International Journal of Innovative Computing, Information and Control
Volume **21**, Number **6**, December **2025**

 ISSN 1349-4198
pp. **1523–1537**

RESEARCH AND APPLICATION OF NEWS ENTITY MODELING METHOD BASED ON DISTRIBUTED REPRESENTATION

Jun Zhou[1,*] and Ce Zhou[2]

[1]Institute of Acoustics
Chinese Academy of Sciences
No. 21, North 4th Ring Road, Haidian District, Beijing 100190, P. R. China
[*]Corresponding author: zhoujun@hccl.ioa.ac.cn

[2]School of Computer Science & Technology
Soochow University
No. 333, Ganjiang East Road, Suzhou 215031, P. R. China
20255227040@stu.suda.edu.cn

Received March 2025; revised July 2025

Abstract. *The Internet hosts numerous public news websites that publish a vast amount of news daily, containing valuable information that has yet to be fully utilized in research. To this end, this paper proposes a media entity modeling and analysis framework based on distributed representation. We first crawl partial news data from the news website "Voice of America" as the research object using a crawler tool, and then identify news themes through topic clustering. Aiming at the challenge of entity modeling in news, three entity distributed representation methods are proposed by drawing on the idea of word distributed representation. An entity association construction method is designed using entity co-occurrence relationships and distributed representations, based on which three entity modeling-related tasks are completed: entity clustering, association information mining, and community discovery. To tackle the issues of unlabeled entity information and difficult model performance evaluation in news, performance indicators for measuring the effectiveness of this task model are designed. The experimental results show that in the entity distance task, the topic distance is 0.187; in the community discovery task, the average relevance is 4.09, verifying the effectiveness of the three entity distributed representation methods and the entity association construction method.*
Keywords: Entity modeling, Distributed representation, Cluster analysis, Association analysis

1. **Introduction.** Under the background of informatization, news, as an important source of open-source intelligence, has received extensive attention from many fields including national security [1,2]. Due to its characteristics such as low acquisition difficulty, wide dissemination range and strong timeliness, news text analysis has always been a research hotspot in text mining. A large number of research works on news text mining have been carried out at home and abroad. One of the most widely applied is to conduct semantic analysis and investor sentiment analysis using financial news to predict market fluctuations and stock price changes [3,4]. In addition, some studies [5] have also explored classification-oriented tasks in multilingual news texts. Further analysis can obtain more valuable in telligence information, which is of positive significance for maintaining national security and regional stability.

Entity information in news text can be obtained through the relevant tools of Chinese named entity recognition, and the relevant work is relatively mature. However, because

DOI: 10.24507/ijicic.21.06.1523

the entities involved in the news have many kinds, a wide range of issues, and the relevant terms do not necessarily exist, it is difficult to carry out large-scale annotation.

In view of the above difficulties, this paper innovatively proposes a method to model and analyze the entities involved in news texts. Inspired by word distributed representation, this method uses the context information of entities to directly train news text to generate distributed representation of entities, and constructs the association between entities involved in news. The experimental results show that the distributed representation of entities can achieve remarkable results in entity clustering, association information mining, community discovery and other tasks, and can meet the requirements of online applications.

There are three main contributions of this paper.

1) Aiming at the difficulty of entity labeling in news texts, this paper innovatively proposes an entity analysis method of news texts based on distributed representation of entities, which can directly analyze news texts, realize modeling and representation of entities involved in news, and be applied in other tasks.

2) Aiming at the problem that the distributed representation of entities and the effect of entity association are difficult to measure, this paper proposes for the first time an evaluation index of the effect of unlabeled entity modeling to measure the entity clustering and community discovery task based on the distributed representation of entities, and demonstrates the effectiveness of the index through experimental results.

3) Using the news entity modeling method proposed in this paper, a large number of key Hong Kong-related entities and their related information were successfully found, such as "Zhifeng Huang", "Ting Zhou", "Guancong Luo" and other related key entities were found by using the institutional entity "Demosisto", which can meet the requirements of online use in practice.

The structure of this paper is as follows: Section 1 introduces the research background, research significance, and the main contributions of the paper; Section 2 reviews the related work on topic clustering, word distributed representation, and community discovery methods; Section 3 elaborates the specific implementation process, including data preprocessing, entity representation learning, entity association construction, and analytical applications; Section 4 verifies the effectiveness of the method through experiments on real-world news data, and provides a detailed analysis and discussion of the results; Section 5 concludes the study and proposes potential directions for future research, such as incorporating graph neural networks to enhance entity modeling.

2. **Related Work.**

The news entity modeling method proposed in this paper is closely related to text topic clustering, word distributed representation and community discovery.

2.1. **Text topic clustering.**

Text topic clustering is a learning task that aims to find texts with similar topics from the corpus. Latent semantic analysis (LSA) is an early approach to identify topic-based semantic relationships between text and words through matrix decomposition, and further represent the topic-based similarity of text through topic vectors [6,7]. Inspired by this, probabilistic latent semantic analysis (PLSA) was proposed, which is a learning method for topic analysis of text collections using probabilistic generation models [8]. The model represents the topic through hidden variables, generates the topic from the text first, then generates the word from the topic, and finally realizes the process of generating the word from the text.

