International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 7696-7705

https://doi.org/10.70517/ijhsa463660

Theme Extraction and Analysis of Modern Chinese Literature Texts Based on Vector Space Modeling

Yiling Sun[1,*]
[1] Faculty of Art and Social Science, National University of Singapore, 119077, Singapore
Corresponding authors: (e-mail: sunyiling2024@163.com).

Abstract Modern Chinese literature contains rich thematic connotations. This study proposes a method for extracting and analyzing themes of modern Chinese literature texts based on vector space model. Firstly, the text is preprocessed, including data cleaning, word splitting and deactivation word removal; then the text is transformed into multi-dimensional vector representation by using vector space model and the text feature weights are calculated by TF-IDF; finally, a two-stage clustering strategy is designed, in which the number of class clusters and centers are estimated by Canopy algorithm, and then fine classification is performed by K-means algorithm. The experimental results show that when the number of topics is set to 7, the model perplexity is the lowest at 6.646, the clustering precision rate reaches 0.81, the recall rate is 0.796, and the F-measure value is 0.802, which is obviously better than other settings of the number of topics. By analyzing 31,226 data of modern Chinese literature, seven major themes are successfully extracted: criticism of nationalism, oppression of feudal rites, enlightenment and salvation, cultural conflict between urban and rural areas, dilemma of women's awakening, writing of war sufferings, and uncertainty of intellectuals. The study shows that the vector space model combined with the optimized K-means algorithm can effectively identify the thematic features in modern Chinese literature and provide data support for literary research.

Index Terms vector space model, text theme extraction, modern Chinese literature, clustering algorithm, Canopy, K-means

I. Introduction

Grasping the theme of a text is an extremely important part of both literary studies and daily reading. In traditional text reading and analysis scenarios, one's access to themes relies mainly on human intuition or logical generalization, which is the most common way when the number of texts is small [1]. However, when it is necessary to extract possible themes from a large number of texts, or to analyze or compare themes contained in numerous works over a period of time, it is often beyond the reach of limited and sometimes unreliable human power [2], [3]. The computer, with its speed, accuracy, powerful processing capabilities and impressive performance in terms of intelligence, has become the researcher's most capable assistant [4]. In fact, the necessity to process a large number of texts with the help of computers has become a reality that language and literature researchers cannot ignore.

Topic analysis, known as “topic modeling” in the field of natural language processing, is currently one of the most interesting trends in this field [5]. On the one hand, algorithms can be used to create a “topic label” for each text, indicating the topic to which the text belongs and the probability of each topic appearing in the text [6], [7]. At the same time, the algorithm creates a “keyword list” for each topic, listing the words that best represent the topic and the probability of occurrence of each word [8], [9]. In this way, the theme analysis technique can identify the potential themes of literary works by analyzing the key words in a large number of texts, and at the same time, it can classify a large number of texts according to the themes like an “automatic classifier” [10]-[12]. This is to get rid of the drawbacks of traditional text analysis that requires manual identification of referential relations within literary texts in order to construct a thematic content model of the text, and to automate the task of literary text theme extraction so that thematic analysis can be performed more quickly and efficiently [13]-[16]. Therefore, computer-assisted exploratory tools oriented to textual analysis are the key to help researchers obtain potential directions and clues for modern Chinese literary textual studies from large corpora, so that they can more effectively utilize their professional knowledge to conduct in-depth examinations [17], [18].

Literature is an important embodiment of the social culture of a specific era, containing rich ideological connotations and values. Modern Chinese literature, as a literary form in the early to mid-twentieth century, reflects the changes and ideological conflicts of Chinese society in that era, and has important research value. Traditional

literary research mainly relies on researchers' subjective reading and textual analysis, which, although in-depth and detailed, is inefficient and difficult to avoid subjective bias when facing a large number of texts. Therefore, how to efficiently and objectively extract thematic features from a large number of literary works has become an important issue in current literary research. With the development of computer technology and natural language processing, the text analysis method based on vector space model provides new ideas for literary research. The vector space model represents the text as a multidimensional vector, which makes the text analysis can be carried out with the help of mathematical methods, and greatly improves the objectivity and reproducibility of the analysis. However, the application of vector space modeling in literary text analysis is not deep enough in the existing research, especially for the special context of modern Chinese literature. Modern Chinese literature has diverse linguistic styles and rich thematic connotations, and how to accurately capture its thematic features through vector space modeling is still a problem to be solved. In addition, clustering algorithms play an important role in text theme analysis, but the traditional K-means algorithm needs to set the number of class clusters in advance, which often lacks scientific basis in practical applications.

