IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2020, accepted March 5, 2020, date of publication March 30, 2020, date of current version April 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984352

Multi-Dimension Topic Mining Based on Hierarchical Semantic Graph Model

TINGTING ZHANG[1], BAOZHEN LEE[1], QINGHUA ZHU[2], XI HAN[3], AND EDWIN MOUDA YE[4]

[1]School of Information Engineering, Nanjing Audit University, Nanjing 210031, China
[2]School of Management and Engineering, Nanjing University, Nanjing 210009, China
[3]School of Business Administration, Guangdong University of Finance and Economics, Guangzhou 510320, China
[4]School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Australia

Corresponding author: Baozhen Lee (bzli@nau.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71673122 and Grant 71972090, in part by the Jiangsu Social Science Found Project under Grant 20WTB007, in part by the Philosophy and Social Science Foundation of Guangdong Province under Grant GD19YTS01, and in part by the Nanjing University Innovation and Creative Program for Ph.D. candidate under Grant CXCY17-09.

ABSTRACT Topic mining of scientific literature can accurately capture the contextual structure of a topic, track research hotspots within a field, and improve the availability of information about the literature. This paper introduces a multi-dimensional topic mining method based on a hierarchical semantic graph model. The main innovations include (1) the hierarchical extraction of feature terms and construction of a corresponding semantic graph and (2) multi-dimensional topic mining based on graph segmentation and structure analysis. The process of semantic graph construction is based primarily on hierarchical feature term extraction, which can effectively reveal the hierarchical structural distribution of feature terms within documents. Our graph model also takes into account the complementarity of content- and context-related feature terms in documents while avoiding the loss of textual information. In addition, the multi-dimensional features of the topic can be mined effectively via an in-depth analysis of the constructed graph, resulting in a quantitative visualization of the many-to-many association between the topic and feature terms. A variety of experiments on existing document datasets demonstrate that the proposed approach is able to outperform state-of-the-art methods in terms of accuracy and efficacy.

INDEX TERMS Topic mining, multi-dimensional topic, hierarchical semantic graph.

I. INTRODUCTION

With the rapid development of database and Web 2.0 technologies, the volume of literature available online is experiencing explosive growth. This is particularly true for scientific research, whose publishing cycle has been greatly shortened. Together with increasingly blurred disciplinary boundaries, this means that scholars are now commonly facing the challenge of extracting relevant and complementary research topics from diverse and large-scale literature resources [1]–[3]. In addressing this challenge, researchers have widely adopted topic mining technology as a means of quickly discovering emerging research hotspots.

Topic mining is a statistical method used to discover latent topics in a series of documents [4]. It can transform documents from a high-dimensional term space to a low-dimensional topic space in which tasks such as text classification and core content extraction can be realized [5]. The detected topic is usually represented by a set of descriptive and collocated keywords/terms, which can help researchers efficiently detect meaningful hotspots in such areas as information retrieval [6]–[8], machine learning [9], [10], and social media modelling [11].

Before the emergence of the topic model, text representation in the field of text mining relied mainly on the vector space model (VSM) [12] or the statistical language model [13] to implement a "text → word" mapping of a document. However, these methods are hard-pressed to fully reveal the rich topic information contained in the text, as they fail to consider the association between the terms in the document and hence neglect important semantic information. To overcome some of these shortcomings, a topic model

The associate editor coordinating the review of this manuscript and approving it for publication was Santhosh Kumar Gopalan.

based on latent semantic analysis was proposed, which uses a soft clustering method to identify text topics represented by latent Dirichlet allocation (LDA) [14], [15]. LDA has been shown to be a successful algorithm because of its ability to mine semantic information from textual data. However, this method requires pre-set parameter values and only reveals potential semantic association between terms. Moreover, the assumed latent topics are of limited relevance, due to the weak correlation between components of the Dirichlet distribution vector, as contrasted with the often multi-dimensional characteristics of topic feature terms. In this paper, we instead use a graph perspective to undertake the mining of such characteristics. The nodes and edges in the graph can clearly reveal the complex relationships between feature terms and effectively highlight the documents' core information [16], [17].

In this method, as in other graph-based topic mining methods, the construction of a document graph is the foundation. Co-occurrence relationships are a common basis for such graph representations [18]–[20]; they measure the content relevance of the document by extracting meaningful terms or concepts from the text [21]. Although it is easy to construct a graph based on co-occurrence, there remains a deficit in contextual recognition. Moreover, such a graph cannot reflect the deeper structural relationships between feature words, especially when complex hierarchical relationships are involved. The key issue in graph-based topic mining is finding a way to cluster nodes with strong semantic relations to represent a topic. Most researchers have attempted this by means of clustering algorithms [22]–[24] and subgraph mining [25], [26]—that is, by means of graph structure analysis. However, these structural-analysis methods generally fail to reflect the multi-dimensional characteristics of terms due to the complexity of word expressions in the context of topic mining. Further methodological refinements are needed to reveal the contribution of each feature term to different topic clusters.

Given the above considerations, we propose a novel hierarchical term graph-based method for topic detection across multiple documents. The main innovations include (1) hierarchical extraction of feature terms and construction of corresponding semantic graphs and (2) topic mining based on graph segmentation and structure analysis. In constructing the graph, we combine co-occurrence relationships with the embedding relationships of the feature terms in the graph. The graph thus constructed can comprehensively represent the mutual relevance of feature terms based on both content and context while avoiding the loss of textual information; the hierarchical extraction of the terms used in the graph can effectively reveal the hierarchical structure to which they are related. In the process of topic mining, we apply graph segmentation and structure analysis, allowing us to further explore the multi-dimensional characteristics of each feature term in the topic and reveal the contribution of each feature term to different topic results. Thus, our method can quantitatively depict the relative strength of the many-to-many association between topics and feature terms. This also means that there can be more than one term under a topic; conversely, a single term can be spread over multiple topics.

The method proposed in this paper is an unsupervised method that can accomplish deep mining of the topics in multiple documents with high potential for further application and expansion. Experimental results show that the proposed method effectively performs topic mining for massive document corpora. Extensive evaluation using existing datasets demonstrates the improved accuracy and efficacy of our method over the present state of the art. The main advantages of our approach lie in its ability to cluster topics more effectively for literature resources by combing multiple relationships, to mine multi-dimensional topic features, and to describe quantitatively the degree of many-to-many association between feature terms and topics. In addition, the method can obtain mining results of different granularity based on different weight thresholds to meet researchers' varying needs.

The remaining sections of this paper are organized as follows. In Section II, we present the current state of the art in topic detection. In Section III, the framework of the proposed approach is outlined, and in Section IV, a hierarchical semantic graph model is presented along with a detailed procedure for building said model. A combined method of graph segmentation and structure clustering for mining multi-dimensional topics is described in Section V. Section VI reports on an experimental evaluation of the proposed semantic graph model and topic mining algorithm. Results are discussed in Section VII, and conclusions are presented in Section VIII.

II. RELATED WORK

The effective extraction of valuable topics from documents has always been a hot topic in text mining. Recently, topic extraction has attracted the attention of researchers who have applied a variety of different algorithms to the problem. Initially, researchers used TF-IDF or VSM [27]–[30] to identify text topics via automatic indexing. These methods map unstructured data to a feature vector space, which is used as the representation of the text for purposes of classification or cluster analysis. They consider the characteristics of words in isolation and ignore their interaction in the text, making them unable to reveal thematically relevant but low-frequency words.

Latent Semantic Indexing (LSI) [31], an early form of the topic model, maps documents from a high-dimensional word space to a low-dimensional latent semantic space. This solves the problems of high dimensionality, feature sparseness, and semantic loss witnessed in the results of VSM processing. Based on the concept of LSI, Hofmann [32] proposed a refinement known as probabilistic latent semantic analysis (PLSA). This is a generative model for representing the words in a given document. However, it cannot provide a distribution over a corpus of related documents. This led in turn to the development of the latent Dirichlet allocation (LDA)

model [14], which can be regarded as a Bayesian model with three levels of hierarchy. In LDA, documents are presented in a textual database which is modelled as a finite mixture over a set of topics. Then, the topics are modelled once more as an infinite mixture over a set of topic probabilities with associated words. LDA has a wide range of uses and achieves effective dimensionality reduction, but it makes the hypothetical topics almost irrelevant, due to the weak correlation between components of the random variable in the Dirichlet distribution—a trait not consistent with many practical scenarios.

At present, graph-based methods have found wide use in topic detection due to their good performance in detecting the association between words [20]. Such methods are important visualization tools, using nodes and edges to reflect words and relationships in the graph and thus facilitating a more interesting and flexible application of topic mining algorithms [18]. Current graph-based topic mining methods mainly focus on two aspects, namely, the selection of the core terms in the graph as the discovered topics and the identification of important associated subgroups in the graph.

The core term-based approach recognizes text topics based primarily on network centrality, which measures the importance of a node in the graph. Hassan and Louiqa [19] modeled their KeyGraph as a network by considering keyword co-occurrence. Betweenness centrality for an edge is fixed at the number of the shortest paths for all pairs of nodes in the KeyGraph; a topic is represented following a procedure to remove intercommunity edges. Jo *et al.* [33] detected topics in large-scale linked document collections based on a citation graph. Their work uses the correlation between term distributions to identify topics; topic scores are measured for each term using the likelihood ratio of binary hypotheses based on a probabilistic description of graph connectivity. Liu *et al.* [34] proposed a generative graph model to mine topic-level direct influence and predict user behaviour, using a propagation and aggregation method based on heterogeneous link information and content information. The navigable topic graph method of Cataldi *et al.* [35] is based instead on temporal and social term evaluation and applied to the detection of emerging topics in tweets. The graph connecting emerging terms with other related keywords is constructed in the form of a directed, node-labeled, and edge-weighted graph for a given time interval. Pons-Porrata *et al.* [36] presented a topic detection system combining both partitional and agglomerative approaches to reveal implicit knowledge; in their study, each discovered topic consists of a set of documents and a summary extracted from these documents based on testor theory.