On this basis, latent Dirichlet allocation (LDA) model is proposed, which is a generative probability model based on Bayesian learning and is also the most influential topic clustering model. It assumes that each text is represented by a multinomial distribution

of topics, each topic by a multinomial distribution of words, and both distributions are Dirichlet distributions. The model first randomly generates a topic distribution for each text, then generates a topic at each position of the text according to the distribution, generates a word at that position according to the word distribution of the topic, and finally generates the whole text [9]. On the basis of LDA, a series of improvements have been proposed by scholars to make LDA applicable to various situations [10-12]. In this paper, LDA model is used to cluster news topics, so as to obtain the topic distribution of each news.

2.2. **Word distributed representation method.** Early word representation methods were represented by one-hot vectors, where each word was represented by a vector of multiple zeros and a single one. Because different words are completely orthogonal, the data is sparse and cannot represent the semantic distribution. Then came the language model based on neural network [13], and the distributed representation of words is a by-product of this model. The key work of word distributed representation is proposed by [14], including two word vector training models: Skip-gram and CBOW, as illustrated in Figure 1.

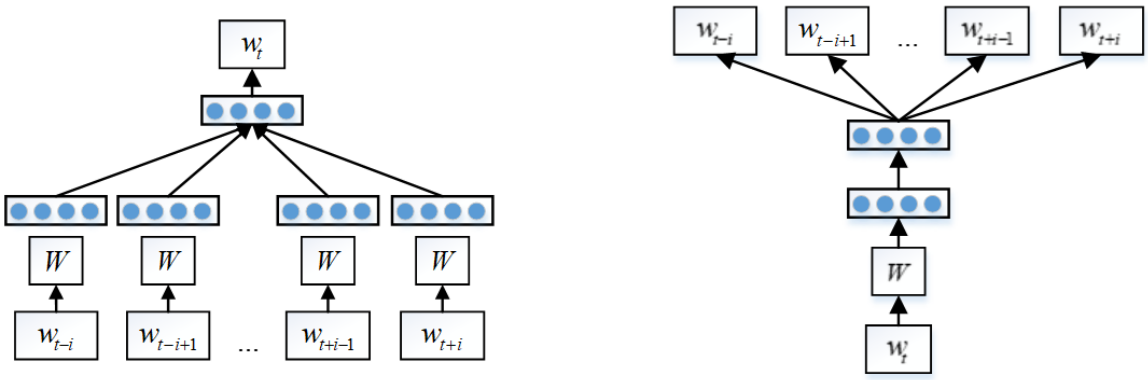

FIGURE 1. CBOW model (left) and Skip-gram (right)

The idea of the two models is to map each word into a vector through the context information. CBOW predicts the word w_i through the information in the context window length k $[w_{i-k}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+k}]$ to achieve the word vector training. On the contrary, Skip-gram model gives the center word to predict possible context $[w_{i-k}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+k}]$. SoftMax function is usually used to calculate the occurrence probability $P(w_c|w_i)$ of w_c given w_i, expressed as

$$P(w_c|w_i) = \frac{e^{w_c^T w_i}}{\sum_{w_j \in v} e^{w_c^T w_j}} \tag{1}$$

The above two word vector representations only consider the co-occurrence information of words in the context. To incorporate the co-occurrence information from the entire corpus, the structure of the Glove model was proposed in [15]. By obtaining the word co-occurrence matrix from the entire corpus, global information was introduced. The final optimization function is as follows:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T w_j + b_i + b_j - \log X_{ij} \right)^2 \tag{2}$$

where X_{ij} represents the co-occurrence times of w_i and w_j, and b_i and b_j are word bias parameters, which need to be obtained through training.

In this work, we adopt three distributed representation methods – Skip-gram, CBOW, and Glove – to train entity embeddings. These entity vectors are subsequently utilized for entity clustering, association information mining, and community discovery tasks.

2.3. **Community discovery.** The purpose of community discovery is to discover the community structure in the network, which can be regarded as a clustering algorithm on the graph structure. There are many algorithms discovered by the community. The first one proposed by Newman and Girvan is the GN algorithm based on the betweenness [16]. Later, algorithms based on label passing [17] and spectral clustering [18] have also been proposed successively. The community found that there was a problem: there was no relevant measurement index until the Louvain algorithm for conceptual row optimization of modularity was proposed in [19]. Modularity is an indicator to measure the level of community division without knowing the real division of the community. It can be used to measure the degree of association within the community, expressed by the formula

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w) \tag{3}$$

where m is the sum of network connection weights; v and w are any two nodes in the network, and A_{vw} is their connection weight; k_v stands for the degree of point v; $\delta(c_v, c_w)$ indicates whether v and w belong to the same community, if they belong to 1, otherwise 0.

In this paper, Louvain algorithm is used to divide the constructed entity relationship network into communities, and the correlation degree of entities in the same community is analyzed to verify the results of community division. The specific optimization steps of Louvain algorithm are as follows:

1) Divide the community for each node, calculate the modularity at this time, then make a node merge with the community of the neighbor node to calculate the modularity gain, find the largest modularity gain merge node, repeat several times until the modularity no longer increases;

2) Merge the new communities into points and repeat the operations in 1) until convergence.

3. **News Entity Modeling Methods and Application Analysis.** In order to obtain the relevant semantic representation information from the context of the entity and reduce the influence of noise as much as possible, this paper uses the distributed representation method of words for training to get the entity representation vector. The distributed representation of words can make full use of the key semantic information of the context in which the words are located, and the distributed representation of entities can be obtained by using this method. The similarity between entities is obtained through distributed representation of entities, and similar entities are further discovered by clustering. On this basis, association information mining is used to actively discover similar entities to achieve deep mining of key information, such as using key information "Zhifeng Huang" to discover similar Hong Kong politicians "Ting Zhou" and "Guancong Luo".

