




Intelligent Text Clustering Analysis of Novels Based on Digital Semantic Similarity Calculation

Xiaoming Sun^{1*} 

¹School of Marxism/School of General Education, Liaoning University of International Business and Economics, Liaoning, China

Corresponding Author: Xiaoming Sun, hexintougao1@163.com

Abstract. Faced with the increasing number of novels on the Internet, how to quickly and effectively retrieve novels that meet the needs of users has become a difficult problem for novel retrieval. This paper explores the clustering analysis of novel texts based on semantic similarity calculation. This paper improves a text clustering algorithm based on semantic similarity and applies it to novel text clustering. The improved algorithm puts forward a feature selection method and a clustering description method suitable for novel texts. In this paper, the algorithm overcomes the problem of insufficient accuracy of text similarity due to incomplete semantic consideration of traditional similarity algorithm, and describes the semantic similarity of synonyms and synonyms more accurately. At the same time, the influence of threshold selection on clustering results is overcome, and the clustering accuracy and time efficiency are improved. Experiments show that the clustering accuracy of this algorithm can reach 94.34%, which is 10.89% higher than that of K-means. The clustering method proposed in this paper has high comprehensiveness, and it can improve the accuracy of text clustering and the efficiency of clustering algorithm. This study provides an idea for novel text clustering analysis.

Key words: Semantic similarity; Fiction; Text clustering

DOI: <https://doi.org/10.14733/cadaps.2024.S16.199-213>

1 INTRODUCTION

Language is not only a tool for human communication, but also a tool for human artistic creation, which is closely related to the construction of human knowledge and cognitive forms. The generalized structure of novel readers' subjective feelings is rooted in the potential psychological space of the text [23]. By discussing the novel text at the lexical level from the psychological perspective, we can get rid of the limitation of the storyline, and show the psychological elements inadvertently revealed in the diction of the work and the psychological space formed. Faced with the increasing number of

novels on the Internet, how to quickly and effectively retrieve novels that meet the needs of users has become a difficult problem for novel retrieval. In addition, text analysis of classic novels can present readers with unique artistic works. Through text analysis, we can analyze the novel concretely [16]. Exploring the psychological dimension rooted in the text in the form of dispersion is also helpful to better understand readers' psychology and personality images in novels, and enrich relevant psychological theories. This meta-level analysis based on text vocabulary has become possible with the emergence of intelligent text analysis [2]. Text clustering is an effective text analysis technology. Text clustering, as a clustering method, can help us find out data rules from massive text data. It is of great significance to analyze the text clustering of novels. By analyzing the text of narrative works and putting forward opinions, it plays a great role in the promotion and dissemination of novels [15]. This will help to sublimate literary theory and help readers appreciate novels. Therefore, the research on intelligent text clustering analysis of novels in this paper has certain theoretical and practical significance.

Text clustering is a typical unsupervised machine learning problem. It divides a text set into several classes, the members of each class have great similarity, while the texts between classes have little similarity. Text clustering is an important means and method of text mining and an important branch of data mining [4]. It has a high capability of automatic document processing. Because the tedious text category labeling problem is avoided in the training process, text clustering becomes an important means in text processing. At present, many text clustering algorithms have been proposed [7]. Typical clustering methods are K-means and K-Medoids based on partition, AGENS and DIANA based on hierarchy, DBSCAN based on density and ant colony based [14]. At present, the most widely used clustering algorithms are K-means, SOM clustering algorithm and its improved algorithm. Of course, hierarchical clustering algorithm and density-based clustering method are still applied in various fields of clustering. According to the goal of clustering, text clustering makes the similarity of themes and contents of texts within a class high, while the similarity of texts between classes is low. In addition, cluster analysis can be used as a preprocessing step for other algorithms such as classification [18]. Firstly, the existing document sets are processed by clustering algorithm, and the cluster structure of the existing document sets is found. Then use these clusters to generate an effective document classifier to classify new documents. The traditional text clustering algorithm ignores the semantic correlation between words in the text, and there are some problems such as unstable clustering results. This paper improves the traditional text clustering algorithm. Based on this, this paper explores the intelligent text clustering analysis of novels based on semantic similarity calculation. Its innovations are as follows:

⊙ This paper improves a text clustering algorithm based on semantic similarity and applies it to novel text clustering. It overcomes the problem of insufficient accuracy of text similarity due to incomplete semantic consideration of traditional similarity algorithm, and describes the semantic similarity of synonyms and synonyms more accurately.

⊙ In this paper, a feature dimensionality reduction algorithm based on concept clustering is proposed, which aims to extract coarse-grained features from the text through concept clustering, so as to reduce the dimension of text representation. This algorithm overcomes the influence of threshold selection on clustering results, and improves clustering accuracy and time efficiency.

The full text is divided into five sections, and the specific organizational structure is as follows: The first section is the introduction of this article. This section first summarizes the research background and practical significance of this topic, and then gives the innovation and organizational structure of the article. The second section summarizes the research status of text clustering, and finally introduces the main research work of this paper. The third section introduces the related knowledge and technology of text clustering in detail. Then, the principle and concrete implementation of the intelligent text clustering analysis algorithm proposed in this paper are given. Finally, the concrete implementation process of applying the algorithm to novel clustering analysis

is described. In the fourth section, the clustering algorithm proposed in this paper is tested and the results are analyzed. The fifth section summarizes the research work and looks forward to the future research direction.

