IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 20, 2019, accepted September 6, 2019, date of publication September 11, 2019,
date of current version September 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940516

Automatic Labeling of Topic Models Using Graph-Based Ranking

DONGBIN HE[1,2,3], MINJUAN WANG[1,2], ABDUL MATEEN KHATTAK[1,4], LI ZHANG[1,2], AND WANLIN GAO[1,2]

[1]College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[2]Key Laboratory of Agricultural Informatization Standardization, Ministry of Agriculture and Rural Affairs, Beijing 100083, China
[3]College of Computer Science and Engineering, Shijiazhuang University, Shijiazhuang 050021, China
[4]Department of Horticulture, The University of Agriculture Peshawar, Peshawar 25120, Pakistan

Corresponding authors: Minjuan Wang (minjuan@cau.edu.cn) and Wanlin Gao (wanlin_cau@163.com)

This work was supported in part by the Project of Scientific Operating Expenses, Ministry of Education of China, under Grant 2017PT19.

ABSTRACT Generated topic label, an alternative representation of topics learned by topic model, is widely used to help the user interpret the topics more efficiently. A major challenge now is to label a discovered topic accurately in an objective way. This article introduces a novel graph-based ranking model (TLRank), to find a meaningful topic label with high Relevance, Coverage, and Discrimination. The model applies a specific strategy that suppresses or enhances the matrix transition probability according to the textual similarity between vertices (sentences) and the characteristics of vertices respectively. Moreover, to boost diversity and enhance performance, TLRank scores the candidate sentences and refrains redundancy of topic labels simultaneously in a single labeling process. In our experiments, the evaluation results showed that the TLRank model significantly and consistently outperformed the prevailing state-of-the-art and classic models in topic labeling task.

INDEX TERMS Graph-based ranking, topic model, automatic labeling, topic label, labeling, LDA.

I. INTRODUCTION

The topic model is a very important technique in natural language processing, such as information retrieval and text mining, and the results of topic discovery are usually a set of high-frequency terms [1]. A user needs to observe all the top terms of topics to scrutinize the discovered results. Unfortunately, it is impossible for a user to fully understand a topic and distinguish between different topics merely due to the high-frequency terms of topics, especially when the user is not familiar with the literature in the corpus [2]. Although the manual topic label is more interpretive and understandable, labeling topics need considerable human labor to review massive data of the corpus. Moreover, it tends to add subjective opinions to the manual label unconsciously.

It is well known that automatic labeling methods to generate meaningful labels for discovered topics assist with topic interpretation. This work received more attention and was very challenging. The existing studies use phrases, summaries, or images to enhance the interpretation of topics, e.g., based on phrases [3]–[12], summaries [2], [13], [14] or images [12], [15]–[17].

However, in most cases, the description of phrases is usually insufficient due to their finite length, and the images can only be used to assist in specific scenes. Automatic summarization methods generate a summary (a paragraph in common) which including the most important information of the document. This is mainly based on two approaches: abstractive [18]–[23] and extractive [2], [13], [24]–[33]. As the abstractive approaches are highly complex, the researchers generally focus more on extractive ones. Thus, automatic summarization could be the preferable choice to assemble the topics label by the extractive sentences manner.

The processing of the existing studies mainly divided into two separate parts, i.e. scoring the sentences and choosing the appropriate ones [30]. These models usually have a lower upper bound of the algorithm. Because they do not take into account the redundancy of generating summary in the ranking process, the ranking results do not really reflect the importance of sentences. Therefore, the accuracy of the ranking order will be weakened [13], [25]. Wan and Wang [2] proposed a submodular optimization model, and Ren *et al.* [13] contributed a sentence regression method based on deep

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

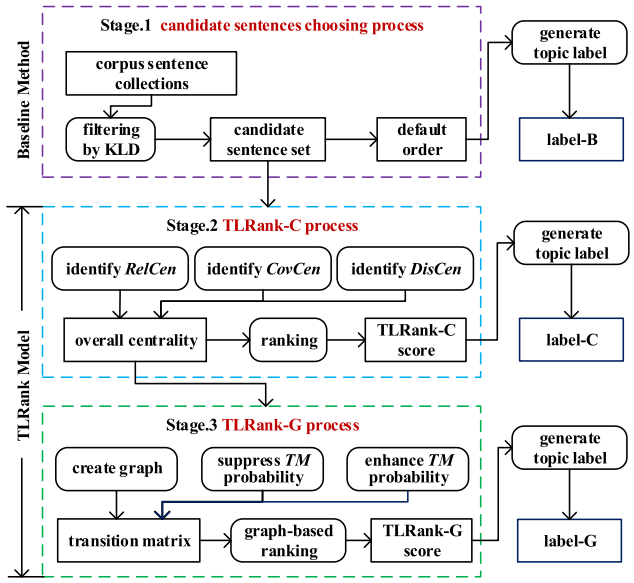

FIGURE 1. An overview of the labeling process of TLRank model.

neural network. Both studies combined sentence scoring and extracting in a single process to work based on greedy algorithm. This is useful for designing efficient approximation algorithms of intractable NP-hard problem. However, it is very expensive.

To the best of our knowledge, we are the first to investigate and propose a novel redundancy-aware graph-based ranking model, TLRank. The purpose is to extract salient sentences from a corpus automatically and generate meaningful labels with minimum redundancy for the discovered LDA-style topics. Since the graph-based ranking algorithm considers the whole picture, rather than relying solely on the local specific vertex information, TLRank would reach the final global optimum effectively and address the issue of high computing costs efficiently. Moreover, the textual summaries generated by TLRank are more accurate, and explicitly in line with the three critical criteria proposed by Mei *et al.* [3] and Wan and Wang [2], namely, high **Relevance**, **Coverage** and **Discrimination** for all topics. The critical part of this paper, i.e. labeling discovered topics process, decomposed into three stages, is shown in Fig. 1.

(1) In order to improve the processing efficiency of the system for each topic, we select the most relevant 500 sentences in the corpus sentence collections according to the Kullback-Leibler Divergence (KLD) proposed by Wan and Wang [2]. Then we set up the candidate sentences sets and fetch the top ones to assemble into a topic label, named label-B. The method of generating label-B is called Baseline in the subsequent experimental section.

TLRank model consists of two processes, i.e. TLRank-C and TLRank-G, corresponding to stage 2 and stage 3 respectively.

(2) In stage 2, the major work is to identify forging the overall centrality in the final three centrality features of each candidate sentence. According to the ranking score based on the overall centrality of sentences, the TLRank-C process generates a topic label (called label-C) for each discovered topic. The most crucial output in this stage is the overall centrality of each candidate sentence. This is also the essential foundation for the work of TLRank-G in the next stage.

(3) In the final stage, candidate sentences can be used as vertices to create a text graph. A transition matrix (*TM* in short) is also established according to overall centrality value.

TLRank-G suppresses or enhances the probability of *TM* with a specific strategy to make it possible in a single graph-based ranking process. This process scores the candidate sentences and refrains the redundancy of generating topic label (called label-G). In the end, the processing boosts the diversity of label-G and further improves the performance of the labeling task.

The experimental results show that the TLRank model can generate the topic label with higher Relevance, Coverage, and Discrimination. Thus it significantly and consistently outperforms the prevailing state-of-the-art and classic models in topic labeling task.

The rest of the paper is arranged as follows: Section 2 presents and discusses the related work. In Section 3 and 4, we introduce our redundancy-aware graph-based ranking model. Section 5 describes the experimental settings and details, followed by the Result and Discussion in Section 6, and finally in Section 7, we present our conclusions and future work.

II. RELATED WORK

The topics discovered by topic modeling technology are usually represented by an ordered list with the highest marginal probabilities terms w_i based on $p_r(w_i|z)$, where z represents LDA-style topic [1]. Aiming to help users understand topics clearly, Blei and Lafferty [34] described a method for visualizing topics extending LDA model to generate the multi-word distribution of topics. Their method found significant n-grams related to a topic, which were then used to help understand and interpret the underlying distribution. Obviously, it is not enough to assist users with the interpretations of topics. Therefore, to help users understand, researchers developed many labeling methods to generate descriptive and meaningful textual labels automatically.

Most of the existing work exploits phrases to label topics. For example, Mei *et al.* [3] uses an unsupervised method to generate meaningful phrases for LDA-style topics. This uses lexical association measures to extract bigram collocations from the document sets modeled by topic-modeled technology. The technique then ranks them according to the KL differences for each topic. Lau *et al.* [4] proposed a supervised method that utilized Wikipedia and top-ranking topic terms to choose candidate phrases and then fetch the best by ranking. Kou *et al.* [5] proposed a framework for labeling topics based on word vectors and letter trigram vectors. The general label, a chunk containing at least one word in the set of top 10 terms, can be found by means of the similarity between the topic

and its vectors. Although, an accurate phrase topic label is efficient and concise, it may not be enough to express the rich topics.

Recently, attention is being paid to the well-researched automatic summarization technology and exploiting it to provide better topic labels that contain richer content than top terms. Basave *et al.* [14] introduced a framework to address the problems that the summarization method had with the external sources when the relevant topics did not exist there. The algorithms of the given framework are independent and only rely on the identification of dominant words in documents related to the topics learned from Twitter. In order to obtain the ideal topic labels with high Relevance, Coverage and Discrimination, Wan and Wang [2] were the first to propose a novel method of two stages base on submodular optimization to generate sentences with the definite length for each discovered topics. Ultimately, the evaluation results showed that the topic labels generated by their method achieved high level and demonstrated that the use of summaries, as topic labels outperformed the use of words and phrases.

Baralis *et al.* [35] proposed a novel and general-purpose summarizer, GRAPHSUM, which represents correlations among multiple terms by discovering association rules, and adopts a graph-based ranking and a specified strategy to choose significant sentences for summarization. Ren *et al.* [13] presented a novel neural network model to identify features of sentences and automatically learn contextual relations between them. The model scores and chooses sentences based on a greedy algorithm in a single process. Although experimental results show that it is as good as that of Wan and Wang [2], the computational costs of both the models are enormous.

