co-occurrence relationship between feature words. Secondly, Latent Dirichlet allocation (LDA) model is used to automatically extract topics and capture the mapping relationship between words and topics, and then a 'word-topic' coupling network is built. Thirdly, based on the 'word-topic' coupling network, we describe the topic intensity evolution over time and measure topic status considering the contribution of feature words to a topic. The concept of topic drifting probability is proposed to identify the evolutionary path. Experimental results conducted on two real-world data sets of "COVID-19" demonstrate the effectiveness of our proposed method.

Keywords Online topics · Evolutionary path · Intensity evolution · Topic status

Introduction

Topic evolution has attracted wide attention in academic fields, which focuses on discovering what and how topics change over time from time-stamped document collections (Zhou et al., 2017; Zhu et al., 2015). And it is effective to identify and track the trend of changes in topics by analyzing the process of topic evolution, which helps to monitor the overview of topic spreading.

However, the concept of topics is usually macro and abstract because topics are extracted from large-scale documents, and it is nontrivial to analyze and illustrate the topic evolution process accurately. Particularly, evolution path analysis aims to present topic drifting and it is very interesting but challenging to intelligently infer where a topic may

✉ Li Qian
15852939335@163.com

[1] School of Management, Nanjing University of Posts and Telecommunications, Nanjing, People's Republic of China

[2] School of Computer Science, University of Southern California, Los Angeles, CA, USA

drift and generate the evolutionary path among topics. Much of the previous work has contributed to topic evolution pathway analysis, mainly in the field of technology, and the popular method is to divide documents into different time slices and then extract topics in each slice. However, this method leads to excessive and fragmented topics, and performs poorly in revealing the direction of topic drifting. Therefore, topic evolution still needs further exploration, especially in the area of web text.

In this paper, we propose a method to study topic evolution from a micro perspective. Firstly, this paper constructs a word network according to the co-occurrence relationship of feature words. Secondly, we utilize the topic model, latent dirichlet allocate (LDA), to extract topics and capture the mapping relationship between words and topics, and then build a 'word-topic' coupling network. Finally, based on the 'word-topic' coupling network, we analyze topic intensity evolution over time, the topic status with the consideration of the contribution of feature words to a topic, and the evolution path identified by topic drifting probability.

This study makes several contributions in the realm of topic spreading. First, we describe the process of topic evolution from a micro perspective, that is, the topic evolution is reflected by the changes of feature words. Second, several definitions have been given to measure the topic intensity evolution and identify the topic status, which mainly use the feature words and their contributions to the topics. It is worth noting that the method of describing topic status on technical literatures is improved to evaluate the evolutionary status in web text. Third, different from similarity calculation between two topics in adjacent time slices, the topic drifting probability is defined to identify evolution paths among topics. And experimental results demonstrate that our method can capture the drift among topics during the process of evolution, which enriches the methods for analyzing topic evolution.

The rest of this paper is organized as follows: "Related work" describes the relevant literature. In "Problem formulation", we formally propose the problem of topic evolution analysis. In "Methodology", we present the framework of analyzing topic evolution based on the 'word-topic' coupling network. Afterward, experiments are performed in "Experiments". Conclusion and suggestions for future work are discussed in Conclusion.

Related work

In this section, we survey the related research work, which covers four aspects: topic detection, topic intensity evolution, topic status analysis and topic content evolution.

Topic detection

Topic detection, also called "topic discovery" or "topic identification", aims to find common topics with consistent semantic meaning in a document set (Stein & Zu, 2004). On one hand, there are several topic models detecting topics based on machine learning from large document collections. Wartena et al. (2008) took into account the co-occurrence of keywords, and then detected topics by clustering keywords. To identify the meaningful research fields, Hurtado et al. (2016) performed association rule analysis on document titles with verbs and stop words removed to detect topics. Chen et al. (2019) proposed a nonparametric model (NPMM) that exploits auxiliary word embeddings of NLP to automatically determine whether a given document belongs to existing topics and then inferred the topic number. Additionally, Wei et al. (2020) discovered topics by detecting distinct communities from a co-word network, where words from a specific community belong to the same and unique topic. However, a word usually corresponds to multiple topics in the real word.

On the other hand, most methods have detected topics based on the probabilistic topic model, such as probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) and LDA (Blei et al., 2003). As a variant or extension of PLSA, LDA is a hybrid probability model that clusters words by maximizing the co-occurrence probability of words, which uses Dirichlet distribution to generate the topics of the documents. LDA model can efficiently detect hidden topics of documents and cluster documents. To represent correlations between topics, Blei et al. proposed the correlated topic model (CTM) which uses the logistic normal distribution to model the latent topic proportions of a document (Blei et al., 2006a).

Topic intensity evolution

Studies on topic evolution start with discovering topic trends in temporal documents (Chen et al., 2017). Several models are describing the intensity evolution of topics, which is manifested in the popularity of the topics in different time slices. Wang et al. (2006) proposed the model of topics over time (TOT) to get the evolution of topic intensity, which associates each topic with a continuous distribution over timestamps, and for each generated document, the mixture distribution over topics is affected by word co-occurrence and the document's timestamp. Jian et al. (2018) processed the collection of documents in different time slices with LDA, which calculated the average value of the topic distribution probability on each blog to determine the mean hotness of a topic. Kim et al. (2020) proposed a method to perform blockchain-specific trend analysis, which can extract meaningful topics involved in the current trends of blockchain.

Most studies used the number of topic documents to represent the topic intensity. Generally, the more the number of blogs that discuss a topic is, the more popular the topic is (Song et al., 2018). Liu et al. (2015) used the frequency accumulation of topics in different periods to predict whether a topic would be popular. Zhao et al. (2014) proposed a method to predict short-term topic trends by calculating the "growth factor" of a topic, and considered that the increasing speed of the number of topic documents affects the "growth factor". However, a document may contain multiple topics, and the contribution of the same feature words to different topics is different, so the contribution of feature words to a topic should be considered in topic intensity evolution.

Topic status analysis

The topic status refers to the stage of a topic in its evolution life cycle. When a topic appears, people may be interested in it. As time goes by, people would pay little attention to it. Thus, the life cycle of topic was studied in some work. Fang et al. (2014) developed a topic life cycle model to calculate the energy value of a topic based on aging theory, considering tweets and users authority related to the topic. To track evolution trend of a topic, Du et al. (2020) proposed a hot topic life cycle model (HTLCM) and divided the HTLCM into five stages, i.e., birth, growth, maturity, decline and disappearance. Although some efforts have been made to construct the life cycle of the topic to detect topic trends, few studies have focused on the status of topics in the process of evolution. Collon et al. (1991) proposed the concepts of centrality and density to characterize research themes based on co-word analysis. Liu et al. (2017) and Wei et al. (2020) used density and centrality to measure the status of research topics in the carbon nanotubes area and the status of communities in the co-word network respectively. The above studies calculate density and centrality of a topic only based on the co-occurrence relationship of feature words within the topic and between different topics. However, different topics may contain the same majority of feature words, but the contribution of these same feature words under different topics must be different. Therefore, the contribution of feature words to a topic should be considered in measuring topic status.

Topic content evolution

Topic content evolution is usually manifested in the difference of feature words in different time slices, and this difference is mainly reflected in semantic relevance (Zhang et al., 1991). Although many previous studies have focused on the content evolution to study information diffusion, the actual mechanics of how topics spread on a global scale is still a mystery to be further solved. Different from the topic content evolution, evolution path analysis aims to present topic drifting. Based on definite cases, some scholars studied the evolution path in the field of technology (Jung & Yoon, 2020; Miao et al., 2020; Wu et al., 2014; Xu et al., 2020). For example, Chen et al. (2018) proposed a method that combines LDA and hidden markov model (HMM) to mine the topics of patents, and then explored the topic intensity and the evolution path between topics. However, few scholars applied evolution path to web text, and it is not easy to trace the evolutionary path in large-scale documents. Blei et al. (2006b) developed a dynamic topic model (DTM) which captures the change of topic content, in which the document collection is divided by time slices and the documents of each slice is modeled with the topic model of a fixed topic number. For modeling sequential time-series data with right granularity, Wang et al (2008) developed the continuous time dynamic topic model (cDTM). Topic content evolutionary graphs from short text streams are generated by an Online version of Conditional random field regularized correlated topic model (CCTM) to capture the evolution path of main topics and related subtopics (Gao et al., 2020). In fact, the evolutionary path can be explained by drifting among topics. Consequently, studying the probability of drifting between topics would be a premise for identifying the pathways.

Most of the above studies divide documents into different time slices and then extract topics in each slice. Unfortunately, some topics usually exist in multiple or even all slices, and this method will lead to excessive and fragmented topics. Furthermore, the division of time slices is very arbitrary, and the method of calculating topic similarity in different slices is poor in revealing the direction of topic drifting.