Subgroup mining approaches search for important subgraphs within the text network based on the graph's structure or node attributes and identify topics by revealing various relationships between terms. Liu *et al.* [37] proposed a weighted graph clustering algorithm with the purpose of community detection based on the concept of density and attractiveness. A user's core degree (Liu *et al.* use social network data in their demonstration) is defined as node weight, with attractiveness then defined as edge weight. Wartena and Brussee [38] clustered keywords without relying on a training set using the k-bisecting clustering algorithm, with the resultant keyword clusters taken as topic descriptions for the Wikipedia articles in their dataset. Zhou *et al.* [22] presented a graph clustering method for detecting community structures in complex networks; their novel similarity method uses degrees of attraction and recommendation to calculate node similarities, with the nodes then clustered under a k-medoids framework. The approach of Zhang *et al.* [39] fuses semantic relations and co-occurrence relations into a term graph on whose basis topics are detected. The maximum of the conjugate conditional probabilities between two terms is taken to represent the relation between them, and cohesive sub-clusters are formed by removing weak links. Chen *et al.* [9] constructed a topic graph in which topics were represented as concept nodes and their semantic relationships obtained using WordNet. In this approach, topic pruning via Markov decision processes allows for topic extraction through community discovery.

Ghoorchian *et al.* [40] introduced a dense topological graph utilizing dimensionality reduction and clustering techniques. The graph is then partitioned into multiple dense subgraphs, each representing a topic. Wang *et al.* [41] proposed a human-oriented algorithm, based on KeyGraph, that can facilitate human perception, comprehension, and even innovation. Ma *et al.* [18] proposed a novel keyword filtering model and graph generation method to detect topics in the keyword graph of micro-blogging data. Their graph generation algorithm transforms text data into a disordered keyword graph, which is then grouped in an orderly manner by a heuristic algorithm. Carusi and Bianchi [42] employed a bipartite graph to model the scholar-journal network, where nodes representing scholars and journals are linked together whenever a scholar has his or her paper(s) published in a scientific journal. Spectral techniques are then used to detect communities within social networks. Mikhina and Trifalenkov[43] used a graph community detection method to deal with the problem of text clustering. The change in modularity, which serves as a metric value after merging each pair of documents, is considered in the clustering process. Hachaj and Ogiela [44] proposed a novel hashtag filtration model and community graph generation approach to detect popular microblog topics. Their filtration model can identify clusters of common interests among user groups and extract the community structure of a network based on heuristic methods. Xuan *et al.* [45] used a probabilistic model for graph mining, where Bernoulli distribution is adopted to model the existence of an edge parameterized by topics of two linked nodes.

Overall, methods based on network centrality represent a great improvement over TF-IDF and VSM-based methods, with representative indicators including centrality and degree of intermediateness. However, such methods have very low algorithmic efficiency for large and complex networks.

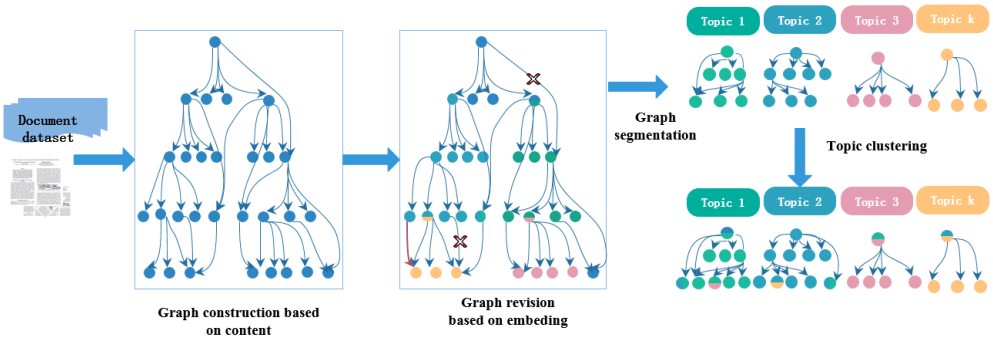

FIGURE 1. Topic mining framework for multiple documents based on semantic graph.

The subgroup-based method for topic mining, which reveals word relationships in a text using nodes and edges in the graph, is a new research idea that merits further exploration.

Current graph-based methods typically analyze a single relationship between terms in the network; there is a dearth of methods that consider multiple term relationships and/or mine the deep hierarchical structures among terms. In this paper, a semantic graph is constructed by combining content and context information based on the hierarchical extraction of feature terms. In addition, due to the contextual complexity of expression and the ambiguity of word boundaries, we must further excavate the multi-dimensional features of words and quantitatively describe the degree of the many-to-many association between a given topic and its feature words.

III. RESEARCH FRAMEWORK

The semantic graph of a document has the characteristics not only of a general network, but of a language network specifically. In such a graph, apart from semantic correlations in content, there are also physical structural correlations among terms. In this paper, we construct a model of multiple overlapping term relations, taking into account the content correlation and the contextual correlation between terms. This allows us to represent the document comprehensively while avoiding the loss of textual information. Multi-dimensional topic mining is then carried out by combining graph segmentation and structure analysis methods.

In this section, we present a general framework for a topic mining process following this strategy. The main steps of topic mining are depicted in Fig. 1 and are described in detail as follows.

Step 1. Document pre-processing The pre-processing stage includes converting the document to a suitable data structure, removing stop words, and segmenting the texts into feature terms. For the documents being analyzed, each document in the dataset is tagged with a unique ID to facilitate later identification and retrieval. Stop words with no semantic significance are then filtered out. Text segmentation in our method entails dividing the documents into a set of feature terms. For this purpose, we use "Jieba," the best available Python module for Chinese word segmentation. Terms, which we take as the basic unit for constructing the semantic graph, are not words in the ordinary sense, but descriptors that represent content features of the document.

Step 2. Semantic graph construction A hierarchical semantic graph is constructed based on the multiple relations between feature terms. As we know, co-occurrence and word2vec consider the correlations between feature terms based on different background information. Co-occurrence considers the co-occurrence relationships between terms at the article level, measuring correlations of content meaning at the level of the document. Word2vec considers the embedding relationship of terms within a local window size, which reflects a finer scale of correlation. The effective combination of the two methods in the semantic graph can thus provide a comprehensive measure of the correlation between terms.

Step 3. Topic mining The aim of topic mining is to identify topics included in the document collection and assign appropriate topics to their respective documents. By exploring and organizing the content of the semantic graph, we attempt to identify topics to enable automatic assembly of diverse pieces of information into manageable clusters. Spectral methods are used for subgraph segmentation, and the contribution of each feature term to each segmentation graph is then calculated based on structural analysis to yield multi-dimensional topics.

As shown in Fig. 1, our approach incorporates multiple types of relationships into a semantic term graph via graph segmentation and structure analysis. Firstly, a hierarchical term graph-based method is adopted to mine co-occurrence relationships and thereby construct the term graph of multiple documents. Then, contextual relevance based on word2vec is adopted to refine the graph. Finally, multi-dimensional topic mining results are detected based on subgraph segmentation and structure analysis.

IV. SEMANTIC GRAPH CONSTRUCTION

This section describes the construction of a hierarchical semantic graph based on term co-occurrence and word2vec relationships. First, the content correlation between feature terms is calculated based on co-occurrence. Second, the hierarchical extraction of feature terms is carried out to

construct a semantic graph based on content correlation. Finally, the contextual correlation between feature terms is calculated based on word2vec, and the previously constructed graph is refined to generate the final semantic graph.

A. TERM CORRELATION BASED ON CO-OCCURRENCE

Feature terms and their correlation relationships play important roles in document analysis. The stronger the correlation by which the terms are associated, the greater their contribution to the document. In this paper, term correlation is defined by the pairwise association of feature terms. To determine the association between any two feature terms in multiple documents, it is necessary to consider the degree of interaction between the two terms at the same time. Departing from the conditional probabilistic representation of one-way dependence, this paper attempts to characterize a more generalized correlation of two-way dependence between term association. In order to characterize the documents comprehensively, we should focus on the association between terms throughout the documents.

The correlations between terms are represented by similarities $P(T_i|T_j)$ and $P(T_j|T_i)$ via the distribution of term frequencies in each paragraph. We use the symbol $r(T_\mathrm{i}, T_\mathrm{j})$ to represent the association between term T_i and term T_j:

$$r(T_i, T_j) = \frac{P(T_i \,|T_j) \cdot P(T_j \,|T_i)}{P(T_i \,|T_j) + P(T_j \,|T_i) - P(T_i \,|T_j) \cdot P(T_j \,|T_i)} \quad (1)$$

where $r(T_\mathrm{i}, T_\mathrm{j})$ is dependent on the degree of interaction (i.e., the degree to which pairwise terms mutually influence each other). $P(T_i|T_j)$ and $P(T_j|T_i)$ represent the mutual conditional dependence between feature terms T_i and T_j. These conditional similarities are, in turn, calculated based on the following formulas:

$$P(T_i \,|T_j) = \frac{b}{b+c} \quad (2)$$

$$P(T_j \,|T_i) = \frac{b}{a+b} \quad (3)$$

Let a represent the number of items with term T_i but not term T_j in the document corpus, c the number of items with term T_j but not term T_i, and b the number of items with both terms. The term association presented here measures the interactions between feature terms across the entire multi-document corpus. A higher value of $r(T_\mathrm{i}, T_\mathrm{j})$ indicates a stronger relationship between T_i and T_j, whereas a lower one represents a weaker relationship.