In this paper, we propose a graph-based ranking algorithm to solve the problem of computing inefficiency, and random walks algorithm using PageRank [36], LexRank [29] and TextRank [37] for reference. Brin and Page [36] proposed the PageRank model, which involves a damping factor that can give a small chance to jump randomly to any vertex, as follows, F

$$p(y) = \frac{1-d}{N} + d \sum_{x \in In(y)} \frac{p(x)}{\#Out(x)} \tag{1}$$

where $p(x)$ is the weight of vertex x, N is the number of vertices in the graph, $In(y)$ indicates the set of vertices (predecessors) that point to vertex y, $Out\ (x)$ is the set of vertices (successors) that vertex x points to, and # represents the count of the set. The d is a low probability damping factor, and it aims to jump to any vertex in the graph with a certain random probability from the current vertex.

With damping factor d, the transition matrix (M) in the PageRank model is $((1\text{-}d)/N)\ E+dM$, so that the value of each element in M is more than 0 and M is a positive matrix. It means that M is both primitive and irreducible. Therefore, the PageRank model is strictly a Markov chain, which enables the random walk algorithm to avoid falling into the local optimum trap, and a unique stationary vector can then be obtained by power method [7], [36], [38]. The above analysis also makes sense for the following Eq. (2, 3).

Erkan and Radev [29] proposed a stochastic graph-based method (LexRank) to measure the importance of sentences. In this paper, we utilized LexRank to generate descriptive labels for the discovered topics.

$$p(y) = \frac{d}{N} + (1-d) \sum_{x \in In(y)} \frac{p(x)}{\#Out(x)_{sim(x,y)>t}} \tag{2}$$

where *sim(x,y)* represents the cosine similarity between vertex x and y, $\#Out\ (x)_{sim(x,y)>t}$ denotes the count of vertex set that x point to y when *sim(x,y)* is above the threshold t, moreover, the weight of each vertex is equally distributed to all its adjacent vertices.

TextRank [37] can be seen to some extent as an improvement of LexRank, and the equation is described as follows.

$$p(y) = (1-d) + d \sum_{x \in In(y)} \frac{edge_{xy}}{\sum_{z \in Out(x)} edge_{xz}} p(x) \tag{3}$$

where $edge_{xy}$ represents the similarity between vertex x and y, it is especially important that vertex x votes to y according to the proportion of $edge_{xy}$ in $\sum_{z \in Out(x)} edge_{xz}$, which denotes the sum of weights of edges from x points to its adjacent vertices.

TextRank does not apply an average strategy to distribute weights in the voting process. Thus, the most representative text can be effectively found.

Referring to Eq. (3), if we let $edge_{xy} = 1$ when the *sim(x, y)* $> t$ (t is a given threshold), and let $edge_{xy} = 0$ in other cases, then the TextRank will become LexRank eventually.

Unlike LexRank and TextRank that both vote based on the similarity between vertices, TLRank does that according to the Relevance with the topic, so it effectively avoids the undesirable impact of noise data [29].

III. PROBLEM DEFINITION

A. TOPIC DISCOVERED BY LDA MODEL

The Latent Dirichlet Allocation [1] model is a Bayesian mixture model for a discrete database on the ''word bag'' hypothesis, as an unsupervised modeling technique, it is widely used in the discovery of latent semantic topics in large-scale text collections. A topic discovered by the LDA modeling approaches is a soft-cluster of weighted terms based on their co-occurrence in the documentation set [39].

In our research, according to Hornik and Grün [40], the number of topics k for an LDA model has to be fixed a-priori, and the estimation methods used to fit the model are the VEM (variational expectation maximization) algorithm and Gibbs algorithm.

A topic θ learned by topic modeling is a probability distribution of terms $\{\boldsymbol{p_\theta}\ (\boldsymbol{w})\}_{\boldsymbol{w} \in \boldsymbol{V}}$ where V is a vocabulary set of the corpus. Besides, for each topic θ, we have $\sum_{w \in V}\ p_\theta\ (w) = 1$ [8].

B. TOPIC LABEL

In this paper, for interpreting the discovered topics, we extract the appropriate sentences to generate a more meaningful and concise topic label for each topic. According to Wan and Wang [2] and the recommended definition of the length of the summary in recent TAC and DUC conferences, topic label length is limited to 250 words.

In general, a good topic label should satisfy the following three important criteria [2], [3]:

(1) Semantically relevant, it is well known that a topic label with a higher **Relevance** to the topic would convey more accurate true meaning to the user.

(2) **Coverage**, higher Coverage of topics could bring integrity and fidelity of information to ensure that users can avoid being misled by incomplete information.

(3) **Discrimination**, if a label has Relevance to many topics at the same time, it will become meaningless because of its poor representativeness of the discovered topics. Thus, a higher Discrimination will result in better topic labels.

The overlap between sentences result in redundancy, considering the finite length of the topic label. So, we select sentences that have smaller similarity with each other and stronger representative to others, to mitigate the redundancy and boost the diversity.

C. TOPIC PSEUDO SENTENCE

A centroid is a set of terms to represent a cluster of document collection and illustrate important statistical characteristics. The discovered topics modeled by the LDA-style approaches can be approximately represented with their major features, i.e. choosing top 500 terms of the discovered topic into a term set $V_{top500\theta}$ [2] to alternatively represent the centroid of a discovered topic.

In addition, since a centroid of the discovered topic and a topic label generated by labeling method have different representations, it is hard to directly evaluate the quality of the generated topic labels by computing their similarity [3]. In order to address this issue, we convert the centroid terms set to a pseudo-sentence, which consists of top 500 topic terms and does not contain syntax and grammar. In this way, we can consider a topic pseudo sentence (TPS) as a discovered topic, and it can be described as follows.

$$TPS(\theta) = \{p_\theta(w)\}_{w\in V_{top500\theta}} \Big/ \left\| \{p_\theta(w)\}_{w\in V_{top500\theta}} \right\| \quad (4)$$

where the $p_\theta(w)$ indicates the probability of word w in the topic θ, $V_{top500\theta}$ represents the top 500 terms set of topic θ.

According to the study on extending term vector to sentence vector [41], we apply weighted additive method [42] to present the centroid of the sentence by tfidf in term vector, so its equation is defined as follows.

$$s = \{tfidf(w)\}_{w\in S} \quad (5)$$

where s represents the centroid of the sentence, and S denotes the words set in the sentence.

In addition, the centroid value of a sentence to a specific discovered topic is to sum all the terms conditional probability in that topic [43]. In the light of Graph Theory and Network Analysis, centrality is an indicator to determine the significance of a vertex in a network [44]. Moreover, Erkan and Radev [29] argue that if a sentence contains more words from the centroid of the cluster, it would be considered more central.

IV. THE TLRANK MODEL

Inspired by the studies [2], [29], [37], we propose a novel automatic summarization method base on the ranking of directed complete graph, which is divided into three stages: (1) choose candidate sentences process; (2) TLRank based on overall centrality process, TLRank-C in short; (3) TLRank based on graph-based ranking process, TLRank-G in short.

A. CANDIDATE SENTENCES CHOOSING PROCESS

In the first stage, we select the most relevant sentences to set up the candidate sentences sets for each discovered topic.

Bigi [45] argues that the KLD method outperforms the conventional methods involving the tfidf method. Using KLD measured similarity between the topic label and TPS is an effective approach in topics labeling task according to Wan and Wang [2]. For the KLD, equation is defined as follows.

$$KLD(TPS(\theta), s) = \sum_{w\in SW\cup V_{TPS}} p_\theta(w) * \log \frac{p_\theta(w)}{tf(w,s)\big/len(s)} \quad (6)$$

where $TPS(\theta)$ denotes the TPS of discovered topic θ, s represents the corresponding topic label or sentence in the corpus, V_{Tps} represents the words set of TPS, SW denotes the words set of sentence s after removing digital, stop words and terms with length less than threshold m (minimum term length), $tf(w, s)$ represents the frequency of w in s, and $len(s)$ denotes the count of SW. According to the strategy introduced by Bigi [45], if a word w is not in SW, the $tf(w, s)/len(s)$ would be replaced with $\min\left(\{p_\theta(w)\}_{w\in V}\right)$ that is the smallest probability of a term in topic θ.

In this stage, we apply the approach of Wan and Wang [2] to extract the top 500 sentences from the current corpus for each discovered topic, and then take them as candidate sentences sets (*CSSets*).

B. THE TLRank-C PROCESS

In this section, we first identify the three central features of candidate sentences, and then forge them to overall centrality. According to the ranking order based on the centrality value of sentences, we generate a topic label for each discovered topic.

1) RELEVANCE CENTRALITY

The Relevance centrality (*RelCen*) is a measurement based on the textual similarity between the candidate sentence and TPS, which is computed using the following equation.

$$RelCen(s, \theta) = \exp(KLD(TPS(\theta), s)^{-1}) \Big/ len(s)^a \quad (7)$$

where s represents the candidate sentence, the exponential smoothing parameter a is used to optimize the result,

To *RelCen*, because of the TPS length being fixed at 500, its effect is ignored in Eq. (7), and only the effect of the length of the candidate sentence s is considered.

2) COVERAGE CENTRALITY

It is quite intuitive that a candidate sentence with good Coverage should contain as many different words as possible, which is very similar to the diverse requirements of multi-document summaries.

In addition, according to the study of Arora and Ravindran [46], the sum of probability of different non-repetitive words in the sentence can well reflect the sentence Coverage to the topic. So the Coverage centrality (*CovCen*) equation is shown as follows.

$$CovCen(s,\theta) = \sum\nolimits_{w\in SW} p_\theta(w)\ tf(w,s) \Big/ len(s)^a \quad (8)$$

where s represents a candidate sentence. In a topic, the extents of coverage of different words vary from each other. The larger $p_\theta(w)$ value a word w has, the stronger representativeness of topic it has. Thus, a sentence s will achieve a higher coverage of the topic if the $CovCen(s,\theta)$ has a higher value. According to the coverage evaluation equation of summary in the study of Wan and Wang [2], we argue that Eq. (8) can be a good criterion to consider the coverage of candidate sentences to the topic. It is fully illustrated by the experimental result in Table 2 in Section VI.

According to Eq. (8), the *CovCen* of any candidate sentence would be certainly greater than 0. Thus, it ensures that our proposed TLRank model converges to a unique vector within finite iteration. The more specific details will be further discussed in Section IV (C).