This study is based on the vector space model, combined with the two-stage clustering strategy, to extract and analyze the themes of modern Chinese literature texts. The study first preprocesses the text, including data cleaning, word segmentation and deactivation removal; then the text is transformed into a multidimensional vector representation using the vector space model, and the feature weights of the text are calculated by the TF-IDF; finally, a two-stage clustering strategy is designed, in which the number of class clusters and centers are estimated by the Canopy algorithm, and then fine classification is carried out by the K-means algorithm. Through experimental verification, the optimal number of topics is determined, the clustering effect is analyzed, and the extracted topics are interpreted. The study aims at constructing an efficient and objective framework for analyzing the themes of modern Chinese literature texts, providing data support for literary research, and exploring the application value of vector space models in humanities research.

II. Research on text topic clustering based on vector space modeling

II. A. Text pre-processing

The data collected through web crawler tools contain a lot of irrelevant and noisy information. Therefore, before clustering the data, the first step is to perform cleaning operations on the collected data. The processing of noisy data is very necessary. After the processing of the noisy data, then text segmentation, word line labeling, and removal of deactivated words are performed for better topic extraction.

Since words are the smallest constituent units that can be used independently, and there are many isolated and sticky texts like Chinese, where there is no obvious pause mark between words to indicate word boundaries, automatic word segmentation and deletion of deactivated words in natural language processing is the first and foremost basic work of text mining, which is an indispensable and important part of many text applications. At present, there are three main categories of Chinese word separation methods, based on string matching, based on understanding, and based on statistics.

After the text segmentation, the words in the sentence will be divided, but some words have no real meaning and appear frequently in the document, these words are called deactivated words, which include connectives, adverbs, prepositions, and auxiliaries, etc. These words not only cause the storage space, but also cause the storage space, which is very important in many text applications. These words not only cause a waste of storage space, but also reduce the efficiency of search in the process of information retrieval, especially in the latter text representation will result in the feature dimension is too high. Therefore, it is necessary to carry out the removal of deactivated words in the process of text preprocessing. These deactivated words are basically entered manually and the generated deactivated words form a deactivation table.

II. B. Text Representation Model

II. B. 1) Vector space model

Vector space modeling (VSM) transforms a traditional corpus into a computer-recognizable word-document matrix, transforming the processing of textual content into operations between vectors in a multidimensional space. The dimensions of the n-dimensional vector space are represented by the feature terms t of all documents, and by calculating the weight of each feature term in a document $T=\{t_1,t_2,\cdots,t_n\}$ on the corresponding dimension of the VSM $w_1,w_2,\cdots,w_n$, to obtain the corresponding multidimensional space vector $\vec{T}=(w_1,w_2,\cdots,w_n)$ for this document [19]. Thus, the set of M documents is combined into a corpus $D=\{T_1,T_2,\cdots,T_M\}$, and the vector space of the document set is shown in Figure 1.

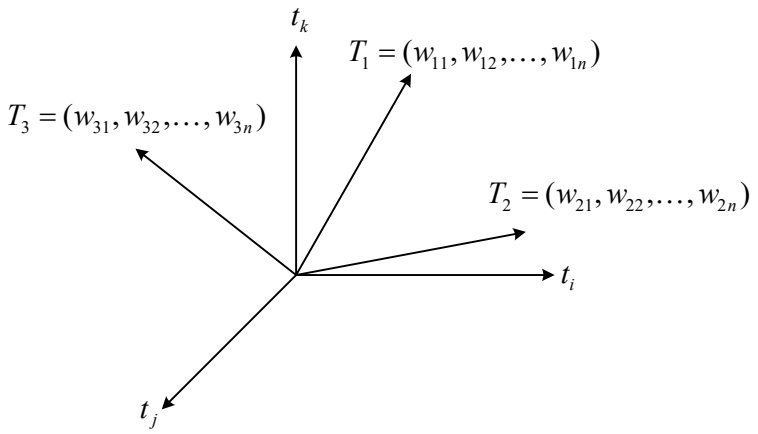

Figure 1: The vector space of the document set

The idea of the VSM model is to use the document-inverse document frequency (TTF-IDF) to calculate to get the similarity between two texts, where TF is the word frequency, which represents the ratio of the number of times a word occurs in a particular text to the total number of words in this document.IDF is the inverse document frequency, which has the value of the total number of documents in the corpus divided by the number of documents in which the word exists, and logarithmizes the result, which is represented by when the number of documents containing a word is low [20]. The weight of each feature item in a particular document corresponding to the VSM can be obtained by calculating the TF-IDF, which is represented below by Eq.