2 RELATED WORK

Pilehvar M T and others proposed a suffix tree clustering algorithm, and applied this technology to the visualization of search results, and achieved good results [12]. Jiang Y et al. improved the K-means by using the maximum distance method to select the initial cluster center and dynamically adjusting the number of categories of clusters [8]. It overcomes the shortcomings of the K-means and improves the stability and practicability of the algorithm. Taieb M et al. explored a new method of text clustering on the basis of text semantic similarity calculation, and proposed two clustering algorithms: iterative semantic clustering algorithm and clustering algorithm based on weighted topic concept map [17]. Peng J et al. Proposed a text clustering algorithm based on semantic list. The algorithm uses semantic similarity to calculate the similarity of text and obtain the semantic relevance of text; The synonym and synonym pointer in the semantic list are used to reduce the redundancy of words and the dimension of text data; Finally, the text is clustered by the partition based clustering algorithm [11]. Based on the effectiveness of neural network language model in semantic feature extraction research, Godoy D et al. explored to encode the paraphrase relationship between concepts in Word Net into a concept corpus; to learn semantic representations of concepts [5]. The swarm intelligence text clustering algorithm proposed by Cao J et al. is based on text similarity. It takes advantage of the fast classification characteristics of K-means, the advantages of ant colony algorithm and simulated annealing algorithm, and avoids their respective shortcomings [3]. Zhang X et al. combined with Quillian's joint concept distance calculation method, and proposed a calculation method of semantic similarity between texts [22]. Experiments show that the algorithm has a higher classification accuracy than previous text clustering algorithms. The algorithm proposed by Xing F Z et al. introduces the maximum and minimum distance algorithm in the initial stage, which makes the initial value setting rational and the initial point distribution uniform. The K-means is added in the algorithm execution process, and the cluster centers obtained each time in the bee colony algorithm search process are locally updated to obtain better cluster centers. This not only speeds up the running pace of the algorithm, but also makes the algorithm more optimal and more robust [19]. Lang Q et al. proposed a new method for swarm-intelligent text clustering based on semantic similarity [9]. This method combines the global search of simulated annealing algorithm and the positive feedback ability of ant colony algorithm.

At present, most clustering algorithms are based on VSM (Vector space model), and their performance and accuracy are not high. Based on this, this paper explores the clustering analysis of novel texts based on semantic similarity calculation. This paper improves a text clustering algorithm based on semantic similarity and applies it to novel text clustering. When calculating the weight in the feature extraction stage of text preprocessing, this paper not only considers the word frequency and document frequency, but also combines the part of speech of words and the position of words in the text. The research shows that this method not only improves the accuracy of text clustering, but also improves the efficiency of clustering algorithm. It overcomes the problem of insufficient accuracy of text similarity due to incomplete semantic consideration of traditional similarity algorithm, and describes the semantic similarity of synonyms and synonyms more accurately.

3 Methodology

3.1 Key Technologies of Text Clustering

The process of dividing a set of abstract or concrete objects into similar object classes is called clustering. Represents the collection of cluster objects. Objects in the same cluster are similar to

each other, but objects in different clusters are different from each other. Clustering results as the basis of classification: text classification is an important direction in text mining, and text classification is to determine the category of new texts according to existing knowledge, so as to achieve the purpose of automatic classification [21]. Firstly, classification needs to have a sample set of a known category in advance, and the sample set of this known category should be trained and studied to learn the description information of the category, and then the unknown data set should be classified. Clustering is that there is no need for example learning, and there are no pre-classified categories. Data can be classified by clustering. Text clustering is the clustering of text data, so that the text data with similar information can be gathered and divided into a class. Text clustering is a process of automatically classifying text objects according to their semantic similarity. The difference between classification and clustering is that classification requires some text sets with class labels as input training sets, while clustering does not. Therefore, classification is guided learning, while clustering is unsupervised learning. When the similarity within the same group is as large as possible, and the similarity between different groups of texts is as small as possible, it indicates that the accuracy of text clustering results is relatively high. Because text clustering does not need training process, nor does it need to manually label the categories of texts in advance, it makes text clustering have high flexibility and automatic processing ability [20]. Text clustering is an important means and method of text mining and an important branch of data mining. The data processed by text clustering is a set of texts in the form of natural language. Because of the unstructured or semi-structured characteristics of text data, it lacks semantic information that can be understood by computers. Therefore, it is necessary to convert the text into a structured form that can be processed by a computer.

Search engine allows users to input search keywords and cluster the retrieved documents; And output a brief description of each different category. Clustering the results returned by the search engine, so that users can quickly locate the required information; Allow users to input search keywords and cluster the retrieved documents; And output a brief description of each different category. In this way, the scope of search can be narrowed, and users only need to pay attention to the topics they are interested in. Text clustering is classified according to the distance of text in nature [6]. In order to make the classification result as reasonable as possible, it is necessary to measure the distance between texts correctly. This requires the definition of some effective measurement indicators for classification. Commonly used statistical indicators include distance and similarity. Text clustering can be roughly divided into three stages: ⊖ Text preprocessing. Word segmentation technology is used to segment the text, and the stop words in the text is filtered. ⊕ Text representation. Standardize text processing, description and representation, feature extraction and feature weight calculation, extract feature words and establish feature space. ⊗ Text clustering. After the text information is processed, the text is scientifically clustered according to a certain metric clustering algorithm. The process of text clustering is shown in Figure 1.