3) DISCRIMINATION CENTRALITY

Referring to Eq. (8), we argue that the importance of a sentence s belonging to the topic θ can be measured by the ratio of the $Cov(s,\theta)$ to $\sum Cov(s,\theta)$. The higher the ratio is, the more important to the topic θ and the more discriminative to other topics the sentence s is. So, the Discrimination centrality (*DisCen*) equation can be written as follows.

$$DisCen(s,\theta) = \frac{\sum_{w\in SW} p_\theta(w)\ tf(w,s)}{\sum_{\theta^*\in U}\sum_{w\in SW} p_{\theta^*}(w)\ tf(w,s)} \quad (9)$$

where U represents a set of all the discovered topics.

4) OVERALL CENTRALITY

The three centrality features of the candidate sentence (*RelCen*, *CovCen* and *DisCen*) are corresponding to the three criteria for evaluating the quality of topic labels (Relevance, Coverage, and Discrimination). The purpose is to identify the three centrality features that aim to find suitable sentences and hence improve the three evaluating scores of the topic labels generated by our method. In order to use a scale value to measure the sentences in the ranking process, we introduce a new concept, overall centrality (*OC*). This combines the three centrality features to estimate the overall quality of sentences and generates summaries. This way, we achieve a satisfactory balance among the three evaluating criteria. Moreover, the overall centrality of sentence s_y (OC_y) equation is defined as follows.

$$OC_y = \alpha RelCen(s_y,\theta) + \beta DisCen(s_y,\theta) + (1-\alpha-\beta)CovCen(s_y,\theta) \quad (10)$$

where α and β are proportion parameter to be empirically set, and it has $\alpha > 0, \beta > 0, \alpha+\beta < 1$.

C. THE TLRank-C PROCESS

In this section, candidate sentences in *CSSets* are used as vertices to create a directed complete graph, and a novel graph-based ranking algorithm is given accordingly. We will accomplish the following tasks. First, a transition matrix (*TM*) is established according to overall centrality value, and then, the transition probability TM_{xy} between vertex-pairs is suppressed according to their relationship and enhanced base on their *Degree* (a characteristic of each vertex). Thence we can ensure that *TM* is the transition matrix of a Markov chain and the random walk algorithm can converge to a stable state after a finite number of iterations. Finally, we generate a topic label for each discovered topic based on the TLRank-G score.

1) DIRECTED COMPLETE GRAPH

Consider $G = (Vertices, Edges)$ to be a directed complete graph, $x, y \in Vertices$, Cand $x \neq y$, C $edge_{xy} \in Edges$ represents the weight of edge from x points to y and its original value can be derived from the overall centrality of vertex y (sentence s_y). Here, it has $edge_{xy} = OC_y$, which means that vertex x votes to y based on the overall centrality value of vertex y. In this graph, the weights of edges are the critical factor in determining the transition matrix, which directly defines the output of the graph-based ranking algorithm. Therefore, it is feasible to modify the weights of edges in the graph to adjust the ranking results.

2) SUPPRESS WEIGHT OF EDGE

In order to refrain the redundancy of the topic label during the generating process, the sentence which has the smallest similarity with those existing in the topic label set should be chosen preferably [2], [25]. Therefore, one of the important tasks of our ranking method is to ensure that the sentence in the back position of the ranked order keeps the similarity as small as possible with the ones ahead.

In this paper, we use the Jaccard distance to measure the similarity between two vertices (sentences) in the graph. In general, for two sentences s_x and s_y, the Jaccard distance equation is defined as follows.

$$similarity\_Jaccard(s_x,s_y) = \cap(s_x,s_y) \Big/ \cup(s_x,s_y) \quad (11)$$

If vertex x (s_x), and y (s_y) in the graph are similar, and it has $OC_x > OC_y$, then the $edge_{xy}$ should be suppressed, and the suppression coefficient bases on the value

of similarity_Jaccard(s_x, s_y). The $edge_{xy}$ which has been suppressed can be described as follows.

$$edge_{xy} = edge_{xy} \Big/ e^{similarity\_Jaccard(s_x, s_y)} \tag{12}$$

3) ENHANCE WEIGHT OF EDGE

If the similarity between two vertices surpasses the threshold t, the *Degree* count of both vertices is increased by 1 [29]. If a vertex has a higher *Degree*, it means that the vertex is similar to much more vertices and has stronger representativeness. Thus, the corresponding sentence should be chosen preferentially into the topic label set.

In order to avoid the effects of sentence length, the *Degree* of vertex should be normalized before using. We exploit the following equation, shown as Eq. (13). Then enhance $edge_{\bullet y}$ which represents a set of all adjacent vertices that point to vertex y to improve the representativeness of vertex y. We derive the following equation, shown as Eq. (14).

$$Degree_y = Degree_y \big/ len(s_y)^a \tag{13}$$

$$edge_{\cdot y} = edge_{\cdot y}\, r^{(Degree_y / \sum_{x \in CSSets} Degree_x)^a} \tag{14}$$

where $Degree_y$ is *Degree* of vertex y, $edge_{\cdot y}$ represents the weights set of all the vertices pointing to vertex y, *CSSets* denotes the candidate sentences sets, $\sum_{x \in CSSets} Degree_x$ represents the sum of *Degree* value of all vertices in graph G, γ indicates a base number parameter to be set empirically. In addition, if the $Degree_y$ is zero, the $edge_{\cdot y}$ would be unchanged.

4) GRAPH-BASED RANKING PROGRESSING

According to Eq. (8-10, 12), $edge_{xy} > 0$ and $edge_{xy} \neq edge_{yx}$ in any conditions. That ensures creating a weighted directed complete graph. Furthermore, it obtains the global optimal results and avoids falling into local optimal issues in the graph-based ranking process. For TLRank-G ranking, the score of a vertex y is defined as follows.

$$p(y) = \sum_{x \in In(y)} \frac{edge_{xy}}{\sum_{z \in Out(x)} edge_{xz}} p(x) \tag{15}$$

Considering Eq. (3), if we let damping factor $d = 0$, TLRank will look like a simplified version of TextRank [37]. In order to figure out why Eq. (15) can eventually converge to a stable value, we will discuss it below.

According to Erkan and Radev [29], we can get the transition matrix from Eq. (15), and it can be written as follows.

$$TM_{xy} = edge_{xy} \Big/ \sum_{z \in Out(x)} edge_{xz} \tag{16}$$

Based on the above analysis, we can consider that $TM_{xy} > 0$ in any case, and therefore it is certain that *TM* (a stochastic transition matrix) must be a positive matrix. Thus, because *TM* is an irreducible and aperiodic Markov chain, our graph-based algorithm will converge to a stable state within finite iterative steps [29]. For this reason, refer to Eq. (1-3), the "damping factor" d is bypassed in TLRank model.

It should be noted that the hyper-parameters involved in all the equations in this paper will be described in detail in the following experimental part, Section V(A).

V. EXPERIMENT

A. EXPERIMENT SETUP

1) DATA SETS

We also used two different public document collections: SIGMOD[1] and APNews[2] [2], [3]. After a series of pretreatment involving stemming, removing stop words, etc., in total there were 3016 and 2246 documents, 22105 and 46043 sentences, 8252 and 25902 vocabularies in the SIGMOD and APNews respectively.

2) TOPIC DISCOVERY

In this paper, all methods are implemented based on R. we utilized both R tools packages: topicmodels and tm[3] to model the two corpora, and use two estimation methods: VEM and Gibbs for the LDA function. For the Gibbs method, it is necessary to fix the parameter $k = 25$ and set the parameters $burnin = 1000$, $thin = 100$, $iter = 1000$. If two estimation methods apply in two corpora, we obtain four different cases, i.e. SIGMOD VEM, SIGMOD Gibbs, APNews VEM and APNews Gibbs. Finally, we learn topics using LDA models in each case.

3) MODEL PARAMETERS

The parameter values in our experiment directly borrowed from previous research or empirically set are $a = 0.85, \alpha = 0.3, \beta = 0.4, \gamma = 2.25, t = 0.2$ and $m = 3$. Especially, the parameter a usually can be adjusted in the range 0-1 according to the experimental dataset.

B. LABELING MODELS

To demonstrate the effectiveness of our approach, some classical and state-of-the-art models are compared experimentally. *CSSets* serve as input to all comparison labeling models, and the outputs of each labeling model is a topic label with the finite length for all discovered topics. All comparison labeling models are listed as follows.

1) BASELINE

The top sentences in the *CSSets* can be directly extracted to generate the topic label (label-B). It means that the Baseline approach only considers the Relevance feature to choose sentences for generating a topic label.

2) LexRank

The LexRank[4] model applies a strategy of vote equally to related vertices in the graph-based ranking process.

[1] A total of 3016 SIGMOD summaries (1975-2018) downloaded from ACM DL, https://dl.acm.org/citation.cfm?id=500080& picked = prox

[2] A total of 2246 APNews articles downloaded from GitHub, https://github.com/Blei-Lab/lda-c/blob/master/example/ap.tgz

[3] all the R tools packages we used in the experiment are downloaded from R Archive Network, https://cran.r-project.org/web/packages/

[4] LexRank Model implemented by lexRankr (R tools package)

It proposes a graph-based automatic summarization method for the first time [29]. Besides, it can also find the most representative sentences to generate topic labels for the discovered topics.

3) TextRank

TextRank[5] is regarded to some extent as an improvement of LexRank. In this case, if two vertices are more similar, they vote more to each other [37].

4) SUBMODULAR

It is a model of automatic labeling topic based on submodular optimization, called Submodular. It scores and chooses sentences based on a greedy algorithm in a single process [2].

5) TLRank-C

As mentioned in Section IV (B), TLRank-C process is the first phase of TLRank, and generates a topic label (label-C) based on over centrality of candidate sentences.

6) TLRank-G

As mentioned in Section IV (C), TLRank-G process is the last phase of TLRank, and generates a topic label (label-G) based on the graph-based scores ranking.

VI. RESULTS AND DISCUSSION

In this section, we firstly evaluated the different labeling methods with automatic measures in four different cases, SIGMOD VEM, SIGMOD Gibbs, APNews VEM and APNews Gibbs. At the end, we evaluated the result via human evaluation. It is to be noted that the length of the topic label is defined as 250 by default.