B. SEMANTIC GRAPH CONSTRUCTION

Representing the document as a graph allows the retention of some important information, such as semantic relationships and internal structures. In this paper, we capitalize on a semantic graph to represent hierarchical relationships among terms in the document. In our graph, nodes represent the terms, and node weights are allocated using $r(T_i, T_j)$ values between the terms. Directed edges represent the relationship between pairs of terms: edge direction records the sequential extraction of term information, and edge weight is allocated to reflect the extent of the relationship between terms. In contrast to the existing graph construction processes, which define term correlation after the graph is created, we construct the graph hierarchically based on a predefined correlation threshold for the feature terms.

The following specific steps are undertaken in the construction of the semantic graph.

(1) Setting the root node (k_i). The root node k_i is regarded as the first target node—the starting point for feature extraction in the graph construction—and is selected based on term frequencies. The root node can be selected either manually based on user preference or automatically according to term frequency. In general, a term that is both high-frequency and highly relevant to the document will be selected as the root node. (2) Setting the weight threshold (w). The weight threshold (w) defines the minimum weight for an edge to exist between two nodes in the graph. In our paper, the weight between two nodes is determined both by the correlations $r(T_i, T_j)$ (hereafter abbreviated as r) and the weight threshold (w). If r for a given pair of terms is below this threshold, then no relationship will be identified between the corresponding two nodes in the graph. In general, the choice of w can be referred to the needs of the specific application being considered. (3) Graph model construction. Based on the parameters set above, the graph is constructed using hierarchical term extraction and snowball sampling. First, k_i is selected as the target node, and the terms which have correlations r with the target node are acquired. Then, terms whose r with node k_i exceeds w are selected as the first-layer nodes in the graph, described as set $\{\mathrm{K}_1 |\ \mathrm{K}_1 \in \text{terms}\}$. Next, the first-layer nodes $\{\mathrm{K}_1\}$ are designated as target nodes. Traversing these one by one, the terms whose r with $\{\mathrm{K}_1\}$ exceeds w are selected as second-layer nodes $\{\mathrm{K}_2 \mid \mathrm{K}_2 \in \text{terms} \ \&\ \mathrm{K}_2 \cap \mathrm{K}_1 = \emptyset\}$. Subsequent layers are selected via the same method until all terms are traversed. Finally, a hierarchical graph is assembled from the terms in their respective layers. Nodes in this graph may have multiple incoming and outgoing edges, and lower-layer nodes will always point to higher-layer nodes. These different layers can express the document's content at different levels of granularity.

G(d) can thus be represented as a directed hierarchical graph made up of a node set *nod.*, an edge set *edg.*, node frequency N, and edge weight E.

$$G(d) = G(nod., edg., N, E) \quad (4)$$

where *nod.* denotes terms, *edg.* indicates term relationships, N represents term frequencies based on the numbers of terms in the whole document, and E shows relationships between the terms as determined by r.

C. SEMANTIC GRAPH REFINEMENT

We use the word2vec model [46] to refine the semantic term graph by considering word embedding information. The essence of this procedure is to combine content correlation and contextual correlation in graph analysis for topic

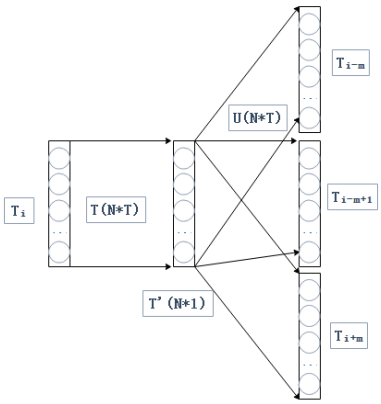

FIGURE 2. Skip-gram of word embedding.

modelling [17], [47]. The procedure contains two main steps, (a) generation of contextual relationships and (b) refinement of the semantic graph.

1) GENERATING THE CONTEXTUAL RELATIONSHIPS

Word embedding is one of the most popular representations of document vocabulary. It is capable of capturing the context of a word within a document. Word2vec can utilize either of two model architectures to produce a distributed representation of words: continuous bag-of-words (CBOW) or continuous skip-gram (CSG) [48]. In CBOW, the model predicts the current word from a window of surrounding context words. The order of context words does not influence prediction. In CSG (Fig.2), the model uses the current word to predict the surrounding window of context words. CSG weighs nearby context words more heavily than more distant context words. In this section, we use CSG to measure the contextual relationships between feature terms.

The objective function of the skip-gram (CSG) model is to use each word to predict all other words in its context as defined by a window around the word. Contextual correlation $c(T_i, T_j)$ is defined as the probability of predicting word T_i given word T_j:

$$c(T_i, T_j) = \frac{exp(v_i, v_j)}{\sum_{T_k \in V} exp(v_k, v_j)} \tag{5}$$

where v_i is the corresponding vector representation of word T_i. In words, such a model says that the probability $c(T_i, T_j)$ is proportional to the dot product of the vectors corresponding to the two words T_i and T_j. With this model, we can find a vector representation for all terms that maximizes the probability of using each term to predict all other terms in a small contextual window.

2) REFINING THE SEMANTIC GRAPH

This contextual information is applied to the semantic graph by using the contextual correlation $c(T_i, T_j)$ to revise the term relationship. Specifically, the correlation threshold k is first set as a standard of comparison. Then, the node pairs in the semantic graph are traversed and pairs lacking an edge are given an edge if they have a high c value ($c > k$). An existing edge between two nodes is deleted if the corresponding c value is low ($c < k$). $\mathrm{R}(T_i, T_j)$ is the final edge weight in the hierarchical semantic graph.

$$\mathrm{R}(T_i, T_j) = \begin{cases} c(T_i, T_j), & \text{if } c(T_i, T_j) > k \& r(T_i, T_j) = 0 \\ 0, & \text{if } c(T_i, T_j) < k \& r(T_i, T_j) < k \end{cases} \tag{6}$$

In sum, the graph model proposed in this paper simultaneously considers the content of terms and the contextual relationship between feature terms. It thus effectively reveals the hierarchical association among terms in the document. Consequently, the graph takes a radioactive tree shape, from which nodes can be extracted repeatedly to determine their relationships.

V. TOPIC MINING FROM THE REFINED SEMANTIC GRAPH

This section describes the use of subgraph segmentation and structure analysis to mine the multi-dimensional topics. First, a spectral technique is used for subgraph segmentation. Then, a structural analysis method is applied to mine multi-dimensional topics for use in the topic clustering.

A. SUBGRAPH SEGMENTATION

The spectral technique, as a method to detect communities in social networks, bears the advantage of strong adaptability to different data distributions. It revolves around a change in representation based on the eigenvectors of a suitable matrix (e.g., the Laplacian of the graph) [42], which enhances the inherent structure of the data, making potential clusters more evident and easier to detect than in the original space.

Subgraph segmentation comprises two main steps. The first step is to compose a graph G. The process of composition entails reconstructing the association matrix R into adjacency matrix W:

$$W_{i,j} = W_{j,i} = \begin{cases} R(T_i, T_j), & w \in edg. \\ 0, & else \end{cases} \tag{7}$$

where association matrix R is expressed as the node weight E in the constructed graph, and W is a symmetric matrix based on the transfer of R.

After the adjacency matrix W is obtained, the degree matrix D must be calculated:

$$D_{i,j} = \begin{cases} \sum_j W_{i,j}, & if\ i = j \\ 0, & else \end{cases} \tag{8}$$

Here, $W_{i,j}$ is an element in the adjacency matrix, and $\sum_j W_{i,j}$ represents the sum of weights of the specific node and its connected nodes in the graph. D is the diagonal matrix in this situation, and the Laplacian matrix (L) can be calculated based on D and W: $\mathrm{L} = D - W$.

The second step is to cut the graph [49]. In this step, the normalized cut algorithm (Ncut) is used to obtain the subgraph segmentation results by finding edges which have the least weight and can balance the size of the cut-out subgraphs.

The goal of Ncut is to cut graph G into k unconnected subgraphs, each with a set of nodes $A_1, A_2, \ldots, A_k$:

$$NCut(A_1, A_2, \ldots, A_k) = arg\ min\ tr(F^T D^{-1/2} L D^{-1/2} F) \quad s.t.\ F^T F = I \tag{9}$$

The optimization goal of the function is to find m eigenvalues that minimize $D^{-1/2}\ LD^{-1/2}$. Then, the GMM algorithm is used to obtain the k subgraph segmentation results ($A_1, A_2, \ldots, A_k$) based on the m eigenvalues. Finally, the k initial topic mining results are generated based on eigenvalue correlation, namely ($C_1, C_2, \ldots, C_k$).

B. TOPIC CLUSTERING

Based on the above, the k initial mining results are further revised with the use of structural analysis, revealing the contribution of specific feature terms to different topics. Owing to the complexity of contextual expression and the ambiguity of boundaries, the feature terms of topics often have multi-dimensional characteristics; i.e., a feature term may belong to multiple topics.