For the word t_i in a certain document d_j, the word frequency of t_i can be expressed as Equation (1):

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

where n_{ij} is the number of occurrences of the word t_i in the document d_j, and $\sum_k n_{kj}$ is the sum of occurrences of all words in the document d_j. I.e:

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

where $|D|$ is the total number of all documents in the corpus, and $|\{j : t_i \in d_j\}|$ is the number of all documents containing the word t_i.

Then we can get equation (3):

$$TF - IDF = tf_{ij} * idf_i \tag{3}$$

II. B. 2) Word Embedding Model (Word2vec)

Word embedding is a collective term for a class of models that vectorize words, and the core idea is to map each word into a dense vector on a low-dimensional space.

Word2vec can be used to map each word in the original corpus to a vector, which is used to represent the relationship between word to word, there are two network structures in Word2vec real, which are CBOW (Continus Bag of Words) and Skip-gram. Both CBOW and Skip-gram can be represented as neural networks consisting of input, mapping and output layers.

Word2vec gives two frameworks, Hierarchical Softmax and Negative Sampling, and this paper focuses on the CBOW and Skip-gram models based on Hierarchical Softmax.

For the CBOW model, the objective function of the model is shown in equation (4) below:

$$L = \sum_{w \in c} \log p(w \mid Context(w)) \tag{4}$$

For the Skip-gram model, the objective function of the model is shown in equation (5) below:

$$L = \sum_{w \in c} \log p(Context(w) \mid w) \tag{5}$$

II. B. 3) Implicit Dirichlet distribution (LDA)

The most widely used topic discovery model is the hierarchical Bayesian model based on bag-of-words theory-implicit Dirichlet Distribution Supervised Topic Model (LDA).The LDA topic model is the classical model used for text representation, which yields a probability distribution of potential topics in each document. In the LDA topic model it is assumed that the document is a mixed distribution over multiple potential implicit topics and the words in the document belong to each topic.

The LDA model uses the Dirichlet distribution as the prior distribution of the polynomial distribution in the model. Given a collection of modern Chinese literature texts, α is the hyperparameter of θ, β is the hyperparameter of φ, z represents the potential topic distribution, θ represents the document-topic probability distribution, φ represents the topic-word probability distribution, and the joint probability distribution of all variables can be obtained:

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \tag{6}$$

In the LDA probabilistic model graph, the parameters α and β are fixed values that can be specified in advance, each word in the document $w_{m,n}$ is data that can be observed in advance, and the unknown implied parameters need to be derived probabilistically.The number of parameters in the LDA model is only related to the number of topics and the number of words, so Gibbs sampling approximation is used to The implied parameters θ and φ are estimated. Then:

$$\theta_{m,k} = \frac{n_m^k + \alpha}{\sum_{k=1}^{K} n_m^k + k\alpha}, \varphi_{k,w} = \frac{n_k^w + \beta}{\sum_{w=1}^{V} n_k^w + w\beta} \tag{7}$$

where V is denoted as the number of words, n_k^w denotes the number of times the word w occurs in the topic k, and n_m^k is denoted as the number of times the topic k occurs in the document m.

Based on the obtained document-topic distribution, the LDA-based topic vector model is obtained, and each text can be represented in the following vector form:

$$V = \{(\theta_1, t_1\theta_1), (\theta_2, t_2\theta_2), \cdots, (\theta_K, t_K\theta_K)\} \tag{8}$$

where K is the number of topics specified in the text collection and $t_i\theta_i$ is denoted as the probability distribution of topics θ_i in the text.