Table 1 summarizes the recent work on topic evolution. The research points of previous studies cover three aspects, topic intensity evolution, topic status, and topic content evolution. As shown in Table 1, most studies are aimed at the literature in the field of technology. Various factors were considered in measuring topic intensity and topic status; the methods of division of time slices and calculation of topic similarity were adopted in most previous studies of evolution path analysis, and it may be suitable for analyzing the evolution process of topics with a long span. However, online hot topics usually last for a short period, only few months at most, and it is hard to analyze the evolution path of topics through division of time slices. Furthermore, topics will drift in the evolution process, which are usually different from the original topics in terms of content, and the correlation

Table 1 Comparison of previous studies on topic evolution analysis

References	Research area	Topic detection	Factors of measuring topic intensity	Factors of measuring topic status	Methods of path evolution
Wei et al., 2020	Parliamentary text	Community detection	N	FC	DTS
Jian et al., 2018	Microblog	LDA	Probability distribution of blog-topic	N	DTS; CTS
Liu et al., 2017	Scientific literature	PLDA	ND	FC	DTS; CTS
Miao et al., 2020	Patents	LDA	CWT	N	DTS
Chen et al., 2018	Patents	LDA	ND	N	DTS; CTS; HMM
Gao et al., 2020	Short text	OCCTM	N	N	DTS; CTS
Current study	Web text	LDA	ND; CWT	FC; CWT	Topic drifting probability

PLDA parallel latent dirichlet allocation, ND number of documents, CWT contribution of words to a topic, FC frequency of co-occurrence, DTS division of time slices, CTS calculation of topic similarity

*The letter N represents ‘without the referred research point’ between topics is poor to detect the drifted topics. In this paper, Topic drifting probability is defined to analyze the evolution path of topics based on the basic topic model, LDA.

Problem formulation

Preliminary definitions and the formal description of the topic evolution problem are given. There are three related tasks to describe topic evolution. The first task is about topic intensity evolution which aims to detect changes in topic over time. The second task is to determine the status of topics. The third task is to identify the evolution path among topics. We formalize the evolution problem as follows.

Topic intensity evolution

After a topic is extracted, the topic intensity is used to quantify topic popularity and changes over time. The number of documents is usually used to measure the intensity of a topic. However, a document may contain several topics, so it is difficult to find sensible ways of assigning documents to a single topic. A topic is represented by some feature words and their contribution (probability distribution) to the topic, so the intensity of a topic can be defined by feature words from a micro perspective.

Definition 1 (*Topic Intensity*) Topic intensity is used to measure the popularity of a topic in a specific time period. Topic intensity evolution shows the changes of intensity over time. Let the feature word set of topic t_i be $V_i = \left\{v_{i1}, v_{i2}, \cdots, v_{im}\right\}$, where v_{ij} represents the j-th feature word of the topic t_i, topic t_i*'s* intensity function can be defined as follows.

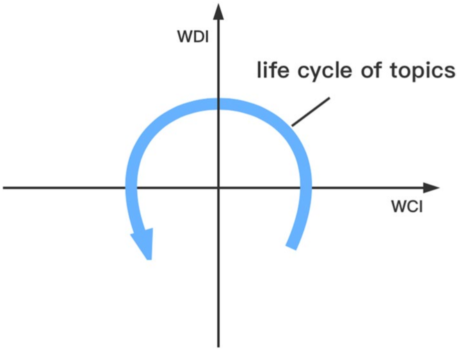
$$thot(t_i, l) = \sum_{k=1}^{m} whot\left(v_{ik}\right) \cdot weight\left(v_{ik}\right) \tag{1}$$

where $thot(t_i, l)$ is the intensity value of t_i at time l; $whot\left(v_{ik}\right)$ refers to the hotness of v_{ik} at time l, and it can be measured by the number of documents containing v_{ik} of topic t_i; $weight\left(v_{ik}\right)$ represents the contribution of v_{ik} to topic t_i.

Topic status

Topic status refers to the stage of the topic in its evolutionary life cycle which includes birth, growth, maturity, decay, and even disappearance. A topic in different status may attract different attention from netizens, which is usually reflected in the varying number of topic-related documents, and even the co-occurrence frequency of feature words. Therefore, centrality and density were proposed to measure the status of research themes based on the co-occurrence relationship of feature words (Callon et al., 1991). However, different feature words contribute differently to a topic, and they play different roles in describing the status of the topic. Furthermore, we cannot ignore the fact that different topics may contain the same majority of feature words, and they are different just because the contribution of the same feature words to different topics is different. Therefore, both the co-occurrence frequency and the contribution of feature words should be considered to analyze the topic

Fig. 1 Analysis of topic status. The x-axis represents the weighted centrality index; the y-axis represents the weighted density index; the coordinate origin is determined by the average of the weighted centrality and density. The direction of the arrow indicates the evolutionary process of the life cycle of topics starting from birth to decay



status. We propose the concepts of weighted density index (WDI) and weighted centrality index (*WCI*) to measure the status of a topic, as shown in the Eqs. 2 and 3:

$$WDI_{t_i} = \frac{1}{n(n-1)} \sum_{s=1}^{n} \sum_{q=1}^{n} F\left(v_{is} \to v_{iq}\right) \cdot weight\left(v_{is}\right) \cdot weight\left(v_{iq}\right)(s \neq q) \tag{2}$$

$$WCI_{t_i} = \frac{1}{n} \sum_{s=1}^{n} F\left(v_{is} \to t_e\right) \cdot weight\left(v_{ik}\right) \cdot weight\left(t_e\right) \tag{3}$$

where $F\left(v_{is} \to v_{iq}\right)$ represents the co-occurrence frequency of the feature word v_{is} and another feature word v_{iq} that are from the same topic t_i; $F\left(v_{is} \to t_e\right)$ represents the co-occurrence frequency of the feature word v_{is} and the feature word t_e of other topics; n denotes the number of feature words of topic t_i;$weight\left(v_{ik}\right)$ represents the contribution of v_{ik} to topic t_i; $weight\left(t_e\right)$ represents the contribution of t_e to other topics.

Topics have different development statuses and trends due to density and centrality. The density represents the connection strength of feature words within a topic, and the greater the Weighted Density Index is, the higher the maturity of the topic is. The centrality represents the connection strength between a topic and other topics. The larger the Weighted Centrality Index is, the more crucial the topic is. When a topic begins to emerge, it usually receives less attention, and its density value is low. If the centrality of a topic is great, it means that it may be derived from other topics or be crucial, so it is potential to develop. As the topic grows and becomes mature, more related documents are published, its density is increasing. When a topic starts to decay, less attention would be paid to it and it is becoming isolated from topics. Therefore, both its intensity and centrality are low. Drawing the idea that a coordinate map was used for identifying the stage of a research topic in the carbon nanotubes area (Liu et al., 2017), we construct a status diagram to analyze the status of topics, as shown in Fig. 1.

Topic content evolution

The evolution of topic content can be shown by the difference of the sequence of feature words over time. From Topic content evolution, it can be seen that the topic content evolution is often computed by topic similarity, which considers co-occurrence of feature words. However, topic similarity cannot effectively reveal the direction of evolutionary path when topic drifting occurs.

A topic is represented by a series of feature words and their contributions to the topic when extracted with LDA model. In the process of topic drifting, two topics definitely contain some common feature words, but their contributions to these two topics are obviously different. If topic t_i is drifting to topic t_j, the contribution of common feature words to t_j is gradually increasing, while the contribution to t_i is gradually decreasing. Hence, the direction of the evolution path can be identified by comparing the contribution of common feature words to two topics. The concept of topic drifting probability is proposed to reveal the direction of evolutionary path.

Definition 2 (*Topic drifting probability*) The topic drifting probability depicts the probability of a topic drifting to another one in topic evolution. It can be calculated by taking into account the contribution of common feature words to topics. Let the set of feature words shared by topics t_i and t_j be $SV = \left\{sv_1, sv_2, \cdots, sv_n\right\}$, where n represents the number of common feature words, the topic drifting probability of $t_i \to t_j$, $P\left(t_i \to t_j\right)$, can be computed as:

$$P\left(t_i \to t_j\right) = \frac{thot\left(t_j, SV\right)}{thot\left(t_j, SV\right) + thot\left(t_i, SV\right)} \tag{4}$$

where $thot\left(t_i, SV\right)$ represents the intensity of the common feature words set SV under topic t_i, and it can be computed with the Eq. (1). Because the sum of $P\left(t_i \to t_j\right)$ and $P\left(t_j \to t_i\right)$ is 1, if $P\left(t_i \to t_j\right)$ is greater than 0.5, $P\left(t_j \to t_i\right)$ is definitely less than 0.5. It is assumed that topic drifting occurs only when the topic drifting probability is greater than 0.5.

Methodology

In this section, a method of analyzing topic evolution based on a 'word-topic' coupling network is proposed, which can effectively explore topic intensity evolution, topic status and evolution path. We provide a framework to describe the approach of topic evolution analysis, including four main parts: data preprocessing, construction of a word network, word-topic mapping, and construction of a topic network.