The input of text clustering is generally a text set containing multiple documents. To cluster text, the first thing to do is to describe the text data mathematically, and the most commonly used mathematical model is VSM. VSM is one of the most widely used and effective text representation methods in recent years. This model is a statistical model with feature items as the basic unit of text representation, which is simple and effective. In this description model, each different word is regarded as one dimension in the feature space, and each text is a vector in the feature space. This description method leads to a very serious problem, that is, high-dimensional sparseness. High-dimensional sparsity has a considerable impact on text clustering, which not only makes text clustering have a high time complexity, but also greatly reduces the performance of text clustering and text classification.

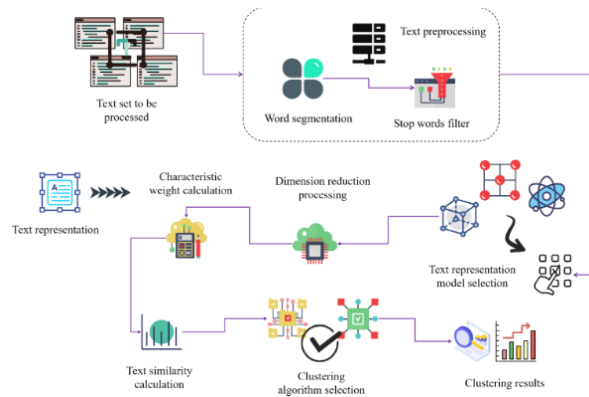


Figure 1: The process of text clustering.

3.2 Text Preprocessing

Before clustering the text set, it is necessary to preprocess the text so that the text information can be clustered reasonably. Text preprocessing technology is a very important link for text mining. It can be said that the quality of pretreatment directly affects the final mining results [1]. At the same time, there are different preprocessing methods for different mining purposes. When extracting feature words from the preprocessed text, the order of words in the text is generally ignored, and it is directly regarded as the form of a collection of words. Text preprocessing is to transform human natural language into information that can be understood by computers. This process includes several steps: word segmentation, stop words removal and feature extraction. After preprocessing, each text will be represented by multiple feature words, and the value of special testimony in a certain text reflects its importance in that text.

(1) The first key problem of preprocessing is word segmentation. Word segmentation is the process of dividing the text according to various character combinations. For Chinese, word segmentation is complicated. Because Chinese characters are connected together except punctuation marks between sentences, and its grammar is flexible [13]. In Chinese texts, words are generally composed of several Chinese characters. The statistical segmentation method uses statistical models to calculate the combination probability between adjacent words in the text. And think that the higher the probability, the closer the semantic relationship between them, and then regard it as a word. This method judges whether the corresponding words exist by matching the adjacent words in the text with the words in the dictionary one by one. Comprehension-based word segmentation makes the computer understand the semantics of the text according to the preset rules, and then makes further word segmentation on this basis.

(2) In the text, there are some words that have little significance for text content identification, which are called stop words in text mining. These words have no practical significance, and they appear frequently in all kinds of texts, which will lead to great errors in the process of feature selection or similarity calculation, and can be regarded as a kind of noise. Going to stop words can reduce the storage space of computers and improve the efficiency of text retrieval. Stop words filtering is realized by establishing a stop words list. Sorting the stop words in the corpus into a word list is the stop word list, that is, the stop word list [23]. It is used to identify words that need to be deleted frequently in the preprocessing process and are unlikely to help the post-processing of the text. The process of using stop words list to filter stop words is very simple, that is, a query process. For each entry, see if it is in the stop words list, and if so, delete it from the entry string.

(3) Feature selection can be divided into unsupervised and supervised methods. The supervision method is to construct evaluation function for feature selection. Evaluate each feature, get a score, and select a certain number of texts with high scores as the feature set of texts. Unsupervised feature selection methods include word frequency, feature enhancement and document frequency. The methods of feature extraction include the following: transforming the original features into fewer new features by mapping or transformation, selecting some of the most representative features from the original features, selecting the most influential features according to the knowledge of experts, and extracting the features with the most category information by mathematical algorithms. After preprocessing such as stem extraction, the entries contained in it are completely coincident. However, if different forms of the same word are considered to be different terms, since there is no morphological intersection between these two sentences, then in the computer's view, these two sentences are completely unrelated.

3.3 Clustering Analysis of Novel Texts Based on Semantic Similarity Calculation

In Chinese text clustering, the calculation of semantic similarity is the core part, which is used to describe the distance between texts. It can be said that the accuracy of similarity calculation between texts reflects the good or bad effect of text clustering to a certain extent. This paper explores how to measure the semantic similarity between the meaning of the target word and the context through the correlation between conceptual annotation and context. At this point, semantic disambiguation is transformed into the process of calculating the semantic similarity between two text fragments. For this process, this paper will adopt the strategy of scoring in turn, that is, each word in the context will score the concept annotation of the target word in Word Net, and the higher the score, the higher the correlation between the context word and the corresponding concept; Finally, the concept with the highest score is regarded as the meaning of the target word in the current context. And its processing process are shown in Figure 2.