The goal of LDA is to find the distribution of topics for each document and the distribution of words in each topic. The following is the LDA definition to generate any text d_j:

II. C.Clustering model construction

For the traditional clustering algorithms need to specify the number of clusters K in advance, this paper proposes a two-stage clustering, the first stage of the Canopy algorithm for “coarse” clustering, the estimated number of data clusters K and cluster center u, as the next eleven stages of clustering algorithms for the initialization of the parameter, i.e., the second stage of the K In the second stage, the K-means clustering algorithm is used for “fine” clustering, in which the u value and the cluster center u are adopted from the clustering results of the first stage.

II. C. 1) Canopy clustering algorithm

Canopy clustering algorithm is simple and efficient, do not need to specify the number of categories of the original data in advance, this feature is very important for today's more and more unlabeled data, do not need to know the distribution of the data in advance, through the algorithm iteratively can estimate the number of categories and cluster centers contained in the data. The Canopy clustering algorithm is schematically shown in Figure 2.

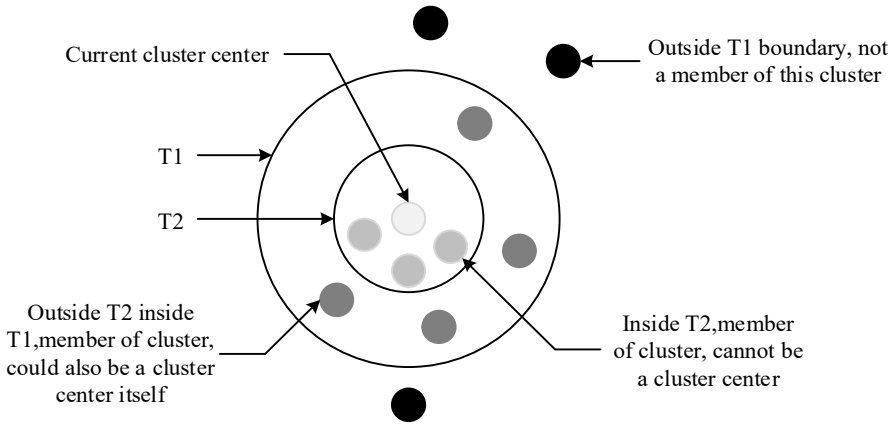

Figure 2: Schematic diagram of Canopy Clustering Algorithm

II. C. 2) K-means clustering algorithm

The K-means algorithm belongs to the clustering method under the category of division, which is widely used because of its simplicity and efficiency. The algorithm needs to specify the cluster centers and the number of clusters before the algorithm can divide them according to the specified number of categories. Assuming that there are N documents, each document can be quantitatively computed by representing it as a space vector V using a vector space model or a distributed word embedding model. The general set of documents can be represented as $X=\{x_i\}, i=1,2.....N$, there are K clusters, each cluster can be denoted as c_k, and the set of clusters can be denoted as $C=\{c_k\}, k=1,2.....k$ [21]. In order to simplify the complexity of the model, each cluster heart u_k is generally used to represent a cluster to calculate the distance between the cluster and the other documents to determine whether the sample belongs to this cluster, and the formula for calculating the cluster heart u_k is shown in Eq:

$$u_k = \frac{1}{|c_k|}\sum_{x_i \in c_k} x_i \tag{9}$$

The K-means clustering algorithm continuously and iteratively optimizes the objective function to find the optimal way of data partitioning, which is often used as the objective function, i.e., the sum of the distances between each document x_i and each cluster c_k, where K is the number of clusters:

$$E(c) = \sum_{k=1}^{K}\sum_{x \in c_k} dist(x_i, c_k) \tag{10}$$

In the field of natural language processing, cosine similarity is usually used as a distance metric and is calculated as shown in Eq:

$$dist(x, x') = 1 - CosineSimilarity(x, x') \tag{11}$$

As can be seen from the clustering process, the number of clusters K in the dataset is determined artificially using a priori knowledge, and then K cluster hearts u_k are randomly initialized to represent each cluster. Each document is continuously assigned to the cluster nearest to it, and the loss function is optimized to readjust the division of cluster hearts u_k and documents x_i until the assignment of documents is not changing, i.e., the model reaches a converged state [22]. As the clustering algorithm is easy to fall into the state of local optimum, different initialization results often produce different clustering results, in order to reduce the volatility of the clustering results, generally choose the average of multiple initialization for the initialization operation of the model.