We use a probability-based method to frame multi-dimensional topic mining as an optimization problem. First, for a specific feature term, we calculate the structural contribution to each initial topic. Then, a specific feature term is assigned to the initial topics with which it has a high association score. It should be noted that in the process of calculating the structural contribution, we consider both the connection (direct relationship) and the intensity (indirect relationship) between the feature terms, rather than simply giving the correlation:

$$p(g|C_i, N) = \frac{f(Pa_{(N)})_g}{f(Pa_{(N)})_G} + \frac{\sum I(N; Pa_{(N)})_g}{\sum I(N; Pa_{(N)})_G} \tag{10}$$

Here, $p(g|C_i, N)$ is the structural score between node N_i and topic C_i, indicating the contribution of N to C_i. $Pa_{(N)}$ represents the parent nodes of node N. The first item in the formula represents the connection between node N and topic C_i; it measures the ratio of $f(Pa_{(N)})_g$ to $f(Pa_{(N)})_G$, where $f(Pa_{(N)})_g$ indicates the number of $Pa_{(N)}$ in C_i and $f(Pa_{(N)})_G$ is the number of $Pa_{(N)}$ in semantic graph G. The second item denotes the intensity between node N and topic C_i, measured as dependence strength between node N and its parent nodes in the graph. This dependence intensity is expressed as:

$$I(N; Pa_{(N)}) = \sum\nolimits_{N_j} R(N, N_j), N_j \in Pa_{(N)} \tag{11}$$

$I(N, \mathrm{Pa}_{(N)})$ is interpreted as the dependence intensity between N and $Pa_{(N)}$, which is obtained through the association between nodes in the semantic graph.

Based on the structure score calculation method described herein, the contribution of each feature term to different topic results can be revealed, and thus multi-dimensional mining of the topic can be carried out. The experimental process can be summarized as follows:

(1) Input initial topic mining set ($C_1, C_2, \ldots, C_k$) from subsection 5.1 as a pending item with k specified.

(2) Select a specific node N from $C_i \in (C_1, C_2, \ldots, C_k)$, then calculate the structure scores between node N and the other $k-1$ members of the topic mining set (other than C_i). For all $k-1$ members of the topic, if $p(g|C_j, N) > \gamma$(the selected threshold), then $N \in C_j$, j is any number of topic categories in $k-1$ categories.

(3) Traverse all nodes in turn and repeat the second step to obtain the k final topic clusters, which are represented as ($\theta_1, \theta_2, \ldots, \theta_k$).

The topic clustering result ($\theta_1, \theta_2, \ldots, \theta_k$) is the final result of multi-dimension topic mining. Each clustering θ_i represents a hypothetical topic, and the terms in each clustering represent the topic features. The topic mining process presented herein allows multiple feature terms to exist under a single topic and for a specific feature term to be scattered over multiple topics, with different correlation strengths.

In addition, this method can represent the importance of feature terms in each topic based on their structure score. For a specific topic θ, the importance score of each node N_i can be calculated. Nodes with higher importance values are selected as the feature terms to represent the topic. The importance score calculation formula is as follows:

$$score(\theta, N_i) = \sum I(N_i; Pa_{(N_i)})_g + f(Pa_{(N_i)})_g \tag{12}$$

Since documents are assigned a likelihood or probability distribution over topics, our model may assign a document to multiple topics, which provides our method with the flexibility to handle multitopic documents. To estimate the probability that document d covers topic θ_j, we would collect all the separate counts of words in document d that belong to each θ_j, then normalize these counts among all k topics:

$$C_{d,j} = \frac{\sum_{w \in \theta_j} c(w, d)}{\sum_{\forall \theta} \sum_{w \in \theta_j} c(w, d)} \tag{13}$$

In the above equation, $C_{d,j}$ indicates the probability that document d covers topic θ_j and $c(w, d)$ means the counts of words in document d that belong to each θ_j. Finally, document d will be assigned to topic z if and only if $C_{d,j}$ exceeds the document-topic similarity threshold.

The topic mining method outlined above constructs a hierarchical semantic graph based on the multiple relations between feature terms, thereby comprehensively characterizing the correlation between these terms. In the process of constructing the semantic graph, we consider the hierarchical extraction of feature terms and effectively reveal the internal hierarchical association among them. In addition, the results of topic mining effectively combine subgraph segmentation with clustering technology, quantitatively depicting the multi-associative strength between topics and feature terms.

VI. EXPERIMENT

The objective of our proposed method is to improve the quality of multi-document topic clustering. To evaluate the effectiveness and performance of our method, we conducted a series of experimental analyses characterized by varying

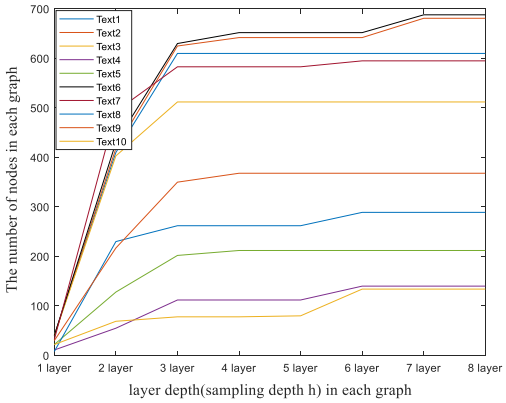

FIGURE 3. The node count at different layer depths.

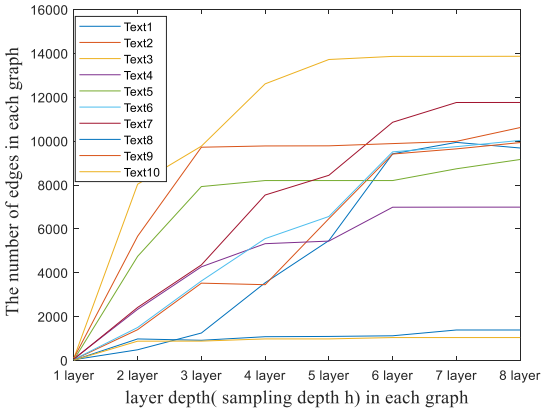

FIGURE 4. The edge count at different layer depths.

parameters, comparison with other methods, and statistical analysis of the experimental results.

The data source we employ comes from the natural language processing group of Fudan University in China and is publicly released (http://www.nlpir.org/wordpress/2017/10/02/). The corpus contains 20 categories and more than 20,000 pieces of text, which can be used for thematic analysis. In the experiments, we randomly select 15 categories and a subset of the documents within each category to cover different domains. In all, 1,576 text items are deployed for performance evaluation.

The topic clustering method as employed in this paper has the following parameters:

In order to compare different layers of the graphs, we calculate the number of nodes in the graph with different depth h. The result indicates that the nodes in graphs constructed with 4 layers have already captured the information content of the document. Therefore, the sampling depth h is set to 4 in this paper. As seen in Fig. 3, the number of terms is stable once w reaches 0.3. Therefore, the weight threshold w (the minimum value of the weight between two nodes) is set to 0.3. The correlation threshold k, used to revise the term relationships in the graph, should be higher than the weight threshold w. We set k to 0.35 based on expert opinion. Also, the score threshold γ, which establishes the minimum structure score between a given node and a topic, is set to 0.5 to retain the main structural information of the document.

A. TOPIC CLUSTERING RESULTS ANALYSIS

Modularity is a common method of measuring the strength of a network community structure. The value of the modularity depends on the community distribution of nodes in the network, which can be used to quantitatively measure the quality of the network community. Generally, the greater the modularity, the better the clustering quality is. The formula for this calculation is as follows:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \tag{14}$$

where A_{ij} is the adjacency matrix that represents the weight of edges between k_i and k_j; m denotes the number of edges in the graph; k_i is the degree of node i; c_i is is the community containing point i; and $\delta = 1$ if $c_i = c_j$, 0 otherwise.

In order to analyze the characteristics of the document graphs at different layer depths and assess their impact on topic clustering, term extraction was carried out hierarchically for the 10 documents at 8 different depths $h = \{1, 2, \ldots 8\}$, and graphs of corresponding depths were constructed based on the extracted feature terms. Since extraction was performed 8 times per document, each document was represented by eight graphs of different layer depths. The quantities of nodes and edges in each graph were counted separately and are reported in Figs. 3 and 4.

In Fig. 3, the count of nodes in each graph increases as graph layer depth (or sampling depth h) grows; the node count stabilizes after reaching 4 layers. The result indicates that the nodes in graphs constructed with 4 or more layers have already captured the information content of the document. In Fig. 4, the count of edges, too, increases as layer depth grows and stabilizes after reaching layer 4. This indicates that edges in the graphs constructed after layer 4 already reflect the structural relationships between terms in the document.

Based on the semantic graph construction in this paper, variations in w will change the number of feature terms extracted from the semantic graph, thus potentially altering the topic clustering results. To analyze the influence of different weights w on the results of topic clustering, we carried out experiments based on several different values of w: [0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1]. Statistical variables such as term count, cluster number, and modularity are calculated for each value of w.

Table 2 reports the results of topic clustering under different weights. Term count is the number of feature terms extracted under different weights, cluster number is the optimal number of clusters, and modularity is the corresponding value when the number of topics is optimal. As seen in the table, when the edge weight is 0.8 (maximum), the number of feature terms is 278 (minimum). These feature terms belong to high frequency terms in documents and thus effectively summarize the document's central themes. The cluster

TABLE 1. Number of documents in the chosen 15 categories.

