Neurocomputing 368 (2019) 11–24

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A text semantic topic discovery method based on the conditional co-occurrence degree

Wei Wei[a,b], Chonghui Guo[b,*]

[a] Center for Energy, Environment & Economy Research, Zhengzhou University, Zhengzhou 450001, PR China
[b] Institute of Systems Engineering, Dalian University of Technology, Dalian 116024, PR China

ARTICLE INFO

Article history:
Received 13 December 2018
Revised 17 August 2019
Accepted 19 August 2019
Available online 27 August 2019

Communicated by Dr Erik Cambria

Keywords:
Text mining
Topic discovery
Semantic information
Conditional co-occurrence degree

ABSTRACT

The topic discovery method, as an effective tool for semantic mining and a key means to extract new features from original text, plays an important role in the field of text mining and knowledge discovery. To solve the problems encountered in traditional topic models, such as the loss of semantic information and the ambiguity of topic concepts, as well as the crossover and coverage among topics, we propose a semantic topic discovery method based on the conditional co-occurrence degree (CCOD_STDM). First, every document is split into multiple subdocuments according to the semantic structure of the document and the independence decision rules. Second, combinatorial words with strong semantic relevance are extracted based on the conditional co-occurrence degree within the subdocuments. Based on these combinatorial words, new subdocuments are formed by feature expansion and content reconstruction. Third, "topic-word" distributions and "document-topic" distributions of new subdocuments are obtained by topic modeling with Gibbs sampling. Finally, "document-topic" distributions of the original documents are obtained by merging new subdocuments' "document-topic" distributions with specific strategies. The numerical experiments are compared with six topic models and two evaluation methods on seven kinds of public corpora, and the experimental results verify the superiority of CCOD_STDM and its efficiency in topic discovery. More importantly, a case study illustrates that the combinatorial words can effectively avoid the polysemy problem and can facilitate the condensation and summary of topics.

1. Introduction

The distillation of knowledge from unstructured data is an extremely challenging task in an era of social media and connectivity [1]. The topic model is a probabilistic statistical model used to discover abstract topics in textual data. It is used to model the topics implied in a corpus by adding a topic layer between the document layer and the word layer. Each document in the corpus is represented as a probability distribution of topics, and each topic is finally represented as a probability distribution of words. A topic model can identify potential topic information in large scale textual data or corpora. It overcomes the unclear relationship between text similarity and topics to some extent and can automatically find some semantic topics between words in massive textual data [2]. Latent Dirichlet allocation (LDA) [3] is one of the most classic topic models. It is derived from the extension of the probabilistic latent semantic indexing [4]. A variety of novel topic discovery methods have been subsequently developed based on the basic theory of the LDA model. Most of the newly proposed topic discovery methods still model the "document-topic" and "topic-word" distributions, which by their nature follow the same basic ideas of LDA. The extended topic discovery methods can be summarized into three categories according to the parameter expansion, introduction of context information and specific task orientation.

As an effective tool for semantic mining and feature extraction for the original text, a topic model plays an important role in document representation, document dimensionality reduction, text classification, text clustering and pattern discovery. As the scale of textual data grows, topic models are constantly evolving to accommodate new demands. Topic models have been widely used in different fields [5,6], such as natural language processing, biomedicine, health care, library and information science, image classification and other fields.

Traditional topic discovery methods are generally based on the Dirichlet process and vector space model (VSM) representation method. However, VSM assumes that the words are independent of each other and it does not consider the associations information between words. In contrast, traditional topic discovery methods

* Corresponding author.
E-mail address: dlutguo@dlut.edu.cn (C. Guo).

https://doi.org/10.1016/j.neucom.2019.08.047
0925-2312/

only consider the co-occurrence of words at a document level, and this relationship of co-occurrence has a negative impact on semantic topic discovery for long documents [7]. In general, a document will cover multiple different topics and describe each specific topic with some words in specific parts of the document in terms of content arrangement. Therefore, the content arrangement of the document with multiple topics has a certain structure and context. Without considering the document context structure and the semantic association among words in the above analysis, the final topics found by traditional topic discovery methods may encounter problems, such as the loss of semantics, concept ambiguity, semantic intersection, coverage and so on.

To combine the semantic-associated information between words and the local semantic structure information of documents together, this study proposes a semantic topic discovery method based on the conditional co-occurrence degree (CCOD_STDM) to model the latent topics in a corpus. In the modeling process, combinatorial words are constructed and used to abstract and summarize the topic of collective words, which can improve the efficiency of topic discovery and the accuracy of results. First, this study uses related theory on semantics to identify and divide the semantic structure of each document, models the correlations between words and generates combinatorial words with stronger correlation according to the modeling results. Then, these combinatorial words are incorporated into the subsequent topic modeling and regarded as new features to improve the topic modeling accuracy and topic semantic summary efficiency. Finally, the effectiveness of the proposed method is verified by combining text classification experiments and topic modeling experiments.

The major contributions of our paper can be summarized as follows:

(1) We propose a text semantic topic discovery method to identify precise topics for long documents.
(2) We adopt an effective method to generate combinatorial words. Depending on the combinatorial words, we design a document reconstruction method to highlight the semantic topic of a subdocument.
(3) We design a reasonable subdocument merging method to compute long documents' topic distributions based on the statistics of different semantic structures from the original long documents.
(4) We conduct extensive experiments on four widely used publicly available datasets. The experimental results show that our method can achieve better performance than other approaches. We also demonstrate the practicality and reliability of our method on the summarization of topics.

The remainder of the paper is organized as follows. Section 2 briefly summarizes the related topic models derived from LDA in three categories. Section 3 describes the proposed method CCOD_STDM in detail. Section 4 presents our experimental results, case study and discussion. Finally, some conclusions are provided in Section 5.

2. Related works

As a generative model, LDA can be easily extended to other models. Next, we will briefly describe the related topic models derived from LDA in three categories.

(1) Parameter expansion

LDA involves two sets of parameter sources, namely, the "document-topic" distribution and the "topic-word" distribution. The extension of the parameters can improve the applicability of topic models. LDA assumes that topics are independent of each other and that the "document-topic" distribution is subject to a polynomial distribution without considering the relevance between topics. Therefore, in the subsequent improvement works, Blei et al. [8] proposed a hierarchical LDA method that regards topics as the nodes in a tree with a hierarchical relationship among the nodes. In 2006, Teh et al. [9] responded to LDA's need to have the number of topics given in advance and proposed a hierarchical Dirichlet processes (HDP) method that can automatically determine the number of topics through previously clustering words. In the same year, Blei et al. [10] proposed the correlated topic model based on LDA by using a covariance matrix to describe the correlation between two topics. Given that correlated topic model is hindered by the issue of only considering the correlation between paired topics, Li et al. [11] proposed a Pachinko allocation model that represents the relationship among topics as a directed acyclic graph. In addition, considering the time spatial and temporal factors, Wang et al. [12] proposed the topic over time model and Blei et al. [13] proposed the dynamic topic model.

(2) Introduction of context information

LDA is carried out on the premise of the VSM representation of the document, which assumes that the different words in a document are independent. Therefore, the training results of LDA will not be affected by the order of words in a document. In subsequent studies, knowledge concepts and contextual information of the text have been gradually introduced by topic models based on a statistical method. In some related studies on syntactic structure, Griffiths et al. [14] proposed the HMM-LDA model by combining the hidden Markov model, which can capture the syntactic structure information of text, with LDA. Boyd-Graber et al. [15] proposed the syntactic topic model based on syntactic structure information. In related studies on expanding the semantic unit of words, Wallach [16] proposed the beyond bag-of-words, Wang et al. [17] proposed the topical n-gram model by introducing the word collocation information, and Gruber et al. [18] proposed the hidden topic Markov model by modeling topics in terms of sentences. Yan et al. [19] proposed a biterm topic model for short texts, the model uses the aggregated patterns in the whole corpus for learning topics to solve the problem of sparse word co-occurrence patterns at document-level. Pennington et al. [20] proposed a log-bilinear regression model, GloVe, and the global vectors in GloVe are trained on global word-word co-occurrence counts and thus makes efficient use of statistics from a corpus. Xiong et al. [21] proposed a word-pair sentiment-topic model for short text reviews, which is used to detect sentiments and topics simultaneously from short texts.

(3) Specific task oriented

Topic models have been widely used in many fields, and the extended topic models about task-oriented applications mainly concern the following three aspects, i.e., text classification, author topic models, sentiment analysis, as well as other fields. In text classification, Mcauliffe and Blei [22] proposed the supervised latent Dirichlet allocation. In sentiment analysis, Mei et al. [23] constructed the topic-sentiment mixture model by dividing the words into two categories, ordinary words and emotional words. In addition, Titov et al. [24] proposed a hybrid model based on text and feature evaluation. Regarding the author topic model, Steyvers et al. [25] proposed the author topics model, and Mccallum et al. [26] proposed the author recipient topic model based on the author topic model. In other application fields, the topic model has also been expanded accordingly. Doyle et al. [27] proposed the Dirichlet compound multinomial LDA model for detecting word bursts in documents. Ramage et al. [28] proposed the labeled LDA, which is a supervised topic model for credit attribution in multi-labeled corpora. Gerrish et al. [29] proposed the document influence model to identify the most influential documents in a

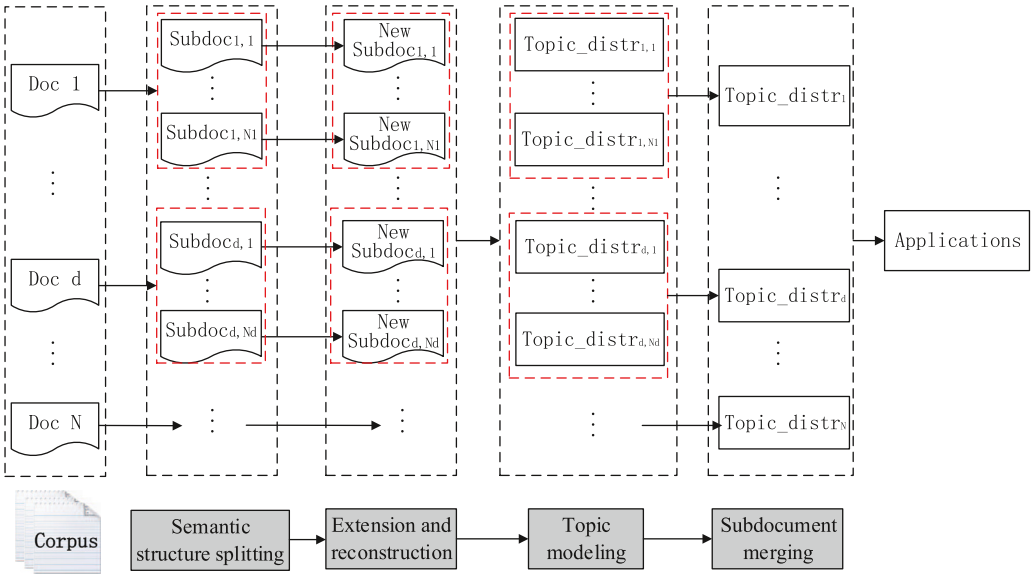

Fig. 1. The operation flow chart of the CCOD_STDM.

document collection or corpus. Yang et al. [30] proposed a topic drift model, which can monitor the dynamicity of authors' writing styles and learn authors' interests simultaneously. Motivated by the topical correspondence between text and mathematical equations observed in scientific documents, Yasunaga and Lafferty [31] proposed the TopicEq model, a joint topic-equation model that generates mathematical equations and their surrounding text.