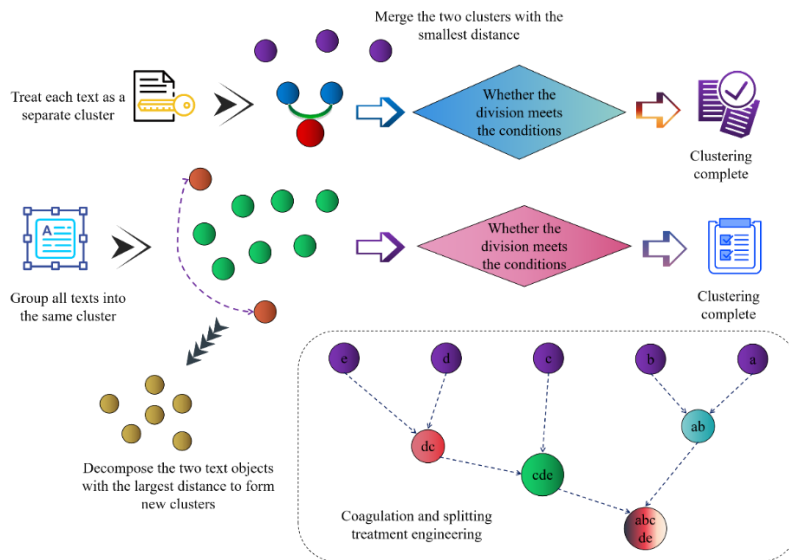


Figure 2: The algorithm of aggregation and splitting and its processing flow chart.

The definition of the semantic space model is as follows: Let D be a collection of texts, each novel document $d \in D$, represented by the eigenfrequency vector as:

$$\vec{d} = (f_1, f_2, f_3, \dots, f_n) \quad (1)$$

Usually the feature is the word in the document, and the matrix of $m \times n$ represents the set of documents; m is the number of documents in D , n is the number of features; the entry (i, j) contains the number of times the feature j appears in the document i . The similarity between two novels d_1 and d_2 is measured by the cosine of the angle between the two vectors:

$$\text{cosine}(\vec{d}_1, \vec{d}_2) = \frac{(\vec{d}_1 \cdot \vec{d}_2)}{\|\vec{d}_1\| \times \|\vec{d}_2\|} \quad (2)$$

Among them, $(\vec{d}_1 \cdot \vec{d}_2)$ represents the dot product of two vectors; $\|\vec{d}_i\|$ represents the modulus of the vector. A cosine value of 0 means the two vectors are uncorrelated, and a cosine of 1 means they are closely related. Given a cluster C in D , the cluster center c is defined as: to the minimum of the cosine sums of the other vectors in C .

$$c = \arg \min_{d_j \in C} \sum_{i=0, \dots, |C|} \text{cosine}(\vec{d}_i, \vec{d}_j) \quad (3)$$

Generally speaking, the weight selection of features should meet two basic requirements: \ominus It should be able to reflect the features of the represented documents. \ominus It is necessary to be able to distinguish the represented document from other documents. Therefore, it can be seen that the importance of an entry is directly proportional to the frequency TF of the entry appearing in the document and inversely proportional to the frequency DF of the document appearing in the training text.

Cosine similarity is to use the cosine value $\cos(A, B)$ of the included angle between two vectors A and B to represent the degree of difference between two novel texts. The closer $\cos(A, B)$ is to 1, the closer the angle between the two vectors is to 0 degrees, and the higher the similarity between the two vectors. The degree of similarity between two texts can be set as $\text{sim}(A, B)$. Then the novel texts A and B represent two vectors in the VSM as shown in Figure 3.

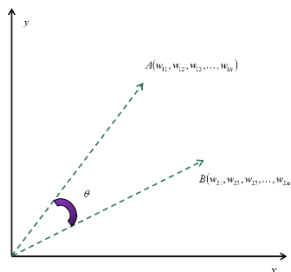


Figure 3: Schematic diagram of two vectors in cosine similarity calculation.

The two vectors in Figure 3 are:

$$A = A(w_{11}, w_{12}, w_{13}, \dots, w_{1n}) \quad (4)$$

$$B = B(w_{21}, w_{22}, w_{23}, \dots, w_{2n}) \quad (5)$$

Then, the VSM is represented by the distance formula between two vectors. Then, the inner product between vectors can be used to calculate:

$$Sim'(A, B) = \sum_{k=1}^n w_{1k} \times w_{2k} \quad (6)$$

Because it is generally necessary to consider the normalization of vectors, it is often expressed by the cosine of the angle between two vectors:

$$Sim(A, B) = \cos \theta = \frac{\sum_{k=1}^n w_{1k} \times w_{2k}}{\sqrt{\sum_{k=1}^n w_{1k}^2 \times \sum_{k=1}^n w_{2k}^2}} \quad (7)$$

Using the relationship between Word Net words to generate features to form vector space dimensions. There are two methods, one is to use Word Net vocabulary classification, and the other is to use Word Net ontology. Both methods can form feature vectors well and keep a low dimension, and they are used to generate the input of clustering algorithm. According to the semantic dictionary, all the related words of keywords are related to keywords, but it is impossible to cluster all these words as the characteristic words of the paper text. Because these related words are not necessarily related to the articles containing this keyword and can not reflect the characteristics of the text, we should screen these related words according to the specific analysis of specific articles. Calculate the weight according to the word frequency of words appearing in the novel text, and the weight calculation formula is as follows:

$$W_k = \frac{f_k}{\sum_{i=1}^n f_i} \quad (8)$$

That is, the weight of a related word is equal to the proportion of its frequency in the novel text to the frequency of all related words in the novel text. Here W_k represents the weight of the k th word; f_k represents the frequency of the word appearing in the text; n represents the number of related words of the keyword in the novel.