Environment: 120	Computer: 123	Education: 111	Transport: 112	Economy: 100
Military: 147	Sports: 120	Medical: 98	Art: 123	Politics: 122
Agriculture: 100	Energy: 63	History: 100	Philosophy: 86	Electronics: 51

TABLE 2. Impact of different W values on the datasets.

Edge weight threshold (w)	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Term count	278	3390	12936	16772	82208	87552	87552	87552
Cluster number	17	35	114	124	188	240	240	240
Modularity	0.5429	0.4898	0.6535	0.5973	0.6412	0.6287	0.6287	0.6287

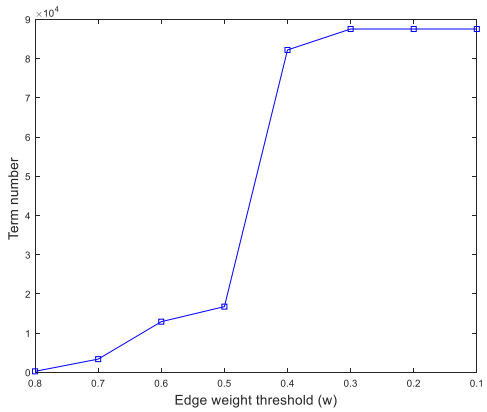

FIGURE 5. Term count at different weight thresholds (w).

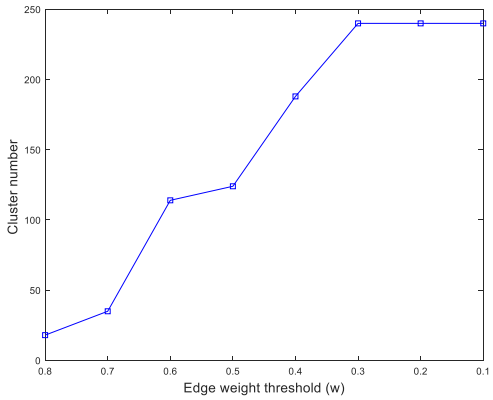

FIGURE 6. Cluster number at different weight thresholds (w).

number here is 17: again, the minimum observed. Its close consistency with the actual category count (15) indirectly illustrates the high cohesion of topic clustering. When the edge weight decreases to 0.3, the number of feature terms reaches its maximum value of 87,552. These terms cover almost all characteristic terms except for stop words. Also, we can see that the values of term count (87,552 in table 2) are constant from 0.1 to 0.3, which means that all terms are extracted (except stop words) when w reaches 0.3. The cluster number here is likewise the largest at 240. The distribution of these topics is more detailed and comprehensive, reflecting a further subdivision of the topics in the dataset's original 15 categories, showing the ability of our method to express the topic of a document at multiple granularities.

In general, as w decreases, the number of feature terms extracted and the number of topic clusters both increase. This shows that the proposed method permits the selection of different weight values based on different sizes of extracted feature terms, making the method highly adaptable to different scales of data.

Figs. 5 and 6 display the trends in term count and cluster number under different weight thresholds (w). As seen in Fig. 5, the increase in term count slows down once w reaches 0.3. At this point, all the feature terms of the documents have already been extracted, and these feature terms cover almost all the thematic information across multiple documents. A similar distribution trend is evident with respect to cluster number, as Fig. 6 shows; with decreasing w, the corresponding distribution of topic clusters also shows a slow upward trend. The experimental results validate the flexibility and practicability of the proposed method. The weight values can be set differently to satisfy the varied topic recognition requirements of practical research.

Due to the diversity and complexity of feature terms, the optimal number of topic clusters may not be unique. We therefore analyze the validity of the proposed method under different numbers of topic clusters, with results shown in Fig. 7.

Fig. 7 shows the modularity of the topic mining results under different topic numbers. A trend resembling a normal distribution is evident. In the range of [225, 240], the modularity values reach their peaks and are consistently greater than 0.6. This shows that the clustering results can be adapted well to the thematic content of the specific dataset used in this study. As a result, we set the topic cluster number in the model to 240, which yields the highest modularity.

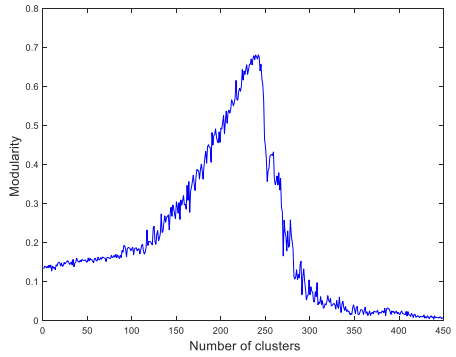

FIGURE 7. Modularity based on different numbers of clusters.

Combining the data from Table 2 and Fig. 7, we find that, based on the semantic graph of this paper, when the edge weight w stabilizes the clustering results, the modularity simultaneously reaches its maximum. This indicates that when w reaches a certain threshold, the content and structure of the extracted feature terms reach a certain stability in the semantic graph. This stability can help the researchers to adapt to various experimental scales and scenarios; different thresholds and cluster numbers can be selected for analysis according to the needs of specific experimental environments.

B. METRICS EVALUATION

For a given dataset, it is a ground truth that a set of reference topics exists, along with a corresponding set of documents per reference topic. Thus, we can use precision, recall, and F-measure to evaluate the accuracy of the topic clustering method. Precision indicates the accuracy with which two documents are identified as belonging to the same topic and recall indicates the accuracy of classifying two documents to different topics.

$$\text{Precision} = \frac{a}{(a+c)}; \quad \text{Recall} = \frac{a}{(a+d)}$$
$$\text{F} - \text{measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

In these formulas, a denotes the frequency of correctly predicting two documents to belong to the same topic, c denotes the frequency of wrongly predicting two documents to belong to the same topic, and d denotes the frequency of wrongly predicting two documents to belong to different topics.

To evaluate the proposed topic detection method, we compared our method with a word occurrence-based approach (WOA) [19], word2vec-based method (word2vec), probabilistic latent semantic analysis (PLSA), and latent Dirichlet allocation (LDA). WOA and word2vec are chosen as clustering-like pattern methods, deemed representative of the clustering approaches, which cluster the topics by term occurrence relationships. PLSA and LDA are examples of topic models, frequently used as text-mining tools for discovery of hidden semantic structures in a body of text.

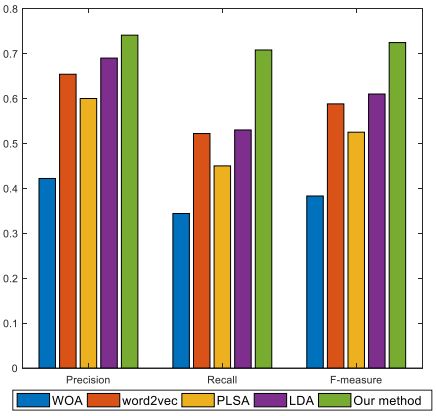

FIGURE 8. Performance of our approach compared to topic clustering methods.

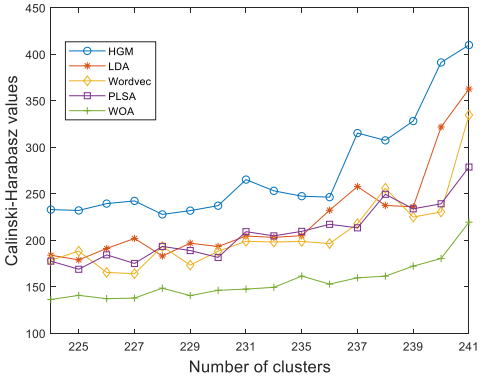

FIGURE 9. Calinski-Harabasz index of different topic clustering methods.

Fig. 8 shows that all four methods can be used to detect topics in our experimental dataset, and that the word2vec-based method and LDA, together with our method, show the best results (> 0.6) in precision. Moreover, our method displays the highest precision, recall, and F-measure, demonstrating the effectiveness of our proposed approach. The WOA method has the worst performance in our experimental dataset—perhaps because this method only considers co-occurrence relationships, ignoring the influence of term frequency.

To evaluate the performance of topic feature extraction, the Caliński-Harabasz index (CHI) is used to measure the quality of our method in term of different numbers of topic clusters. The CHI can be used to evaluate the model if the ground truth labels are not known; it is a kind of estimate that can help us choose the proper clustering number, where a higher CHI indicates a model with better-defined clusters. The formula for this index is:

$$CHI = \frac{tr(B_k)}{tr(W_k)} \times \frac{m-k}{k-1} \tag{15}$$

where B_K is the between-group dispersion matrix, W_K is the within-cluster dispersion matrix, and m is the number of points in the data. CHI determines the optimal number

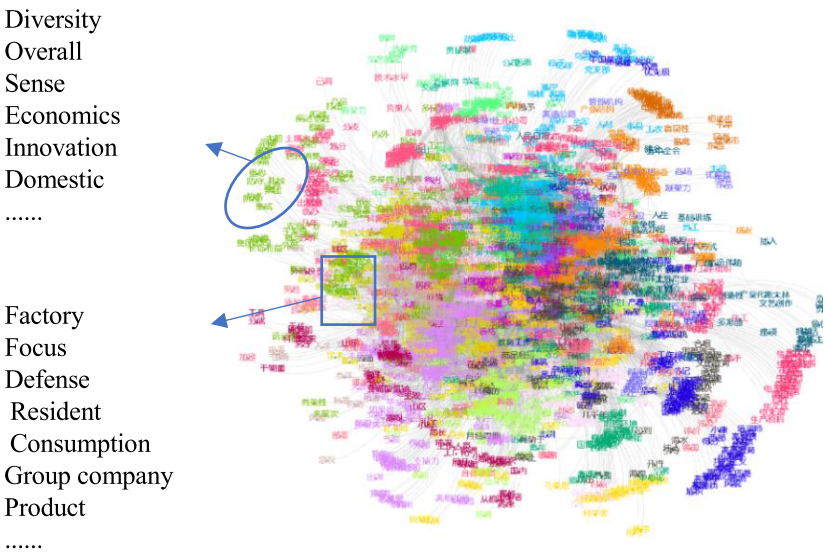

FIGURE 10. Partial visualization of semantic graph based on spectral technique (high-frequency terms).

of clusters by observing the semantic exclusiveness between topics and the semantic consistency within topics.