The entities involved in news often have some correlation relationship. By observing the co-occurrence of entities in news, we can find that entities that often appear together and have similar context description semantics tend to have strong correlation. Therefore, the correlation between entities can be modeled through the co-occurrence frequency of entities and the similarity of the distributed representation of entities. For example, the context descriptions of "Zhifeng Huang" and "Ting Zhou" often appear at the same time and have similar semantics, indicating that the two have a strong correlation. The more

the number of entity cooccurrences, the closer the entity distributed representation, the higher the entity similarity. The two can be used to model the entity association, and then discover the potential community in the entity association network. As entities in the same community are often highly related, certain characteristics of the community can be obtained according to certain important entities in the community. For example, the presence of entities such as "Socialist Action" and "Demosisto" indicates that the community is likely to be related to key people and organizations involved in Hong Kong.

The news entity modeling process is divided into five parts: data preprocessing, news topic analysis, entity distributed representation and merging, entity association construction and entity analysis. The specific modeling process is shown in Figure 2.

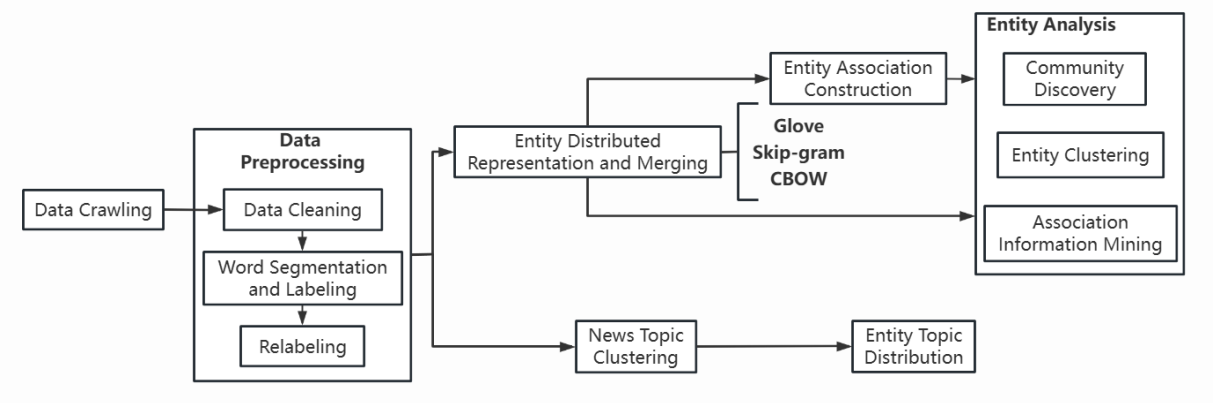

FIGURE 2. News entity modeling process

3.1. **Data preprocessing.** Data preprocessing is divided into three steps: data cleaning, word segmentation and labeling, and re-labeling.

1) Data cleaning

First, clean the text by removing special symbols. Next, to ensure the text is substantive and unique, discard sentences shorter than six words and repeated sentences, as very short sentences (e.g., "In summary" and "All in all") often lack specific entity information and offer limited analytical value. Finally, convert traditional Chinese characters to simplified ones using a word list for consistency and easier subsequent processing.

2) Word segmentation and labeling

In this paper, the open source toolkit pyltp of HIT University is used for word segmentation, parts-of-speech tagging and named entity recognition to remove punctuation and get entity information of text. In order to improve the accuracy of words and labels, a dictionary is created for the entities and words that are easily misdistinguished in the experiment to provide prior knowledge to improve the accuracy of the model.

3) Re-label processing

Considering that some entities cannot be accurately identified every time, this paper uses the results of the first successful identification of entities to re-label the entities involved in the text to improve the recognition rate of entities. The specific process is to save the results of the first named entity recognition, and then search and match in the text after word segmentation and labeling. When a word matches the entity identified for the first time but is not marked as an entity, it is re-annotated in the text, and finally the fully annotated result is obtained.

In this paper, LDA model is used for topic clustering analysis of news, and the number of topic clusters is set as M through experiments. After the clustering, a probability

distribution of a topic is generated for each news. If the probability distribution is the largest, a topic is assigned to each news and all news is divided into M topics.

Given the frequency of all entities appearing in the news of a certain topic, the distribution of all entities appearing in the news topic can be obtained, and the distribution of entity topics is represented by a vector of length M. For example, if an entity appears in four news items, one of which appears in the news item belonging to topic T_0, and three of which appear in the news item belonging to topic T_{M-1}, then the topic distribution vector of the entity is $\left[\frac{1}{4}, \ldots, \frac{3}{4}\right]$.

3.2. **Entity distributed representation and merging.** This paper adopts three models based on word distributed representation: Skip-gram, CBOW and Glove. Consistent with the traditional distributed representation of words, the distributed representation of words and entities is finally obtained by co-training the entities and words in the text, and the similarity degree of entities is measured by the cosine similarity of the entity vector. Negative sampling is used in the training process to reduce the computational complexity [20]. Negative sampling uses the score between word pairs to represent the correlation, aiming to make the target word have a large score between the words in the window and the random words in the vocabulary list, and the optimization goal is expressed as the maximization of the following formula:

$$\log \sigma\left(v_{w_o}^T v_{w_i}\right) + \sum_{i=1}^{n} E_{w_i \sim P_N(w)}\left[\log \sigma\left(-v_{w_i}^T v_{w_i}\right)\right] \tag{4}$$

where w_i is the target word, w_o is the word in the target context, $w_i \sim P_N(w)$ represents the result of random sampling in the word list, and n is the number of negative samples extracted in each training.