Besides, the importance of the full neural network structure for learning semantic topics of texts has been called into question. In terms of learning vector representation of words or documents, Mikolov et al. [32] proposed the word2vec model, Dai et al. [33] established the doc2vec model and Bojanowski et al. [34] constructed the fastTest model. By leveraging the flexibility and representation power of neural networks, recent works [35,36] have proposed neural topic models, and some researchers [37,38] employ neural variational inference to train topic models. In particular, Mikolov and Zweig [39] incorporate the topic vector of a pre-trained LDA model into a recurrent neural network language model, Dieng et al. [40], Lau et al. [41] and Wang et al. [42] have trained neural topic and language models jointly.

Most of the above LDA based studies do not emphasize additional consideration of textual context and the semantic relationship between words, and the neural-network based topic models are trained with large amount of textual data. In addition, the topic modeling process often uses single words rather than more complete semantic words as the semantic unit. Furthermore, coupled with the semantics complexity of a long text structure, the above topic models often encounter problems, such as semantic loss, concept ambiguity, semantic intersection and coverage. In this paper, we mainly focus on the semantic topic modeling of long documents. In addition to textual statistics, we focus on the semantic structure of the document: the semantic association and conditional dependency among words in the process of topic modeling.

3. Model

The semantic topic discovery method based on conditional co-occurrence degree (CCOD_STDM) mainly contains four stages: (1) document semantic structure splitting, (2) feature expansion and subdocument reconstruction based on the conditional co-occurrence degree, (3) topic modeling, and (4) subdocument merging. The flow chart of the CCOD_STDM operations is shown in Fig. 1 and the relevant theoretical methods in each stage are described in detail next.

Table 1
Key symbols used in the CCOD_STDM model.

Symbol	Description
K	the number of topics
N	the number of documents in corpus D
N_d	the number of natural paragraphs in document d
S_k^d	the kth paragraph of document d
com_{ij}^s	the co-occurrence frequency of word i and word j in subdocument s
$ccodm_{ij}^s$	the conditional co-occurrence degree of word i to word j in subdocument s
V	the feature word set of new corpus $\tilde{D}$
M	the number of subdocuments in new corpus $\tilde{D}$
w_n^m	the nth word in subdocument m
z_n^m	the topic associated with the nth word in subdocument m
n_m	the word frequency in subdocument m
α	the Dirichlet priors to the multinomial distribution θ for $\tilde{D}$
β	the Dirichlet priors to the multinomial distribution φ
θ_m	the topic distribution of subdocument m
φ_k	the multinomial distribution of words specific to the topic k

For reading convenience, we summarize and explain the key symbols in the paper, as shown in Table 1.

3.1. Document semantic structure splitting

In the process of document semantic structure splitting, the granularity of the semantic structure is an important factor affecting the accuracy of subsequent experiments. The granularity of the semantic structure can be divided into four categories [43]: block, sentence, paragraph and document. Since the natural paragraph is a basic linguistic structure, it represents a coherent semantic text segment. In addition, the size of the natural paragraph is moderate, and it contains more vocabulary information than sentences and blocks. Document is usually treated as the basic semantic structure directly in real applications. However, the semantics of a document usually cover multiple topics or multiple sub-aspects of a topic. Once the semantic topic structure of a document can be accurately separated, the efficiency and precision of subsequent text processing tasks will increase accordingly.

According to the theory of semantics [44], words often appear along with other related words with specific semantic relationship. For example, “student” often accompanies “homework”, “learning”, “education”, “school” or other words in text. Language is a reflection of objective reality, if there are related objects in the objective world, the relationship between these objects will inevitably be reflected in the language [44]. Extending the above theory, the natural paragraphs in text are the reflection of human thinking and processes. The normal and abnormal, logical and illogical relationships that humans understand can be reflected in text.

The natural paragraph, as an organizational structure in text, can exert specific pragmatic and semantic functions. The natural paragraph also provides the means to understand the coherence of text structure. The distribution of natural paragraphs in a document is a reflection of the textual style, which has a unique pragmatic function. In addition, the content within a natural paragraph will describe one single topic intensively, which is different from the relatively scattered topics described in a long document. Furthermore, considering the paragraph as an independent semantic structure has achieved great success in related applications. Hearst et al. [45] have verified that the searching results obtained by using paragraphs are more accurate than the searching results obtained by the entire document. Later, Hearst [46] proposed an algorithm that uses changes in patterns of lexical repetition as the cue for the segmentation of single text into a contiguous, nonoverlapping and multi-paragraph subtopic structure. By constructing textual networks with consideration of the relationship between words and paragraphs, Landauer et al. [47] achieved good results with semantic information. Dai et al. [33] implemented the embedded representation of the document with the paragraph vector, and obtained high precision results in experiments. In this paper, the natural paragraph provides the granularity for splitting the semantic structure of the document, that is, each document is divided into several independent subdocuments according to the natural paragraph.

However, treating a paragraph with fewer words as an independent subdocument may increase the error in subsequent topic modeling. Therefore, in order to speed up the experimental processing, it is necessary to determine whether the natural paragraphs can be split into subdocuments with an independent semantic structure. In addition, each paragraph without an independent semantic structure will be merged into the next paragraph. The independence decision rules are defined as,

$$P(S_k^d) = \begin{cases} 1, & n(S_k^d) > [\xi] \\ 0, & n(S_k^d) \le [\xi] \end{cases}, \ d = 1, 2, \ldots, N \, ; k = 1, 2, \ldots, N_d \qquad (1)$$

where $n(S_k^d)$ is the number of words in S_k^d, $P(S_k^d)$ is the probability that the paragraph S_k^d is retained and split into an independent subdocument. ξ is an alterable threshold that is related to the square root of the average number of words in paragraphs in each document,

$$\xi = \sqrt{\sum_{k=1}^{N_d} n(S_k^d)/N_d} \qquad (2)$$

and $[\xi]$ represents the rounding operation for ξ.

It is worth noting that this strategy of processing short paragraphs may have a very small impact on the subsequent topic modeling results for some corpora, but to a certain extent, it can reduce the experimental complexity and speed up the experimental processing. In addition, the design here is also to facilitate the subsequent application of the method on other large corpora.

3.2. Feature expansion and subdocument reconstruction

After splitting the document’s semantic structure, each document is split into several subdocuments. According to semantics theory [44], the words in same subdocument mainly state a common topic jointly. The meaning of a single word is relatively extensive and ambiguous, so relying on only a word cannot accurately express the semantic content of the text. However, when the words appear together or words are combined into longer phrases, the semantic relationship of the text can be expressed more accurately. For example, when “apple” appears alone, it can be understood as a kind of fruit or the name of a mobile phone, so the concept of “apple” is ambiguous. When the words “screen” or “Steve Jobs” appear together with “apple”, the semantic topic of the word “apple” as an apple phone is readily determined. When the words “sweet” or “fresh” are combined with “apple”, we can quickly grasp the topic is about fruit. Therefore, this study will expand the features of subdocuments with the word combination method to determine the topic of subdocuments effectively.

To determine the document’s key semantic information effectively, using the word co-occurrence method alone may not achieve precise results. For example, if “apple” and “fresh” coexist, it is also hard to distinguish whether the semantic information is “apple is very fresh” or “fresh fruit contain apples”. In addition, if the co-occurrence frequency of “apple” and “sweet” is the same as the co-occurrence frequency of “apple” and “fresh” in a document, it is difficult to determine whether the semantic topic of the document focuses more on “apple sweet” or more on “apple fresh”. To determine the microscopic focus of text semantics, this study models the dependence of two words with the conditional co-occurrence degree method [48]. According to the conditional co-occurrence degree of a pair of words, new combinatorial words are generated and treated as the new features to expand and represent original subdocuments. It is worth emphasizing that the terminology combinatorial words refers to the combination of two words into one longer phrase, which is identified and constructed by considering the co-occurrence and dependencies between two words. The specific process is as follows.

The first step is calculating the conditional co-occurrence degree between word pairs in subdocument s. To be more practical, we have made some improvements on the original method. Taking $ccodm_{ij}^s$ as the conditional co-occurrence degree of word i to word j in subdocument s, and $ccodm_{ji}^s$ as the conditional co-occurrence degree of word j to word i, $ccodm_{ij}^s$ and $ccodm_{ji}^s$ are defined by

$$ccodm_{ij}^s = com_{ij}^s / com_{jj}^s, \ i \neq j \qquad (3)$$

$$ccodm_{ji}^s = com_{ij}^s / com_{ii}^s, \ i \neq j \qquad (4)$$

where $com_{ij}^s = \min\{com_{ii}^s, com_{jj}^s\}$, and com_{ii}^s is the frequency of word i in subdocument s.