Framework

Figure 2 illustrates the framework of our proposed approach. First, web text is collected and the feature words from the documents are extracted. Second, based on the co-occurrence relationship of feature words in the same document, a network of feature words is constructed. Third, the number of topics is determined according to Topic Coherence, and LDA model is applied to discover topics from the documents. Fourth, a 'word-topic' coupling network is constructed according to the mapping relationship between words and topics. Then, we calculate the intensity of all topics to generate the topic intensity evolution over time, analyze topic status to detect which stage a topic is in, and calculate topic drifting probability to identify the pathway. Finally, the visualization method is used to demonstrate the topic trends and the pathway.

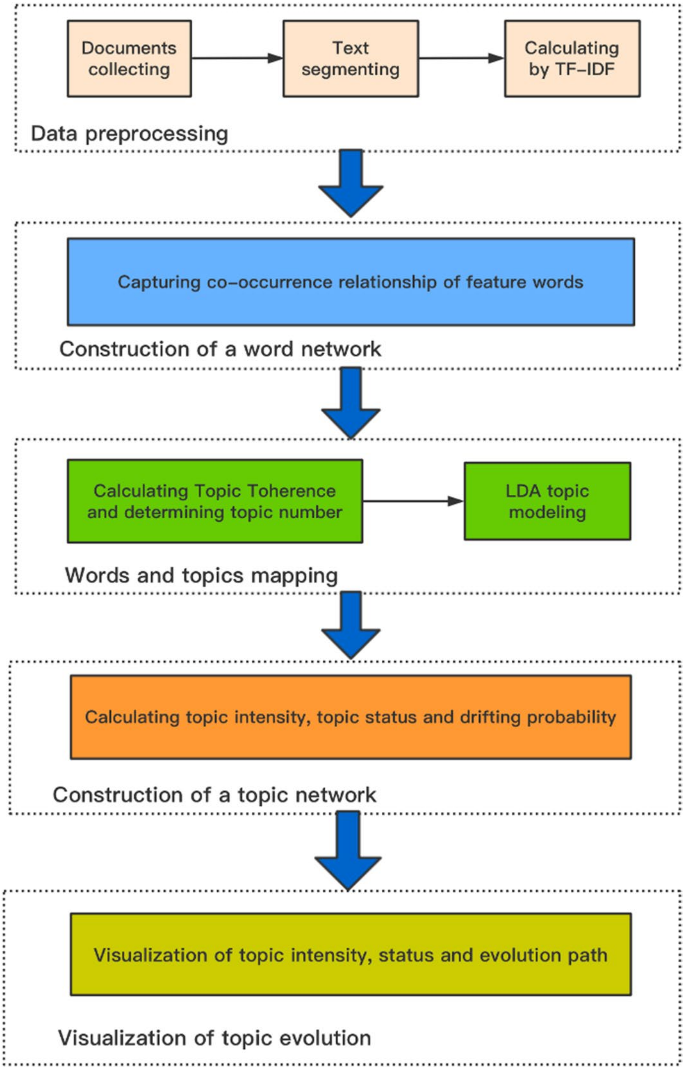

Documents collecting
Text segmenting
Calculating by TF–IDF
Data preprocessing
Capturing co–occurrence relationship of feature words
Construction of a word network
Calculating Topic Toherence and determining topic number
LDA topic modeling
Words and topics mapping
Calculating topic intensity, topic status and drifting probability
Construction of a topic network
Visualization of topic intensity, status and evolution path
Visualization of topic evolution

Fig. 2 Framework of our approach. The process of analyzing topic evolution is briefly summarized

Data preprocessing

Two real data sets are used to study topic evolution in this paper. One is a Chinese data set from Sina Weibo. The other data set is from an openly available database, and the English-language tweet data has been augmented by state-of-the-art named entity recognition algorithms to facilitate access for other researchers.

For Sina Weibo data set, the feature words are extracted from texts because they have definite semantics and reflect topic content. At first, regular expressions are defined to remove non-text symbols in each document. A tool of Chinese text segmentation, Jieba, is applied to segment the texts. In the view of types of vocabularies, feature words are restricted to such words as nouns, person name, location name and organization name, etc., and we remove the words including adverbs, pronouns, onomatopoetic words. Since web text is often colloquial and contains some informal emoticons and cyberwords such as 'u1s1' and 'jimei', they are also removed. A list of stop words (https://github.com/goto456/stopwords) is used to remove some pronouns and non-literal symbols in the text. Then, the weight of feature words is calculated by TF-IDF (Manning et al., 2008) method, which can filter some common words, and retain important words that can provide more meaningful information. These measures can save storage space and improve the quality of the word network. Furthermore, we remove the words of "COVID-19" and "epidemic", because such high-frequency feature words appear in most documents and are not meaningful for distinguishing topics.

Construction of a word network

When extracted from documents, a topic is composed of some feature words. A topic emerges and grows when the feature words contained in the topic appear in many documents within a short period. Sometimes, a topic may be triggered by other topics. Therefore, we can describe the topic evolution from the micro perspective of the change of feature words. A word network can provide an overview of feature words, such as the relationship between feature words and attributes of words, which help to understand topics better.

In this paper, a word network can be represented as a graph $G_v = \{V, E_v, A_v, W_v\}$, where V is the node set of feature words; E_v is the edge set, and an edge e_{ij} represents that there exists a co-occurrence relationship between feature words v_i and v_j, which can be expressed as $E_v = \left\{ e_{ij} \middle| \left(v_i, v_j\right), v_i, v_j \in V \right\}(i \neq j)$; A_v is the set of attributes for nodes that includes hotness and the contribution to a topic; W_v is the weight set of edges, and w_{ij} can be measured by the document count of co-occurrence of v_i and v_j. Since the weights are symmetric, the graph G_v is undirected.

'Word-topic' mapping

After finishing data preprocessing, LDA is applied to extract topics. LDA is a typical statistical topic model. Its basic idea is assuming that implicit semantic structure of a document consists of a set of interrelated topics, and topics are composed of a set of words; assuming that words are generated by the probability distribution of topics, and each topic is represented by the words and their probability distribution on the topic (Wu et al., 2021).

In this paper, we use the genism module of python program to implement LDA. To make the words generated by LDA correspond to the feature words in the word network, the vector of feature words in the word network is used to represent each document. Therefore, we obtain the feature words and their contributions in different topics, and also get the mapping relationship between words and topics.

Before extracting topics with LDA, the number of topics, the parameter K, must be determined, but LDA method itself cannot generate the optimal number of topics. Blei et al. (2012) proposed Perplexity to determine the number of topics, but it is easy to generate topics with large similarities. Teh et al. (2004) proposed Hierarchical Dirichlet Processes, which used a non-parametric model to obtain the number of topics, but this method is low efficiency and it is difficult to guarantee the iteration accuracy for

Table 2 The run times of each step on two data sets

Data sets	Dada preprocessing	Construction of a word network	'Word-topic' mapping	Construction of a topic network
Sina Weibo dataset	24 min	19 min	27 min	14 min
Twitter dataset	–	64 min	76 min	31 min

large-scale text analysis. Topic Coherence scores a single topic by measuring the degree of semantic similarity between high scoring words in the topic, which helps distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference (Stevens et al., 2012). Considering the generalization ability of the model and topic quality, we use Topic Coherence to generate the optimal number of topics and penalize the excessive number of topics on the premise that the topic similarity is small.

Construction of a topic network

A topic network is constructed to visualize the process of topic evolution, which helps users understand the evolution of topics more comprehensively and accurately. In this paper, a topic network is represented as $G_t = \{T, E_t, A_t, W_t\}$. Where, T is the node set, and each node corresponds to a topic; E_t is the edge set, and an edge e_{ij} represents that there exists a drifting relationship from topic t_i to topic t_j, which can be expressed as $E_t = \left\{ e_{ij} \middle| (t_i, t_j), t_i, t_j \in T \right\}(i \neq j)$; A_t is the attribute set of nodes, and to visualize the trend of topic evolution, the attributes are defined as the intensity and status; W_t is the weight set of edges, and an element w_{ij} refers to the probability of topic t_i drifting to t_j which can be measured by Eq. (4). Here the graph G_t is directed.

To examine the time efficiency of the proposed approach, the above steps are run on a Macintosh laptop equipped with a 2.8 GHZ CPU and 8G RAM in the later experimental data sets. Since the time of each run fluctuates slightly, we run each step 10 times and take the average time for recording, and the results are listed in Table 2. Twitter data has been augmented by state-of-the-art named entity recognition algorithms, so we don't perform the step of preprocessing on it. It can be seen that the step of 'word-topic' mapping is the most time-consuming among all steps as the result of the topic extraction by LDA topic model.

Experiments

To verify the effectiveness of the proposed approach, we choose COVID-19 as a case study because it is an event of public health, and most of us are involved and are familiar with the relevant topics on it. Experiments are conducted on two real-world data sets about "COVID-19" to explore topic intensity, topic status and evolution path. The effectiveness of the method is verified by designing comparative experiments.