Through the experiment of entity distributed representation, it is found that some entity results are misidentified, such as "Carrie Lam Cheng Yuet-ngor" in the news may be identified as "Carrie Lam Cheng Yuet-ngor", "Carrie Lam Cheng Yuet", "Carrie Lam", etc., and personal names and place names may be confused. These errors are often caused by incorrect segmentation, and the error entity is because similar contexts often have similar entity representations, and there is a certain possibility that they exist in the same text. In the entity merging part of this paper, entity representation results obtained by Skip-gram model training are adopted. When the string representation of an entity recognition result is a substring of another entity recognition result, the two results are considered to belong to the same entity if either of the following two rules is met:

Rule 1: Cosine similarity of two entities is > 0.80;

Rule 2: The cosine similarity of two entities is > 0.75 and appears in at least one news text at the same time.

In this paper, some entities with incorrect recognition due to word segmentation errors are added to the dictionary and the model is retrained, and the results are continuously optimized through several iterative experiments.

3.3. **Entity association construction.** The co-occurrence between the two entities indicates that there may be some correlation between the entities. This paper chooses the co-occurrence times of the two entities in the same paragraph to model the correlation between the entities, because there are too few entities appearing in a single sentence, and the entities appearing in the same news may not be related. The more times of inter-entity co-occurrence, the greater the similarity of entity vectors and the stronger the correlation between entities. Use the following formula to construct the relationship between two entities:

$$rel(w_1, w_2) = \begin{cases} 0 & \text{if } rel'(w_1, w_2) < \theta_3 \\ rel'(w_1, w_2) & \text{if } rel'(w_1, w_2) \geq \theta_3 \end{cases} \tag{5}$$

$$rel'(w_1, w_2) = \min(\log_2(co\_times + 1) * sim(w_1, w_2)/2, \theta_2) \tag{6}$$

$$sim(w_1, w_2) = \max\left(\frac{w_1 * w_2}{\|w_1\| * \|w_2\|} - \theta_1, 0\right) \tag{7}$$

where *co_times* represents the number of co-occurrences of entities w_1 and w_2; θ_1 indicates that only when the degree of similarity is higher than a certain threshold is there considered to be a relationship between entities; θ_2 represents the maximum value of the relation; the function of θ_3 is to remove some weak associations, which can reduce many unimportant edges to facilitate analysis results. In the process of analyzing the results, we can adjust the size of θ_3 to make the number of entity associations generated by the distributed representation of the three models close, so as to compare the results found by the community.

3.4. **Entity analysis.** Entity clustering is the clustering of trained entity vectors. By analyzing the clustering results, the distributed representation of entities is measured and similar entities are found. Association information mining is to use entity vector to associate input information and output similar entities and their related descriptions. The community finds that the entity association obtained can be used to divide entities into several communities, and the entity association within the community is stronger.

1) Entity clustering

The cosine similarity between entity vectors quantifies inter-entity proximity. Leveraging distributed entity representations, we implement k-means clustering and assess its effectiveness through intra-class topic similarity analysis. To mitigate noise, only the entities whose length is greater than 1 and less than 10 are clustered. Finally, the results of three entity representation models are generated respectively.

2) Association information mining

This paper uses the distributed representation of the generated entities and words to screen out K entities or words whose similarity topK of the vector representation of the input seeds in the thesaurus as candidates, and selects the part with high similarity as the input associated content. There may be some hidden information in the text that has not been discovered. In order to find some hidden relevant information as much as possible, this paper draws on snowball's idea [21] to design the algorithm and uses the training results of Skip-gram model. The specific flow of the algorithm is as follows:

(a) Search for all possible identification results found by the seed entity during the entity merge stage, and use all possible identification cases as candidates for input search. For example, if you search for "Carrie Lam Cheng Yuet-ngor", the search list that may be added is ["Carrie Lam Cheng Yuet-ngor", "Carrie Lam Cheng Yuet", "Carrie Lam"] and other relevant entity identification results;

(b) Each element in the search candidate is judged. If it is an entity, the words with the similarity Top50 in the model dictionary and the similarity degree greater than the threshold value 0.80 are extracted and added to the keyword list;

(c) Identify all entities from the keyword list, and search the entities in the Top30 words of similarity for each entity;

(d) Search the new entity Top50 words and the existing keyword list intersection, overlapping number is greater than 12 to join the list; If the number of words in the keyword list is less than 24, the overlap number is greater than half of the number of words in the keyword list.

(e) Repeat steps (c) and (d) until the elements in the keyword list no longer change.

Through steps (c), (d) and (e), you can find the complete list of keywords in the seed information, and then discover the key information in the news.

3) Community discovery

In the part of entity association construction, three kinds of entity association relationships can be obtained based on different distributed representation of entities. In this paper, Louvain algorithm is used to realize community discovery task based on modularity optimization, and finally three kinds of community division results are obtained.

4. Experimental Results and Analysis.

4.1. Experimental data.

This paper uses the news text of Voice of America, a public news website, as the research object, using "China" and "Hong Kong" as the keywords, and randomly crawled 2,162 news articles. There are 27,679 paragraphs and 86,597 sentences in the news corpus, of which the longest sentence length is 653, and the average sentence length is about 31.36. Using pyltp open source toolkit process to the news text, the final size of the word list is 44158, and the number of entity recognition is 9610.