Then, in order to highlight the semantic strength of these combinatorial words in the subdocument, we change the weight of these combinatorial words after obtaining the conditional co-occurrence degree between words. In one paragraph, the higher the word frequency is, the more important it is to some extent, and the new_ccodm obtained by the high word frequency is also relatively important. Therefore, the new weight of these combinatorial words $new\_ccodm_{ij}^s$ can be calculated by using the conditional co-occurrence degree of the original two words and multiplying by the difference between the sum of the two words’ frequency in the subdocument and λ,

$$new\_ccodm_{ij}^s = ccodm_{ij}^s \cdot \left(com_{ii}^s + com_{jj}^s - \lambda\right), \ i \neq j \qquad (5)$$

where λ is used to control the degree of prominence for these combinatorial words, such that the smaller the value is, the greater

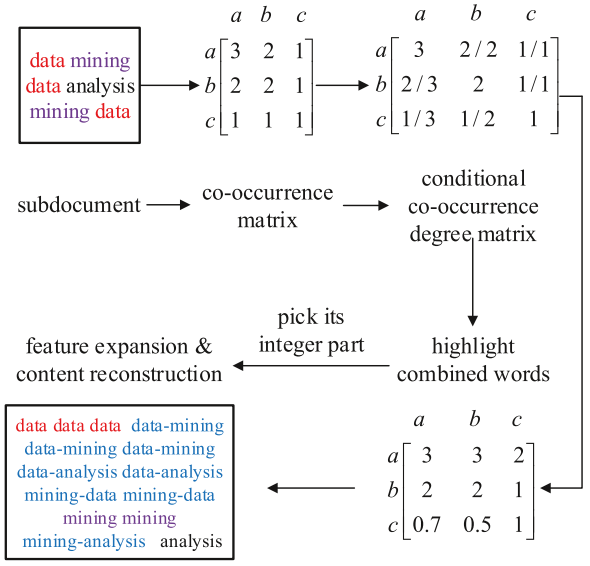

Fig. 2. An example diagram for feature expansion and subdocument reconstruction.

the degree of prominence, conversely, the larger the value is, the smaller the degree of prominence. In addition, in the subsequent experiments, we take λ as the median of the word frequency of all words in the corresponding paragraph.

Finally, according to these combinatorial words and their new weight, we obtain a new corpus consisting of new subdocuments by employing subdocument feature expansion and subdocument content reconstruction. In the process of reconstruction, these combinatorial words are repeated according to the integer part of their weights. The reconstructed subdocuments are composed of the original words and some new combinatorial words. At this time, the order of the words is no longer considered.

To describe the subdocument feature expansion and subdocument content reconstruction process clearly in this section, we consider an example to illustrate the detailed process, as shown in Fig 2. There are three different words in a natural paragraph of a document, namely "data", "mining" and "analysis". In addition, we take "a" as "data", "b" as "mining", "c" as "analysis" and $\lambda = 2$. First, the conditional co-occurrence degree matrix of the subdocument is calculated by the co-occurrence matrix. Then, these combinatorial words are highlighted according to formula (5). Finally, depending on the integer part of the combinatorial words' weight, the subdocument features are expanded and the subdocument content is reconstructed. From the content of the final reconstructed subdocument, it can be clearly ascertained that these combinatorial words, such as "data mining", "mining data" and "data analysis" are the key content. Moreover, it can be clearly distinguished that "data mining" is more important than "mining data", and "data analysis" is more important than "analysis data" in this example.

3.3. Topic modeling

After the above operations, topic modeling is performed on the new corpus to represent all subdocuments as topic distributions and all topics as word distributions. LDA [3], assumes that each document can be represented by a mixed distribution of multiple topics, and each topic can be represented by a polynomial distribution of all words in a corpus. Similarly, in this study, after the original document is separated by the semantic structure, it can also be assumed that the subdocument consists of multiple topics, and each topic is formed by a polynomial probability distribution of all words. However, it should be noted that the subdocument is originated from a paragraph of the original document, the multiple topics distribution of the subdocument is extremely asymmetrical, i.e., the value of a topic in the distribution will be much higher than other topics.

Suppose that corpus D consists of N documents initially. After the processes of splitting the semantic structure, feature expansion and subdocument reconstruction, a new corpus $\tilde{D}$ consists of M subdocuments and K latent topics in total, and let V be the feature word set of new corpus $\tilde{D}$ which is composed of the original words in D and these new combinatorial words. At the same time, $\theta_m (m = 1, 2, \ldots, M)$ is the topic distribution of subdocument m that obeys the Dirichlet distribution. φ_k is the multinomial distribution of words specific to the topic k, $k = 1, 2, \ldots, K$.

There are two alternating loop generation processes in the topic model for generating the subdocument m: (1) Select a θ_m randomly from $p(\theta|\alpha)$ that satisfies the Dirichlet distribution to generate the topic of the subdocument, and (2) generate every word w_n^m in subdocument m depending on $p(w_n^m|\theta_m, \varphi)$. For a subdocument m, containing a word sequence $\mathbf{w}^m = \left(w_1^m, w_2^m, \ldots, w_{n_m}^m\right)$ consisting of n_m words and a vector $\mathbf{z}^m$ consisting of K topics, whose generating probability is

$$p(\theta_m, \mathbf{z}^m, \mathbf{w}^m|\alpha, \beta) = p(\theta_m|\alpha) \prod_{n=1}^{n_m} p(z_n^m|\theta_m) p(w_n^m|z_n^m, \beta) \tag{6}$$

The generation process of the entire corpus can be described as follows.

(1) For each topic $k(k = 1, 2, \ldots, K)$, draw a multinomial φ_k from a Dirichlet prior β, $\varphi_k \sim Dir(\beta)$;
(2) For each subdocument $m(m = 1, 2, \ldots, M)$, draw a multinomial θ_m from a Dirichlet prior α, $\theta_m \sim Dir(\alpha)$;
(3) For each word $w_n^m (n = 1, 2, \ldots, n_m)$ in subdocument m,
  i) Draw a topic $z_n^m = k$ from $Multinomial(\theta_m)$;
  ii) Draw a word w_n^m from $Multinomial(\varphi_k)$.

Therefore, integrating the θ_m of formula (6) and adding the sum to topic z_n^m, the marginal distribution of the subdocument m is

$$p(\mathbf{w}^m|\alpha, \beta) = \int p(\theta_m|\alpha) \left(\prod_{n=1}^{n_m} \sum_{z_n^m} p(z_n^m|\theta_m) p(w_n^m|z_n^m, \beta) \right) d_{\theta_m} \tag{7}$$

The probability distribution of the entire corpus can be obtained by multiplying each subdocument's marginal distribution,

$$p(\mathbf{z}, \mathbf{w}|\alpha, \beta) = \prod_{m=1}^{M} \int p(\theta_m|\alpha) \left(\prod_{n=1}^{n_m} \sum_{z_n^m} p(z_n^m|\theta_m) p(w_n^m|z_n^m, \beta) \right) d_{\theta_m} \tag{8}$$

The topic model imitates the process of document generation to extract latent topics in the corpus. In the case of only observation variables such as feature words, the latent topic (also called hidden variable) of each word can be obtained by backward reasoning. There are two key variables in the model, i.e., polynomial distributions φ_k and θ_m, and the values of these two variables can be estimated by the variational method [3], Gibbs sampling [49] or expectation maximization [50]. In this study, we use the Gibbs sampling method.

After determining the joint probability distribution of the corpus, Gibbs sampling is used to obtain the samples, and the detailed processes of Gibbs sampling can be seen in LDA [3]. We only show the final Gibbs sampling formula.

$$p(z_i = k|\mathbf{z}_{\neg i}^m, \mathbf{w}^m) \propto \frac{n_{k,\neg i}^{(v)} + \beta_v}{\sum_{v=1}^{V} n_{k,\neg i}^{(v)} + \beta_v} \cdot \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k} \tag{9}$$

The topic discovery process of the new corpus is illustrated by a graph model, as shown in Fig. 3.

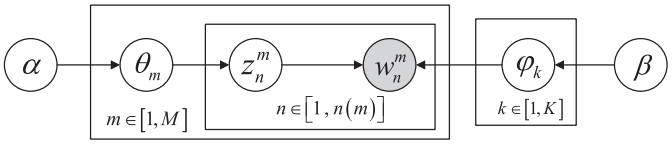

Fig. 3. Graphical model representation for the topic discovery process of the new corpus.

3.4. Subdocument merging

After topic modeling, each subdocument is represented by a probability distribution with several potential topics, and each potential topic is a probability distribution formed by all words that occur in the new corpus. Next, we focus on calculating the topic distributions of the original documents. In combination with traditional writing habits, if a large number of words in a document are used to state a topic, then the topic will be more important than another topic stated with a small number of words. Therefore, the topic distribution of the original document can be calculated by using the proportion of the total word frequency in the semantic structure to the total word frequency of the original document.

Suppose that each subdocument is represented as a topic distribution $\mathbf{z}_i = \left(z_i^1, z_i^2, \ldots, z_i^K\right)$ $(i = 1, 2, \ldots, N_d)$ with K latent topics, and z_i^k is the weight of the kth topic in distribution $\mathbf{z_i}$. The ratio of the total word frequency in the original subdocument obtained by splitting the semantic structure to the total word frequency of the original document is $a_1, \ldots, a_{N_d}$, where $a_i = n_{d_i}/n_d$, n_d is the total word frequency of document d, and n_{d_i} is the total word frequency of the ith subdocument in d. The topic distribution of document d can be calculated by

$$\mathbf{z}(d) = \sum_{i=1}^{N_d} a_i \cdot \mathbf{z}_i = \left(\sum_{i=1}^{N^d} a_i z_i^1, \sum_{i=1}^{N^d} a_i z_i^2, \ldots, \sum_{i=1}^{N^d} a_i z_i^K \right) \tag{10}$$

It is worth noting that subdocument merging is not required for the topic modeling task, but it is required to obtain the original documents document-topic distribution.