As reported in Fig. 5, the modularity was found to be highest when the cluster number was in the interval [225, 240]. Here, therefore, we only compare the performance of different topic clustering methods within this interval.

Fig. 9 shows a comparison of Calinski-Harabasz (CHI) results for each topic mining method. The CHI value of the proposed method is higher than that of other methods based on the number of topic clusters in the interval, which shows that the purity of the topics is higher and further illustrates the validity of the topic number selection method in this paper.

C. TOPIC DISTRIBUTION STATISTICS

The construction of the semantic graph relies mainly on the hierarchy of feature terms in the document, which can effectively reveal the hierarchical structural distribution among feature terms. Fig. 10 presents the visualization of parts of the semantic graph results based on the spectral technique with a higher term frequency.

In Fig. 10, we see that the semantic graph constructed in this paper can effectively extract feature terms related to different topics (square), as well as feature terms within a single topic (ellipse). Thus, the diversity of feature terms can be further excavated, which is a boon to the effectiveness of topic mining. In addition, the constructed semantic graph can maximize the extraction of feature terms belonging to the same topic. Therefore, conducting topic mining based on the semantic graph as proposed here can increase the speed of topic clustering and reduce the computational complexity of the model.

The semantic graph was divided into certain subgroups by the proposed topic mining method, with each subgroup denoting one candidate topic. According to the topic mining results presented in this paper, a topic can contain more than one feature term, and a feature term may belong to multiple topics.

In Table 3, we show six relatively independent example topics, each with a set of representative feature terms from the topic mining result. For example, topic 1 relates to the topic of ''Art'', topic 2 relates to ''Education'', 3 to ''Computer'', 4 to ''Economy'', 5 to ''Medical'', and 6 to ''Sport''. In order to analyze the relative importance of each feature term to the topic in detail, we select the feature terms with high importance values in various topics and conduct a ranking analysis.

Importance scores were calculated using the formula introduced in the section on topic clustering, and the top 10 nodes were selected as terms representative of their topic. The scores reflect a column vector normalization and are shown in Table 3.

Table 3 reports the top ten terms detected for each topic cluster, along with their importance scores. As seen in the table, the first major topic cluster, related to art, was tagged with ''form'', ''creation'', ''culture'', ''features'', ''art'', and so forth. We can discern that topic 1 likely involves the creation and form of art. Topic 4, related to economics, may refer to financial policy; its top terms include ''Securities'', ''Quality'', ''Fiscal policy'', etc. Topic 5, which evidently concerns medicine, may address a relevant situation in medical administration, since top terms include ''expert'', ''State Council'', and ''medical administration''. These experimental results show that the proposed topic mining method is effective in

TABLE 3. Feature terms and importance score for each topic result.

Feature terms in **topic 1**	Score
Effect	0.00580
Content	0.00549
Development	0.00548
Form	0.00540
Creation	0.00534
Culture	0.00532
Times	0.00515
Features	0.00506
Art	0.00503
China	0.00497

Feature terms in **topic 2**	Score
Talent	0.00506
Cultivate	0.00504
University	0.00504
Learn	0.00500
School	0.00499
Student	0.00493
Society	0.00487
Represent	0.00487
Create	0.00487
College students	0.00487

Feature terms in **topic 3**	Score
Development	0.00641
Operating system	0.00607
Systematic	0.00602
Provide	0.00596
Data	0.00569
Program	0.00535
Operating	0.00533
Design	0.00523
Introduction	0.00513
Information	0.00503

Feature terms in **topic 4**	Score
Securities	0.00401
Quality	0.00400
Merge	0.00387
Rule	0.00384
Management	0.00377
Social security	0.00355
Department	0.00355
Resource	0.00352
Aggregate demand	0.00352
Fiscal policy	0.00352

Feature terms in **topic 5**	Score
Expert	0.0159
State Council	0.0159
China	0.0156
Business	0.0156
Medical Administration	0.0155
Government	0.0154
Life safety	0.0154
Phenomenon	0.0154
State-operated	0.0154
Institute	0.0154

Feature terms in **topic 6**	Score
Coach	0.00466
Basic training	0.00434
Skill	0.00429
Opinion	0.00421
Fill in	0.00411
College	0.00408
Teaching mode	0.00406
College students	0.00404
Member	0.00404
Teaching reform	0.00396

calculating the contribution of feature terms. By calculating the structural score of each feature term, we can further reveal the contribution of feature terms to the clustering results of different topics.

Table 4 shows six additional multidimensional example topics randomly selected from the topic mining results in this paper. Importance scores for each node were calculated as before, and the top 15 nodes were selected as terms representative of the topics. The scores follow a column vector normalization.

As seen in Table 4, some feature terms can be attributed to multiple topics. When this happens, the intensity of attribution may also vary. For example, Topics 1 and 2 both contain the feature terms ''technology'', ''China'', and ''education'', which indicates that both topics concern problems in education. A detailed analysis shows that Topic 1 involves the cultivation of talents in education, whereas Topic 2 is about the development of education. Similarly, both Topic 3 and Topic 4 contain the feature terms ''market'' and ''price''; Topic 3 concerns product price, whereas Topic 4 has to do with stock trading. Topic 4 and Topic 6 are two relatively independent topics, but both are related to ''enterprise''. Topic 4 is about enterprise product pricing, and Topic 6 relates to enterprise taxation.

The experimental results show that the proposed method can effectively mine multi-dimensional features of topics and quantitatively depict the many-to-many association between topics and feature terms. That is to say, a topic can contain more than one term, and one term can be scattered across different topics.

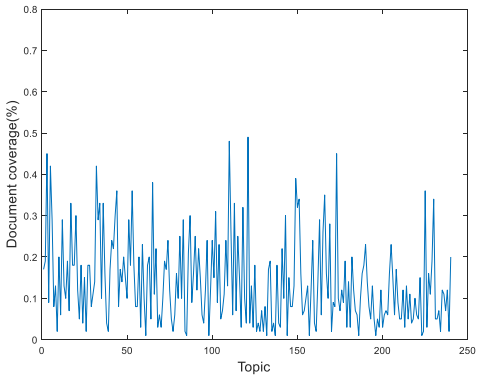

FIGURE 11. Distribution of document coverage across all topics.

Finally, we analyze the document coverage of the 240 topics. The degree of coverage between the 240 topics and the document dataset is calculated based on the probability that document d covers topic θ. Results are shown in the Fig. 11.

Fig. 11 shows the document coverage percentage in every topic. The x-axis represents topics, and the y-values are the topics' respective document coverage. The coverage percentages lie between [0, 0.47], which are obtained based on the actual distribution of document data. According to the trend of the data distribution, the coverage percentage for most topics is between 0.05 and 0.2, which indicates that the content of most documents may be cross-topic. A few topics have a substantially higher coverage percentage ($>$0.3), which suggests that these topics may have very high generality.

TABLE 4. Feature terms and importance score in each topic result.

Feature terms in **topic 1**	Score
Talent	0.013
Culture	0.00942
Ideological quality	0.00716
Profession	0.0063
Restructuring	0.00615
Technology	0.00607
Cultivate	0.00558
Consumer	0.00541
Cultural education	0.00427
China	0.00427
Education department	0.00421
Teacher	0.0042
Education	0.00419
Cultural quality	0.00416
Layout	0.00412

Feature terms in **topic 2**	Score
Awareness	0.0122
Strategy	0.012
Thinking	0.00875
Information flow	0.00829
China	0.00685
High morality	0.00684
Technology	0.00581
High tech	0.00578
Education	0.00559
High level	0.0052
Field	0.00498
Atmosphere	0.00497
Leadership team	0.00455
Appearance	0.00445
Develop	0.00445

Feature terms in **topic 3**	Score
Information	0.0114
Internet	0.0107
Distribution	0.0106
Human resources	0.00817
Commodity	0.00799
China Telecom	0.00669
Optimization	0.0066
Price	0.00558
Market	0.00557
Retail	0.00546
Demand	0.00495
Price reduction	0.00491
Nature	0.00489
Configuration	0.00398
Gap	0.00396

Feature terms in **topic 4**	Score
Transaction cost	0.0105
Professional	0.00777
Price index	0.00771
Transaction	0.00655
Market	0.00652
Consumption	0.0063
Commodity	0.00546
Price	0.00546
Enterprise	0.00545
Value	0.00488
Learn	0.00486
Tuition	0.00486
Immediate interest	0.004
Foreign trade	0.004
Public company	0.004

Feature terms in **topic 5**	Score
Professional	0.0103
Fair	0.00765
Society	0.00764
Bond	0.00613
Share	0.0061
Benefit	0.00514
Statistics	0.00512
Economic	0.00479
Purchase power	0.00475
Transaction	0.00469
Market	0.00468
Internet	0.00442
Stock	0.00442
Price	0.0044
Quality	0.00434

Feature terms in **topic 6**	Score
Wholesale	0.0103
Taxation	0.0074
Management	0.00737
Consumption	0.00604
Technology	0.006
Tax system	0.00599
Tax authority	0.00593
Tax policy	0.00503
Effect	0.005
Fluidity	0.00499
Channel	0.00466
Tax bureau	0.0046
Enterprise	0.0046
Environment	0.00455
Cost	0.00394

On the whole, the results of topic mining in this paper are in line with the actual distribution of document data. It reflects the distribution of topics in the context of the current specific documents.