4.2. Experimental evaluation method.

The analysis of the results of the model is a key and difficult point: First, the news corpus used is unlabeled, and the associated information of entities is artificially constructed by using experience, and the indicators of traditional tasks are not applicable in this task; Secondly, some entities do not have corresponding Baidu entries, and cannot obtain detailed information to assist judgment. To solve these problems, this paper selects the relevant results containing "Zhifeng Huang" as the research object, and uses manual scoring as the assistance to construct topic similarity index and association index to analyze the clustering results and community discovery results respectively. The specific index construction method is as follows.

1) Manual scoring indicators

We recruited 20 postgraduate students majoring in natural language as volunteers. They manually evaluated the correlation between entities within the class, gave scores ranging from 1 to 5 for each of the three results, and finally obtained a comprehensive score. The scoring is based on the prior knowledge of the relevant entities by the volunteers and the text descriptions of the positions of each entity in the source corpus. When a direct judgment cannot be made, Baidu is used to search for relevant information about the entities to assist in the manual judgment.

2) Topic similarity index

News topic clustering can assign related topics to each news. Generally, topic distribution can be used to measure the topic similarity of news. It can also be considered that entities with strong similarity have similar topic distribution, and the topic similarity of entities in each clustering result can be measured by the difference of the topic distribution of entities, which can be used as an indicator of the entity clustering result. Suppose that the entity topic representations within the class are $[x_1, \ldots, x_M]$, and the average distance between the topics within the class is used to measure the similarity. The smaller the distance, the higher the similarity of the topic. The formula is as follows:

$$dist = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j>i}^{M} \|x_j - x_i\|^2 \tag{8}$$

In addition, the proportion of entity topics in a cluster can also be used as an indicator to measure clustering. The topic with the largest proportion in the clustering results is taken as the topic distribution of the entity clustering results. The larger the value, the more concentrated the topics in the clustering results are.

3) Entity association indicators

As the largest Chinese search engine, Baidu has the largest number of Chinese resources, and Baidu's search for the simultaneous occurrence of two entities can reflect the relationship between the two entities. It can be found that "Zhifeng Huang" and "Ting Zhou" appear on the same webpage many times, but not on the same webpage as "Hua Situ", which also reflects that in reality, "Zhifeng Huang" and "Ting Zhou" are closely related to each other as members of "Demosisto", but not with "Hua Situ".

Build an association between entities based on the number of times two entities appear together in Baidu searches. For entities w_1 and w_2, assuming that the co-occurrence times of the two in the web page is N, the correlation degree of the two can be expressed as

$$rel(w_1, w_2) = \log_2(N) \tag{9}$$

In addition, the search times of entity pairs are 0, indicating that the association between entities is poor, and the number of 0 occurrences in the statistical results can also be used as a result to measure the degree of association between entities.

Taking "Zhifeng Huang", a key figure, as the research object, 25 entities were randomly selected from the community results generated by three distributed representations to form 300 entity pairs, and the average connection between the 300 entity pairs was obtained by Baidu search as the measurement result. As a control, three combinations of 25 entities were randomly selected from all entities to illustrate the effect of community discovery.

4.3. **Analysis of experimental results.** First of all, the LDA model is used to cluster news directly, the cluster number is set to 8, and the Gibbs sampling method is used to group topics. Finally, the results of the news clustering clustering are obtained. The top 10 keywords for clustering are shown in Table 1.

Then Skip-gram, CBOW and Glove models were used to train entities and word vectors with a dimension of 50, respectively. In the training process, the model adopted a window

Table 1. Top 10 topic clustering keywords. This table lists the most representative keywords for each topic obtained through LDA topic modeling on the news corpus.

Clustering result	Top 10 topic keywords
Topic 1	Thailand, department of corrections, complaint, correction, chocolate, prison, milk tea, Sheng Zhao, eat, the Mekong River
Topic 2	Detection, trade union, the whole people, joint production, labor, go to work, connect, New York City, virus detection, strike
Topic 3	Bookstore, Varro, publish, Na, documentary, film, Mulan Hua, Bo Li, Zhiyong Xu, writer
Topic 4	Death penalty, Dalai Lama, commit, minor, Tibet, Peng Li, Rinpoche, diary, Lama, Panchen Lama
Topic 5	Hong Kong, senator, Zhifeng Huang, legislature, democracy, movement, Beijing, police, China, election
Topic 6	Myanmar, India, Iran, Iraq, Afghanistan, official, North Korea, attack, Israel, happen
Topic 7	President, justice, Supreme Court, Kim, Democratic Party, Ginsburg, Trump, nominate, Republican Party, Senate
Topic 8	China, America, human rights, Xinjiang, country, problem, international, government, president, Uygur

size $k = 10$ and negative sampling times $n = 10$. Using the Skip-gram model to merge entities, 835 entity pairs are obtained.

Then the entity association is constructed by using the entity vector similarity and the entity co-occurrence in the text. In the process of entity association construction, the parameters $\theta_2 = 1$, $\theta_3 = 0.1$. In order to make the number of relationships generated by the three methods close, the method $\theta_1 = 0.6$ based on the CBOW model and the method $\theta_1 = 0.65$ based on the Skip-gram and Glove models generate 37428, 37152 and 37820, respectively. Considering the entity relationship generated based on the Skip-gram model, the entity whose relationship with the entities "Demosisto", "Socialist Action" and "Sunflower" is greater than 0.15, the relationship diagram is shown in Figure 3, where the number of edges is 64 and the number of nodes is 177.