A. EVALUATION OF LABELING METHODS

Our evaluation metrics mainly followed the framework proposed by Wan and Wang [2], applying four measures as follows:

1) RELEVANCE

For each labeling method, we computed the Relevance (using KLD method) between the topic labels and corresponding TPS in the four different cases and then averaged the Relevance of all topics in each case. The results are presented in Table 1 revealed that TLRank-G had the lowest value in case of SIGMOD Gibbs and APNews VEM, and second only to the minimum in other cases, while the gap between them was minimal [2].

In particular, we found that the Relevance of TLRank and Baseline were very close in the Gibbs estimation mode, and the Baseline method even had a weak lead in the case of APNews Gibbs. The reason can be referred to Fig. 4(D). It is interesting that although the *RelCen* feature is the only criteria for the Baseline method to choose sentences, TLRank still has the optimal Relevance value under the most cases of different topic labels lengths. This is because TLRank does its best to improve the Relevance through effectively controlling the redundancy and boosting the diversity of the topic label. This issue is further discussed in following Section VI (D).

[5]TextRank Model implemented by textrank (R tools package)

TABLE 1. Relevance of topic label; the average of KL divergence between topic label and discovered topic.

	SIGMOD VEM	SIGMOD Gibbs	APNews VEM	APNews Gibbs
Baseline	0.813665	2.288893	0.401759	**0.996711**
LexRank	**0.634523**	2.553759	0.431941	1.109661
TextRank	0.705769	2.759404	0.535805	1.265098
Submodular	0.814613	3.315231	0.633104	1.57551
TLRank-C	0.698133	2.212693	0.338974	1.006217
TLRank-G	0.696214	**2.204864**	**0.333005**	1.000132

In general, according to the experimental results, the TLRank model outperforms other methods and is stable and available in all four cases.

2) COVERAGE

For each labeling method, we computed the Coverage -the ratio of the top 20 topic terms in the corresponding topic label- for each topic in the four different cases. Then we fetched the average, min, and max from the Coverages of all topics in each case. The reason that choosing the top 20 terms instead of 500 was that the top 20 terms were more significant than the rest and we paid more attention towards the more representative top ones [2].

The results presented in Table 2 show that our model performed better than the others in most cases. Even in the worst cases of SIGMOD VEM and APNews Gibbs, TLRank ranks second and very close to the first. One possible reason is that because the length of the topic label is limited to 250, the number of sentences contained in the topic label is relatively small that indirectly limits the space to improve Coverage by the TLRank. Another reason could be that of the topic label where the Coverage is computed by the top 20 terms of each topic, rather than by the 500 terms applied to compute the *CovCen* of candidate sentence. Therefore, the existing deviation may be reasonable, and it is expected that the accidental decrease of Coverage is due to the combination of these factors as mentioned above. This issue of Coverage is further discussed in Section VI (D) that follows.

3) DISCRIMINATION

For each labeling method, we computed Discrimination (the cosine similarity between the topic label and the corresponding TPS) for each topic in the four different cases. Then we fetched the average, min, and max from the Discriminations of all topics in each case [2].

As obvious from Table 3, Submodular and TLRank are the best among the labeling methods. Further, the Discrimination value of TLRank is much closer to the Submodular when TLRank ranks second. However, the gap gets obviously wider with others when TLRank ranks first.

TABLE 2. Coverage of topic label; the mean, min, and max ratio of the top 20 topic terms in the corresponding topic label.

	SIGMOD VEM			SIGMOD Gibbs			APNews VEM			APNews Gibbs		
	Mean	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max
Baseline	0.676	0.550	0.800	0.700	0.600	0.850	0.662	0.450	0.850	**0.704**	0.450	0.900
LexRank	**0.744**	0.600	0.900	0.626	0.400	0.800	0.616	0.350	0.800	0.624	0.400	0.850
TextRank	0.708	0.450	0.900	0.566	0.300	0.800	0.548	0.200	0.750	0.570	0.250	0.800
Submodular	0.614	0.250	0.850	0.496	0.050	0.750	0.382	0.050	0.700	0.462	0.100	0.800
TLRank-C	0.718	0.550	0.900	**0.708**	0.550	0.850	0.734	0.550	0.900	0.702	0.450	0.900
TLRank-G	0.722	0.550	0.900	0.706	0.550	0.850	**0.742**	0.550	0.900	0.702	0.450	0.900

TABLE 3. Discrimination of topic label; the mean, min, and max cosine similarity between the topic label and corresponding discovered topic (OR TPS).

	SIGMOD VEM			SIGMOD Gibbs			APNews VEM			APNews Gibbs		
	Mean	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max	**Mean**	Min	Max
Baseline	0.09937	0.00275	1.00000	0.04665	0.00636	0.67238	0.04768	0.00026	0.69225	0.02291	0.00000	0.26283
LexRank	0.10032	0.00325	0.62277	0.06258	0.00254	0.64343	0.05343	0.00000	0.46150	0.03477	0.00000	0.28211
TextRank	0.13545	0.00789	0.62671	0.05618	0.00400	0.36151	0.06715	0.00000	0.78909	0.03045	0.00000	0.38393
Submodular	**0.06596**	0.01917	0.24103	0.05077	0.01118	0.19281	**0.03788**	0.00182	0.34464	0.03314	0.00285	0.16205
TLRank-C	0.07152	0.01477	0.24125	0.04651	0.00564	0.36907	0.03820	0.00019	0.49470	**0.02181**	0.00000	0.19026
TLRank-G	0.07136	0.01523	0.24517	**0.04621**	0.00473	0.21219	0.03790	0.00020	0.50407	0.02214	0.00000	0.20545

TABLE 4. Duplication of topic label; for all topic-label-pairs, count the number of duplicated sentences and sum them up to the "Dup Number", accumulate the number of sentences contained in all topic labels to the "Total", and the "Dup Ratio" indicate the ratio of "Dup Number" against "Total".

	SIGMOD VEM			SIGMOD Gibbs			APNews VEM			APNews Gibbs		
	Dup Number	Total	**Dup Ratio**	Dup Number	Total	**Dup Ratio**	Dup Number	Total	**Dup Ratio**	Dup Number	Total	**Dup Ratio**
Baseline	572	76	7.52632	2	195	**0.01026**	70	164	0.42683	2	183	0.01093
LexRank	1398	283	4.93993	88	312	0.28205	202	271	0.74539	50	295	0.16949
TextRank	1790	328	5.45732	60	373	0.16086	596	384	1.55208	58	371	0.15633
Submodular	28	209	**0.13397**	6	270	0.02222	16	257	**0.06226**	0	180	**0.00000**
TLRank-C	426	406	1.04926	4	238	0.01681	60	198	0.30303	2	196	0.01020
TLRank-G	452	410	1.10244	4	260	0.01538	68	202	0.33663	4	213	0.01878

In addition, comparing with the Baseline method, we found that the improved effectiveness of our approach in the case of Gibbs was much lesser than that in the case of VEM. The reason can be seen from Table 5. In the same corpus, between different topics learned in the case of Gibbs, the similarity is much smaller than that in the case of VEM. In other words, VEM preserves the connections between topics, while Gibbs pays more attention to the differences between topics. Therefore, it left very limited room for TLRank to improve the Discrimination when the discovered topics are quite different from each other in the case of Gibbs.

According to Table 5, the topics of APNews are more differentiated than those of SIGMOD, which is consistent with our intuition. News corpus is more divergent than collections of scientific literature. So the clustering centers naturally will be more diverse. As evident from Table 3, in the case of APNews Gibbs, TLRank provides the best result without too much effort, though all other methods also perform well. That shows that our method is effective and stable in terms of improving the Discrimination between topic labels.

4) DUPLICATION

The label length and noisy data can easily interfere with the use of similarity to measure the Discrimination between topic labels. In this regard, the present study proposes an intuitive method to measure the Discrimination quality of the topic label considering the number of duplicated sentences in topic label set for each topic.

The method is described in the following steps: (1) for all topic-label-pairs, count the number of duplicated sentences and sum them up to the given ''Dup Number''; (2) accumulate the number of sentences contained in all topic labels to the ''Total''; (3) let ''Dup Ratio'' value is that ''Dup Number'' divided by ''Total'', if the value is big, it means that there is serious confusion between different topic labels generated by this method, and user can hardly capture the difference between topic labels of each topic.

As evident from Table 4, the ''Total'' value of the Baseline is lower than other methods in all four cases. The fixed length of the topic label depicts that the Baseline method prefers longer sentences. A comparison of Table 1, 2, 3, and 4 shows that if the topic label contains fewer sentences, it commonly has a terrible Relevance, Coverage, and Discrimination. In addition, for the same corpus, the ''Dup Ratio'' is smaller in the case of Gibbs estimation mode, which is consistent with the information in Table 5.

Compared with other approaches, LexRank and TextRank always have the highest values of ''Dup Number'' under all

TABLE 5. Summary of the Jaccard similarity among topic with TPS (top 500 terms).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
APNews VEM	0.1086	0.1820	0.2114	0.2158	0.2469	0.4085
APNews Gibbs	0.0363	0.0579	0.0707	0.0749	0.0870	0.1574
SIGMOD VEM	0.3928	0.4368	0.4556	0.4522	0.4684	0.5038
SIGMOD Gibbs	0.0493	0.0695	0.0788	0.0785	0.0881	0.1198

cases. It suggests that both methods do not have to deal with the redundancy in the labeling process, at the same time. The highest ''Total'' value suggests that they prefer the shorter sentences. Generally, for the ''Dup Number'' in Table 4, TLRank is always second to the best and for the ''Dup Ratio'', it is the best except for the Submodular model.

According to the above experimental results, our method is the best, practical, stable, and outperforms the contrast methods.

5) MANUAL EVALUATION OF TOPIC LABELS

We evaluated the results of four different labeling methods (LexRank, TextRank, Submodular, and TLRank) via manual evaluation in the two different cases, SIGMOD VEM and SIGMOD Gibbs.

According to Wan and Wang [2] and Kou *et al.* [5], we had four human annotators who manually score the topic labels generated. In addition to offering the collection of relevant documents that is necessary to help understand the topics, we provide each human annotator with the top 20 terms of each topic and corresponding topic labels generated by different methods. It should be noted that the topic labels we provide to each annotator are anonymous. The annotators do not know which labels were generated by which methods.

Besides, we require each annotator to consider the following three aspects when they are scoring the topic labels: the Relevance between the label and the corresponding topic, the Coverage of the topic label, the Discrimination between the different topic labels. It needs annotators to score the topic labels in three aspects separately, and then average the three scores for each topic label.