III. Experimental validation

III. A. Performance testing

III. A. 1) Text Topic Vector Generation

Vector space modeling generates probability models for text generation through corpus learning. Unlike text clustering methods that compute text distance and compute region density, each text has a topic probability

distribution model, which allows class clusters to be divided by the topic probability tendency of the text. Topic models are more capable of highlighting the structure of the text due to their ability to organize a series of texts into expressions of keywords under several topics, and these expressions are usually easier to understand and highly interpretive.

In this paper, the processed modern Chinese literature texts are used as model inputs, and the Gibbs sampling algorithm is used to sample and train the model. The first step in model training is to set the number of topics, and determine the number of topics selected by the topic model through the perplexity degree, the perplexity degree is used to measure the clustering effect of the topic model, which can be interpreted as the model's uncertainty about which topic a certain article belongs to, and the lower the perplexity degree is, the better the clustering effect is. The perplexity is shown in Figure 3, which calculates the perplexity of the model when the number of topics is (0,150], and selects the number of topics with the lowest perplexity.

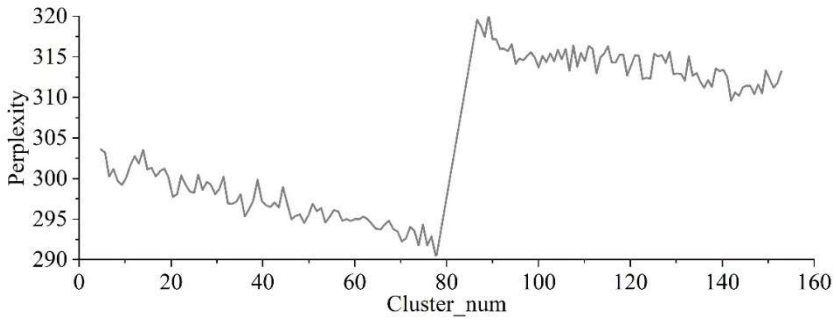

Figure 3: Degree of confusion

The model output results are shown in Table 1. After the model training, the output results are saved in a file for easy use.

Table 1: Model output

Name	File content	Text format
model_parameter.dat	Model parameter	Parameter name: value
wordidmap.dat	The vocabulary and serial number, mainly used as the topn time query	Word: id number
model_twords.dat	Below the topN key	Topic number: list of key words
model_tassgin.dat	The topic number of the word language allocation in the vocabulary table	The word id number: topic number
model_theta.dat	The topic probability distribution of each document is a document for each line	The document belongs to the topic 1 probability, topic 2 probability, topic 3 probability, theme k probability
model_phi.dat	The probability distribution of the words in the subject, each line is a topic	The word word word 1 probability, word 2 probability, word word N probability

III. A. 2) Experimental analysis and comparison

In this paper, the proposed method uses a clustering algorithm to calculate the contour coefficient to evaluate the text vector representation method. The contour coefficient describes the clustering effect by combining the degree of cohesion and separation of class clusters, and the value ranges between [-1,1], and the larger the contour coefficient is, the better the clustering effect is. The data used in the experiment are 31226 data related to modern Chinese literature, and the text of modern Chinese literature is represented in text vectorization, and the k-means clustering model is used to compare and analyze the clustering effect of the improved textual representation with other vectorization methods.

After vectorized expression of all modern Chinese literature texts in the corpus, the clustering model is used to calculate the contour coefficients, in which the range of the number of class clusters is [1,180], the contour coefficients under each number of class clusters are calculated, and the comparison of the contour coefficients is shown in Fig. 4, and the number of class clusters is larger in the contour coefficients values between [20,70] and [120,160], which indicates that the proposed approach in this paper has a better clustering effect in this value range of clustering effect is better.

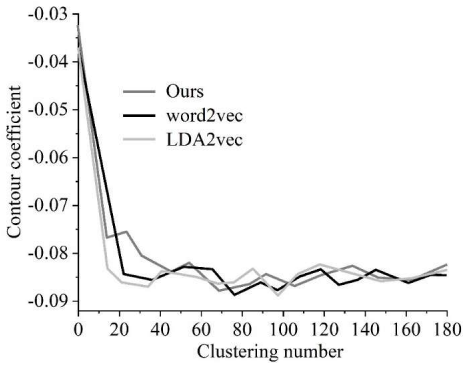

Figure 4: Contour coefficient comparison

In the process of event evolution analysis, 122 data experiments under the topic of modern Chinese literature were selected from the corpus, and the contour coefficients of the corpus for the topic of modern Chinese literature are shown in Figure 5. The method in this paper has significantly better clustering effect when the number of clusters is less than 100.