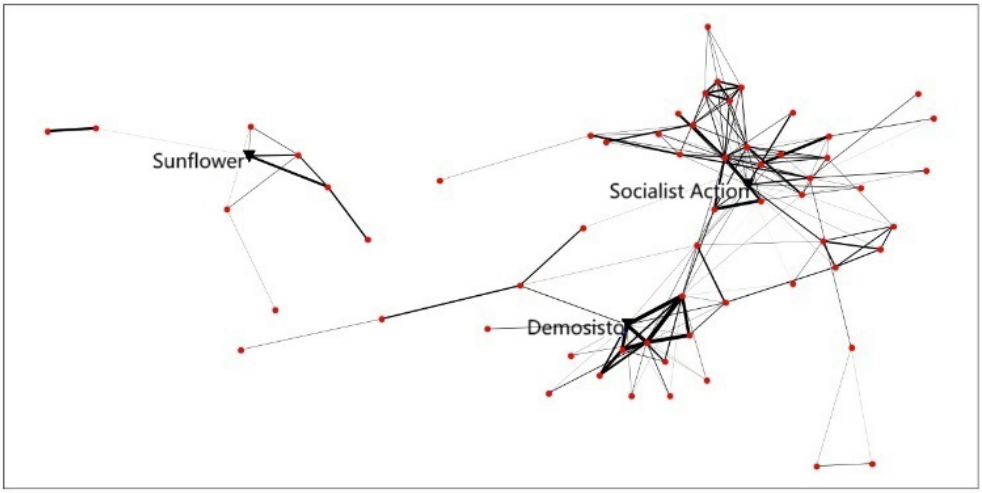

FIGURE 3. Entity relationship graph based on Skip-gram model. There are some closely connected nodes centered around "Sunflower", "Demosisto" and "Socialist Action", forming a structure similar to a community.

In Figure 3, there are some closely connected nodes centered around "Sunflower", "Demosisto", and "Socialist Action", forming a structure similar to a community, which indicates that the community structure is significantly present in the constructed relationship diagram. Among them, there is a certain connection between the members of the Hong Kong organizations "Demosisto" and "Socialist Action", while there is no obvious connection with Taiwan's "Sunflower". Observing the chaotic Hong Kong organization "Demosisto", the entity "Demosisto" and its main members "Zhifeng Huang", "Ting Zhou", "Guancong Luo", and "Langyan Lin" are closely connected to each other, while the entity "Socialist Action" and its main members "Wenyuan Wu", "Guoxiong Liang", "HaoMing Huang", and "Jiancheng Zeng" are also closely connected. The entity "Sunflower" is closely connected to its member "Feifan Lin" and the "Student Movement", which indicates that the constructed relationship weights can reflect real-world relationships.

The relationship graphs generated by the three methods were analyzed to identify the most influential entities in each graph. This paper used PageRank algorithm to calculate the influence of each entity [22], ultimately obtaining a numerical representation of the influence of each entity. For the top 20 influential entities selected by the three methods, an effectiveness filter was applied to eliminating those ranked in the top 20 but actually invalid (such as ambiguous references or incorrectly identified entries). Finally, the person and place entities with analytical value were extracted, as shown in Table 2.

TABLE 2. Top 20 influential persons and place names obtained using the PageRank algorithm. Entities with no analytical value were removed from the original ranking to retain only meaningful person and location names.

Method	Character entity	Geographical name entity
Skip-gram	Zhifeng Huang, Guoxiong Liang, Brookings, Guancong Luo, Carrie Lam Cheng Yuet-ngor	Russia, United Kingdom, Taiwan, Germany, Xinjiang, China, America, Afghanistan, Beijing, Europe
CBOW	Wuer Kaixi, Ling Chai, Chaohua Wang, Kennedy, Rinpoche	Colombia, Shanghai, Nanjing, Los Angeles, Mexico, Indiana, Maryland, Yunnan
Glove	Robert, Chaohua Wang, Gang Liu, Guancong Luo, AJ Ren, John	America, China, Beijing City, Shandong, United Kingdom, Belgium, Hebei, Germany

It can be seen from the results that there are big differences in the influential entities in the association results generated by different methods, indicating that there are big differences in the results of different methods. Finally, entity clustering, association information mining and community discovery tasks are carried out based on the trained entity vector and entity association. The number of entities participating in the cluster is 9171, the number of clusters is set to 100, and the cluster center is randomly initialized. Ten classes were randomly selected from the generated clustering results for visual display using t-SNE, and the results were shown in Figure 4. It can be seen that the clustering effect of entity vector is obvious, and the distinction between different clusters is obvious.

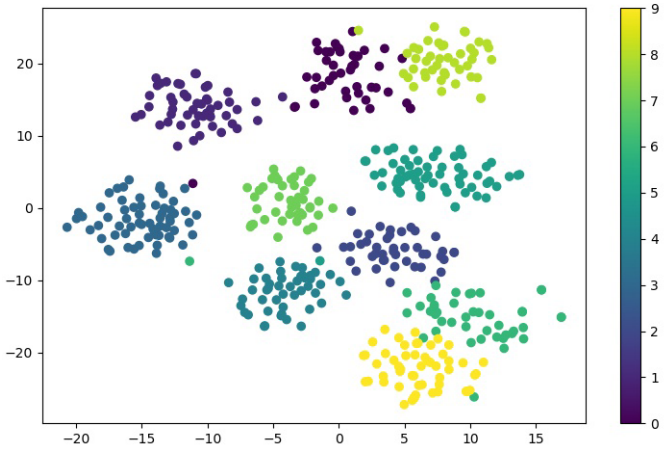

FIGURE 4. Clustering effect

Taking the category of the key figure "Zhifeng Huang" as the research object, the evaluation results of the three methods of entity clustering are shown in Table 3.