3.5. Summary of the algorithm

Integrating the theory of the above four stages, the implementation process of the semantic topic discovery method based on the conditional co-occurrence degree is shown in Algorithm 1. Assuming that each document has an average of $\bar{N}_d$ paragraphs and $\bar{l}$ words, and the number of iterations of Gibbs sampling is N_{iter}, then, the total time complexity of the Algorithm 1 can be calculated by $O(N\bar{N}_d) + O(N\bar{l}^2) + O(N_{iter}NK\bar{l}^2) + O(NK\bar{N}_d)$. Since $\bar{N}_d$ is far less than $\bar{l}$, so the time complexity of the Algorithm 1 is approximately equal to $O(N_{iter}NK\bar{l}^2)$.

Algorithm 1 Semantic Topic Discovery Method based on Conditional Co-Occurrence Degree (CCOD_STDM).

Input: Corpus D formed by N documents, the number of latent topics K, hyperparameters α and β of Dirichlet distribution.
Output: $\mathbf{\Theta}_{N\times K}$ topic distributions of the documents, $\mathbf{\Phi}_{K\times V}$ distributions of the K topics
The Main Procedure:
for $d = 1 : N$ **do**
  Split d into several independent subdocuments by formula (1) and (2).
  for $dd = 1 : N_d$ **do**
    Calculate the conditional co-occurrence degree between every two words by formula (3) and (4), highlight the semantic information between every two words by formula (5), reconstruct new subdocuments depending on the rounding weight of the semantic information.
  end for
end for
repeat
  for $k = 1 : K$ **do**
    Draw a multinomial φ_k from a Dirichlet prior β, $\varphi_k \sim Dir(\beta)$;
  end for
  for $m = 1 : M$ **do**
    Draw a multinomial θ_m from a Dirichlet prior α, $\theta_m \sim Dir(\alpha)$;
    for $n = 1 : n_m$ **do**
      Draw a topic $z_n^m = k$ from $Multinomial(\theta_m)$;
      Draw a word w_n^m from $Multinomial(\varphi_k)$.
    end for
  end for
until Gibbs sampling converges.
Calculate $\mathbf{\Phi}_{K\times V}$ by the "topic-word" co-occurrence frequency matrix obtained by counting each word and its corresponding topic in the new corpus, calculate the topic distribution of each subdocument by counting the topic frequency in the subdocument, calculate $\mathbf{\Theta}_{M\times K}$ by aggregating the topic distribution of these M subdocuments.
for $d = 1 : N$ **do**
  Calculate $\mathbf{\Theta}_{N\times K}$ by formula (10).
end for

4. Experiments

In this section, we first introduce the experimental corpora, the setup of the comparative experiments and the evaluation metrics applied to evaluate the semantic topic discovery method. Next, we report the experimental results of the seven topic discovery methods based on seven different corpora. In addition, we further discuss the practical application effect of CCOD_STDM in topic discovery with a case study at the end of the experiments.

4.1. Experimental preparation

To evaluate the effectiveness of our semantic topic discovery method comprehensively, we applied our method on four Chinese corpora and three English corpora. The Chinese corpora are collected and sorted by Fudan University and Sougou Lab. The Fudan corpus contains 20 categories and nearly 20,000 documents in total. We randomly select three categories and 1040 documents from each category in this corpus as the large corpus I, and select three categories and 310 documents from each category in this corpus as the small corpus II. The Sougou corpus contains nine categories and 17,910 documents in total, and from this corpus, for an additional comparison, we randomly chose four categories and 1040 documents from each category as another large corpus III, and randomly chose four categories and 250 documents from each category as another small corpus IV. Besides, the English corpora are large movie review dataset and Reuters-21578. We chose two standard versions from large movie review dataset, one is randomly select 1000 reviews from each polarity of Large Movie Review Dataset v1.0 [51] (Abbreviated as: LMRD v1.0); the other is the Polarity Dataset v2.0 [52]. For Reuters-21578, we select four categories with a large number of documents, i.e., acq (932), crude (473), earn (990) and grain (466). The detailed experimental data are described in Table 2.

In the data preprocessing stage, only stopword removal is conducted for English corpora. For Chinese corpora, we use the

Table 2
Description for classification experimental corpora.

Source	Corpus number	Number of documents	Number of categories
Fudan University	I	3120	3
	II	930	3
Sougou Lab	III	4160	4
	IV	1000	4
LMRD v1.0	V	2000	2
Polarity Dataset v2.0	VI	2000	2
Reuters-21578	VII	2861	4

famous Chinese word segmentation system, ICTCLAS [53], to process text corpus. After implementing word segmentation and POS tagging, stopword removal and POS filtering are employed as two essential and preliminary effective approaches to reduce the feature dimension. In this paper, we remove the terms that occur frequently but have less meaning based on an extended Chinese stopword list, which was built by the Institute of Computing Technology, Chinese Academy of Science. Then, noun terms that can well reflect the semantics of the document to some extent are chosen [7], as the candidate features in the whole experiment.

To evaluate the new model, CCOD_STDM was compared with two kinds of models. One is the classical topic model, which includes the LDA [3] and hierarchical Dirichlet process (HDP) [9]. The other is the famous and state-of-the-art word vector representation method, which includes the word2vec [32], the fastText model [34] and the GloVe model [20]. All these methods are implemented by calling the open source algorithm packages provided by corresponding authors in Python. HDP is a multi-layered mixture model that can automatically determine the appropriate number of mixture components needed, and since components are shared across groups, statistical strength across groups is also shared. Word2vec is trained by multi-layer neural network technology on a large amount of textual data [54], and each word in the corpus is represented as a finite dimensional vector. For FastText model, skipgram model is used, and each word is represented as the sum of these representations of a bag of character n-grams. GloVe model is a new global log-bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. For the three word vector models, each dimension of the vector can be regarded as a specific potential topic for a word, and the topic distribution of a document can be obtained by the average weighted sum of the entire amount of word vectors in the document. Besides, to evaluate the effectiveness of the feature expansion method for the combinatorial word with the conditional co-occurrence degree method (CCOD), the mutual information (MI) method is used for feature expansion instead of CCOD. We define this contradistinctive method as MI_STDM, which is the same as CCOD_STDM except for the feature expansion method. The calculation for the mutual information $MI(i, j)$ between word i and word j in subdocument s is defined by

$$MI(i,j) = p(i,j) \cdot \log \frac{p(i,j)}{p(i)p(j)} \tag{11}$$

where $p(i, j)$ is the co-occurrence probability of word i and word j in subdocument s, and $p(i)$ is the occurrence probability of word i in subdocument s.

This study first uses different topic discovery methods to represent the "document-topic" distribution of each document in the corpus, then performs the classification experiment on the documents in each corpus, and finally illustrates the superiority of different topic discovery methods by evaluating the classification results. For the evaluation of topic models, it is difficult to perform an intrinsic evaluation as there is inherently no 'correct' answer.

Table 3
Contingency table with regard to binary classification.

		Standard Judgment	
		Related	Unrelated
Model Prediction	Related	A_i	B_i
	Unrelated	C_i	D_i

However, from the view of the "document-topic" distribution, a good performance topic model can effectively represent documents with similar semantic themes as vector forms with similar "document-topic" distributions. In addition, document categorization can effectively aggregate documents with similar topics according to the document-topic distributions. Furthermore, in the study [3], the authors conducted a set of comparison experiments on document classification with the topic-based document representation method and the VSM-based document representation method, and they suggested that the topic-based representation provided by LDA may be useful as a fast filtering algorithm for feature selection in text classification. Therefore, the document classification method can indirectly evaluate the effect of the topic models [55].

Of the classification approaches, random forest [56] is considered one of the most robust and accurate methods among all well-known algorithms. Random forest has a wide range of applications because of its good stability and generalization ability [56]. In this study, to classify the experimental data, we utilize 200 trees and use five-fold cross-validation to obtain the classification accuracy.

To evaluate the effectiveness and reliability of classification results, we use the classical precision, recall and $F1$ measure for the testing corpus [57]. Precision is defined as the ratio of correct assignments by the system divided by the total number of the system's assignments. Recall is the ratio of correct assignments by the system divided by the total number of correct assignments. The overall $F1$ measure score of the multi-classification result can be computed in two manners, namely, macro-average and micro-average. In this paper, we chose the two ways as evaluation methods. In the macro-average manner, the precision and recall for each category are calculated first with the evaluation measures of the binary classification. After calculating the precision and recall for each category, the $F1$ measure $F1_i$ is calculated by the weighted harmonic mean of precision P_i and recall R_i.

$$F1_i = 2 \times P_i \times R_i/(P_i + R_i), P_i = A_i/(A_i + B_i), R_i = A_i/(A_i + C_i) \tag{12}$$

where A_i, B_i, C_i and D_i are the number of the corresponding knowledge points, as defined in Table 3.

We assume that there are C object classes in the experimental corpus. Therefore, the overall macro-averaged precision, recall and the $F1$ measure for the testing corpus, denoted as $Macro\_P$, $Macro\_R$ and $Macro\_F1$, respectively, can be calculated as follows:

$$Macro\_P = \frac{1}{C}\sum_{i=1}^{C} P_i \tag{13}$$

$$Macro\_R = \frac{1}{C}\sum_{i=1}^{C} R_i \tag{14}$$

$$Macro\_F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i \tag{15}$$

Additionally, the micro-averaged of precision ($Micro\_P$), recall ($Micro\_R$) and the $F1$ measure ($Micro\_F1$) for the testing corpus can be calculated as follows:

$$Micro\_P = \frac{\sum_{i=1}^{C} A_i}{\sum_{i=1}^{C} (A_i + B_i)} \tag{16}$$

Table 4
The classification evaluation results of seven methods on the large-scale Fudan corpus I under different numbers of topics.