A novel text is composed of many keywords, and putting these keywords together can construct a vector reflecting the information of the text. Let $A_k = \{w_{k1}, w_{k2}, w_{k3}, \dots, w_{ki}, \dots, w_{kk}\}$ represent the

word set whose number of words is k ; w_{ki} represents the i th word in A_k . The similarity matrix of two novel texts A_k and A_p is denoted as $S_{kp} = (s_{ij})$, where:

$$s_{ij} = \text{WORDSIM}(w_{ki}, w_{pj}) \quad (9)$$

$$i = 1, 2, 3, \dots, k \quad j = 1, 2, 3, \dots, p \quad (10)$$

Most of the similarities between the words in the two texts are relatively low. Only the similarity between the words in one text and the words with the highest similarity in the other text is considered. Considering the synonyms and synonyms in the text, this paper overcomes the shortcomings of traditional cosine distance measurement and editing distance measurement in similarity calculation, which only consider the same words and ignore that most words are synonyms and synonyms.

The length of documents in the text set to be processed is different, which is lack of comparability. In order to express the novel text features more effectively and make the calculation convenient, it is usually necessary to normalize the novel text feature vectors. The processed weight function is:

$$w_{ik} = \frac{TF_{ik} \times \log\left(\frac{n}{n_k} + 0.01\right)}{\sqrt{\sum_{k=1}^m \left(TF_{ik} \times \log\left(\frac{n}{n_k} - 0.01\right) \right)^2}} \quad (11)$$

In this way, the feature vectors of n documents form a vector space:

$$w = [w_1^T, w_2^T, \dots, w_n^T]^T = (w_{ik})_{n \times m} \quad (12)$$

A characteristic word in TCUSS method is characterized by three elements: word, word frequency and the number of noun meanings of the word. This method is more suitable for the calculation of semantic similarity between words. However, the number of noun meanings of words is mainly used to solve the problem of polysemy, that is, semantic disambiguation. In this paper, polysemy has been eliminated in the process of feature word selection, so the ternary list is improved to make it more suitable for the algorithm in this paper. The improved concept list representation is shown in the formula:

$$D = \{(w_1, f_1, s_1), (w_2, f_2, s_2), \dots, (w_n, f_n, s_n)\} \quad (13)$$

Where w_i is a word that occurs in a document. f_i is the number of times w_i appears in the document. s_i is the node where the word w_i is located in Word Net, represented by the Synset_offset attribute of the node. This set of 8-bit decimal numbers indicates the offset of the concept node in the file, that is, the position of synset in the dictionary file. This item is added mainly for the convenience of calculating semantic similarity. Every word in the corpus will have a unique vector in the vector space to correspond to it, and the more relevant words in the context in the

corpus will be closer in the vector space. In this paper, by integrating semantic disambiguation algorithm based on continuous word vector and feature dimension reduction algorithm based on concept cluster, a text clustering algorithm based on continuous word vector and concept cluster is realized, which can not only ensure the accuracy of text clustering, but also effectively improve the efficiency of clustering algorithm.

4 RESULT ANALYSIS AND DISCUSSION

The experimental data of this paper is mainly from the open platform of Chinese natural language processing of Chinese Academy of Sciences. Download 100 documents from CNLP website as test data. The running environment is Windows operating system; 16 gigabytes of memory; The programming tool is Visual. The purpose of this paper is to compare and analyze the performance and quality of several clustering methods in the same sample set. In comparison, an algorithm based on semantic similarity and the classical K-means, which this paper mainly refers to, are selected. In Chinese text clustering, text preprocessing is the first step, in which the word segmentation technology is quite mature. Table 1 shows the accuracy of several common word segmentation methods.

Algorithm	Accuracy
THULAC_lite	0.912
ANSJ_ToAnalysis	0.931
ANSJ_NlpAnalysis	0.927
HanLP_Standard Tokenizer	0.924

Table 1: Accuracy comparison of several common word segmentation methods.

Considering the accuracy of word segmentation and the accuracy of part-of-speech tagging, this paper chooses ANSJ method to segment the text.

In the pre-processing part of the text, the process of extracting text features is an important step in the pre-processing part. In this paper, the weight of words in the text is calculated by combining the part of speech and the position of words in the text, and the 10 words with the largest weight are selected as the keywords of the text, which achieves good results. Accuracy and recall are used to evaluate the performance of the algorithm. The number of iterations of the algorithm is 1000. The experimental results of precision of different algorithms are shown in Figure 4. The experimental results of different recall rates are shown in Figure 5.