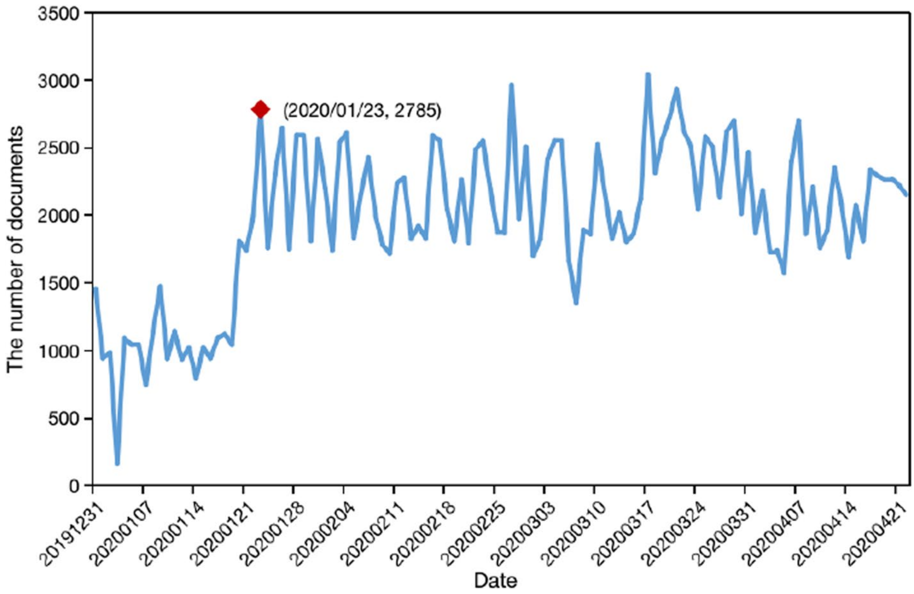


Fig. 3 The daily number of documents (Sina Weibo dataset). The horizontal axis represents the date, and the vertical axis represents the daily number of documents. The red marker point refers to the number of documents corresponding to January 23rd

Data collecting

Two real data sets are used to study topic evolution in this paper. The information about these data sets is detailed as follows:

Sina weibo dataset

The experimental data is crawled from Sina Weibo, which is one of the most popular online social platforms in China. The set of documents is from two hot defined topics of "COVID-19" and "epidemic", and regular expressions are used to guarantee that the content conforms to the "#topic#" format. Each document is composed of release time and content. We obtain a data set of 270,158 documents which are collected from December 31, 2019 (The time point of experts arriving in Wuhan and formally intervening in the investigation) to April 22, 2020 (The time point of China assisting more than 150 countries and international organizations). The daily number of documents of the Sina Weibo dataset is shown in Fig. 3. As can be seen from the figure, the number of documents on January 23 increased sharply and after that maintained a high level because Wuhan Epidemic Prevention and Control Headquarters issued Announcement No. 1, and Wuhan was closed on the same day.

Twitter dataset

The dataset used in this paper is available on GitHub at: lopezbec/COVID19_Tweets_Dataset (Lopez et al., 2021). The dataset contains 3,639,122 related tweets from January

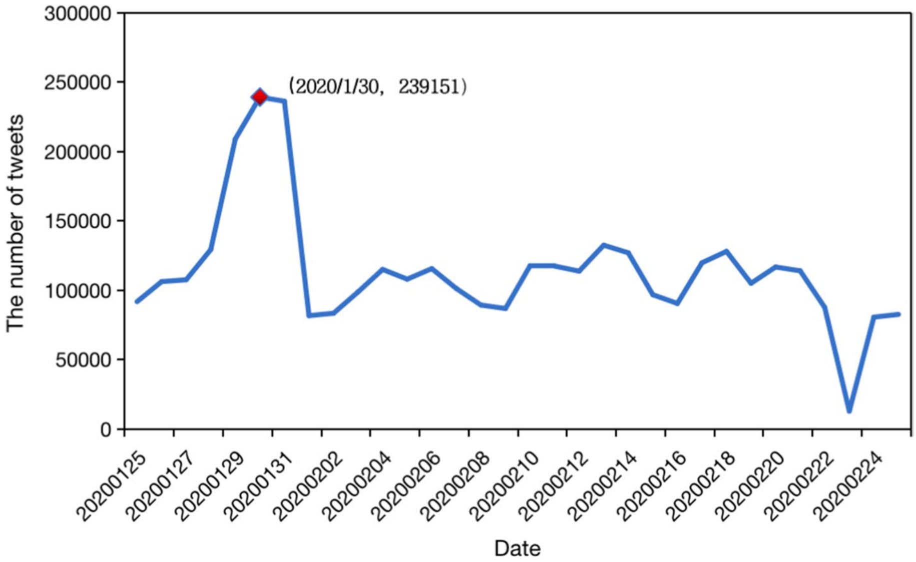


Fig. 4 The daily number of tweets (Twitter dataset). The horizontal axis represents the date, and the vertical axis represents the daily number of tweets. The red marker point refers to the number of tweets corresponding to January 30th

25 to February 25, 2020, when the World Health Organization (WHO) regional director for Europe issued a public statement outlining the importance of being ready at the local and national levels for detecting cases, testing samples and clinical management, thus indicating the potential for a worldwide pandemic. Figure 4 shows the number of English tweets per day, which peaks on January 30, 2020, the day the WHO has declared the outbreak of a novel coronavirus as a public health emergency of international concern.

Construction of 'word-topic' coupling network

For Sina Weibo dataset, 706 feature words are eventually retained through the data preprocessing of "Data preprocessing". According to the method of "Construction of a word network", the word network is established with 469,136 edges as shown in Fig. 5.

The optimal topic number is determined according to the Topic Coherence when extracting topics from the documents. For all textual documents, the topic coherence is calculated individually with different topic numbers as inputs to train LDA model, and the curve of Topic Coherence is shown in Fig. 6. It can be seen that as the number of topics increases, the Topic Coherence gradually increases and tends to be stable. The topic number of seven is the first point that the curve becomes stable. Considering the risk of overfitting with a larger number of topics, we finally set the number of topics as seven.

Seven topics are detected by LDA model, and each topic is represented by the top 15 feature words ranked by their contributions. We name each topic by summarizing the feature words. Since the crawled documents are in Chinese, the feature words of each topic are translated into English and listed in Table 3. Then, a topic network is constructed to visualize the process of topic evolution, as shown in Fig. 7. Each topic can drift toward another one via an evolution path, and it is easy to understand the information spreading process.

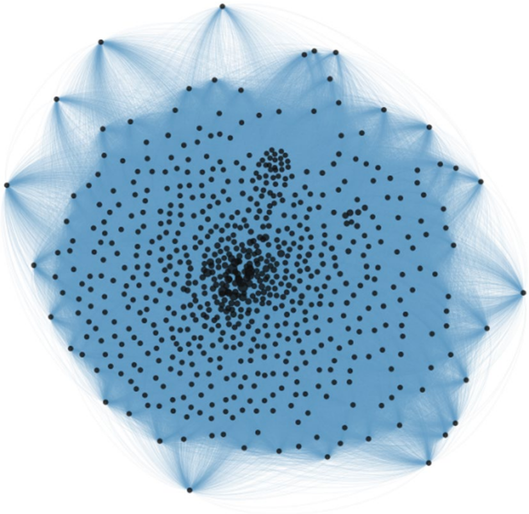


Fig. 5 Visualization of the whole word network (Sina Weibo dataset). The thickness of an edge represents the frequency of co-occurrence between nodes. The node size indicates the hotness of the feature word

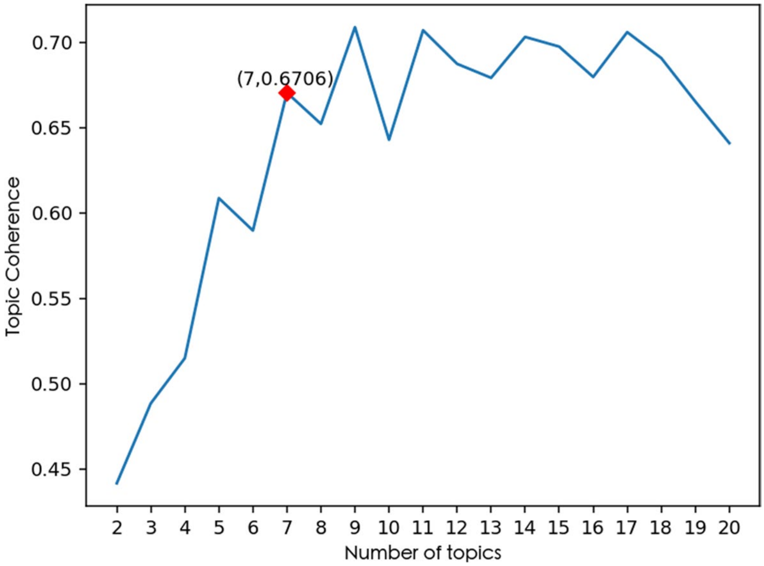


Fig. 6 Topic coherence of different numbers of topics. The horizontal axis represents the number of topics, and the vertical axis represents the value of topic coherence. The red marker point in the figure refers to the number of topics determined