	Model	Number of topics											
		5	10	15	20	50	70	100	120	150	170	200	250
Macro_F1 (%)	LDA	79.17	92.36	91.15	91.01	89.76	91.74	92.01	90.90	89.78	90.82	89.71	89.76
	word2vec	68.44	76.73	80.62	84.53	85.08	86.84	86.67	86.80	87.22	86.53	86.75	87.45
	HDP	54.90	62.40	64.67	64.87	78.12	75.49	78.55	84.32	77.95	85.07	87.79	84.13
	MI_STDM	80.21	84.01	88.40	86.33	90.62	91.68	89.50	89.31	86.51	88.82	87.03	86.54
	GloVe	77.52	90.17	93.00	90.27	92.48	93.84	93.70	94.42	93.92	94.21	93.98	93.79
	fastText	87.53	90.18	91.37	92.76	93.21	92.17	93.70	93.77	93.93	93.94	93.48	93.88
	CCOD_STDM	81.26	88.88	93.46	93.29	93.23	93.01	93.01	92.01	90.55	92.10	92.18	92.02
Micro_F1 (%)	LDA	79.28	92.42	91.05	89.85	91.85	92.07	93.03	90.98	89.88	90.92	89.78	89.88
	word2vec	69.27	77.43	81.05	84.63	85.18	86.90	86.75	86.85	87.28	86.63	86.82	87.52
	HDP	56.45	65.67	64.47	66.33	77.98	74.95	78.42	84.57	78.13	85.37	87.77	83.8
	MI_STDM	80.60	84.25	88.70	86.47	90.78	91.78	89.72	89.50	86.73	89.02	87.22	86.88
	GloVe	77.52	90.22	93.00	90.33	92.50	93.85	93.72	94.43	93.93	94.22	94.00	93.8
	fastText	87.58	90.20	91.40	92.78	93.22	92.18	93.72	93.78	93.93	93.95	93.50	93.88
	CCOD_STDM	81.52	88.93	93.50	93.32	93.27	93.03	93.03	92.12	90.7	92.17	92.28	92.10

$$Micro\_R = \frac{\sum_{i=1}^{C} A_i}{\sum_{i=1}^{C} (A_i + C_i)} \tag{17}$$

$$Micro\_F1 = \frac{2 \cdot Micro\_P \cdot Micro\_R}{Micro\_P + Micro\_R} \tag{18}$$

To improve the general significance and reduce the error of each experiment, the experiments in each group are repeated 10 times and the final macro-average and micro-average are calculated by the average of the 10 experiments in each group.

4.2. Experimental results and analysis

The text classification evaluation results of the seven different methods for the large-scale Fudan corpus (corpus I) are shown in Table 4. On the whole, CCOD_STDM has obvious advantages in text classification tasks. Specifically, when the number of topics ranges from 15 to 50, CCOD_STDM has an absolute advantage over the other six methods. When the number of topics is 15, the *Macro_F*1 and *Micro_F*1 of the CCOD_STDM reaches 93.46% and 93.5%, respectively, which is approximately 2%–29% higher than the first four methods under the same conditions. As the number of topics increases, the results of the entire set of classification methods gradually improve and stabilize. However, CCOD_STDM still has a small advantage. While for Glove and fastText, when the topic number is larger than 100, they have a stable and outstanding performance than other methods, which indicates that these two word vector models have certain advantages when represented words by higher-dimensional vectors. In addition, the classification results of HDP method are relatively inferior when the number of topics is small. However, as the number of topics increases, the performance of HDP improves. MI_STDM's performance is not as good as that of CCOD_STDM, indicating that it is better to perform feature expansion through CCOD than through MI.

The results on the large-scale Sougou corpus (corpus III) are shown in Table 5. In the case with a relatively low number of topics, CCOD_STDM has obvious advantages compared with other methods, especially for these traditional topic models. However, fastText performs its strong robustness for each topic numbers on this large corpus, which indicates that it has an absolute hegemonic position in the field of word representation, which also indicates that the morphology of words has a positive influence on word representation compared with word2vec. However, interpretability of the results is still the shortcoming of these neural-network based methods. Except for some performance of the GloVe method is not better than or similar to CCOD_STDM, some of the performance is higher than CCOD_CCOD, which indicates that word co-occurrence plays a momentous role in word or text representation. In addition, as the number of topics increases, the classification effects of LDA, word2vec and CCOD_STDM gradually become stable and similar. For *Macro_F*1 (*Micro_F*1), the variances of the three methods are significantly similar as the *p*-value of the variance similarity test for LDA and CCOD_STDM is 0.1879 (0.1405), and for word2vec and CCOD_STDM is 0.6964 (0.9147), which are all larger than 0.05. Whats more, the p-value of the *t*-test for LDA and CCOD_STDM is 0.238 (0.2034), and for word2vec and CCOD_STDM is 0.0776 (0.1013), which are also larger than 0.05. It shows that the average *Macro_F*1 and *Micro_F*1 of the three methods are significantly similar in the 95% confidence interval. However, in most cases, CCOD_STDM still has subtle advantages over the two methods in terms of specific averages. MI_STDM achieves a relatively better performance when the topic number ranges from 10 to 50. However, when the topic number is larger than 70, it has a rapid decline. Among these topic discovery models, HDP has a performance similar to corpus I, which may be directly related to the small number of subject categories in the original data.

The experimental evaluation results of seven topic discovery methods for the small-scale Fudan corpus (corpus II) are shown in Table 6. CCOD_STDM performs the best among the all methods with different evaluation metrics. As the number of topics changes, CCOD_STDM shows strong stability compared with other methods, with *Macro_F*1 and *Micro_F*1 fluctuating between 78% and 85%. The word2vec method based on the multi-layer neural network also has high stability, and its *Macro_F*1 and *Micro_F*1 fluctuate at approximately 75%. Due to the small-scale corpus, the performance of the GloVe method and fastText method are not good, especially for fastText, the average values in the whole topics numbers are much lower than CCOD_STDM. For LDA, it has good performance when the number of topics is relatively small, but as the number of topics increases, the experimental evaluation results gradually decrease. LDA's poor performance is closely related to the size of the corpus. When the topic number is large, the topic distributions of the corpus are relatively sparse, which leads to a decrease in classification accuracy. The process of CCOD_STDM, however, expands subdocument features and reconstructs subdocuments to highlight the semantic words prior to representing documents, so CCOD_STDM still has better stability even with a large number of topics in a small-scale corpus. However, the MI-based feature expansion method cannot achieve the same performance with the CCOM approach. In addition, the experimental evaluation results of HDP in the small-scale corpus gradually increases as the number of topics increases, which is a characteristic closely related to the principles of HDP. In particular, HDP first clusters words in the corpus and then constructs the topic distribution for documents according to the clustering results. However, the clustering results

Table 5
The classification evaluation results of seven methods on the large-scale Sougou corpus III under different numbers of topics .

	Model	Number of topics											
		5	10	15	20	50	70	100	120	150	170	200	250
Macro_F1 (%)	LDA	54.80	74.53	77.03	77.29	83.32	85.71	83.32	82.08	81.70	81.07	81.08	80.75
	word2vec	64.36	73.50	80.04	80.72	80.22	79.31	79.62	79.95	79.91	81.31	79.21	79.37
	HDP	44.23	39.01	36.04	52.37	56.70	54.64	55.53	59.91	66.78	71.00	72.95	69.51
	MI_STDM	46.63	73.43	75.16	75.68	77.24	71.20	66.38	61.78	62.12	61.21	61.39	58.88
	GloVe	78.08	84.07	86.67	86.04	85.26	85.71	86.56	86.14	85.19	86.22	85.90	85.64
	fastText	81.04	84.96	87.16	87.99	88.71	89.12	88.55	88.39	88.79	88.76	88.61	88.52
	CCOD_STDM	67.03	77.19	81.00	84.69	86.15	85.92	84.20	85.58	85.09	82.84	82.38	81.41
Micro_F1 (%)	LDA	54.84	74.65	76.95	76.92	83.42	85.75	83.14	81.91	81.51	80.92	80.84	80.69
	word2vec	64.49	73.45	80.45	81.29	80.75	79.94	80.20	80.44	80.45	81.78	79.82	79.90
	HDP	45.74	43.95	39.85	52.44	56.04	55.99	57.58	61.86	67.34	70.91	73.09	69.76
	MI_STDM	47.61	73.38	75.05	75.66	77.22	71.51	67.39	64.11	63.84	63.24	63.45	61.78
	glove	77.88	84.01	86.66	86.00	85.22	85.68	86.52	86.12	85.15	86.20	85.88	85.60
	fasttext	80.92	84.91	87.06	87.95	88.64	89.08	88.48	88.34	88.72	88.71	88.54	88.44
	CCOD_STDM	67.99	77.74	80.91	84.74	86.15	85.88	84.10	85.60	85.10	82.88	82.26	81.45

Table 6
The classification evaluation results of seven methods on the small-scale Fudan corpus II under different numbers of topics.

	Model	Number of topics											
		5	10	15	20	50	70	100	120	150	170	200	250
Macro_F1 (%)	LDA	78.12	80.98	82.39	77.68	80.38	72.93	73.23	74.44	74.64	75.87	70.40	73.55
	word2vec	69.83	78.99	74.72	76.86	75.00	77.10	76.77	75.66	74.73	75.30	73.45	72.77
	HDP	46.15	42.46	57.78	52.39	66.16	63.54	63.94	63.32	63.90	66.55	74.27	74.61
	MI_STDM	55.81	69.69	72.72	75.39	71.95	73.75	70.80	68.68	70.63	63.37	68.86	65.32
	GloVe	66.14	70.90	74.94	75.04	73.81	75.60	75.30	73.68	74.60	73.81	72.04	73.46
	fastText	66.97	66.00	67.27	65.69	71.67	71.65	70.57	71.12	72.72	75.03	72.41	73.93
	CCOD_STDM	79.22	82.39	84.28	79.31	80.98	83.50	80.80	83.87	77.78	80.92	82.27	78.59
Micro_F1 (%)	LDA	77.78	81.11	82.22	76.67	80.00	72.22	72.78	73.89	73.33	74.44	68.33	72.22
	word2vec	68.89	78.33	74.44	76.67	73.89	76.11	75.56	73.89	73.33	74.44	72.22	72.22
	HDP	46.11	41.67	55.56	52.56	64.44	61.11	61.67	60.00	61.67	62.78	74.44	72.78
	MI_STDM	56.39	69.83	72.72	75.28	71.94	73.44	70.78	68.22	70.50	63.44	68.78	65.39
	glove	67.39	72.39	75.72	75.56	73.89	76.22	75.44	74.28	74.78	74.33	72.33	73.78
	fasttext	69.06	68.33	70.06	67.83	73.22	73.22	72.00	72.72	74.28	76.06	73.78	75.06
	CCOD_STDM	79.44	82.56	83.89	79.44	81.11	83.33	80.56	83.89	77.78	80.56	82.22	78.33

Table 7
The classification evaluation results of seven methods on the small-scale Sougou corpus IV under different numbers of topics.