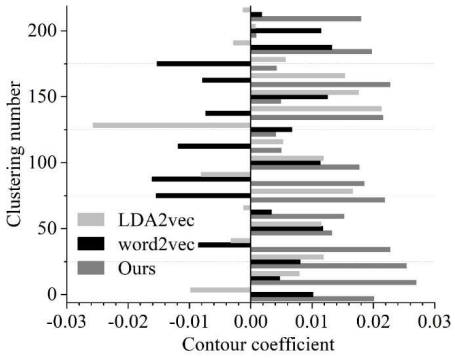

Figure 5: The contour coefficient of China's modern literature

III. B. Analysis of clustering effects

III. B. 1) Optimization Theme Validation Experiments

Set the number of topics in the vector space model as 5, 6, 7 and 8, and take the results of the mixed topic model as the initial center of mass and class group of the k-means clustering algorithm, find the size of the perplexity degree, and carry out the topic clustering of the text of modern Chinese literature, and the results of the verification of the optimal topics are shown in Table 2. From the table shows the standard results of clustering under different number of topics, when the number of topics is at the optimal, the training accuracy of the model as well as the F-measure value is the highest, and the perplexity is the smallest.

Table 2: Optimal theme validation

Topic number	Degree of confusion	Accuracy rate	Recall rate	F-measure
5	7.94	0.671	0.608	0.68
6	7.568	0.705	0.646	0.672
7	6.646	0.81	0.796	0.802
8	7.842	0.763	0.65	0.695

III. B. 2) Effect of clustering system realization

Through the above series of experiments, the number of classes in the collection of modern Chinese literature text topics is set to be 7, and the vector space model and k-means algorithm are used to process the data for clustering, and the results of the clustering of modern Chinese literature text topics are shown in Table 3. Each category contains about 400 modern Chinese literature text data. The words with the highest probability are selected as the theme words, and finally, the themes of modern Chinese literary texts outputted by the vector space model are: criticism of nationalism, oppression of feudal rites, enlightenment and salvation, urban-rural cultural conflict, the dilemma of women's awakening, the writing of war sufferings and the uncertainty of intellectuals.

Table 3: Chinese modern literary text theme clustering results

Numbering	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Data volume	399	377	413	392	409	386	420
Subject matter	National criticism	Feudal etiquette oppression	Enlightenment and salvation	Urban and rural cultural conflict	Women's wake	War suffering	Intellectual uncertainty

Output the content of the theme in each category, divide the categories according to the size of similarity, sort the probability of the words in the theme, and select 8 words for display, and select the word with the first probability as the theme word of the category. The distribution of theme words in modern Chinese literature texts is shown in Table 4. As can be seen from the table, the theme word “intellectuals' uncertainty” has the highest probability of 0.1136, which shows the confusion and loneliness of intellectuals and their response to the problems of the times.

Table 4: The distribution of Chinese modern literary text

Theme	Title Distribution			
Cluster 1	National Criticism (0.1046)	Numbness (0.1053)	Ignorance (0.1019)	Servile (0.0988)
	Hypocrisy (0.0965)	Conservatism (0.0953)	Self-Deception (0.0933)	Cowardice (0.0898)
Cluster 2	Feudal Etiquette Oppression (0.1069)	Family (0.1036)	Marriage (0.0984)	Oppression (0.0973)
	Ethics (0.0966)	Cramp (0.0936)	Revolt Against (0.0957)	Paternity (0.0872)
Cluster 3	Enlightenment And Salvation (0.0926)	Science (0.0896)	Democracy (0.0868)	Reason (0.0841)
	Patriotism (0.0826)	Solidarity (0.0795)	Education (0.075)	Liberal Personality (0.0744)
Cluster 4	Urban And Rural Cultural Conflict (0.0899)	Countryside (0.0869)	Tradition (0.0822)	Nature (0.0796)
	Nostalgia (0.0765)	Change (0.0765)	Simplicity (0.0728)	Oppression (0.0697)
Cluster 5	Women's Wake (0.1085)	Awakening (0.0986)	Marriage (0.0958)	Oppression (0.0923)
	Independence (0.0889)	Struggle (0.0875)	Gender (0.082)	Freedom (0.7954)
Cluster 6	War Suffering (0.0996)	Suffering (0.0968)	Wound (0.0902)	Death (0.0882)
	Displaced (0.0869)	Struggle (0.0836)	Hero (0.0808)	Human Nature (0.0782)
Cluster 7	Intellectual Uncertainty (0.1136)	Confusion (0.0995)	Loss (0.0966)	Ideal (0.0927)
	Reality (0.0895)	Dissimilation (0.0885)	Solitude (0.0849)	Struggle (0.08)