It can be seen from the results that the average topic distance of the three methods in clustering is smaller than the average topic distance of all entities. Although most entities belong to topic 8, the topic assigned by the entities similar to "Zhifeng Huang" obtained by the three methods is topic 5, and the top keywords of topic 5 are "Hong Kong", "Senator", "Democracy", "Movement", etc., which shows that the generated distributed representation has a good effect. According to the results of manual evaluation, topic distance and maximum topic distribution, the method based on Skip-gram model has the

TABLE 3. Evaluation results of model entity clustering task

Clustering result	Method	Human evaluation result	Topic distance result	Max topic distribution
	Skip-gram	$\mathbf{4.24 \pm 0.16}$	**0.187**	**91.4% (Topic 5)**
	CBOW	3.47 ± 0.23	0.926	63.6% (Topic 5)
	Glove	3.82 ± 0.21	0.416	74.8% (Topic 5)
All entities		–	1.114	55.8% (Topic 8)

best effect, and the method based on CBOW model has the worst effect. Analyzing the reasons, the reasons for the best results based on Skip-gram model may be as follows:

1) Compared with CBOW, which uses the mean value training represented by contextual word vectors, Skip-gram and Glove train each combination in the context, with more training times and more adequate training;

2) Glove made use of the number of co-occurrences of entities in the news during the training process, and the more the number of co-occurrences of entities, the more similar it is. However, in fact, the number of occurrences of entities is quite different, and whether the entities appear together can better explain the similarity between entities than the number of co-occurrences of entities. Therefore, compared with Skip-gram, Glove introduces noise, resulting in performance degradation.

Since the distributed representation clustering generated by the Skip-gram model has a good effect, the entities "Demosisto", "Ting Zhou" and "Carrie Lam Cheng Yuet-ngor" are used as association information mining by using the training results of the Skip-gram model, and some of the results are shown in Table 4. Search "Demosisto" to find key figures, affiliated organizations and people in Hong Kong public Records; Search "Ting Zhou" to get similar entities and their organizations, English names and other key information; Search "Carrie Lam Cheng Yuet-ngor" for relevant Hong Kong politicians. It can be seen from the results that some entity information related to input information can be found after snowball operation, and when the number of entities in the results

TABLE 4. Association information mining results based on snowball. This table shows the entities discovered from initial inputs using the snowball algorithm, where the first search results are expanded through iterative similarity-based mining.

Input information	First search partial results	Snowball search partial results
Demosisto	Ting Zhou, Langyan Lin, Guancong Luo, Wuzi, Jiawei Yuan, Zhifeng Huang, Ivan Lam and others	Kaidi Zhu, Yongkang Zhou, Wenluo Li, Jialang Zheng, Aohui Ce, Zhifeng Xu, Zongze Li, Student Union and others
Ting Zhou	Guancong Luo, Demosisto, Langyan Lin, Zhifeng Huang, Haiyi Xiqu, Agnes Chow, November 23	Zhuoxuan Ao, Yongkang Zhou, Facebook
Carrie Lam Cheng Yuet-ngor	Yinquan Zeng, ZhengYing Liang, Jianzong Zhang, Huanguang Lin, Ruohua Zheng, Zhenghua Fu, Jianhua Dong, Hong Kong government and others	Ip Lau Suk-yee, Jiachao Li, Maobo Chen

meeting the requirements in the first round is higher and the results are more accurate, the more new entities will be obtained after snowball operation. It can be found that the result of snowball operation is closely related to input information.

Finally, community discovery was carried out based on the entity relationship constructed by Skip-gram, CBOW and Glove models, and the number of communities obtained were 927, 427 and 722, respectively. The results of the community where the key figure "Zhifeng Huang" is located are analyzed, and the evaluation results of the community discovery are shown in Table 5.

TABLE 5. Community discovery task evaluation results

Community construction method	**Manual evaluation results**	**Average relevance**	**Number of unrelated entity pairs**
Skip-gram	3.78 ± 0.19	2.50	28
CBOW	3.95 ± 0.23	2.62	22
Glove	$\mathbf{4.19 \pm 0.21}$	**4.09**	**0**
Random sampling 1	2.38 ± 0.52	1.57	129
Random sampling 2	3.12 ± 0.33	2.18	69
Random sampling 3	2.31 ± 0.48	0.94	187

Compared with the results of three random extraction, it can be seen that the entities in the community found by the three community construction methods have obvious correlation, and the community discovery effect is better. Through the analysis of the results of manual evaluation, average correlation degree and number of unassociated entities, it is found that the Glove model is the best community construction method, and the Skip-gram model is the worst, but compared with the results obtained by random extraction, the three community construction methods have obvious effects. Analyzing the reasons, the Glove model used the co-occurrence matrix of the corpus in training, and introduced the correlation information embodied in the corpus to some extent, so it performed better in the community discovery task. During training, the CBOW model can simultaneously learn all the context information around the central word. In the learning process, because the neighboring entity information is not directly used, the result is poor in the clustering task. However, the introduction of association is equivalent to the introduction of the information of the neighboring entity, so the performance is better. The Skip-gram model directly uses the relationship between neighboring entities in the training process, but compared with the CBOW model, the context information contained by the entity vector is inferior to that contained by the CBOW model because the context features utilized by the Skip-gram model are less than that of the CBOW model. Therefore, the entity vector performs better in entity clustering and worse in community discovery tasks than the CBOW model.