	Model	Number of topics											
		5	10	15	20	50	70	100	120	150	170	200	250
Macro_F1 (%)	LDA	63.17	66.66	64.62	68.84	68.03	69.69	68.82	71.77	69.72	66.91	65.16	63.53
	word2vec	60.09	66.37	70.63	74.94	76.38	76.38	78.58	75.21	77.03	76.80	76.21	77.83
	HDP	48.33	46.77	49.06	52.91	56.83	60.42	59.07	53.58	60.30	60.92	58.75	66.84
	MI_STDM	64.43	67.51	71.67	73.29	75.91	72.74	77.44	72.10	71.82	69.72	70.26	72.71
	glove	52.10	68.36	71.84	70.88	75.92	79.75	78.28	75.86	77.33	76.07	76.49	76.85
	fasttext	74.62	77.56	79.94	78.46	80.30	80.75	80.04	80.36	80.75	80.76	80.36	79.39
	CCOD_STDM	66.00	68.81	71.00	76.93	76.48	81.02	80.40	79.13	78.06	76.81	76.68	77.04
Micro_F1 (%)	LDA	63.00	66.25	64.50	69.00	67.75	69.75	68.50	71.25	69.50	66.50	64.75	63.25
	word2vec	59.50	65.50	70.50	74.50	76.00	76.00	78.50	75.00	77.00	76.50	75.50	77.50
	HDP	48.50	47.00	48.00	53.00	57.00	60.00	58.00	52.50	60.00	60.50	59.00	67.00
	MI_STDM	65.70	67.43	71.00	74.40	76.35	73.15	77.60	72.65	72.25	70.35	70.80	73.25
	glove	52.50	68.75	72.00	70.90	75.90	79.05	78.28	76.00	77.40	76.13	76.60	76.90
	fasttext	74.80	77.50	79.95	78.70	80.40	80.90	80.10	80.45	80.80	80.75	80.45	79.45
	CCOD_STDM	66.00	68.50	71.00	77.00	76.00	81.00	80.35	79.00	78.00	76.50	76.50	77.00

of words are poor due to the sparse and scattered distribution of words in the small-scale corpus, which degrades HDP's performance when the number of topics is relatively small. When the number of topics increases to a certain extent and meets the number of potential clusters of words, such as when the number of topics ranges around 250, HDP achieves better experimental results.

The experimental evaluation results of seven topic discovery methods for the small-scale Sougou corpus (corpus IV) are shown in Table 7. Compared with LDA and HDP, CCOD_STDM has obvious advantages and its average *Macro_F*1 and *Micro_F*1 are approximately 9% and 17% higher than LDA and HDP, respectively. Moreover, the average *Macro_F*1 and *Micro_F*1 increased by approximately 2% compared with word2vec and GloVe on the whole. In detail, when the number of topics ranges from 5 to 150, CCOD_STDM has certain advantages over word2vec and GloVe, when the number of topics is between 170 and 250, the three methods perform equally well. On the small-scale corpora, the overall performance of the LDA method is not good, with an average *Macro_F*1 and *Micro_F*1 of approximately 66%. In addition, when the number of topics reaches a certain level, the experimental classification accuracy of LDA begins to gradually decrease. These phenomena reflect the instability characteristics of LDA on the small-scale corpus. MI_STDM has a similar performance to

Table 8
The classification evaluation results of seven methods on the LMRD v1.0 corpus V under different numbers of topics .

	Model	Number of topics											
		5	10	15	20	50	70	100	120	150	170	200	250
Macro_F1 (%)	LDA	50.51	50.06	48.76	54.62	49.92	63.78	56.64	55.69	53.72	52.60	51.73	51.79
	word2vec	52.04	50.05	50.93	50.54	51.02	52.67	51.17	49.40	51.30	48.96	50.53	48.68
	HDP	53.84	53.12	50.19	51.65	51.66	48.19	43.15	55.47	51.99	41.09	41.60	43.22
	MI_STDM	51.30	53.48	52.04	46.84	52.41	52.33	50.46	52.86	49.14	47.92	48.32	47.62
	GloVe	55.28	53.58	51.06	53.53	50.56	49.20	49.17	51.01	52.01	50.06	51.16	52.09
	fastText	55.12	57.59	59.06	54.22	58.76	58.12	59.38	60.55	59.94	58.03	59.37	59.91
	CCOD_STDM	51.52	53.23	58.45	57.25	60.89	64.70	61.38	57.74	57.33	57.72	55.05	53.99
Micro_F1 (%)	LDA	50.52	50.08	48.98	54.65	50.08	63.80	56.88	55.75	53.75	52.65	51.85	52.05
	word2vec	52.05	50.08	51.00	50.55	51.05	52.70	51.18	49.42	51.32	48.98	50.58	48.78
	HDP	53.87	54.15	52.25	52.60	52.95	53.00	52.12	59.98	59.00	51.82	52.98	53.32
	MI_STDM	51.32	53.50	52.05	46.85	52.60	52.35	50.60	52.88	49.15	47.95	48.35	47.65
	GloVe	55.35	53.70	51.15	53.82	50.60	49.48	49.55	51.35	52.22	50.32	51.32	52.32
	fastText	55.40	57.65	59.33	54.55	58.90	58.38	59.70	60.65	60.05	58.20	59.60	60.08
	CCOD_STDM	51.55	53.32	58.48	57.42	60.92	64.72	61.42	57.80	57.38	58.20	55.18	54.00

Table 9
The classification evaluation results of seven methods on the Polarity Dataset v2.0 corpus VI under different numbers of topics.

	Model	Number of topics											
		5	10	15	20	50	70	100	120	150	170	200	250
Macro_F1 (%)	LDA	51.43	56.22	54.61	56.00	59.40	58.65	58.94	55.05	50.96	53.26	53.25	50.74
	word2vec	50.06	51.58	50.22	50.81	51.45	50.36	49.21	50.60	49.35	50.61	49.87	51.35
	HDP	50.60	57.09	45.18	37.88	38.80	41.16	39.75	38.19	40.16	39.14	37.49	43.91
	MI_STDM	50.89	53.92	54.10	52.57	55.40	54.67	56.07	50.44	50.00	50.25	51.31	50.35
	GloVe	54.88	55.71	55.99	57.81	56.44	59.42	57.81	58.72	56.95	57.40	57.21	55.47
	fastText	54.94	60.93	62.08	63.50	62.01	62.02	62.50	62.13	60.86	62.77	62.22	62.24
	CCOD_STDM	53.43	58.54	59.12	59.93	63.30	61.10	61.92	58.99	56.66	57.74	55.99	50.70
Micro_F1 (%)	LDA	51.48	56.25	54.65	56.02	59.50	58.88	58.95	55.08	50.98	53.27	53.35	50.80
	word2vec	50.52	51.88	50.50	51.30	51.75	50.48	49.28	50.62	49.42	50.62	49.88	51.35
	HDP	50.98	59.45	52.15	48.40	52.08	52.25	52.08	51.90	53.20	52.18	51.75	54.00
	MI_STDM	50.90	53.92	54.12	52.65	55.42	54.68	56.08	50.48	50.12	50.25	51.32	50.38
	GloVe	54.98	55.75	56.08	57.85	56.50	59.45	57.82	58.73	56.98	57.45	57.25	55.50
	fastText	55.10	61.05	62.18	63.60	62.12	62.10	62.52	62.15	60.92	62.80	62.25	62.28
	CCOD_STDM	53.45	58.60	59.13	59.95	63.38	61.28	62.05	59.03	56.68	57.75	56.02	50.85

Table 10
The classification evaluation results of seven methods on the Reuters-21578 corpus VII under different numbers of topics.

	Model	Number of topics											
		5	10	15	20	50	70	100	120	150	170	200	250
Macro_F1 (%)	LDA	52.39	57.42	73.75	74.48	82.78	86.20	87.38	86.51	82.59	82.38	78.03	77.44
	word2vec	56.40	62.58	61.38	64.51	65.47	66.21	65.01	65.12	63.27	62.89	63.66	63.45
	HDP	42.93	45.42	44.76	53.75	56.39	63.91	59.06	61.18	62.87	46.48	65.87	47.80
	MI_STDM	53.72	55.76	65.35	71.84	70.03	62.56	53.84	57.45	51.72	53.50	53.47	51.40
	GloVe	67.66	78.45	80.05	74.08	85.55	83.95	84.61	83.81	83.03	86.47	85.88	84.42
	fastText	88.44	89.01	90.89	91.41	92.17	91.17	90.88	91.06	91.57	91.42	91.07	91.06
	CCOD_STDM	72.43	79.49	86.53	88.46	92.62	90.97	90.59	87.61	86.52	82.50	85.46	81.86
Micro_F1 (%)	LDA	56.67	60.98	74.77	75.01	83.60	86.67	87.14	86.70	82.42	82.39	77.56	76.72
	word2vec	60.02	65.66	65.13	67.94	68.31	68.56	67.43	67.40	66.23	66.08	65.99	65.17
	HDP	50.88	52.44	51.50	58.38	60.47	68.14	62.78	63.24	66.14	53.52	67.55	52.34
	MI_STDM	57.19	61.16	68.03	74.52	73.99	68.37	58.71	61.94	56.40	58.14	59.06	56.28
	GloVe	70.21	80.23	82.34	76.85	86.73	85.14	85.58	85.28	84.60	87.08	87.03	85.93
	fastText	89.33	90.44	91.30	92.18	92.29	91.83	91.48	91.66	92.22	91.86	91.75	91.65
	CCOD_STDM	72.21	80.15	86.48	88.03	92.36	91.12	90.89	88.40	87.66	83.92	86.76	82.78

CCOD_STDM when the topic number is less than 50, but as the number of topics increases, CCOD_STDM's performance is even better. The accuracy of HDP continues to increase as the number of topics increases to some extent. Furthermore, CCOD_STDM performs not better than fastText in most circumstances. However, when the topic number ranges from 70 to 120, the performance of CCOD_STDM is similar to fastText, and sometimes even better than fastText, which indicates that CCOD_STDM is competitive compared to the state-of-the-art fastText method in the field of word representation and text classification.