The t-SNE algorithm is used to reduce the dimensionality of the collection of modern Chinese literature texts, and the k-means algorithm is used to cluster the reduced dimensionality data, and the final visualization of the clustering results is shown in Figure 6. From the figure, it can be seen that the data in the class group are more closely spaced and the boundaries are clear, and the data clustering effect is good.

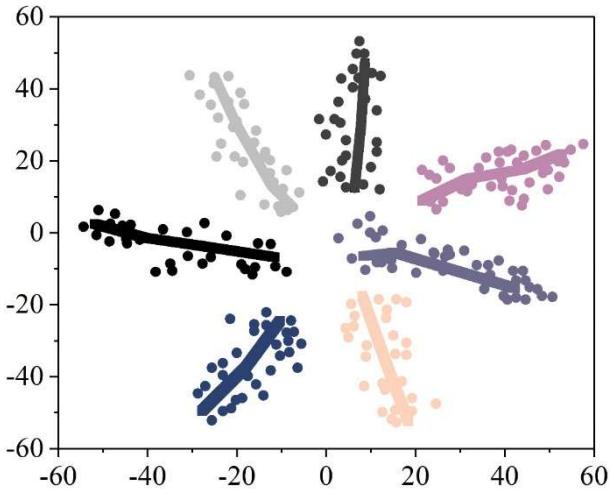

Figure 6: Final clustering results visualization

In summary, this paper uses the vector space model to optimize the k-means algorithm and applies it to the text clustering of modern Chinese literature, and the final distribution of the class groups is better than the single model. Through the comparative analysis of the three aspects, the accuracy and F-measure value of the vector space model to deal with the data collection are higher than that of the single model, the data in the class clusters are more closely separated, and the boundary between the classes is clear, that is, there is a single class division of each data point, and there is no fuzzy and chaotic situation of the division of the data points. Although the time spent on the optimized k-means method of the mixed topic model is relatively large, the model performs better than the single model when the amount of modern Chinese literature text data is small. Therefore, the model designed in this paper can achieve a better effect of modern Chinese literature text clustering.

IV. Conclusion

The vector space model combined with the optimized K-means clustering algorithm shows remarkable effects in the extraction and analysis of themes in modern Chinese literature texts. Experiments show that when the number of themes is set to 7, the model performance is optimal, the perplexity is reduced to 6.646, the clustering accuracy rate is as high as 0.81, and the F-measure value reaches 0.802. By analyzing 31,226 data on modern Chinese literature, seven major themes were successfully extracted: criticism of national character, oppression of feudal rites, enlightenment and salvation, cultural conflict between urban and rural areas, the dilemma of women's awakening, the writing of war sufferings, and the uncertainty of intellectuals. Among them, the highest probability of occurrence of the theme word “intellectuals' uncertainty” reaches 0.1136, reflecting the confusion and exploration of the intellectuals in modern literature in the face of the change of the times. The two-stage clustering strategy effectively solves the limitation that the traditional K-means algorithm needs to preset the number of class clusters, and the introduction of Canopy algorithm enables the model to automatically estimate the optimal number of class clusters. The t-SNE dimensionality reduction visualization results show that the data of various clusters are closely spaced and well defined, which verifies that the clustering effect is good. The experimental results prove that the method proposed in this study not only optimizes the efficiency and accuracy of text theme extraction, but also provides an objective and quantitative analytical tool for the study of modern Chinese literature, and expands the new ideas of digital humanities research.