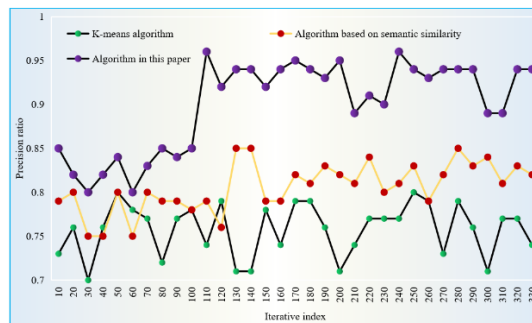


Figure 4: Experimental results of precision of different algorithms.

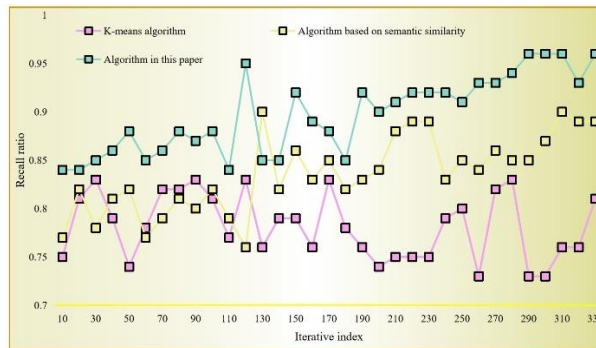


Figure 5: Experimental results of recall rate of different algorithms.

It can be seen that the precision and recall of this algorithm are at a high level. F is the harmonic average of accuracy and recall. Sometimes we think that the weight of accuracy and recall is the same with F1-measure, but sometimes we think that accuracy will be more important because of different scenarios, so we need to adjust the weight ratio. In this paper, the F value is also used for many experiments. Comparing this algorithm with K-means and K-Medoes algorithm, the results are shown in Figure 6.

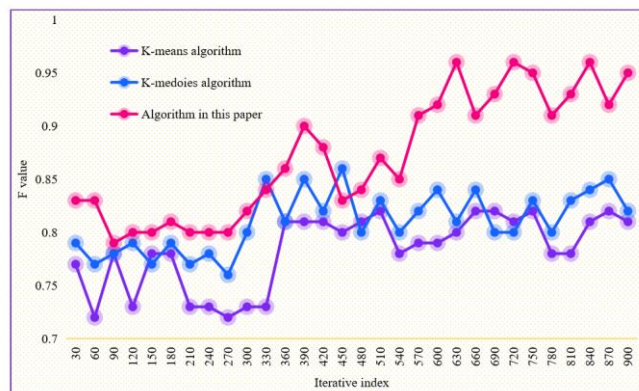


Figure 6: Comparison results of F values of different algorithms.

According to the results in the figure, compared with K-means and K-Medoes algorithm, the F value of this algorithm is better. The algorithm was tested 40 times. Table 2 shows the comparison results of clustering accuracy, and Table 3 shows the comparison results of clustering recall.

Algorithm	1~10	11~20	21~30	31~40
K-means	0.769	0.798	0.754	0.779
K-medoes	0.802	0.811	0.809	0.814
AGENS	0.752	0.746	0.779	0.763
Algorithm in this paper	0.953	0.941	0.965	0.948

Table 2: Comparison results of clustering accuracy.

Algorithm	1~10	11~20	21~30	31~40
K-means	0.795	0.792	0.788	0.795
K-medoids	0.843	0.837	0.846	0.841
AGENS	0.813	0.819	0.807	0.811
Algorithm in this paper	0.961	0.957	0.941	0.959

Table 3: Comparison results of cluster recall rate.

Through the comparison of the above results, it can be seen that the novel text clustering algorithm adopted in this paper is far higher than K-means clustering algorithm, K-Medoes algorithm and other algorithms, both in clustering accuracy and clustering recall. This verifies the effectiveness of this method.

There are many evaluation criteria of clustering quality, such as fuzzy matrix, entropy, overall similarity and classification accuracy. Among them, the method of classification accuracy is more intuitive and easier to calculate. Therefore, this paper uses this method to evaluate the clustering quality, and the correctness of the classification is checked by manual analysis. Clustering accuracy reflects the degree to which similar text units and dissimilar text units are merged into the same class, and reflects the ability to distinguish different topics. The higher the clustering accuracy, the more concentrated the content in each class. Figure 7 shows the clustering accuracy of different algorithms.

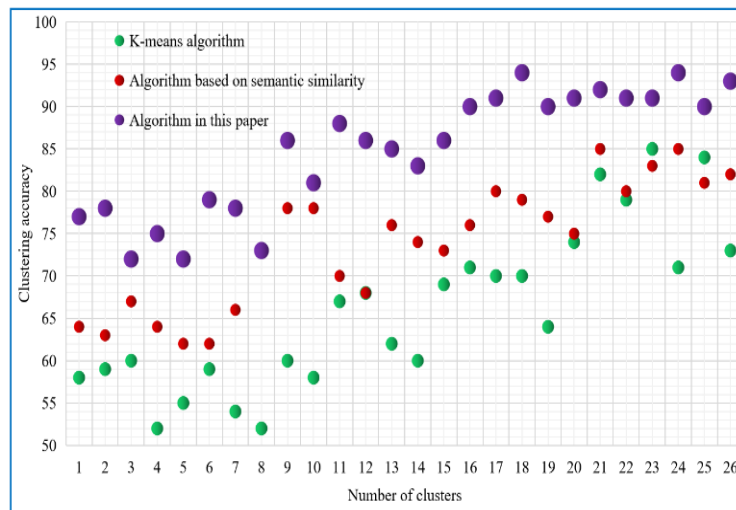


Figure 7: Clustering accuracy obtained by different algorithms.