The experimental evaluation results of seven topic discovery methods on English corpora (corpus V, VI and VII) are shown in Tables 8–10, respectively. From the whole results, we can find that CCOD_STDM performs better than LDA, word2vec, HDP and MI_STDM, and in most cases with different number of topics, CCOD_STDM is better than GloVe. But for fastText, it performs best in most cases. Specifically, in Table 8, CCOD_STDM performs not good when the number of topic is relatively small. With the number of topic ranging from 20–100, CCOD_STDM performs the best compared with whole methods. However, when the number of topic is larger than 120, fastText performs the best. In Table 9, CCOD_STDM is better than GloVe when the number of topic ranging from 10–170. When the number of topic ranging from 50–100, the performance of CCOD_STDM is similar to fastText,

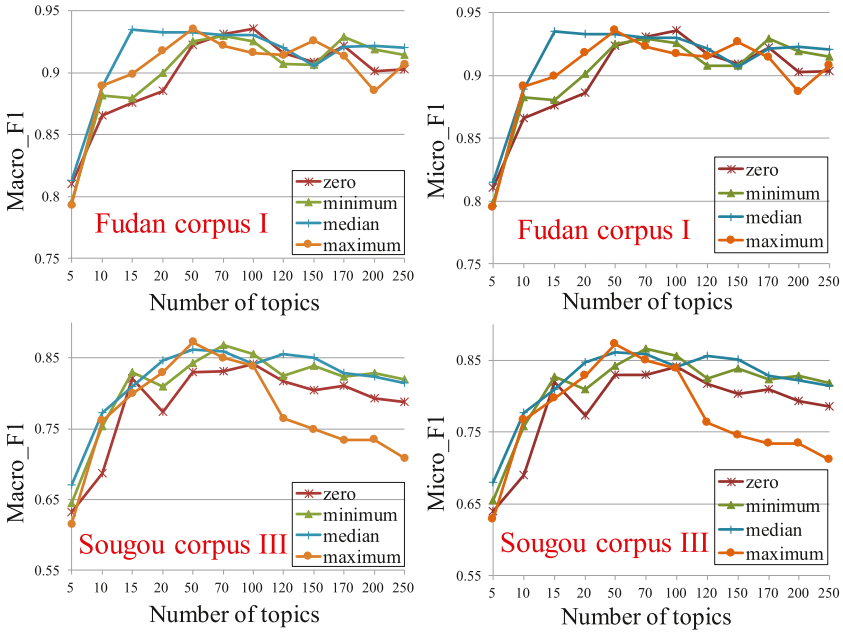

Fig. 4. The classification evaluation results of λ under different conditions on the large-scale Fudan corpus I and Sougou corpus III.

Table 11
"Topic-word" distributions obtained by CCOD_STDM with 5 topics on corpus II.

Topic.1	Topic.2	Topic.3	Topic.4	Topic.5
system	aerospace	environment	environment	computer
control	diagram	water	standard	keyword
diagram	university	**waste-water**	test	diagram
system-control	system	concentration	**environment-test**	x
model	Beijing	liquid	technology	control
parameter	data	industry	work	**single-thread**
algorithm	China	**water-disposal**	method	system
matrix	engine	pollution	check	key
system-research	technology	**environment-pollution**	equipment	influence
network	structure	disposal	quality	form
function	**space-technology**	biology	engineering	algorithm
system-method	**image-diagram**	film	product	value
system-model	aviation	bacterium	measure	systems
equation	information	oxygen	project	paper
controller	science	condition	**test-equipment**	machine
design	satellite	craft	temperature	signal
system-design	Shanghai	company	condition	data
structure	**aviation-company**	**environment-management**	metering	database
system-structure	analysis	technology	personnel	action
system-theory	flight	**environment-protection**	**check-personnel**	object

even when the number of topics is about 50, CCOD_STDM is better than fastText. For CCOD_STDM, GloVe and fastText, the comparative analysis of these three methods in Table 10 is similar to the analysis in Table 9. When the number of topic ranging from 5 to 150, CCOD_STDM is better than GloVe, and when 50–100, CCOD_STDM is similar to fastText, even when the number of topics is about 50, CCOD_STDM is better than fastText.

In order to evaluate the parameter λ in the combinatorial words methodology, we take zero, minimum, median and maximum of the word frequency of all words in the corresponding paragraph as the value of λ, respectively. And then compare the classification results of this parameter with *Macro_F*1 and *Micro_F*1 measure, which are illustrated in Fig 4. In addition to the unstable effect of the maximum strategy on Sougou corpus, the other three strategies are similar. While from the overall experimental results of these two large-scale corpus, the median strategy is relatively stable.

4.3. Case study

The classification experimental results in Section 4.2 have verified the validity of CCOD_STDM in topic discovery modeling on seven corpora. The "topic-word" distribution results obtained by different topic discovery methods are manually summarized to verify the practicality and reliability of CCOD_STDM in this section. In the classification experiment of Section 4.2, when the number of topics is small, LDA and CCOD_STDM achieve similar classification accuracies on corpus II. Therefore, in the next step we will extract the "topic-word" distribution of corpus II obtained by different topic discovery methods when the number of topics is 5 and then judge the quality of the model intuitively by observing the aggregation of words in different topics. The corresponding "topic-word" distribution results of CCOD_STDM, LDA and HDP are shown in Tables 11–13, respectively, and only the top 20 words of each topic in each "topic-word" distribution are retained.

Table 12
"Topic-word" distributions obtained by LDA with 5 topics on corpus II.

Topic.1	Topic.2	Topic.3	Topic.4	Topic.5
environment	system	aerospace	data	environment
water	control	diagram	application	standard
amount	x	measure	computer	test
diagram	model	engine	diagram	product
concentration	matrix	test	technology	work
form	algorithm	aviation	information	management
waste water	function	method	method	company
influence	parameter	machine	function	industry
environment	method	error	software	equipment
disposal	key	check	date	production
action	systems	temperature	Shanghai	China
research	controller	signal	satellite	plan
soil	network	velocity	author	country
content	condition	system	aerospace	development
substance	equation	work	design	personnel
liquid	theorem	analysis	structure	quality
condition	control	variety	program	automation
engineering	literature	time	database	America
film	state	axis	network	technology
time	structure	accuracy	procedure	author

Table 13
"Topic-word" distributions obtained by HDP with 5 topics on corpus II.

Topic.1	Topic.2	Topic.3	Topic.4	Topic.5
system	technology	algorithm	diagram	environment
method	diagram	network	system	system
diagram	method	structure	method	management
algorithm	test	research	function	state
environment	information	value	control	technology
model	company	literature	application	diagram
parameter	property	function	network	equipment
target	data	x	features	work
control	aluminum	controller	technology	Event
value	meeting	author	software	production
modular	research	biology	structure	research
thread	database	technology	form	procedure
time	value	model	sulfur	program
fault	application	neural	value	pollution
technology	condition	application	research	enterprise
procedure	plan	method	computer	industry
structure	computer	sulfur	procedure	water
water	form	parameter	characteristic	method
state	time	amount	slope	plan
data	amount	matrix	author	form

In the "topic-word" distribution results obtained by CCOD_STDM, as shown in Table 11, five topics were obtained and summarized in total: "system control", "aerospace technology", "environmental pollution", "environmental test" and "computer". These expanded combinatorial words (bold fonts in Table 11) that appear in each topic, such as "system-control", "system-research", "space-technology", "aviation-company", "water-disposal", "environment-test", "test-equipment" and "single-thread", can assist in the summary of semantic topics effectively. Due to the strong conditional dependence in the corpus of these combinatorial words, they can be highlighted to display the topic semantic information more intuitively after the processing of CCOD_STDM. When it is difficult to summarize and determine the semantic topic formed by each group of collected words, these highlighted combinatorial words can be used as a reliable and intuitive reference standard to summarize the semantic topic effectively to a certain extent, as well as representing and illustrating the semantic of each topic directly. Compared with the LDA results, as shown in Table 12, these top words in the topic distributions are roughly the same as those in the CCOD_STDM topic distributions. However, while a plausible topic that solely relies on a set of aggregated words to summarize the topic semantics is easily obtained, it is difficult to accurately determine the semantic information of a topic without knowing the logical combination order between words and the emphasis of text content. In particular, since the "topic-word" distributions obtained by HDP, as shown in Table 13, are relatively confusing and much semantic cross-information between topics exists, it is difficult to clearly summarize the potential topic semantics based on the aggregated word set.

5. Conclusion

The quality of topic discovery plays an important role in effective text mining and machine learning, and it is the foundation of knowledge organization. The process of traditional topic modeling that lacks the document context structure and the semantic association among words may yield final topics characterized by several problems, such as the loss of semantics, concept ambiguity, semantic intersection and coverage, and so on. In response to these problems, we conducted a related study that addresses the semantic topic discovery method and proposed a novel semantic topic discovery method based on the conditional co-occurrence degree (CCOD_STDM). CCOD_STDM mainly contains four stages: 1) splitting the semantic structure of the document, 2) feature expansion and subdocument reconstruction based on the conditional co-occurrence degree, 3) topic modeling, and 4) subdocument merging. To verify the effectiveness of CCOD_STDM, we conducted several controlled experiments with seven different methods on seven different corpora. For each corpus, seven topic discovery methods are used to obtain the "document-topic" distributions, and documents in the corpus are represented by the topic distributions. In addition, classification experiments are performed on the basis of the representation for each corpus. Experimental results show that our method performs much more satisfactory and stable compared with LDA, HDP, word2vec and MI_STDM. Even compared with the state-of-the-art word vector models, such as GloVe and fasTtext, CCOD_STDM is still competitive and dominant in terms of relatively small-scale corpora and some certain number of topics. At the end of the experiment, we evaluate the effectiveness of parameter λ in the combinatorial words methodology, and find that the median strategy is relatively stable compared with other three strategies. Furthermore, we conducted a case study on corpus II by using three topic discovery methods, manually summarizing the semantic topic for these "topic-word" distributions. We find that CCOD_STDM identifies more relevant words in the same topic and that the combinatorial words can help us summarize the topic conveniently.