The analysis above affirms that the method proposed in this paper can not only detect topics more effectively from multiple documents across different categories but also mine important topics by leveraging weight threshold relations. Our method constructs its hierarchical term graph by combining mutually complementary relations to reveal the hierarchical distribution of feature terms. In addition, the effective combination of graph segmentation and clustering technology in this paper leads to high performance in topic clustering.

VII. DISCUSSION

The experimental results show that the proposed topic mining method can effectively identify clusters of multi-dimensional topics from documents through in-depth analysis of multi-semantic association and the hierarchical structure of feature terms. Compared with existing topic mining algorithms, our method shows higher precision, recall, and F-measure, all of which illustrate its effectiveness. Extensive empirical results show that the new method significantly outperforms the state-of-the-art methods in terms of accuracy and efficacy. As demonstrated in the experiment section, the key advantages of our approach are as follows. (1) It can cluster topics more effectively to track research hotspots within the literature by combining multiple relationships: both co-occurrence and contextual. (2) It can mine multi-dimensional features of topics and reveal the degree of many-to-many association among feature terms and topics. (3) It can mine important topics at different levels of granularity by leveraging changes in edge weight threshold, meeting the needs of human researchers who seek to perceive the most significant topics.

The main innovations of this study are twofold: (1) hierarchical extraction of feature terms and construction of a corresponding semantic graph, and (2) multi-dimensional topic mining based on structure score. The semantic graph model

constructed can not only effectively reveal the hierarchical structure of the feature term distribution but also comprehensively represent the many-to-many correlation strength between topic and feature terms by taking into account the multiple correlations among feature terms.

Most methods of graph construction consider an edge to exist between two nodes if there is a correlation between two feature terms [19], [35]. The graph construction in this paper departs from the existing research by relying on hierarchical extraction of nodes based on a specific threshold, ensuring that the extracted feature terms will be highly dependent on each other. Moreover, unlike the single relationship between feature terms [19], [38], [50] used in the existing research, the correlation measurement for feature terms in this paper takes into account both content and context, comprehensively representing the correlation between feature terms by synthesizing different background information.

Traditional topic mining methods can be traced back to text clustering algorithms [29], such as VSM [23], [30]. However, such algorithms generally rely on distance calculations and ignore semantic information. PLSA [32] and LDA [14], [51], as examples of latent semantic indexing, can expand the semantic coverage of identified topics. However, the hypothetical topics are almost unrelated to each other due to the weak correlation among the random vectors of the Dirichlet distribution. In addition, existing methods of subgraph mining, such as network centrality [19], cannot reveal the multilevel features of words in a complex context. In this work, we explore the multi-dimensional topic features of documents by revealing the multiple semantic relationships between feature terms. The importance score of each feature term is calculated based on the combination of graph segmentation and structure analysis, which improves the results of topic clustering. This method can thus effectively represent the contribution of each feature term to different topic results.

There are some limitations to the present line of research. In the process of graph construction, the root node, term weight w, and regulatory variables k, λ, γ were all chosen manually, and each of these may have some influence on the experimental results. Meanwhile, the topic mining method in this paper considers only information presented in text format. In future research, we can use the structural topic model to incorporate external information such as author, journal categories, formulas, and charts for further analysis. In addition, the experiment reported above was implemented on a single corpus; there will undoubtedly be new challenges when applying the method to external data sources such as social media to simulate a real-life use scenario.

VIII. CONCLUSION

Scientific literature, as an important carrier of knowledge, contains an abundance of topics, each with information value. In this paper, we present a hierarchical term graph approach to integrate content relations and context relations for multi-dimension topic mining. Our method merges multiple relations into a hierarchical semantic graph and detects topics based on this graph. The proposed method can not only detect topics more effectively by fusing multiple relations but also mine different granularities of important topics via changes in the weight threshold. In addition, mining the strength of many-to-many association between topic and feature terms can effectively characterize the contribution of each feature term based on topic clusters. Extensive experiments using existing document datasets demonstrate that our method generates cohesive clusters and achieves the best outcome, driven by the quality of the topic clusters. The evaluation proves the efficacy of our method over state-of-the-art methods. Therefore, the proposed method is well suited to monitor and track new research hotspots and provide decision support for scientific research and policymaking.

REFERENCES

[1] L. Zhang, Z. Wu, Z. Bu, Y. Jiang, and J. Cao, "A pattern-based topic detection and analysis system on chinese tweets," *J. Comput. Sci.*, vol. 28, pp. 369–381, Sep. 2018, doi: 10.1016/j.jocs.2017.08.016.

[2] M. Chong, "Sentiment analysis and topic extraction of the Twitter network of #prayforparis," in *Proc. 79th ASIS&T Annu. Meeting*, Oct. 2016, p. 133, doi: 10.1002/pra2.2016.14505301133.

[3] J. Sang and C. Xu, "Browse by chunks: Topic mining and organizing on Web-scale social media," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 7S, no. 1, pp. 1–18, Oct. 2011.

[4] Y. Liang, Y. Liu, C. Chen, and Z. Jiang, "Extracting topic-sensitive content from textual documents—A hybrid topic model approach," *Eng. Appl. Artif. Intell.*, vol. 70, pp. 81–91, Apr. 2018, doi: 10.1016/j.engappai.2017.12.010.

[5] C. Bouveyron, P. Latouche, and R. Zreik, "The stochastic topic block model for the clustering of vertices in networks with textual edges," *Statist. Comput.*, vol. 28, no. 1, pp. 11–31, Jan. 2018, doi: 10.1007/s11222-016-9713-7.

[6] H.-G. Kim, S. Lee, and S. Kyeong, "Discovering hot topics using Twitter streaming data: Social topic detection and geographic clustering," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining ASONAM*, Niagara Falls, ON, Canada, Apr. 2013, pp. 1215–1220.

[7] Q. Chen, X. Guo, and H. Bai, "Semantic-based topic detection using Markov decision processes," *Neurocomputing*, vol. 242, pp. 40–50, Jun. 2017, doi: 10.1016/j.neucom.2017.02.020.

[8] S. Bhutada, V. V. S. S. S. Balaram, and V. V. Bulusu, "Semantic latent Dirichlet allocation for automatic topic extraction," *J. Inf. Optim. Sci.*, vol. 37, no. 3, pp. 449–469, May 2016, doi: 10.1080/02522667.2016.1165000.

[9] P. Chen, N. L. Zhang, T. Liu, L. K. M. Poon, Z. Chen, and F. Khawar, "Latent tree models for hierarchical topic detection," *Artif. Intell.*, vol. 250, pp. 105–124, Sep. 2017, doi: 10.1080/02522667.2016.1165000.

[10] K. Hashimoto, G. Kontonatsios, M. Miwa, and S. Ananiadou, "Topic detection using paragraph vectors to support active learning in systematic reviews," *J. Biomed. Informat.*, vol. 62, pp. 59–65, Aug. 2016, doi: 10.1080/02522667.2016.1165000.

[11] A. Rafea and N. A. GabAllah, "Topic detection approaches in identifying topics and events from arabic corpora," *Procedia Comput. Sci.*, vol. 142, pp. 270–277, 2018, doi: 10.1080/02522667.2016.1165000.

[12] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975, doi: 10.1145/361219.361220.

[13] D. Gildea and T. Hofmann, "Topic-based language models using EM," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, 1999, pp. 1–4. [Online]. Available: http://http.icsi.berkeley.edu/ftp/global/pub/speech/papers/euro99-emlm.pdf

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003, doi: 10.5555/944919.944937.

[15] D. Li, B. He, Y. Ding, J. Tang, C. Sugimoto, Z. Qin, E. Yan, J. Li, and T. Dong, "Community-based topic modeling for social tagging," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. CIKM*, Toronto, ON, Canada, 2010, pp. 1565–1568.