In this paper, each topic is represented by some feature words. To describe the topic evolution in detail, the word network is colored according to the mapping relationship between words and topics, as shown in Fig. 8. If some feature words make a little contribution to the topic, we can pay little attention to them. Thus, we color those feature words whose contributions sum up to 80% for each topic. If a feature word is contained in multiple topics, it will be colored with the topic in which the feature word is with the largest contribution. We use Gephi to display the network, and ForceAtlas2 is used for network spatialization

Table 3 The top 15 ranked feature words of 7 topics (Sina Weibo dataset). The abbreviation No. represents the topic identification number

No	Feature words	Name
1	Prevention; Control; Measures; Service; Community; Management; Leading group; Emergency; In order; The public; Residents; Disinfection; Staying home; Materials; School	Community prevention and control
2	Masks; Vaccines; Research; Transmission; Zhong Nanshan; Drugs; Team; Clinical trials; Biology; Antibodies; Disinfection; Infection; Experts; Clinical; Medical	Drug test
3	Patient; Cases of disease; Hospital; Test; Discharge; Nucleic acid; Report; Symptom; Cause; Positive; Viral; Medical institution; Antibody; CDC; Sample	Virus detection
4	Wuhan; China; Fight; Medical staff; Supplies; Medical team; Heroes; Protect; Doctor; Cooperation; Frontier; Police; Success; Assistance; Nurse	Fight for epidemic
5	Influence; Economy; Company; Enterprise; Market; Loan; Production; Industry; Price; Employee; Consumption; Shock; Investment; Tourism; Unemployment	Epidemic impact
6	United States; Global; Japan; Trump; United Kingdom; Italy; WHO; News; Korea; France; Russia; Spain; Measures; Europe; World	International epidemic
7	Court; Guardian; Illegal; Public welfare; Foundation; Evidence; Case; Law; Epidemic area; Monitoring; Institution; Lawyer; Judicial; Trial; Fair	Laws

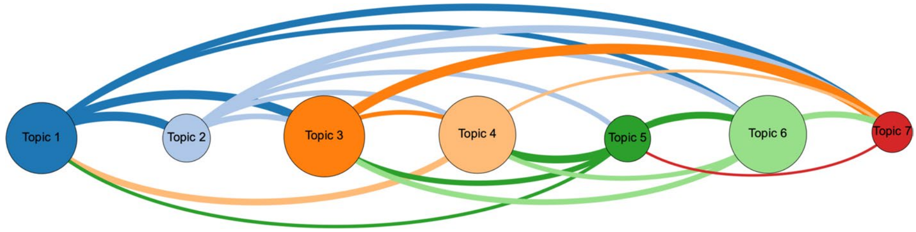


Fig. 7 The topic network with topic drifting probability greater than 0.5 (Sina Weibo dataset). Thicker edges represent the larger topic drifting probability between nodes. The node size indicates the topic intensity. Different colors represent different topics, and the color of the edge is determined by the color of the source node

which is a force-directed layout algorithm. As long as it runs, the nodes repulse and the edges attract (Jacomy et al., 2014). From Fig. 8, it is easy to see the distribution of topics at the feature words level. Some topics gather in the middle area of this network and are

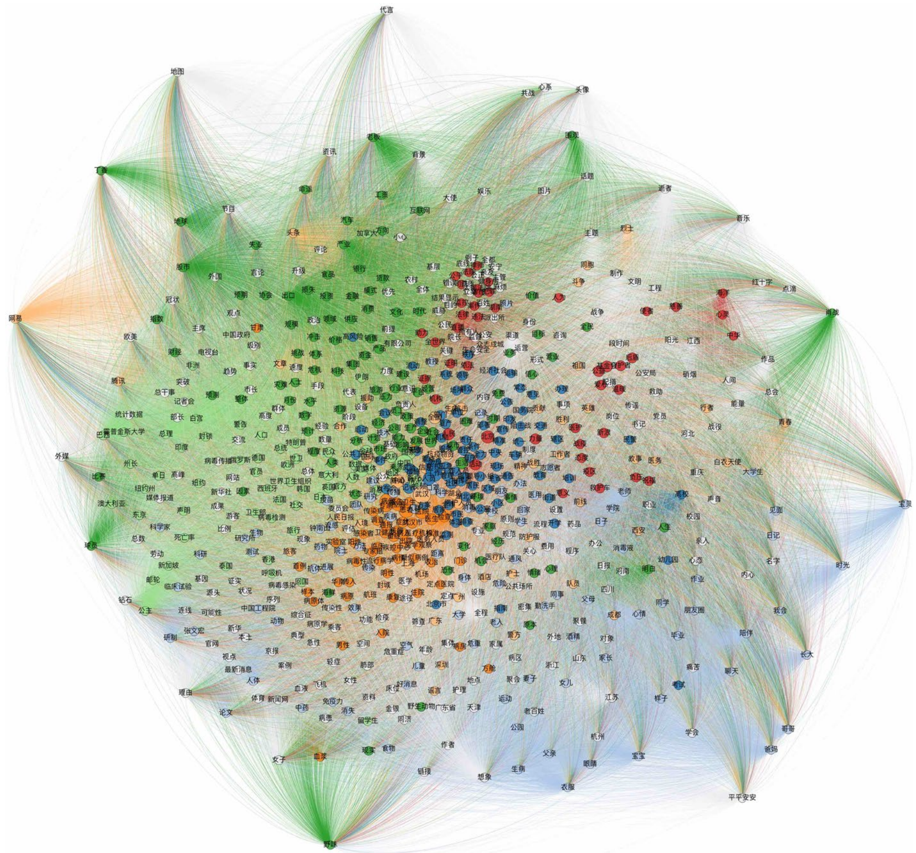


Fig. 8 Coloring the whole word network. The thickness of edges represents the co-occurrence frequency between the nodes. The node size indicates the hotness of the feature word. The nodes of the same color represent they are in the same topic

interpenetrating, such as Topic 1, Topic 4, Topic 5 and Topic 6. It shows that the co-occurrence frequency of the feature words from these topics is high, and the probability of these topics drifting from one to another is high.

For Twitter dataset, a 'word-topic' coupling network with 674,173 nodes is constructed in the same way. The optimal topic number is determined to be 5, and the feature words of each topic are listed in Table 4. Then, a topic network is constructed to visualize the process of topic evolution, as shown in Fig. 9.

Topic intensity evolution

The topic intensity of each day is calculated according to Eq. (1), and then topic intensity evolution can be described by the percentage changes of each topic's intensity over time. For Sina Weibo dataset, as shown in Fig. 10, Topic 3 receives widespread attention in January, and its intensity is very high; as time goes by, the intensity is gradually declining. Actually, in January 2020, the event of COVID-19 appearing in the Wuhan caused uproar, and netizen started a heated discussion about the threaten of the epidemic and virus detection; in February, virus detection was still a hot topic, but more people were concerned about the subsequent prevention measures and other topics.

Two critical junctures of "Laws" (Topic 7) are detected, including January 21, 2020, and April 14, 2020 (Fig. 10). The first one is in the early stage of the outbreak of the domestic epidemic. Hubei Red Cross announced that the donated goods would be used for epidemic prevention and then raised doubts about the unfair assignment. The second one is due to several newly imported cases and criminal liability which would be imposed on those who concealed their exit history or failed to declare their health status truthfully at the time of entry. Also, "Fight for epidemic" (Topic 4) peaks on April 4, 2020 (Fig. 10), although the popularity had been weakening until then. On that day, national leaders of China attended mourning activities in Beijing, and the half-mast ceremony was held in Tiananmen Square.

To describe the topic intensity evolution in detail, this paper divides the period into three parts, which are from December 31, 2019, to January 21, 2020, from January 22 to March 10, and from March 11 to April 22. We find that the intensity of some topics that were mature and prominent in the early stages experienced a downward trend over time. In the first stage, "Virus detection" (Topic 3) and "Fight for epidemic" (Topic 4) aroused the public's greater attention. In response to the rapid spread of the epidemic, National Health and Construction Commission announced that it would start providing the detection and diagnosis kits for suspected cases, but the detection reagents were still in short supply. In the second period, "Community prevention and control" (Topic 1), "Virus detection" (Topic 3) and "Fight for epidemic" (Topic 4) maintained a high intensity, while Topic 6 of "International epidemic" showed a significant increase in intensity. At this stage, it is worth noting that the number of new local cases was gradually decreasing, while new imported cases were increasing and the epidemic indicated a trend of globalization. In the third stage, "International epidemic" (Topic 6) was the most popular topic. Fourteen hospitals in Wuhan closed one after another; the epidemic overseas was serious and many regions had adopted closed measures.