GloVe and fastText are the famous methods in the field of word representation, so their good performance under large corpus conditions is unquestionable, especially for the fastText method. However, from the perspective of topic modeling, the interpretability of the results obtained by GloVe and fastText is still insufficient. CCOD_STDM can be used to summarize textual semantic topics and to visually interpret the generated topical results. In addition, CCOD_STDM still has some problems to be improved. How to extract valid co-occurrence word pairs hidden in the corpus, which will have an important improvement on the final topic modeling results. It has been proved that neural network plays an important role in word representation, and how to conduct the deep topic modeling with multi-layer neural network is the focus of my future research.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China [Grant Numbers 71771034, 71421001], the Scientific and Technological Innovation Foundation of Dalian (2018J11CY009) and the Science and Technology Program of Jieyang (2017xm041). Besides, We are very grateful to Dr. Xiangzhu Meng and Master Menglin Lu for guiding us to do the related experiments on GloVe, fastText and word2vec. We would like to thank the anonymous reviewers for their constructive comments on this paper.

References

[1] A. Hussain, E. Cambria, Semi-supervised learning for big social data analysis, Neurocomputing 275 (2018) 1662–1673.
[2] D. Blei, L. Carin, D. Dunson, Probabilistic topic models, IEEE Signal Process. Mag. 27 (6) (2010) 55–65, doi:10.1109/MSP.2010.938079.
[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[4] T. Hofmann, Probabilistic latent semantic analysis, in: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.
[5] Z. Huang, W. Dong, L. Ji, H. Duan, Outcome prediction in clinical treatment processes, J. Med. Syst. 40 (1) (2016) 1–13, doi:10.1007/s10916-015-0380-6.
[6] J. Boyd-Graber, Y. Hu, D. Mimno, et al., Applications of topic models, Found. Trends Inf. Retr. 11 (2–3) (2017) 143–296, doi:10.1561/1500000030.
[7] R. Alix, C. Jean-Philippe, P. Bearman, Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014, Proc. Natl. Acad. Sci. USA 112 (35) (2015) 10837–10844.
[8] D.M. Blei, T.L. Griffiths, M.I. Jordan, The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies, in: Proceedings of the Advances in Neural Information Processing Systems, 2004, pp. 17–24.
[9] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical Dirichlet Processes, 101, Publications of the American Statistical Association, 2006, pp. 1566–1581.
[10] D.M. Blei, J.D. Lafferty, A correlated topic model of science, Ann. Appl. Stat. 1 (1) (2007) 17–35, doi:10.1214/07-AOAS114.
[11] W. Li, A. Mccallum, Pachinko allocation: DAG-structured mixture models of topic correlations, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 577–584.
[12] X. Wang, A. Mccallum, 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, ACM, 2006, pp. 424–433.
[13] D.M. Blei, J.D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 113–120.
[14] T.L. Griffiths, M. Steyvers, D.M. Blei, J.B. Tenenbaum, Integrating topics and syntax, in: Proceedings of the Advances in Neural Information Processing Systems, 2005, pp. 537–544.
[15] J.L. Boyd-Graber, D.M. Blei, Syntactic topic models, in: Proceedings of the Advances in Neural Information Processing Systems, 2009, pp. 185–192.
[16] H.M. Wallach, Topic modeling: beyond bag-of-words, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 977–984.
[17] X. Wang, A. Mccallum, X. Wei, Topical n-grams: Phrase and topic discovery, with an application to information retrieval, in: Proceedings of the IEEE International Conference on Data Mining, IEEE, 2007, pp. 697–702.
[18] A. Gruber, Y. Weiss, M. Rosen-Zvi, Hidden topic Markov models, in: Proceedings of the Artificial Intelligence and Statistics, 2007, pp. 163–170.
[19] X. Yan, J. Guo, Y. Lan, X. Cheng, A Biterm topic model for short texts, 2013, pp. 1445–1456, doi:10.1145/2488388.2488514.
[20] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[21] S. Xiong, K. Wang, D. Ji, B. Wang, A short text sentiment-topic model for product reviews, Neurocomputing 297 (2018) 94–102, doi:10.1016/j.neucom.2018.02.034.
[22] J.D. Mcauliffe, D.M. Blei, Supervised topic models, in: Advances in Neural Information Processing Systems, 2008, pp. 121–128.
[23] Q. Mei, X. Ling, M. Wondra, H. Su, C. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, in: Proceedings of the 16th International Conference on World Wide Web, ACM, 2007, pp. 171–180, doi:10.1145/1242572.1242596.
[24] I. Titov, R. McDonald, A joint model of text and aspect ratings for sentiment summarization, in: Proceedings of the Association for Computational Linguistics with the Human Language Technology Conference, 2008, pp. 308–316.
[25] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, Probabilistic author-topic models for information discovery, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 306–315.
[26] A. Mccallum, A. Corrada-Emmanuel, X. Wang, The author-recipient-topic model for topic and role discovery in social networks: experiments with Enron and academic email, Comput. Sci. Depart. Faculty Publ. Ser. (2005) 1–16.
[27] G. Doyle, C. Elkan, Accounting for Burstiness in topic models, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 281–288.
[28] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2009, pp. 248–256.
[29] S. Gerrish, D.M. Blei, A language-based approach to measuring scholarly impact., in: Proceedings of the International Conference on Machine Learning, 10, Citeseer, 2010, pp. 375–382.
[30] M. Yang, X. Chen, W. Tu, Z. Lu, J. Zhu, Q. Qu, A topic drift model for authorship attribution, Neurocomputing 273 (2018) 133–140, doi:10.1016/j.neucom.2017.08.022.
[31] M. Yasunaga, J. Lafferty, Topiceq: a joint topic and mathematical equation model for scientific texts, arXiv:1902.06034 (2019).
[32] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[33] A.M. Dai, C. Olah, Q.V. Le, Document embedding with paragraph vectors, arXiv:1507.07998 (2015).
[34] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Lingu. 5 (2017) 135–146, doi:10.1162/tacl_a_00051.
[35] H. Larochelle, S. Lauly, A neural autoregressive topic model, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 2708–2716.
[36] Z. Cao, S. Li, L. Yang, W. Li, H. Ji, A novel neural topic model and its supervised extension, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 2210–2216.
[37] Y. Miao, E. Grefenstette, P. Blunsom, Discovering discrete latent topics with neural variational inference, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2410–2419.
[38] A. Srivastava, C. Sutton, Autoencoding variational inference for topic models, arXiv:1703.01488 (2017).
[39] T. Mikolov, G. Zweig, Context dependent recurrent neural network language model, in: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), IEEE, 2012, pp. 234–239.
[40] A.B. Dieng, C. Wang, J. Gao, J. Paisley, Topicrnn: a recurrent neural network with long-range semantic dependency, arXiv:1611.01702 (2016).
[41] J.H. Lau, T. Baldwin, T. Cohn, Topically driven neural language model, arXiv:1704.08012 (2017).
[42] W. Wang, Z. Gan, W. Wang, D. Shen, J. Huang, W. Ping, S. Satheesh, L. Carin, Topic compositional neural language model, arXiv:1712.09783 (2017).
[43] H.L. Hammer, A. Bai, P. Engelstad, et al., Improving classification of tweets using word-word co-occurrence information from a large external corpus, in: Proceedings of the 31st Annual ACM Symposium on Applied Computing, ACM, 2016, pp. 1174–1177.
[44] R. Jackendoff, Semantics and Cognition, MIT Press, 1983.
[45] M.A. Hearst, C. Plaunt, Subtopic structuring for full-length document access, in: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1993, pp. 59–68.
[46] M.A. Hearst, Texttiling: segmenting text into multi-paragraph subtopic passages, Comput. Lingu. 23 (1) (1997) 33–64.
[47] T.K. Landauer, D. Laham, M. Derr, From paragraph to graph: latent semantic analysis for information visualization, Proc. Natl. Acad. Sci. 101 (suppl 1) (2004) 5214–5219.
[48] W. Wei, C. Guo, J. Chen, T. Lin, L. Sun, Ccodm: conditional co-occurrence degree matrix document representation method, Soft Comput. (5) (2017) 1–17.
[49] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (suppl 1) (2004) 5228–5235.
[50] T. Minka, J. Lafferty, Expectation-propagation for the generative aspect model, in: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 2002, pp. 352–359.
[51] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 142–150.
[52] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004, pp. 1–8, doi:10.3115/1218955.1218990.
[53] Q. Liu, H. Zhang, H. Yu, X. Cheng, Chinese lexical analysis using cascaded hidden Markov model, J. Comput. Res. Devel. 41 (8) (2004) 1421–1429.
[54] T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing, IEEE Comput. Intell. Mag. 13 (3) (2018) 55–75.
[55] G. Xun, Y. Li, G. Jing, A. Zhang, Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts, in: Proceedings of the 23rd ACM SIGKDD International Conference, 2017.
[56] Q. Zhou, H. Zhou, T. Li, Cost-sensitive feature selection using random forest: selecting low-cost subsets of informative features, Knowl. Based Syst. 95 (2016) 1–11.
[57] F. Ren, M.G. Sohrab, Class-indexing-based term weighting for automatic text classification, Inf. Sci. 236 (2013) 109–125.

Wei Wei received the B.S. degree from the School of Mathematics and Information Science, Henan University, in 2012, and the Ph.D. degree in Institute of Systems Engineering from the Dalian University of Technology, in 2018. He is a lecturer of the Center for Energy, Environment & Economy Research, Zhengzhou University. His research interests include text mining, machine learning, and artificial intelligence.

Chonghui Guo received the B.S. degree in mathematics from Liaoning University, in 1995, the M.S. degree in operational research and control theory, and the Ph.D. degree in Institute of Systems Engineering from the Dalian University of Technology, in 2002. He is a professor of the Institute of Systems Engineering, Dalian University of Technology. He was a postdoctoral research fellow in the Department of Computer Science, Tsinghua University. His interests include data mining and knowledge discovery.