Regarding the manual evaluation of topic labels generated, previous works mainly adopt the Likert scale, a common rating format of evaluation or scoring [47]. There are two main approaches in common use: the scale of five points [48] and four points [2], [5], [49]. Our scale of four points is described in Table 6. The scores of topic labels generated range from 0 to 3, with 0 representing the worst, and 3 the best. Besides, the floating-point number is allowed as a score.

Finally, we average the scores across all topics from the four annotators. The overall scores of the topic label in the cases of SIGMOD VEM and SIGMOD Gibbs are shown in Table 7. According to the experimental results, the higher

TABLE 6. Our Likert scale of four points.

Score	Description
3	The topic label is perfect.
2	The topic label is reasonable.
1	The topic label is semantically related to the topic, but it is not a good label.
0	The topic label is the worst.

TABLE 7. Manul evaluation, the average of scores between topic label and discovered topic in the cases of sigmod vem and sigmod Gibbs.

	SIGMOD VEM	SIGMOD Gibbs
LexRank	1.567	1.288
TextRank	1.476	1.148
Submodular	1.650	1.094
TLRank	**1.662**	**1.458**

the scores are, the better the topic labels are. It shows that our method outperforms the contrast methods in all cases.

Furthermore, it seems that TLRank is more appropriate to label the discovered LDA-style topics when using the VEM estimation method. This issue will be further discussed in the following Section VI (D).

6) TOPIC COHERENCE AND TOPIC LABEL EVALUATION

The quality of topic labels generated is affected by the discovered topics. If a topic has rich content, the topic label generated will appear extensive, and the consistency between the different sentences in the label will be lower.

Generally, we can apply coherence measures on each topic to measure the quality of the topic label generated [50]. According to Lau and Baldwin . [51], there is a significant impact of cardinality hyper-parameter on topic coherence. In our study, we apply the approach from Wan and Wang [2] and set the cardinality as 500, so the topic label generated will be more diverse.

If a topic label has a higher semantic Relevance with the discovered topic, it can be better to represent and explain the topic. The topic label generated could clearly and accurately convey the true meaning to the audiences and help them understand the topic.

Higher Coverage brings integrity and fidelity of information. The topic label with a higher Coverage can ensure that the audiences do not misunderstand or even distorts the real intention of the topic.

There are not only differences but also intrinsic connections among the topics, which are inherent in the content carried by most documents in the same corpus. Especially for scientific literature (SIGMOD), there is certain inheritance and evolution relationship among some topics due to the crossover and development of disciplinary knowledge.

Generally, we find that there exist some connections between the topics themselves (e.g., common foundation or evolutionary relationship), and the connections are usually expressed by overlapping of a certain number of top topic

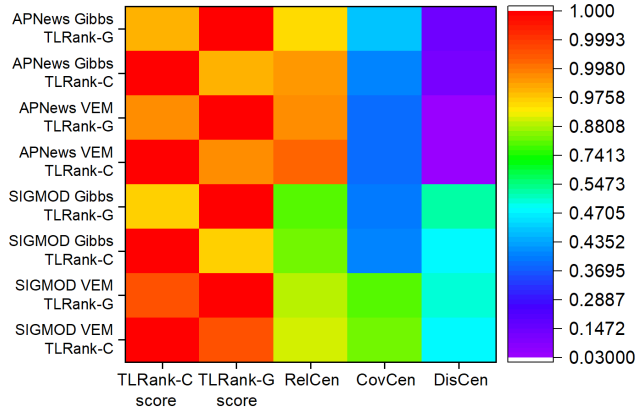

FIGURE 2. The average value of the Pearson correlation coefficient between TLRank score and the three centrality features (*RelCen*, *CovCen*, *DisCen*) of candidate sentences for all topics.

terms, which have different probability distributions in the corresponding topics. Therefore, the more duplications in topic labels generated between different topics, the more likely it is to confuse the meaning of each topic and lead to poor representativeness of the topic label.

B. TLRANK SCORE VS. RELCEN, COVCEN AND DISCEN

Here, the TLRank-C score and TLRank-G score are separate outputs of the two phases of TLRank, called TLRank score briefly.

To some extent, the Pearson correlation coefficient can be used to reflect the degree of correlation between the candidate sentence features and the output of the two stages of TLRank. As shown in Fig. 2, the TLRank score has a strong positive linear correlation with *RelCen* and *CovCen*. In particular, for the SIGMOD corpus, there is a certain degree of correlation between TLRank score and *DisCen*, while for APNews corpus, there is no correlation between them.

We know that the Pearson coefficient can only reflect a linear correlation. Hence, the scatter plots were employed to visualize the relationship between TLRank feature space and the centrality features of candidate sentences. The aim was to explore how to use TLRank score to find the sentences with high Relevance, Coverage and Discrimination in the case of APNews VEM. The reason to choose the case of APNews VEM was that in Fig. 2, the R-value between *DisCen* and TLRank-C score was the lowest.

As shown in Fig. 3(A, D), for most of the vertices, a higher TLRank-C score has a higher *RelCen*. The two types values have a linear positive correlation distribution. Besides, according to Fig. 3(B, E), most of the vertices with higher TLRank-C score show a positive correlation with their *CovCen* values.

In this experiment, the k is set to 25 for LDA discovery [1]. If a sentence belongs to each topic equally, the probability of it to represent each topic is equal to 0.04 (1/25). In Fig. 3(C), most of the vertices (sentences) are scattered irregularly on the left and right sides with 0.04 axis, and

TABLE 8. The amplitude of changes in ranking position of candidate sentences after TLRank-G processing. the column with the heading unchanged shows the number of sentences without position change.

	Min.	1st Qu.	Median	3rd Qu.	Max.	Unchanged
APNews VEM	-63	-7	1	6	45	30
APNews Gibbs	-136	-13	2	15	110	17
SIGMOD VEM	-14	-2	0	2	12	103
SIGMOD Gibbs	-75	-8	0	7	72	26

TABLE 9. The top 20 terms of a topic in the case of APNews Gibbs.

Index	Topic terms				
1	million	company	said	bank	billion
6	offer	will	corp	new	business
11	inc	plan	sale	share	also
16	firm	financi	manag	stock	invest

R^2 =–0.00164. It means that there is no correlation between them. In Fig. 3(F), the average of *DisCen* of any sentence can only be 0.04, so the correlation is meaningless.

From the above facts, the sentence with high TLRank scores mostly has high *RelCen* and *CovCen*. It means using TLRank score can distinguish significant sentences effectively. As obvious from Fig. 3(A, B, C), the TLRank model tends to choose the critical sentences in the upper right, while ignoring the useless sentences in the lower left. However, this remarkable ability is still limited due to the fact that no apparent correlation relationship exists between the TLRank-C score and *DisCen* feature of candidate sentences and the Discrimination of topic label is still primarily limited to a large extent by the quality of discovered topics.

Finally, according to Fig. 3 and Tables 1 to 4, in comparison with other labeling methods, the TLRank model can generate a topic label with high Relevance, Coverage, and Discrimination based TLRank score without directly using the RelCen, CovCen, and DisCen of the candidate sentences.

C. TLRank-C VS. TLRank-G

According to Fig. 2, there is a strong correlation between TLRank-C score and TLRank-G score. Meanwhile, it can be observed intuitively from Fig. 4 that for most vertices, the ranking position of TLRank-G score does not change much compared with that of TLRank-C score. This is also confirmed by the data in Table 8.

To further illustrate the improvement effect of TLRank-G on TLRank-C results, an example is given below.

For convenience, unless the user wants to see all 500 terms of a topic, only the top 20 terms of this topic are shown in Table 9. In addition, it should be noted that all these terms are stemmed words.

TLRank-C and TLRank-G generate a topic label of 100 words length for this topic, denoted as label-C and label-G respectively. The details are shown in Table 10 and Table 11.

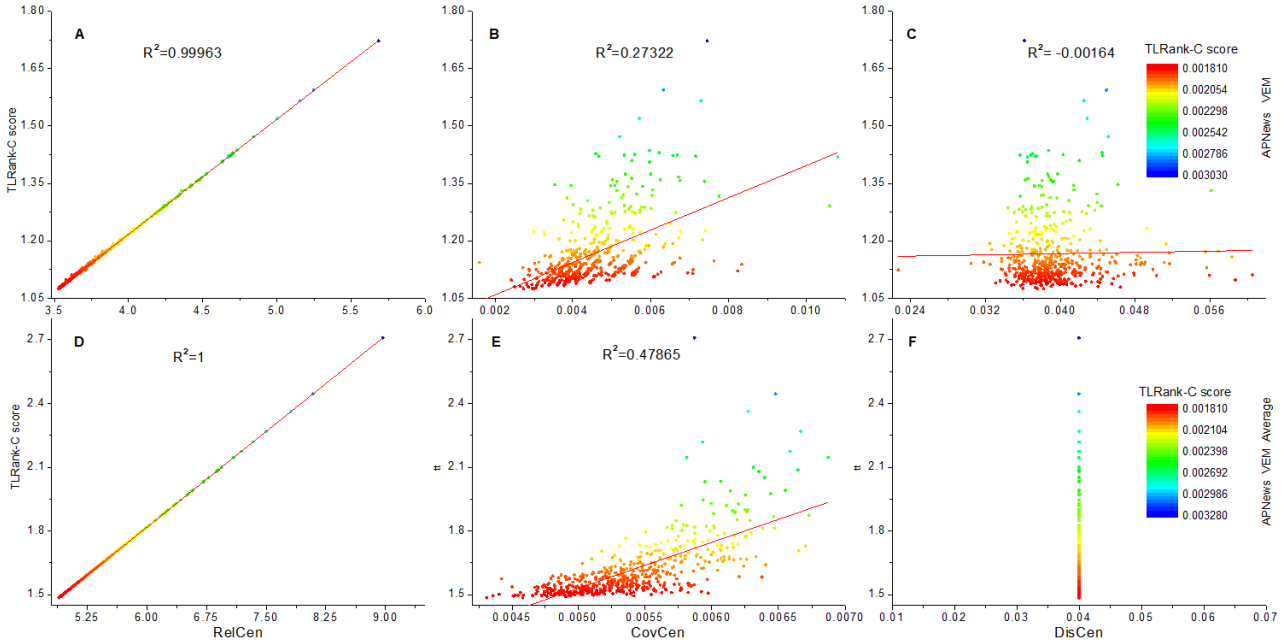

FIGURE 3. TLRank score vs. *RelCen*, *CovCen*, and *DisCen*. Each vertex represents a sentence in *CSSets* (candidate sentences sets for all discovered topics), the color depth represents the value of the TLRank-G score; A, B, C in the case of APNews VEM under a specific topic, D, E, F in the case of APNews VEM Average (average value of all topics).