The Fig. 11 presents the proportion of intensity per day for all topics of the Twitter dataset. It can be seen that the popularity of "Cases and death" (Topic 2) and "Preventive measures" (Topic 5) show an increasing trend. Since the first case of COVID-19 was reported in Wuhan, China, in December 2019, the number of infections and deaths worldwide has increased rapidly. There were several fluctuations in the intensity

Table 4 The top 15 ranked feature words of 5 topics (Twitter dataset)

No	Feature words	Name
1	Gov; Trump; Announce; United states; White house; Notice; USA; Washington; Coronavirus; California; York; Georgia; Close; Mike Pence; Breaking	US States authorities' announcements
2	Confirm; Report; Rise; Philippines; Death; Case; Outbreak; South; Italy; Lockdown; Korea; Total; Official; Country; Flu	Cases and death
3	CDC; Hongkong; Japan; Diamond Princess; Cluster; Japanese; Ship; Cruise; Passenger; America; Infect; News; Biden; Virus; China	Diamond Princess cruise ship outbreak
4	China; Wuhan; Coronavirus; Chinese; Covid-19; CDC; Hongkong; Coronavirusoutbreak; Corona; Hubei; Beijing; Trump; Corona virus; Outbreak; Cemocrats	China and Wuhan outbreak
5	Spread; Lockdown; France; Mask; US; Patient; Australia; Vietnam; Singapore; Hospital; Test; Discharge; Nucleic acid; Report; Viral; Medical institution	Preventive measures

No topic identification number

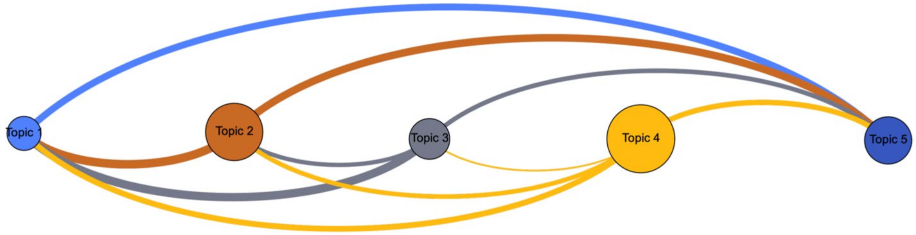


Fig. 9 The topic network with topic drifting probability greater than 0.5 (Twitter dataset). Thicker edge represents the larger topic drifting probability between nodes. The node size indicates the topic intensity. Different colors represent different topics, and the color of the edge is determined by the color of the source node

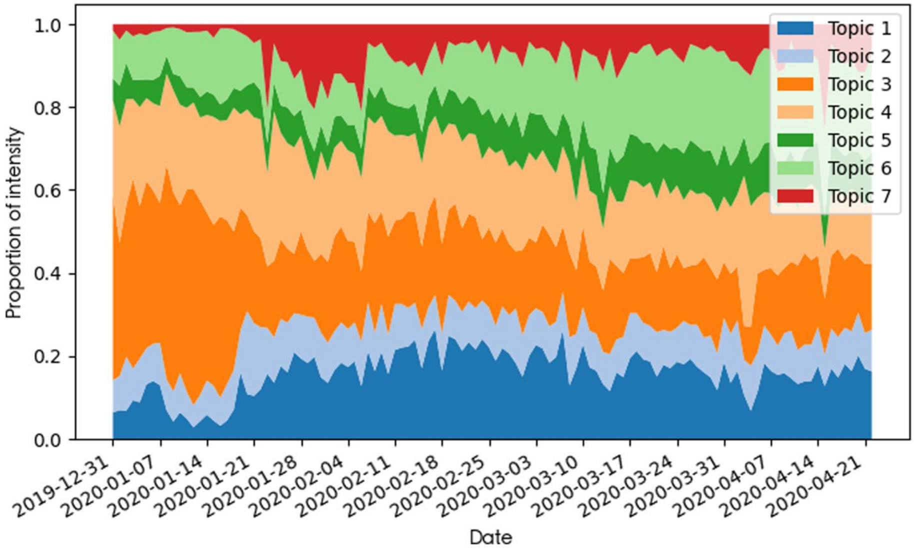


Fig. 10 Topic intensity evolution over time (Sina Weibo dataset). The horizontal axis represents date, and the vertical axis represents the proportion of intensity

of “cases and death” (Topic 2), with a peak on February 2 due to the Philippine Department of Health notified a case of death, the first outside China. “China and Wuhan outbreak” (Topic 4) has a strong prevalence up to February 8, 2020, while “Cases and death” (Topic 2) predominates from February 23 to February 25. In addition, “Diamond Princess cruise ship outbreak” (Topic 3) peaked on February 5, when a cluster of infections was reported on the Diamond Princess cruise ship. Similarly, “US States authorities’ announcements” (Topic 1) peaked on February 5, the day the White House asked U.S. scientists to investigate the origin of COVID-19.

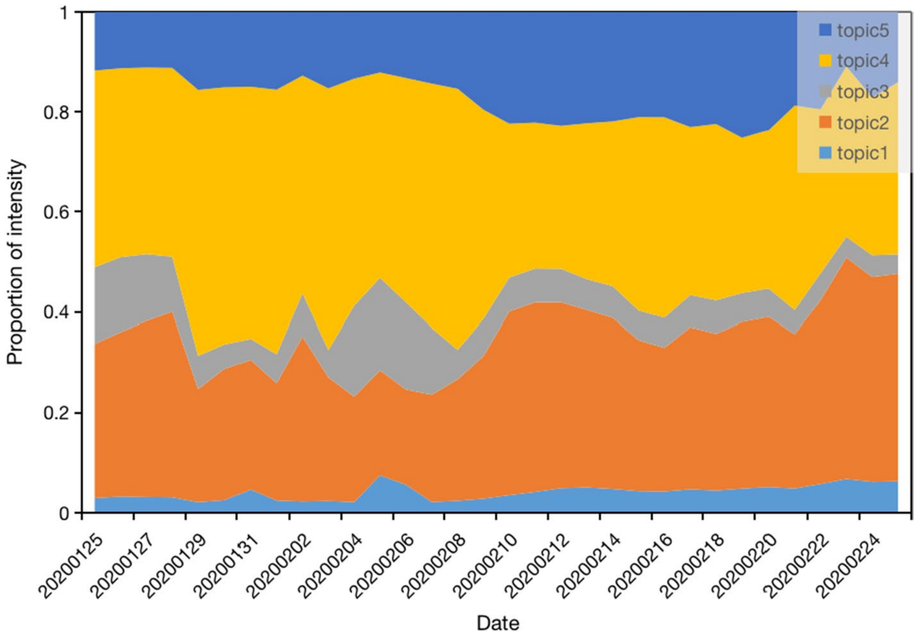


Fig. 11 Topic intensity evolution over time (Twitter dataset). The horizontal axis represents date, and the vertical axis represents the proportion of intensity

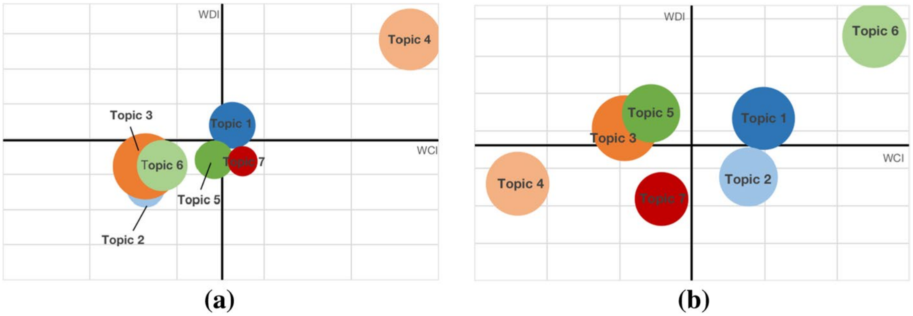


Fig. 12 Analysis of topic status (Sina Weibo dataset). The horizontal axis indicates the weighted centrality index, and the vertical axis indicates the weighted density index. The size of the dot is determined by its topic intensity. Different colors represent different topics. **a** represents the topic status from December 31, 2019 to January 7, 2020; **b** represents the topic status from March 17, 2020 to March 24

Topic status

For Sina Weibo dataset, taking the documents from December 31, 2019 to January 7, 2020 and March 17 to March 24 for example, the topic status is measured by Eqs. 2 and 3, and the results are shown in Fig. 12. As seen in Fig. 12a, in the early period of the epidemic, most of the topics are undeveloped and weakly connected to other topics, but "Fight for epidemic" (Topic 4) is a crucial topic in this period. At the beginning of

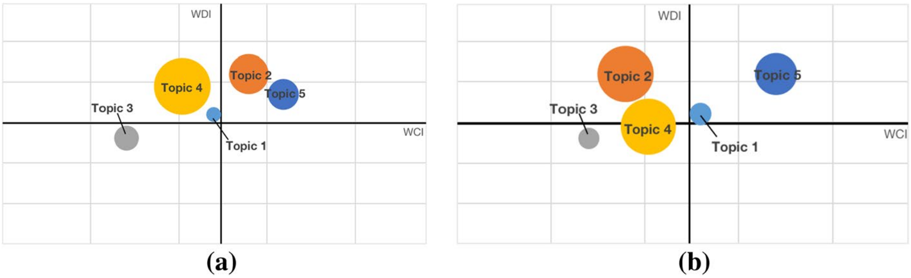


Fig. 13 Analysis of topic status (Twitter dataset). The horizontal axis indicates the weighted centrality index, and the vertical axis indicates the weighted density index. The size of the dot is determined by its topic intensity. Different colors represent different topics. **a** represents the topic status from January 29, 2020 to February 8; **b** represents the topic status from February 19 to February 25

the epidemic, most of the topics are beginning to emerge, so the density and centrality are low. "Community prevention and control" (Topic 1) and "Epidemic impact" (Topic 5) have not attracted much attention from the netizens and undeveloped, but they have potential to grow because of their higher centrality value. During the period, the patient suffered from health problems and the public feared the long-term effects of the epidemic. So people will be more involved in the epidemic prevention and the discussion of its impacts, and Topic 1 and Topic 5 will gradually grow. As shown in Fig. 12b, Topic 1 and Topic 5 develop in the later stages of the epidemic, and Topic 1 becomes increasingly crucial and mature. "International epidemic" (Topic 6) has been mature because of its high density and centrality. It is much likely that the topic would burst after this time period. In the real world, after observing that the imported cases increase over time, people did concern more about the globalization of the epidemic. Besides, Topic 4 is in the status of decay and people no longer paid more attention to this topic in this period. It is because we have obtained remarkable achievements in prevention and control through the joint efforts.