References

[1] Vitullo, A. (2022). Displacement in young adult literature: A thematic analysis. Children's Literature in Education, 53(3), 296-312.

[2] Manoliu, M. N. (2015). THEME AND THEMATIC ANALYSIS. International Journal of Communication Research, 5(1).

[3] Lian, J., Yordchim, S., Sudmuk, C., Aghaei, B., & Lieungnapar, A. (2024). The Themes Analysis in Hemingway’s Novels Based on Corpus Techniques. Journal of Roi Kaensarn Academi, 9(10), 1602-1612.

[4] Jayady, S. H., & Antong, H. (2021). Theme Identification using Machine Learning Techniques. Journal of Integrated and Advanced Engineering (JIAE), 1(2), 123-134.

[5] Chu, K. E., Keikhosrokiani, P., & Asl, M. P. (2022). A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts. Pertanika Journal of Science & Technology, 30(4), 2535-2561.

[6] Li, D., Wu, K., & Lei, V. L. (2024). Applying Topic Modeling to Literary Analysis: A Review. Digital Studies in Language and Literature, 1(1-2), 113-141.

[7] Erlin, M. (2014). The location of literary history: topic modelling, network analysis, and the German novel, 1731–1864. Distant readings: topologies of German culture in the long nineteenth century, 55-90.

[8] Robledo, S., & Zuluaga, M. (2022). Topic modeling: Perspectives from a literature review. IEEE Access, 11, 4066-4078.

[9] Uglanova, I., Gius, E., Karsdorp, F., McGillivray, B., Nerghes, A., & Wevers, M. (2020). The Order of Things. A Study on Topic Modelling of Literary Texts. CHR, (18-20), 2020.

[10] Navarro-Colorado, B. (2018). On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry. Frontiers in Digital Humanities, 5, 15.

[11] Rha, L., & Silver, S. (2021). Topic Modeling and Analysis: Comparing the Most Common Topics in 19th-Century Novels Written by Female Writers. Aresty Rutgers Undergraduate Research Journal, 1(3).

[12] Hui, H. (2025). Improving topic modeling for literary studies: a hybrid model combined with Word2Vec visualization in the case of Robinson Crusoe. Digital Scholarship in the Humanities, fqaf002.

[13] Schröter, J., & Du, K. (2022). Validating topic modeling as a method of analyzing sujet and theme. Journal of Computational Literary Studies, 1(1).

[14] Gillings, M., & Hardie, A. (2023). The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. Digital Scholarship in the Humanities, 38(2), 530-543.

[15] Jafery, N. N., Keikhosrokiani, P., & Asl, M. P. (2022). Text analytics model to identify the connection between theme and sentiment in literary works: A case study of Iraqi life writings. In Handbook of research on opinion mining and text analytics on literary works and social media (pp. 173-190). IGI Global Scientific Publishing.

[16] Qin, Z., Cong, Y., & Wan, T. (2016). Topic modeling of Chinese language beyond a bag-of-words. Computer Speech & Language, 40, 60-78.

[17] Omar, A. (2021). Identifying themes in fiction: A centroid-based lexical clustering approach. Journal of Language and Linguistic Studies, 17(S1), 580-594.

[18] Yan, Y., & Liu, T. (2024). Mining Thematic Trends in Chinese Literature Using Text Mining Technology. International Journal of Multiphysics, 18(3).
[19] Vipin Jain& Kanchan Lata Kashyap. (2024). Enhanced word vector space with ensemble deep learning model for COVID-19 Hindi text sentiment analysis. Multimedia Tools and Applications,84(9),1-22.
[20] Wang Feifei,Liu Jingyuan & Wang Hansheng. (2021). Sequential Text-Term Selection in Vector Space Models. Journal of Business & Economic Statistics,39(1),82-97.
[21] Hongbin Wu,Jiajia Zhang,Chen Luo & Bin Xu. (2019). Equivalent Modeling of Photovoltaic Power Station Based on Canopy-FCM Clustering Algorithm..IEEE Access,7,102911-102920.
[22] Giyasettin Ozcan. (2018). Unsupervised Learning from Multi-Dimensional Data: A Fast Clustering Algorithm Utilizing Canopies and Statistical Information. International Journal of Information Technology & Decision Making,17(3),841-856.