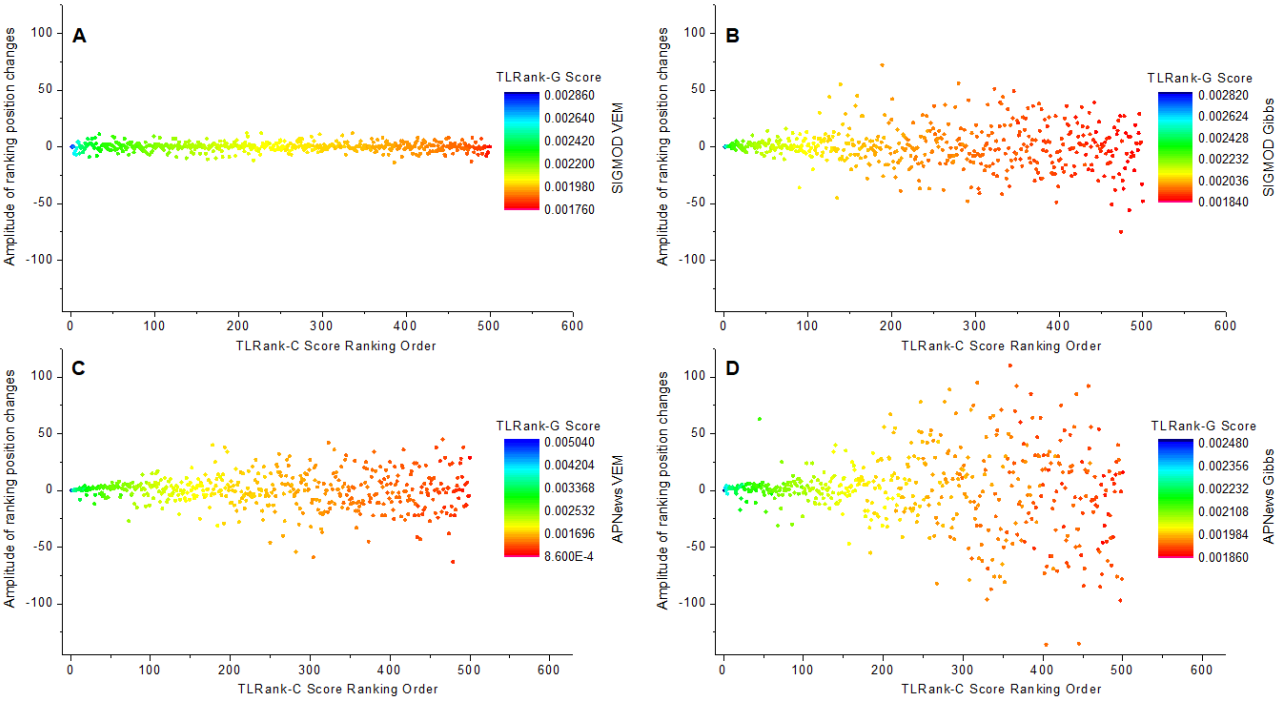

FIGURE 4. The vertices represent 500 candidate Sentences in *CSSets* under a specific topic, The X-axis denotes the order of the sentences ranked by TLRank-C score; and the Y-axis presents the amplitude of ranking position changes, where the negative value indicates forward displacement and the color depth represents the value of the TLRank-G score. The A, B, C, D denote the ranking position change of TLRank-C to TLRank-G score after the graph-based ranking process in the cases of SIGMOD VEM, SIGMOD Gibbs, APNews VEM, and APNews Gibbs respectively.

Fig. 4 reveals that for most of the top 100 vertices, the ranking positions have small changes, and cluster together on both sides of the axis. In particular, as shown in Fig. 4 (B, C, D), for a vertex out of top 100, if the ranking changes significantly, there may be two reasons. One could be that it is greatly affected by the TLRank model, and the other could

TABLE 10. The 100 words-length topic label generated by TLRank-C under a specific topic in the case of APNews Gibbs, where the column "Label-C Sentences" shows all the sentences in label-C, column "C" presents the index of sentences in label-C, column "G" denotes the index of sentences in label-G and # means that the current sentence is not contained in label-G.

C	Label-C Sentences	G
1	The **company said** Thursday the groups, CDI Holdings **Inc**. and DC Acquisition **Corp**., will pay $22.50 per **share** to acquire about 6.95 **million share**s of American Health **stock**.	1
2	He **said** the **company** had **offer**ed to inject $500 **million** last November through a **stock** repurchase **plan**.	2
3	Florida investor Paul Bilzerian has acquired 2 **million share**s of the Hadson **Corp**. as part of Hadson's agreement to purchase HRB Holdings **Inc**., a Singer Co. division, **company** officials **said** Monday."	#
4	In contrast, Blockbuster has been expanding by 400 to 500 stores a year since Huizenga bought the **company** in 1987 for $19 **million**...	#

TABLE 11. The 100 words-length topic label generated by TLRank-G under a specific topic in the case of APNews Gibbs.

G	Label-G Sentences	C
1	The **company said** Thursday the groups, CDI Holdings **Inc**. and DC Acquisition **Corp**., will pay $22.50 per **share** to acquire about 6.95 **million share**s of American Health **stock**.	1
2	He **said** the **company** had **offer**ed to inject $500 **million** last November through a **stock** repurchase **plan**.	**2**
3	He **said** Hadson acquired the **company** for $137 **million** in cash and 2 **million share**s of common **stock**.	#
4	The **company said** its second-quarter loss came on **sale**s of $198.4 **million**, and compared with earnings of $1.6 **million**, or 20 cents per **share**, on **sale**s of $49.6 **million** in the same period a year earlier.	#
5	Florida ...	**3**

be the cumulative effect of change caused by the vertices that rank ahead of it. It seems that there is a certain distortion situation in Fig. 4. However, because of the limitation to the topic label length, TLRank always ignores those interfering sentences with the lower ranking position.

Besides, as obvious from Fig. 4(A), the ranking position of 103 vertices has not changed (see Table 8), while the amplitude of ranking position changed and the rest vertices varied within 15. In Fig. 4(A) and Fig. 4(B), it can be seen that for the different LDA model approaches, the improved effect of the TLRank-C is still different even in the same corpus.

According to Table 10 and 11, label-C and label-G contain 4 and 5 sentences respectively, and the 3rd and 4th sentences of both labels are different. By comparing the contents of both labels, it can be seen that label-G has more top 20 terms, e.g. the word "sale" in label-G does not appear in label-C. Therefore, it seems that label-G has a higher Coverage than label-C.

Moreover, due to the restriction of topic label length, only the first word is reserved in the 5th sentence of label-G. If the label length is allowed to increase, the sentence will contain more words and get completed. In particular, according to Table 4 and Table 8, TLRank-G prefers to use more sentences

TABLE 12. Comparison of relevance, coverage, and discrimination of the 100 words-length topic labels (label-C and label-G) under a specific topic in the case of APNews Gibbs.

	Relevance	Coverage	Discrimination
label-C	1.53469	0.45000	0.01471
label-G	1.48780	0.50000	0.01456

TABLE 13. Comparison the number of sentences contained in the 100 words-length topic labels (label-C and label-G) under a specific topic in the case of APNews Gibbs.

	Min.	1st Qu.	Median	**Mean**	3rd Qu.	Max.
label-C	1.00	3.00	3.00	**3.36**	4.00	5.00
label-G	2.00	3.00	4.00	**3.68**	5.00	5.00

to generate topic labels than TLRank-C. Thus, it enables to reduce sentence overlap and boost the diversity.

After the TLRank-G processing, the sentences with original ranking 6th and 5th rise to 3rd and 4th in label-G, while the sentences with original ranking 3rd dropped to 5th in label-G.

Compared with the TLRank-C dealing with the label-C, TLRank-G adjusts the sentences of label-G to refrain the redundancy, and improve its Relevance, Coverage, and Discrimination. The details are provided in Table 7.

Therefore, combined with the analysis in the first two sections, the suppression and enhancement strategy of TLRank-G is reasonable and effective, the ranking results are stable and available, and the improvement to TLRank-C results is sufficient and significant.

D. COMPRESSION RATIO VS. VEM AND Gibbs

According to the analysis in the previous section, the number of critical sentences screened by the TLRank score is limited, as topic label length increases. How to choose more sentences is a severe test of the stability of the TLRank model.

To further understand the impact of compression ratio on the quality of topic label generating, we present a set of values: LenSet = {50, 100, 150, 200, 250, 500, 1000, 2000, 3000, 4000, 5000}. In case of different lengths in the LenSet, we plot line diagrams to visualize the difference between TLRank and Baseline method in terms of Relevance, Coverage, and Discrimination, and to observe the improvement of TLRank over Baseline, as shown in Fig. 5 and 6.

In particular, in the case of Gibbs, the improvement level of TLRank against Baseline in terms of three indicators is in an irregular fluctuation state with the length changes. However, in the case of VEM, except for three negative cases (exceptions), the TLRank always has the best performance. Additionally, the overall improvement effect is still significant, which shows that our method is efficient and stable.

The anterior two exceptions occur when the topic label length is equal to 1000 (see Fig. 5(B)), and is equal to 4000 (see Fig. 6(B)), while the Coverage of TLRank method is overtaken by Baseline with a weak advantage.