For Twitter dataset, taking the tweets from January 29, 2020 to February 8 and February 19 to February 25 for example, the results are shown in Fig. 13. Topics have different development status and trends due to the different density and centrality. As can be seen in Fig. 13a, "US States authorities' announcements" (Topic 1) is not developed and has few connections to other topics, while "China and Wuhan outbreak" (Topic 4) has been mature. Until then, the epidemic in China had been underway for about a month, while the cumulative number of confirmed cases in the U.S. was just more than 10 on February 3, and Trump claimed the situation domestically was not serious. "Cases and death" (Topic 2) and "Preventive measures" (Topic 5) are crucial topics with high centrality. In fact, in response to the increasing number of cases and deaths worldwide, more and more countries were taking preventive measures. For example, France, Singapore and other countries banned people from China entering the country. The topic status of 3 weeks later is shown in Fig. 13b. "Cases and death" (Topic 2) has been mature, while "China and Wuhan outbreak" (Topic 4) is in the status of decay. It is because the number of confirmed cases outside China continues to rise, and the first confirmed case of Africa was found on February 15.

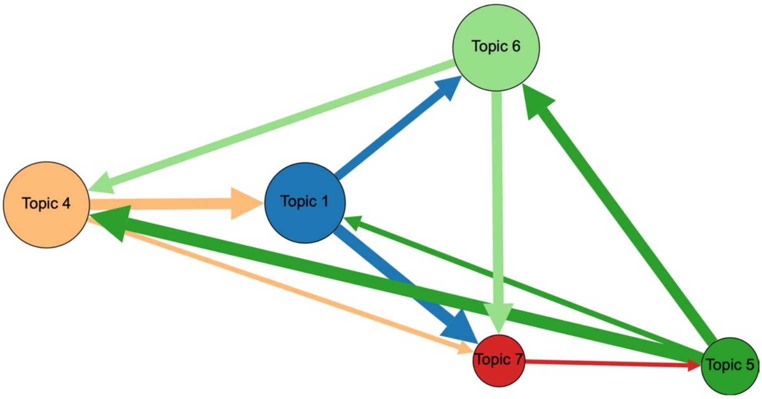


Fig. 14 Analysis of main paths of topic evolution (Sina Weibo dataset). Thicker edge represents the larger topic drifting probability. The node size indicates the topic intensity. Different colors represent different topics, and the color of the edge is determined by the color of the source node

Topic evolution path

For Sina Weibo dataset, several main evolution paths are extracted out from the whole 'word-topic' coupling network (Fig. 7 and Fig. 8), as shown in Fig. 14. There exist two evolutionary paths, including Topic 4→Topic 1→Topic 6 and Topic 4→Topic 1→Topic 7→Topic 5. First, for "Fight for epidemic" (Topic 4), the probability of drifting toward "Community prevention and control" (Topic 1) is the largest. Actually, we query the documents and find that as the epidemic developed, more and more measures were taken to fight for epidemic, such as community prevention and control. Second, among these topics, the topic with the largest cumulative drifting probability is Topic 1, which mainly drifts toward "International epidemic" (Topic 6) and "Laws" (Topic 7). With the increase of the number of imported cases, the attention paid to domestic "Community prevention and control" in the early stages gradually drifted to "International epidemic". This indicates that the globalization of the epidemic may attract more attention in the next period. Moreover, Topic 1 drifts toward Topic 7 because there exists hot debate about the unfair distribution of supplies by the Red Cross in the topic of "community prevention and control", and the government was discussing relevant laws against this situation. Third, there is an evolution path from Topic 7 to Topic 5 which indicates people discuss more about the impact of the epidemic and less about the laws. The relevant laws had been well established, and the government began to address issues after the epidemic: vaccination experiments and other prophylactic measures were carried out; 17 provinces clarified the time when the next semester would begin; many institutions in Beijing released some information about recruiting college graduates in the next 2 years. Besides, we can find that the topic drifting probability of Topic 5 to Topic 4 and Topic 6 is high because the impact of the epidemic is closely related to the domestic and international epidemic. All of these verify the value of analyzing the evolution paths.

For Twitter dataset, Fig. 15 depicts the topic evolution paths. It is worth noting that "China and Wuhan outbreak" (Topic 4) drifts towards the other four topics, which indicates that Topic 4 is a crucial topic, and it was also quite popular during that period since the Wuhan epidemic received great public attention worldwide. For example, "China and Wuhan outbreak" (Topic 4) drifts toward "Cases and death" (Topic 2), and the fact

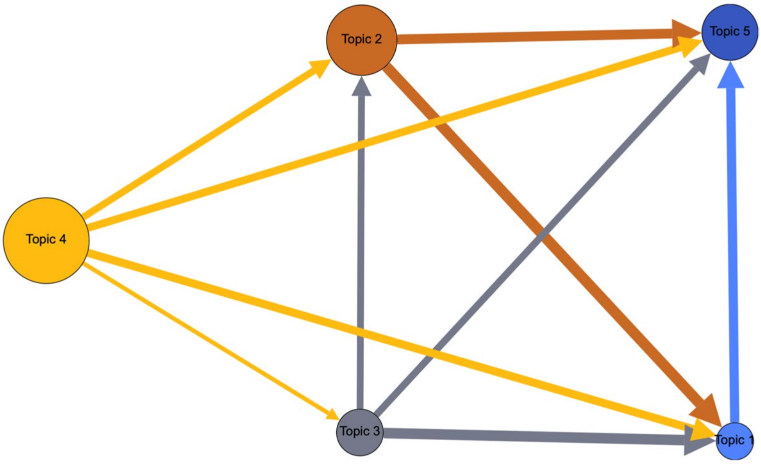


Fig. 15 Analysis of main paths of topic evolution (Twitter dataset). Thicker edge represents the larger topic drifting probability. The node size indicates the topic intensity. Different colors represent different topics, and the color of the edge is determined by the color of the source node

is that the number of confirmed cases in China declined, but the number of global confirmed cases continued to rise. Then, many countries took measures to avoid the spread of the epidemic, so the public began to focus on "Preventive measures" (Topic 5). Besides, "China and Wuhan outbreak" (Topic 4) drifts toward "Diamond Princess cruise ship outbreak" (Topic 3) because of the increasing concern on clustered infections outside China. Specifically, forty-four Americans on the Diamond Princess cruise ship were infected with COVID-19. As a result, the CDC (US) announced on February 24, 2020, that businesses, schools and other places in the United States might need to be locked down. Therefore, several topic evolution paths can be generated, such as Topic 4→Topic 2→Topic 5 and Topic 4→Topic 3→Topic 1.

Comparative experiment

At present, there is no standard criteria for evaluating topic evolution or topic intensity evolution. There is also lack of effective quantitative comparison methods for the topic evolution results, instead they are subjectively evaluated by human users on a case-by-case basis (Zhou et al, 2017). From the "Related work", the popular method of topic evolution is by calculating the similarity between topics in different time slices. We compare the performance of our method with the method of calculating cosine similarity between topics with 4 weeks per time slice (as Baseline). For each time slice, the topic coherence is used to determine the number of topics and LDA is applied to extract the topics. If the topic similarity is greater than a specific threshold, then it denotes that there is an evolutionary relationship between topics. Thus, the topic evolution path diagram is obtained, as shown in Fig. 16.