As shown in the figure, the accuracy of this algorithm is higher than that of the comparison algorithm. This is mainly due to the fact that the algorithm in this paper is optimized for the novel text, and the feature selection part can better highlight the theme of the article and provide a better basis for clustering. Therefore, the clustering effect applied to the novel text set has certain advantages over the other two algorithms widely used in text clustering, and it has achieved good results in novel text clustering. Figure 8 shows the time efficiency comparison results of different algorithms.

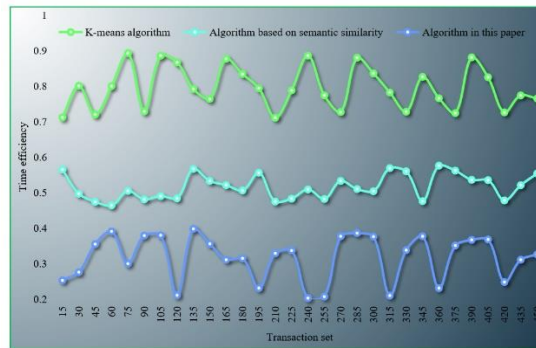


Figure 8: Comparison of time efficiency of different algorithms.

According to the information in the figure, the time consumption of semantic-based clustering algorithm is obviously higher than that of this algorithm. The reason is that semantic-based clustering requires frequent calculation of semantic similarity between words, and the calculation of semantic similarity is more complicated than the mathematical operation between vectors. In this paper, the algorithm first uses K nearest neighbor distance to sort the objects, then distinguishes the sequences with different densities by quantiles, and finds the corresponding optimization. According to the optimization threshold, the density clustering method is used to cluster the objects, which effectively improves the clustering efficiency.

The experimental results in this section prove that the clustering accuracy of this algorithm can reach 94.34%, which is 10.89% higher than that of K-means. The clustering results of this paper are basically consistent with the categories of documents, and are not disturbed by noise. The clustering method proposed in this paper has high comprehensiveness, and it can improve the accuracy of text clustering and the efficiency of clustering algorithm.

5 CONCLUSIONS

Text clustering plays an important role in many text mining and information retrieval systems. This technology can improve retrieval performance and find similar texts. At present, there are more and more novels on the Internet, so how to quickly and effectively search out the novels that meet the needs of users has become a research hotspot. This paper explores the intelligent text clustering analysis of novels based on semantic similarity calculation. Because of the similarity between entity relation extraction and text classification, this paper introduces a feature selection algorithm in text classification to solve the problem of excessive dimension of feature space in entity relation extraction based on feature vector. In this paper, when calculating the weight in the feature extraction stage of text preprocessing, we not only consider the word frequency and document frequency, but also combine the part of speech of words and the position of words in the text. The algorithm combines graph theory for clustering analysis, avoiding the restriction of clustering shape by the algorithm. The word frequency of feature words in the whole cluster and the amount of information contained in Word Net are used to measure the weight of feature words. Select some feature words with significant weight to cluster description. It overcomes the problem of insufficient accuracy of text similarity due to incomplete semantic consideration of traditional similarity algorithm, and describes the semantic similarity of synonyms and synonyms more accurately. Experiments show that the clustering accuracy of this algorithm can reach 94.34%, which is 10.89% higher than that of K-means. It not only improves the accuracy of text clustering, but also improves the efficiency of clustering algorithm. However, only a small number of texts are selected for testing.

If the number of texts increases, whether the clustering results are satisfactory or not, and whether the clustering algorithm needs to adjust the parameters, further research is needed.