5. **Closing Remarks.** This paper proposes an approach to modeling news entities. Firstly, the news data of "Voice of America" was obtained by using web crawlers as the research object, and three distributed representation methods of news entities were proposed based on the distributed representation of words, and the learned entity vectors were used for three tasks: entity clustering, association information mining and community discovery, and the experimental results showed that the distributed representation of entities had a good effect. In this paper, we also propose an evaluation method for the clustering results and community findings for this task, test the effects of the three

models and analyze the reasons. Experimental results show that the results based on the Skip-gram model perform well in the clustering results, while the results based on the Glove model have more advantages in the community discovery task. Using the above conclusions, key entities and related groups can be identified from the media, and association information mining can be used to obtain more information about key entities. In practical application, we have found such politicians as "Zhifeng Huang", "Ting Zhou", "Guancong Luo", and "Zhiying Li", as well as related organizations such as "Demosisto", "Student Movement", "Sunflower", and "Next Media", which prove the effectiveness of this method.

At present, this method only uses some semantic information without deeper analysis. Next, we will use graph neural network and graph embedding method to model related entities. In addition to the co-occurrence of entities in the text, the relationship between entities is further considered, and the relationship diagram and attribute diagram of entities are constructed. The final research goal is to use the entity-related text to directly mine the entity-related information, and obtain the in-depth analysis results of the entity through the entity modeling and association analysis, and directly form the analysis report of the relevant sensitive people.

Acknowledgment. This work is partially supported by The Youth Innovation Promotion Association of the Chinese Academy of Sciences (E1291902), Jun Zhou (2021025). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] A. Hossain, M. Karimuzzaman, M. M. Hossain et al., Text mining and sentiment analysis of newspaper head-lines, *Information*, vol.12, no.10, 414, 2021.

[2] H. Kim and M. Park, Discovering fashion industry trends in the online news by applying text mining and time series regression analysis, *Heliyon*, vol.9, no.7, 2023.

[3] N. Kanungsukkasem and T. Leelanupab, Financial latent Dirichlet allocation (FinLDA): Feature extraction in text and data mining for financial time series prediction, *IEEE Access*, vol.7, pp.71645-71664, 2019.

[4] D. J. Pyo and J. Kim, News media sentiment and asset prices in Korea: Text-mining approach, *Asia-Pacific Journal of Accounting & Economics*, vol.28, no.2, pp.183-205, 2021.

[5] R. Phann, C. Soomlek, P. Janyoi and P. Seresangtakul, Multi-class text classification on Khmer news articles using deep learning models with optimal hyperparameters, *ICIC Express Letters*, vol.18, no.6, pp.541-550, 2024.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas et al., Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, vol.41, no.6, pp.391-407, 1990.

[7] M. Mahdikhani and P. Meena, Metaverse applications and supply chain innovation: In-sights from text mining, *Journal of Innovation & Knowledge*, vol.9, no.4, 100591, 2024.

[8] T. Hofmann, Probabilistic latent semantic analysis, *The 15th Conference on Uncertainty in Artificial Intelligence (UAI)*, vol.99, pp.289-296, 1999.

[9] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.

[10] M. Hoffman, F. Bach and D. Blei, Online learning for latent Dirichlet allocation, *Advances in Neural Information Processing Systems*, vol.23, 2010.

[11] S. Moghaddam and M. Ester, ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews, *Proc. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.665-674, 2011.

[12] J. Yin and J. Wang, A Dirichlet multinomial mixture model-based approach for short text clustering, *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.233-242, 2014.

[13] Y. Bengio, R. Ducharme and P. Vincent, A neural probabilistic language model, *Advances in Neural Information Processing Systems*, vol.13, 2000.

[14] T. Mikolov, Efficient estimation of word representations in vector space, *Computer Science*, 3781, 2013.
[15] J. Pennington, R. Socher and C. D. Manning, Glove: Global vectors for word representation, *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.1532-1543, 2014.
[16] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Physical Review E*, vol.69, no.2, 026113, 2004.
[17] U. N. Raghavan, R. Albert and S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, *Physical Review E – Statistical, Nonlinear, and Soft Matter Physics*, vol.76, no.3, 036106, 2007.
[18] U. V. Luxburg, A tutorial on spectral clustering, *Statistics and Computing*, vol.17, pp.395-416, 2007.
[19] V. D. Blondel, J. L. Guillaume, R. Lambiotte et al., Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, vol.2008, no.10, P10008, 2008.
[20] T. Mikolov, I. Sutskever, K. Chen et al., Distributed representations of words and phrases and their compositionality, *Advances in Neural Information Processing Systems*, vol.26, 2013.
[21] E. Agichtein and L. Gravano, Snowball: Extracting relations from large plain-text collections, *Proc. of the 5th ACM Conference on Digital Libraries*, pp.85-94, 2000.
[22] L. Page, *The Pagerank Citation Ranking: Bringing Order to the Web*, Technical Report, 1999.

Author Biography

Jun Zhou obtained his Ph.D. degree in Computer Software and Theory from the Institute of Information Engineering, Chinese Academy of Sciences, China, in 2017. He is now working as a researcher in the Institute of Acoustics, Chinese Academy of Sciences, China. His main research interest is natural language processing such as semantic retrieval and large language model.

Ce Zhou obtained his B.Eng. degree in Computer Science and Technology from Southwest University of Science and Technology, China in 2025. He is studying for a master's degree at Soochow University, China. His main research interests include deep learning.