It can be seen in Fig. 16 that there are several common topics existing in different time slices. There is no doubt that the baseline method can identify the correlation between the common topics in adjacent time slices, except that between "3-Epidemic impact" and "4-Epidemic impact". It is because of the change in the semantics of the topic over time, and "3-Epidemic impact" refers to the impact of domestic epidemic, while "4-Epidemic

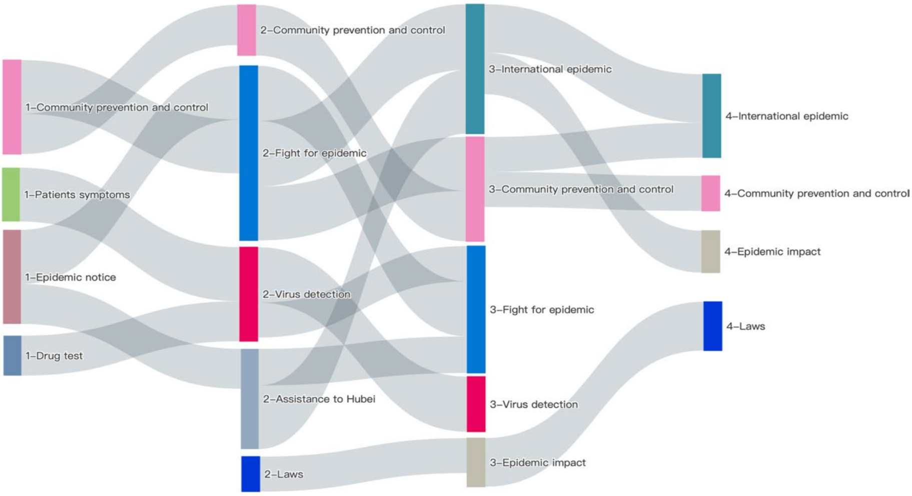


Fig. 16 Topic evolution using cosine similarity. (threshold = 0.38). Thicker edge represents greater topic similarity

impact" refers to the impact of global epidemic. Furthermore, the baseline method can also discover the evolution path between different topics. For example, "2-Fight the epidemic" drifts to "3-Community prevention and control", and "3-Community prevention and control" drifts to "4-International epidemic". These evolution paths are also detected by our approach, as shown in Fig. 14. Hence, the evolutionary path identified by the baseline method reflects the semantic similarity between topics, and it denotes that the topic is a continuation of another topic (Zhou et al., 2017). However, the baseline method is unable to discover this kind of evolutionary paths when the drifted topic is quite different from the original topic in terms of content. For example, as shown in Fig. 14, the path of "Community prevention and control" drifting to "Laws" is not shown in Fig. 16. Therefore, our approach performs better in capturing the drift among topics in the process of evolution.

Conclusion

Topics extracted from large-scale documents are macro and abstract, so it is hard to analyze and illustrate the topic evolution process accurately. Different from most previous studies which divide documents into different time slices and then extract topics in each slice, a method based on a 'word-topic' coupling network is proposed to analyze the process of topic evolution from a micro perspective, which can reflect topic evolution through the change of feature words. The method combines intensity, evolution status, and evolution path to explore topic evolution. First, when calculating the topic intensity, we consider not only the number of documents but also the contribution of feature words to the topic. Second, the concepts of weighted density index (WDI) and weighted centrality index (WCI) are proposed to analyze the status of a topic in its evolution process. Third, we emphasize that the topic drifting probability can effectively reveal the direction of the evolutionary path. Experiments conducted on real-world data sets demonstrate that our method could describe the evolution process of topic intensity, determine topic status and depict evolutionary path among topics.

In addition to describe the topic evolution, our method is potentially useful for topic forecasting. Hot topics in the future might be inferred with the help of topic drifting. Besides, in the case of a new topic emerging, feature words in the word network would be the first to change, so a new topic could be detected by monitoring the change of feature words. We will explore this direction in the future.

Acknowledgements This work is supported by the National Natural Science Foundation of China [grant numbers 71874088, 71704085]; the Cultivation Base of Excellent Innovation Team in Philosophy & Social Sciences in Jiangsu Universities [grant number 2017ZSTD022]; Postgraduate Research & Practice Innovation Program of Jiangsu Province [grant number KYCX20_0840].

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Blei, D., & Lafferty, J. (2006a). Correlated Topic Models. *Neural Information Processing Systems, 18*, 147.

Blei, D. M., & Lafferty, J. D. (2006b). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning. 113–120

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics, 22*(1), 155–205.

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Journal of Informetrics, 11*(4), 1175–1189.

Chen, J., Gong, Z., & Liu, W. (2019). A nonparametric model for online topic discovery with word embeddings. *Information Sciences, 504*, 32–47.

Chen, W., Lin, C., Li, J., & Yang, Z. (2018). Analysis of the evolutionary trend of technical topics in patents based on lda and hmm: Taking marine diesel engine technology as an example. *Journal of the China Society for Entific and Technical Information, 37*, 731–742.

Du, Y., Yi, Y., Li, X., Chen, X., Fan, Y., & Su, F. (2020). Extracting and tracking hot topics of micro-blogs based on improved latent dirichlet allocation. *Engineering Applications of Artificial Intelligence, 87*, 103279.

Fang, M., Chen, Y., Gao, P., Zhao, S., & Zheng, S. (2014). Topic trend prediction based on wavelet transformation. In 2014 11th Web Information System and Application Conference. 157–162. IEEE

Gao, W., Peng, M., Wang, H., Zhang, Y., Han, W., Hu, G., & Xie, Q. (2020). Generation of topic evolution graphs from short text streams. *Neurocomputing, 383*, 282–294.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1–2), 177–196.

Hurtado, J. L., Agarwal, A., & Zhu, X. (2016). Topic discovery and future trend forecasting for texts. *Journal of Big Data, 3*(1), 7.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE, 9*(6), e98679.

Jian, F., Yajiao, W., & Yuanyuan, D. (2018). Microblog topic evolution computing based on LDA algorithm. *Open Physics, 16*(1), 509–516.

Jung, S., & Yoon, W. C. (2020). An alternative topic model based on common interest authors for topic evolution analysis. *Journal of Informetrics, 14*(3), 101040.

Kim, S., Park, H., & Lee, J. (2020). Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: a study on blockchain technology trend analysis. *Expert Systems with Applications, 152*, 113401.

Liu, W., Deng, Z. H., Gong, X., Jiang, F., & Tsang, I. W. (2015). Effectively predicting whether and when a topic will become prevalent in a social network. In Proceedings of the National Conference on Artificial Intelligence

Liu, Z., Wang, X., & Bai, R. (2017). Construction and empirical research on multi-dimensional topic evolution analysis model. *Information Studies: Theory & Application, 3*, 18.

Lopez, C. E., & Gallemore, C. (2021). An augmented multilingual Twitter dataset for studying the COVID-19 infodemic. *Social Network Analysis and Mining, 11*(1), 1–14.

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Miao, Z., Du, J., Dong, F., Liu, Y., & Wang, X. (2020). Identifying technology evolution pathways using topic variation detection based on patent data: A case study of 3D printing. *Futures, 118*, 102530.

Song, Y., Li, A., & Quan, Y. (2018). Topics' popularity prediction based on ARMA model. In Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence. 68–72

Stein, B., & Zu Eissen, S. M. (2004). Topic identification: Framework and application. In Proceedings of the International Conference on Knowledge Management. 522–531

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 952–961

Wang, X., & McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 424–433

Wang, C., Blei, D., & Heckerman, D. (2008). *Continuous time dynamic topic models. In Uncertainty in Artificial Intelligence*. Helsinki.

Wartena, C., & Brussee, R. (2008). Topic detection by clustering keywords. In 2008 19th International Workshop on Database and Expert Systems Applications. 54–58. IEEE

Wei, L., Jiamin, W., & Jiming, H. (2020). Analyzing the topic distribution and evolution of foreign relations from parliamentary debates: A framework and case study. *Information Processing & Management, 57*(3), 102191.

Whye Teh, Y., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). Sharing clusters among related groups: Hierarchical Dirichlet processes. In NIPS'04 Proceedings of the 17th International Conference on Neural Information Processing Systems. 1385–1392

Wu, H., Yi, H., & Li, C. (2021). An integrated approach for detecting and quantifying the topic evolutions of patent technology: A case study on graphene field. *Scientometrics, 126*(8), 6301–6321.

Wu, Q., Zhang, C., Hong, Q., & Chen, L. (2014). Topic evolution based on LDA and HMM and its application in stem cell research. *Journal of Information Science, 40*(5), 611–620.

Xu, H., Winnink, J., Yue, Z., Liu, Z., & Yuan, G. (2020). Topic-linked innovation paths in science and technology. *Journal of Informetrics, 14*(2), 101014.

Zhang, Y., Mao, W., & Lin, J. (1991). Modeling topic evolution in social media short texts. In 2017 IEEE International Conference on Big Knowledge (ICBK). 315–319. IEEE

Zhao, J., Wu, W., Zhang, X., Qiang, Y., Liu, T., & Wu, L. (2014). A short-term trend prediction model of topic over Sina Weibo dataset. *Journal of Combinatorial Optimization, 28*(3), 613–625.

Zhou, H., Yu, H., & Hu, R. (2017). Topic evolution based on the probabilistic topic model: A review. *Frontiers of Computer Science, 11*(5), 786–802.

Zhu, J., Li, X., Peng, M., Huang, J., Qian, T., Huang, J., Liu, J., Hong, R., & Liu, P. (2015). *Coherent topic hierarchy: A strategy for topic evolutionary analysis on microblog feeds. International Conference on Web-Age Information Management*. Springer.