Xiaoming Sun, <https://orcid.org/0009-0004-1534-5033>

REFERENCES

- [1] Al-Smadi, M.; Jaradat, Z.; Al-Ayyoub, M.: Paraphrase Identification and Semantic Text Similarity Analysis in Arabic News Tweets Using Lexical, syntactic, and Semantic Features, *Information Processing & Management*, 53(3), 2017, 640-652. <https://doi.org/10.1016/j.ipm.2017.01.002>
- [2] Cambria, E.; Song, Y.; Wang, H.: Semantic Multidimensional Scaling for Open-Domain Sentiment Analysis, *Intelligent Systems, IEEE*, 29(2), 2014, 44-51. <https://doi.org/10.1109/MIS.2012.118>
- [3] Cao, J.; Wu, Z.; Wu, J.: SAIL: Summation-bAsed Incremental Learning for Information-Theoretic Text Clustering, *IEEE Transactions on Cybernetics*, 43(2), 2013, 570-584. <https://doi.org/10.1109/TSMCB.2012.2212430>
- [4] Chandrasekaran, D.; Mago, V.: Evolution of Semantic Similarity—A Survey, *ACM Computing Surveys*, 54(2), 2021, 1-37. <https://doi.org/10.1145/3440755>
- [5] Godoy, D.; Rodriguez, G.; Scavuzzo, F.: Leveraging Semantic Similarity for Folksonomy-Based Recommendation, *IEEE Internet Computing*, 18(1), 2014, 48-55. <https://doi.org/10.1109/MIC.2013.26>
- [6] Hui, Z.; Wang, D.; Li, W.: A Semantics-Based Method for Clustering of Chinese Web Search Results, *Enterprise Information Systems*, 8(1), 2014, 147-165. <https://doi.org/10.1080/17517575.2013.857793>
- [7] Jiang, Y.; Bai, W.; Zhang, X.: Wikipedia-Based Information Content and Semantic Similarity Computation, *Information Processing & Management*, 53(1), 2016, 248-265. <https://doi.org/10.1016/j.ipm.2016.09.001>
- [8] Jiang, Y.; Zhang, X.; Yong, T.: Feature-Based Approaches to Semantic Similarity Assessment of Concepts Using Wikipedia, *Information Processing & Management*, 51(3), 2015, 215-234. <https://doi.org/10.1016/j.ipm.2015.01.001>
- [9] Lang, Q.; Pan, X.; Liu, X.: A Text-Granulation Clustering Approach With Semantics for E-Commerce Intelligent Storage Allocation, *IEEE Access*, 8, 2020, 164282-164291. <https://doi.org/10.1109/ACCESS.2020.3021421>
- [10] Martinez-Gil, J.: CoTO: A Novel Approach for Fuzzy Aggregation of Semantic Similarity Measures, *Cognitive Systems Research*, 40(dec.), 2016, 8-17. <https://doi.org/10.1016/j.cogsys.2016.01.001>
- [11] Peng, J.; Zhang, X.; Hui, W.: Improving the Measurement of Semantic Similarity by Combining gene Ontology and Co-Functional Network: a Random Walk Based Approach, *BMC Systems Biology*, 12(2), 2018, 109-116. <https://doi.org/10.1186/s12918-018-0539-0>
- [12] Pilehvar, M. T.; Navigli, R.: From Senses to Texts: An all-in-one Graph-Based Approach for Measuring Semantic Similarity, *Artificial Intelligence*, 228, 2015, 95-128. <https://doi.org/10.1016/j.artint.2015.07.005>
- [13] Saif, A.; Ab Aziz M, J.; Omar, N.: Reducing Explicit Semantic Representation Vectors using Latent Dirichlet Allocation, *Knowledge-Based Systems*, 100,2016, 145-159. <https://doi.org/10.1016/j.knosys.2016.03.002>
- [14] Shafiee, F.; Shamsfard, M.: Similarity Versus Relatedness: A Novel Approach in Extractive Persian Document Summarisation, *Journal of Information Science*, 44(3), 314-330. <https://doi.org/10.1177/0165551517693537>
- [15] Shirakawa, M.; Nakayama, K.; Hara, T.: Wikipedia-Based Semantic Similarity Measurements for Noisy Short Texts Using Extended Naive Bayes, *IEEE Transactions on Emerging Topics in*

- Computing, 3(2), 2017, 205-219. <https://doi.org/10.1109/TETC.2015.2418716>
- [16] Steiner, T.; Verborgh, R.; Gabarro, J.: Clustering Media Items Stemming from Multiple Social Networks, *Computer Journal*, 58(9), 2015, 1861. <https://doi.org/10.1093/comjnl/bxt147>
- [17] Taieb, M.; Aouicha, M. B.; Hamadou, A. B.: Ontology-Based Approach for Measuring Semantic similarity, *Engineering Applications of Artificial Intelligence*, 36(nov.), 2014, 238-261. <https://doi.org/10.1016/j.engappai.2014.07.015>
- [18] Wang, J. H.; Zuo, X. L.; Zuo, W. L.: Word Semantic Similarity Measurement Based on Evidence Theory, *Acta Automatica Sinica*, 41(6), 2015, 1173-1186.
- [19] Xing, F. Z.; Cambria, E.; Welsch, R. E.: Intelligent Asset Allocation via Market Sentiment Views, *Computational Intelligence Magazine*, 13(4), 2018, 25-34. <https://doi.org/10.1109/MCI.2018.2866727>
- [20] Yang, S.; Lu, W; Yang, D.: KeyphraseDS: Automatic Generation of Survey by Exploiting Keyphrase Information, *Neurocomputing*, 2017, 224, 58-70. <https://doi.org/10.1016/j.neucom.2016.10.052>
- [21] Zamora, J.; Mendoza, M.; Allende, H.: Hashing-Based Clustering in High Dimensional Data, *Expert Systems with Applications*, 62, 2016, 202-211. <https://doi.org/10.1016/j.eswa.2016.06.008>
- [22] Zhang, X.; Liu, C.: Image Annotation Based on Feature Fusion and Semantic Similarity, *Neurocomputing*, 149(11), 2015, 1658-1671. <https://doi.org/10.1016/j.neucom.2014.08.027>
- [23] Zhu, G.; Iglesias, C. A.: Exploiting Semantic Similarity for Named Entity Disambiguation in Knowledge Graphs, *Expert Systems with Applications*, 101, 2018, 8-24. <https://doi.org/10.1016/j.eswa.2018.02.011>