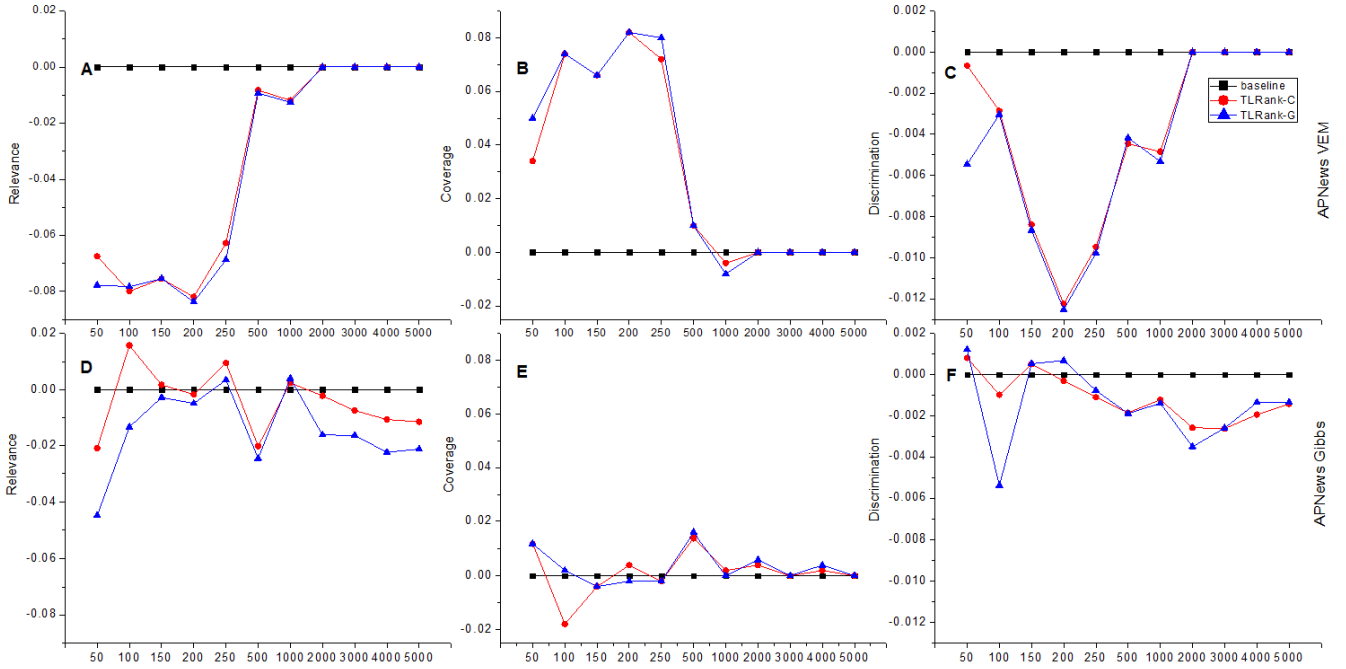

FIGURE 5. The difference between TLRank (TLRank-C and TLRank-G) and Baseline method in terms of Relevance, Coverage, and Discrimination; A, B, C is in the case of APNews VEM, D, E, F is in the case of APNews Gibbs.

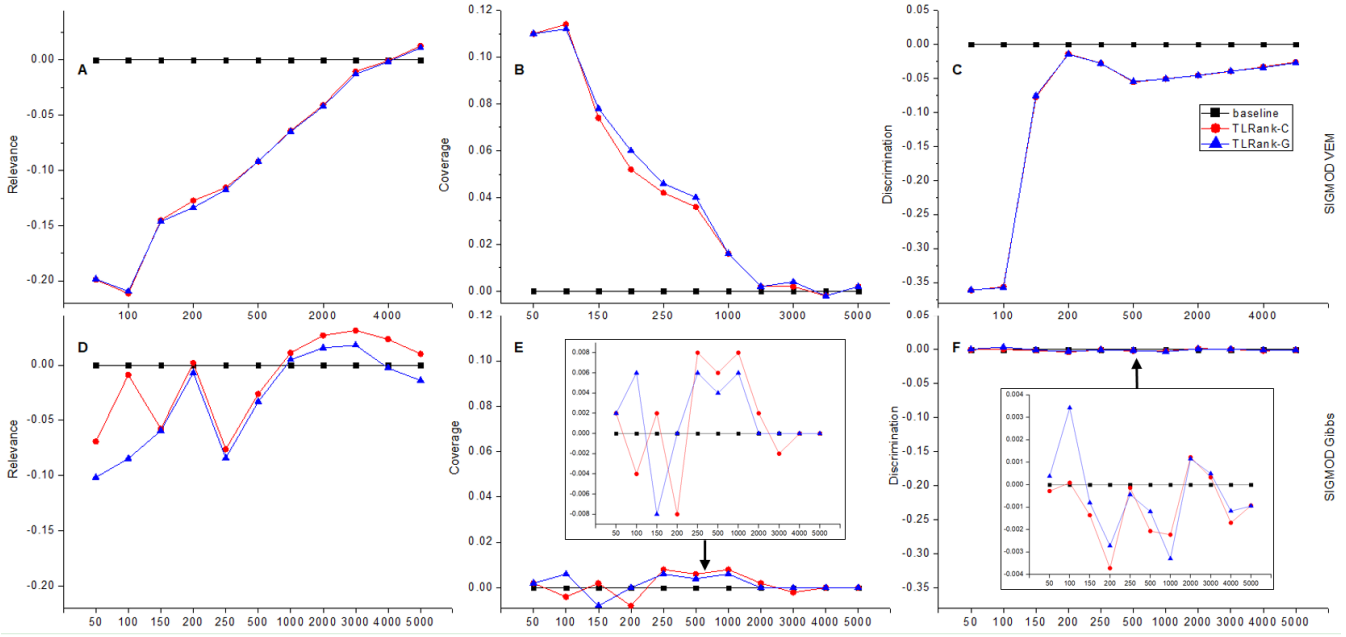

FIGURE 6. The difference between TLRank (TLRank-C and TLRank-G) and Baseline method in terms of Relevance, Coverage, and Discrimination; A, B, C is in the case of SIGMOD VEM, D, E, F is in the case of SIGMOD Gibbs.

It is found that, as the length of the topic label increases, the Coverage value of TLRank increases, but the increasing trend keeps decreasing. According to Fig. 3(B), the vertices with high TLRank score scatter in the upper right section and possess a high corresponding *CovCen* value, though there are not many of these ones. However, as the vertices' TLRank scores gradually decrease, these tend to gather to the lower left section, while their *CovCen* values differentiate seriously. In addition, the topic terms set used in the *CovCen* computing process is different. The Coverage of a topic label corresponds to Top 20 topic terms, while the *CovCen* of a sentence corresponds to Top 500 topic terms. So a certain deviation is inevitable besides the similar issue mentioned in Section VI (A)-2. Therefore, as more and more sentences

are involved in the topic label, the advantage of TLRank continuously falls, the Coverage of TLRank gradually tends towards the Baseline, and even lowers than the Baseline in some cases occasionally.

The 3rd exception occurs in Fig. 6(A), the Relevance of TLRank surpasses that of Baseline when the topic label length is 5000. Fig. 6(A) shows that contrary to our intuition, Baseline only considers *RelCen* centrality while choosing a sentence. However, its Relevance indicator is not as good as TLRank. There are two reasons for this. One is that Baseline does not deal with the redundancy of topic label and the other is that the improvement room of redundancy suppression by TLRank decreases with the increase in topic label length.

The topic label generated by Baseline (label-B) contains the least number of sentences (see the Total from Table 4), and the sentences in label-B are more similar. Therefore, label-B has a relatively high redundancy as mentioned in Section VI (A)-1. It is redundancy that limits the diversity and drops the possibility of joining more top 20 topic terms with label-B.

According to Fig. 4(A) and Table 8, TLRank-G has the smallest change of ranking position of sentences under the case of SIGMOD VEM. Thus its efforts to refrain the redundancy is the weakest among all four cases. Consequently, as the topic label length increases, compared with Baseline (see Fig.6 (A)), the improvement room of TLRank to restrain redundancy gradually decreases, and its advantage of Relevance gradually diminishes, eventually tending towards Baseline, or even higher than Baseline occasionally.

Furthermore, as per Fig.5 and Fig.6, the TLRank model provides outstanding performance, better applicability, and stability when labeling the LDA-style topics learned by VEM estimation method. However, using Gibbs estimation method to discover topics, the improvement in Baseline method is still evident in most circumstances but lacks stability. Therefore, it can be concluded that TLRank is more appropriate to label the discovered LDA-style topics when using the VEM estimation method. Of course, TLRank still has practical significance in the case of Gibbs estimation method, because its results are very close to the best.

VII. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a novel graph-based ranking model, TLRank, to generate a topic label for each discovered topic to help the user understand discovered topics clearly. Experiments done on two corpora and subsequent analyses demonstrate that our method significantly and consistently outperforms the prevailing state-of-the-art and classic models in topic labeling tasks.

In summary, there are three key contributions of this paper. (1) We proposed a novel model to generate topic labels with high Relevance, Coverage, and topics-inter Discrimination. (2) To the best of our knowledge, we are the first to investigate exploiting the strategy of suppressing and enhancing matrix transition probability to restrain the topic label redundancy and boost its diversity. (3) In a single graph-based ranking process, it improves the efficiency and accuracy of scoring and extracting the candidate sentences to assemble a topic label.

In this study, only the morphological similarity was used to compute the Relevance of candidate sentences. The effects of context, coreference, and discourse information have not been considered. Therefore, there is still room to further improve the existing scoring and extracting techniques in order to cater for selecting candidate sentences.

In future research, we will try to use the deep neural network to identify discriminative features to represent sentences and capture contextual relations among sentences. Then combine with existing surface features to extract the more apposite sentences and further improve the performance of topic labeling tasks.