[16] I. Ali and A. Melton, "Graph-based semantic learning, representation and growth from text: A systematic review," in *Proc. IEEE 13th Int. Conf. Semantic Comput. (ICSC)*, Newport Beach, CA, USA, Jan. 2019, pp. 118–123.
[17] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowl.-Based Syst.*, vol. 151, pp. 78–94, Jul. 2018, doi: 10.1016/j.knosys.2018.03.022.
[18] T. Ma, Y. Zhao, H. Zhou, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "Natural disaster topic extraction in sina microblogging based on graph analysis," *Expert Syst. Appl.*, vol. 115, pp. 346–355, Jan. 2019, doi: 10.1016/j.eswa.2018.08.010.
[19] H. Sayyadi and L. Raschid, "A graph analytical approach for topic detection," *ACM Trans. Internet Technol.*, vol. 13, no. 2, p. 4, Dec. 2013, doi: 10.1145/2542214.2542215.
[20] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. CIKM*, Glasgow, Scotland, 2011, pp. 1031–1040.
[21] S. S. Sonawane and P. A. Kulkarni, "Graph based representation and analysis of text document: A survey of techniques," *Int. J. Comput. Appl.*, vol. 96, no. 19, pp. 1–8, 2014, doi: 10.5120/16899-6972.
[22] H. Zhou, J. Li, J. Li, F. Zhang, and Y. Cui, "A graph clustering method for community detection in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 551–562, Mar. 2017, doi: 10.1016/j.physa.2016.11.015.
[23] S. R. Fitriyani and H. Murfi, "The K-means with mini batch algorithm for topics detection on online news," in *Proc. 4th Int. Conf. Inf. Commun. Technol. (ICoICT)*, Bandung, Indonesia, May 2016, pp. 1–5.
[24] D. Yan, E. Hua, and B. Hu, "An improved single-pass algorithm for chinese microblog topic detection and tracking," in *Proc. IEEE Int. Congr. Big Data (BigData Congress)*, San Francisco, CA, USA, Jun. 2016, pp. 251–258.
[25] S. Yang, Q. Sun, H. Zhou, Z. Gong, Y. Zhou, and J. Huang, "A topic detection method based on KeyGraph and community partition," in *Proc. Int. Conf. Comput. Artif. Intell. ICCAI*, 2018, pp. 30–34.
[26] C. Böhm, G. Kasneci, and F. Naumann, "Latent topics in graph-structured data," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. CIKM*, Maui, HI, USA, 2012, pp. 2663–2666.
[27] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Inf. Process. Manage.*, vol. 39, no. 1, pp. 45–65, Jan. 2003, doi: 10.1016/S0306-4573(02)00021-3.
[28] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern-taxonomy extraction for Web mining," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Sep. 2004, pp. 242–248.
[29] P. H. Adams and C. H. Martell, "Topic detection and extraction in chat," in *Proc. IEEE Int. Conf. Semantic Comput.*, Santa Clara, CA, USA, Aug. 2008, pp. 581–588.
[30] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decis. Support Syst.*, vol. 48, no. 2, pp. 354–368, Jan. 2010, doi: 10.1016/j.dss.2009.09.003.
[31] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990, doi: 10.1002/(SICI)1097-4571(199009)41:63.0.CO;2-9.
[32] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, Burlington, MA, USA: Morgan Kaufmann, Jul. 1999, pp. 289–296, [Online]. Available: https://arxiv.xilesou.top/ftp/arxiv/papers/1301/1301.6705.pdf
[33] Y. Jo, C. Lagoze, and C. L. Giles, "Detecting research topics via the correlation between graphs and texts," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD*, Aug. 2007, pp. 370–379.
[34] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. CIKM*, 2010, pp. 199–208.
[35] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in *Proc. 10th Int. Workshop Multimedia Data Mining MDMKDD*, Jul. 2010, p. 4, doi: 10.1145/1814245.1814249.
[36] A. Pons-Porrata, R. Berlanga-Llavori, and J. Ruiz-Shulcloper, "Topic discovery based on text mining techniques," *Inf. Process. Manage.*, vol. 43, no. 3, pp. 752–768, May 2007, doi: 10.1016/j.ipm.2006.06.001.
[37] R. Liu, S. Feng, R. Shi, and W. Guo, "Weighted graph clustering for community detection of large social networks," *Procedia Comput. Sci.*, vol. 31, pp. 85–94, 2014, doi: 10.1016/j.procs.2014.05.248.
[38] C. Wartena and R. Brussee, "Topic detection by clustering keywords," in *Proc. 19th Int. Conf. Database Expert Syst. Appl.*, Turin, Italy, Sep. 2008, pp. 54–58.
[39] C. Zhang, H. Wang, L. Cao, W. Wang, and F. Xu, "A hybrid term–term relations analysis approach for topic detection," *Knowl.-Based Syst.*, vol. 93, pp. 109–120, Feb. 2016, doi: 10.1016/j.knosys.2015.11.006.
[40] K. Ghoorchian, S. Girdzijauskas, and F. Rahimian, "DeGPar: Large scale topic detection using node-cut partitioning on dense weighted graphs," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 775–785.
[41] H. Wang, F. Xu, X. Hu, and Y. Ohsawa, "IdeaGraph: A graph-based algorithm of mining latent information for human cognition," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Manchester, U.K., Oct. 2013, pp. 952–957.
[42] C. Carusi and G. Bianchi, "Scientific community detection via bipartite scholar/journal graph co-clustering," *J. Informetrics*, vol. 13, no. 1, pp. 354–386, Feb. 2019, doi: 10.1016/j.joi.2019.01.004.
[43] E. K. Mikhina and V. I. Trifalenkov, "Text clustering as graph community detection," *Procedia Comput. Sci.*, vol. 123, pp. 271–277, Jan. 2018, doi: 10.1016/j.procs.2018.01.042.
[44] T. Hachaj and M. R. Ogiela, "Clustering of trending topics in microblogging posts: A graph-based approach," *Future Gener. Comput. Syst.*, vol. 67, pp. 297–304, Feb. 2017, doi: 10.1016/j.future.2016.04.009.
[45] J. Xuan, J. Lu, G. Zhang, and X. Luo, "Topic model for graph mining," *IEEE Trans. Cybern.*, vol. 45, no. 12, pp. 2792–2803, Dec. 2015, doi: 10.1109/TCYB.2014.2386282.
[46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[47] C. Wu and B. Wang, "Extracting topics based on Word2 Vec and improved jaccard similarity coefficient," in *Proc. IEEE 2nd Int. Conf. Data Sci. Cyberspace (DSC)*, Shenzhen, China, Jun. 2017, pp. 389–397.
[48] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.
[49] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, "Less is more: Sparse graph mining with compact matrix decomposition," *Stat. Anal. Data Mining*, vol. 1, no. 1, pp. 6–22, Feb. 2008, doi: 10.1002/sam.102.
[50] M. Shams and A. Baraani-Dastjerdi, "Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction," *Expert Syst. Appl.*, vol. 80, pp. 136–146, Sep. 2017, doi: 10.1016/j.eswa.2017.02.038.
[51] H. Xu, F. Zhang, and W. Wang, "Implicit feature identification in chinese reviews using explicit topic mining model," *Knowl.-Based Syst.*, vol. 76, pp. 166–175, Mar. 2015, doi: 10.1016/j.knosys.2014.12.012.

TINGTING ZHANG was born in Jilin, China, in 1988. She received the B.S. degree in management and the M.S. degree in management science and engineering from the Jiangsu University of Science and Technology, China, in 2012 and 2015, respectively. She is currently pursuing the Ph.D. degree with the School of Engineering Management, Nanjing University, China. She has several publications in journals and currently hosts a project about data analysis. Her research interests include data analysis, information resource management, optimization research, and user behavior analysis.

BAOZHEN LEE was born in Xiyang, Shanxi, China, in 1975. He received the B.S. degree in geography science from Shanxi University, Taiyuan, China, in 2000, the M.S. degree in management science and engineering from Southwest Petroleum University, Chengdu, in 2002, and the Ph.D. degree in management science and engineering from Nanjing University, Nanjing, China, in 2008.

From 2002 to 2005, he was a Lecturer with the Public Management Department, Shandong University. From 2008 to 2017, he was an Assistant Professor with the School of Economic and Management, Jiangsu University of Science and Technology. From 2010 to 2011, he was a Visiting Scholar with the Sonic Lab, Northwestern University. From 2014 to 2015, he was a Visiting Scholar with Media Lab, MIT. Since 2017, he has been a Professor with the Research Center of Government Auditing Big Data, Nanjing Audit University. He is the author of two books and more than 60 articles. He holds three patents. His research interests include data analytics, optimization research and decision science in business or public management. He is a Reviewer of the journal *European Journal of Operational Research* and *Information Sciences*.

QINGHUA ZHU is currently a Professor with the School of Information Management, Nanjing University. He is a member of the Association for Information Science and Technology (ASIST) and the Association for Information Systems (AIS). His research interests include social media, human information behavior, and health informatics. His works have been published in journals, such as *Scientometrics*, *Online Information Review*, *Information Systems Frontiers*, *Aslib Journal of Information Management* and other journals and conference proceedings, such as iConference, International Conference on Information Systems (ICIS), and Pacific Asia Conference on Information Systems (PACIS).

XI HAN was born in Dengzhou, Henan, China, in 1984. He received the B.S. degree from Northeast Normal University, in 2007, the M.S. degree from Sun Yat-sen University, in 2013, and the Ph.D. degree in information science from Nanjing University, in 2019. Since 2019, he has been a Teacher with the School of Business Administration, Guangdong University of Finance and Economics. He has published articles in journals such as the Journal of Librarianship and Information Science, Social Science and Medicine, Computers in Human Behavior. His research interests focus on big data and business analytics.

EDWIN MOUDA YE is currently pursuing the Ph.D. degree with the University of South Australia. He is a member of the Association for Information Science and Technology (ASIS&T) and Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR). His research interests include information behaviour, user preference, and decision processes. His works have been published and presented at ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), Pacific Asia Conference on Information Systems (PACIS), Association for Information Science and Technology Annual Meeting (ASIS&T), amongst others.