REFERENCES

[1] D. M. Y. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[2] X. Wan and T. Wang, "Automatic labeling of topic models using text summaries," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Aug. 2016, pp. 2297–2305.
[3] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2007, pp. 490–499.
[4] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2011, pp. 1536–1545.
[5] W. Kou, F. Li, and T. Baldwin, "Automatic labelling of topic models using word vectors and letter trigram vectors," in *Proc. AIRS*. Cham, Switzerland: Springer, 2015, pp. 253–264.
[6] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, Feb. 2013, pp. 465–474.
[7] S. Bhatia, J. H. Lau, and T. Baldwin, "Automatic labelling of topics with neural embeddings," 2016, *arXiv:1612.05340*. [Online]. Available: https://arxiv.org/abs/1612.05340
[8] N. Aletras and M. Stevenson, "Labelling topics using unsupervised graph-based methods," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, Jun. 2014, pp. 631–636.
[9] M. Allahyari and K. Kochut, "Automatic topic labeling using ontology-based topic models," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 259–264.
[10] Z. Li, J. Li, Y. Liao, S. Wen, and J. Tang, "Labeling clusters from both linguistic and statistical perspectives: A hybrid approach," *Knowl.-Based Syst.*, vol. 76, pp. 219–227, Mar. 2015.
[11] A. Alokaili, N. Aletras, and M. Stevenson, "Re-ranking words to improve interpretability of automatically generated topics," 2019, *arXiv:1903.12542*. [Online]. Available: https://arxiv.org/abs/1903.12542
[12] N. Aletras, T. Baldwin, J. H. Lau, and M. Stevenson, "Representing topics labels for exploring digital libraries," in *Proc. 14th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Sep. 2014, pp. 239–248.
[13] P. Ren, Z. Chen, Z. Ren, F. Wei, J. Ma, and M. de Rijke, "Leveraging contextual sentence relations for extractive summarization using a neural attention model," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 95–104.
[14] A. E. C. Basave, Y. He, and R. Xu, "Automatic labelling of topic models learned from Twitter by summarisation," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, Jun. 2014, pp. 618–624.
[15] N. Aletras and A. Mittal, "Labeling topics with images using a neural network," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2017, pp. 500–505.
[16] N. Aletras and M. Stevenson, "Representing topics using images," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2013, pp. 158–167.
[17] I. Sorodoc, J. H. Lau, N. Aletras, and T. Baldwin, "Multimodal topic labelling," in *Proc. Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 2, Apr. 2017, pp. 701–706.

[18] A. Khan, N. Salim, and Y. J. Kumar, "Genetic semantic graph approach for multi-document abstractive summarization," in *Proc. 5th Int. Conf. Digit. Inf. Process. Commun. (ICDIPC)*, Oct. 2015, pp. 173–181.
[19] G. C. V. Vilca and M. A. S. Cabezudo, "A study of abstractive summarization using semantic representations and discourse level information," in *Proc. Int. Conf. Text, Speech, Dialogue*. Cham, Switzerland: Springer, 2017, pp. 482–490.
[20] J. Tan, X. Wan, and J. Xiao, "Abstractive document summarization with a graph-based attentional neural model," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Jul. 2017, pp. 1171–1181.
[21] P. K. Rachabathuni, "A survey on abstractive summarization techniques," in *Proc. Int. Conf. Inventive Comput. Inform. (ICICI)*, Nov. 2017, pp. 762–765.
[22] H. T. Le and T. M. Le, "An approach to abstractive text summarization," in *Proc. Int. Conf. Soft Comput. Pattern Recognit. (SoCPaR)*, Dec. 2013, pp. 371–376.
[23] S. Song, H. Huang, and T. Ruan, "Abstractive text summarization using LSTM-CNN based deep learning," *Multimedia Tools Appl.*, vol. 78, no. 1, pp. 857–875, Jan. 2019.
[24] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and W. Houfeng, "Learning summary prior representation for extractive summarization," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, vol. 2, Jul. 2015, pp. 829–833.
[25] P. Ren, F. Wei, Z. Chen, J. Ma, and M. Zhou, "A redundancy-aware sentence regression framework for extractive summarization," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers*, Dec. 2016, pp. 33–43.
[26] A. Jain, D. Bhatia, and M. K. Thakur, "Extractive text summarization using word vector embedding," in *Proc. Int. Conf. Mach. Learn. Data Sci. (MLDS)*, Dec. 2017, pp. 51–55.
[27] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2017.
[28] R. Alguliyev, R. Aliguliyev, and N. Isazade, "A sentence selection model and HLO algorithm for extractive text summarization," in *Proc. IEEE 10th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2016, pp. 1–4.
[29] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, Dec. 2004.
[30] P. Ren, Z. Chen, Z. Ren, F. Wei, L. Nie, J. Ma, and M. De Rijke, "Sentence relations for extractive summarization with deep neural networks," *ACM Trans. Inf. Syst.*, vol. 36, no. 4, p. 39, 2018.
[31] D. Cao and L. Xu, "Analysis of complex network methods for extractive automatic text summarization," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Oct. s2016, pp. 2749–2756.
[32] D. Miller, "Leveraging BERT for extractive text summarization on lectures," 2019, *arXiv:1906.04165*. [Online]. Available: https://arxiv.org/abs/1906.04165
[33] Y. Liu, "Fine-tune BERT for extractive summarization," 2019, *arXiv:1903.10318*. [Online]. Available: https://arxiv.org/abs/1903.10318
[34] D. M. Blei and J. D. Lafferty, "Visualizing topics with multi-word expressions," 2009, *arXiv:0907.1013*. [Online]. Available: https://arxiv.org/abs/0907.1013
[35] E. Baralis, L. Cagliero, N. Mahoto, and A. Fiori, "GraphSum: Discovering correlations among multiple terms for graph-based summarization," *Inf. Sci.*, vol. 249, pp. 96–109, Nov. 2013.
[36] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.
[37] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 404–411.
[38] E. Seneta, *Non-Negative Matrices and Markov Chains*. New York, NY, USA: Springer, 2006.
[39] N. A. Sanjaya, M. L. Ba, T. Abdessalem, and S. Bressan, "Harnessing truth discovery algorithms on the topic labelling problem," in *Proc. 20th Int. Conf. Inf. Integr. Web-Based Appl. Services*, Nov. 2018, pp. 8–14.
[40] K. Hornik and B. Grün, "topicmodels: An R package for fitting topic models," *J. Stat. Softw.*, vol. 40, no. 13, pp. 1–30, 2011.
[41] 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.*, 2013, pp. 3111–3119.
[42] J. Mitchell and M. Lapata, "Composition in distributional models of semantics," *Cognit. Sci.*, vol. 34, no. 8, pp. 1388–1429, 2010.
[43] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Inf. Process. Manage.*, vol. 40, no. 6, pp. 919–938, Nov. 2004.
[44] M. Newman, *Networks: An Introduction*. Oxford, U.K.: Oxford Univ. Press, 2010.
[45] B. Bigi, "Using Kullback–Leibler distance for text categorization," in *Proc. 25th Eur. Conf. IR Res.* Berlin, Germany: Springer, 2003, pp. 305–319.
[46] R. Arora and B. Ravindran, "Latent Dirichlet allocation based multi-document summarization," in *Proc. 2nd Workshop Anal. Noisy Unstructured Text Data*, Jul. 2008, pp. 91–97.
[47] I. E. Allen and C. A. Seaman, "Likert scales and data analyses," *Qual. Prog.*, vol. 40, no. 7, pp. 64–65, 2007.
[48] M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, and J. Han, "Automatic construction and ranking of topical keyphrases on collections of short documents," in *Proc. SIAM Int. Conf. Data Mining*, 2014, pp. 398–406.
[49] A. Gourru, J. Velcin, M. Roche, C. Gravier, and P. Poncelet, "United we stand: Using multiple strategies for topic labeling," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2018, pp. 352–363.
[50] J. Velcin, A. Gourru, E. Giry-Fouquet, C. Gravier, M. Roche, and P. Poncelet, "Readitopics: Make your topic models readable via labeling and browsing," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 5874–5876.
[51] J. H. Lau and T. Baldwin, "The sensitivity of topic coherence evaluation to topic cardinality," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2016, pp. 483–487.

DONGBIN HE received the B.S. degree in computer and its application from the Hebei University of Technology, Tianjin, China, in 1996, and the M.S. degree in computer software and theory from Inner Mongolia University, Hohhot, China, in 2006. He is currently pursuing the Ph.D. degree with the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. He is also an Associate Professor with the College of Computer Science and Engineering, Shijiazhuang University. His research interests include the Internet of Things technologies, data mining, and natural language processing.

MINJUAN WANG received the Ph.D. degree from the School of Biological Science and Medical Engineering, Beihang University, under the supervision of Prof. Liu, in 2017. She was a Visiting Scholar with the School of Environmental Science, Ontario Agriculture College, University of Guelph, from 2015 to 2017. She is currently a Postdoctoral Fellow with the School of Information and Electrical Engineering, China Agricultural University. Her research interests include bioinformatics and the Internet of Things key technologies.

ABDUL MATEEN KHATTAK received the Ph.D. degree in horticulture and landscape from the University of Reading, U.K., in 1999. He was a Research Scientist in different agricultural research organizations before joining Agricultural University Peshawar, Pakistan, where he is currently a Professor with considerable experience in teaching and research, Department of Horticulture. He has conducted academic and applied research on different aspects of tropical fruits, vegetables, and ornamental plants. He was also with Alberta Agriculture and Forestry, Canada, as a Research Associate, and with the Organic Agriculture Centre of Canada, as a Research and Extension Coordinator (for Alberta province). There he helped in developing organic standards for greenhouse production and energy saving technologies for Alberta greenhouses. He is also a Visiting Professor with the College of Information and Electrical Engineering, China Agricultural University, Beijing. He has published 55 research articles in scientific journals of international repute. His research interests include greenhouse production, medicinal, aromatic and ornamental plants, light quality, supplemental lighting and temperature effects on greenhouse crops, aquaponics, and organic production.

LI ZHANG is currently pursuing the Ph.D. degree with the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. She conducted research on deep learning for classification, object recognition, tracking, detection, and semantic segmentation for the vision system of agricultural robot. She also studies image/video processing techniques, including enhancement and denoising.

WANLIN GAO received the B.S., M.S., and Ph.D. degrees from China Agricultural University, in 1990, 2000, and 2010, respectively. He is currently the Dean of the College of Information and Electrical Engineering, China Agricultural University. He has published 90 academic articles in domestic and foreign journals, and among them, over 40 are cited by SCI/EI/ISTP. He has written two teaching materials, which are supported by the National Key Technology Research and Development Program of China during the 11th Five-Year Plan Period, and has written five monographs. Moreover, he holds 101 software copyrights, 11 patents for inventions, and eight patents for new practical inventions. His research interests include the informationization of new rural areas, intelligence agriculture, and the service for rural comprehensive information. He is also a Principal Investigator of over 20 national plans and projects, a member of the Science and Technology Committee of the Ministry of Agriculture, a member of the Agriculture and Forestry Committee of Computer Basic Education in Colleges and Universities, a Senior Member of the Society of Chinese Agricultural Engineering, and so on.