Retraction

Retracted: Text Clustering and Economic Analysis of Free Trade Zone Governance Strategies Based on Random Matrix and Subject Analysis

Mathematical Problems in Engineering

Received 13 September 2023; Accepted 13 September 2023; Published 14 September 2023

Copyright © 2023 Mathematical Problems in Engineering. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

1. Discrepancies in scope
2. Discrepancies in the description of the research reported
3. Discrepancies between the availability of data and the research described
4. Inappropriate citations
5. Incoherent, meaningless and/or irrelevant content included in the article
6. Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article’s content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable.

We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

Research Article

Text Clustering and Economic Analysis of Free Trade Zone Governance Strategies Based on Random Matrix and Subject Analysis

LingYan Meng

Universiti Pendidikan Sultan Idris (UPSI), Faculty of Management and Economics, Tanjong Malim, Perak Darul Ridzuan 35900, Malaysia

Correspondence should be addressed to LingYan Meng; menglingyan@yangdong-edu.com

Received 15 June 2022; Revised 26 July 2022; Accepted 3 August 2022; Published 30 August 2022

Copyright © 2022 LingYan Meng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The steps of generating basic data by the LDA model and calculating text by the weighted algorithm have a good effect on text clustering. In this paper, the LDA topic model is used to effectively improve the accuracy of strategy text clustering. FTZ economics text clustering simulates FTA economics text data and economic data, imports economics and economic figures and word lists, and uses the traditional vector space model for factor representation. After that, the text vectors are independent of each other, ignoring the semantic relationship, which affects the clustering analysis results. A Chinese text clustering algorithm based on semantic clustering is proposed. Based on the principle of cooccurrence and semantic relevance of words, the algorithm uses the collocation vector of feature words to construct semantic clustering; find the document vector with embedded semantic information. Finally, document vectors with embedded semantic information are used. Finally, $K$ vector is used for cluster analysis. The simulation analysis in this paper shows that the economic growth of the free trade zone is the largest under the economics guidance, which can reach 15%.

1. Introduction

Economics text is a special type of text, which is a process document produced in the process of economics formulation and implementation, and mainly contains three levels of content: official documents such as laws, rules, and regulations promulgated by authorities or administrative organs at all levels in the form of documents; research, consultation, or resolution materials formed by economics makers or political leaders in the process of economics formulation; speeches, reports, and comments related to the process of economics implementation. The economics text is an important part of collaborative economics research. Economics text is an important tool and vehicle for collaborative economics research [1].

The original proposal of text clustering is based on the well-known clustering hypothesis: documents of the same class are more similar, while documents of different classes are less similar [2]. Text clustering uses the concept of similarity to divide the text into meaningful clusters and thus perform clustering operations on samples, with the aim of speeding up text retrieval and improving retrieval accuracy. With the rapid development of machine language learning technology, economics text research is mostly done by technical means [3], text clustering is the most important tool for text retrieval, and text clustering has become an important direction in the field of economics text research. Since economics texts are characterized by a large amount of data, strict specification and data diversity, and the connotations of economics words vary greatly in different contexts [4], traditional clustering methods cannot solve the problems of sparse data and the semantics behind the implied words, and it is even more impossible to effectively determine the existence of synonyms and polysemous words in economics texts [5]. In actual economics texts, words with the same meaning often have multiple economics terms, so
Semantic analysis becomes an important part to improve the accuracy of clustering, and the tool of semantic analysis is to build a topic model. Text clustering is an effective tool for text information analysis, which can be used to my relevant information by coalescing similar data. Similarity calculation between data [6] is one of the key techniques of clustering, and an effective similarity measure can help users to assign data to several classes. However, the lack of correlation and fusion between data when processing large-capacity, high-dimensional, and semantically complex Chinese text data [7] leads to the low reliability of clustering results, so how to effectively express the complex text semantics is still a problem that needs to be solved. In the process of clustering analysis, in the text representation stage, the traditional vector space model (VSM) is used to convert the unidentifiable text data in the document into the form of vectors, usually using the word frequency-based VSM representation method, without considering the influence of deep features on text clustering [8], which has two defects: (1) the weight of each word in the document in the vector. The weight of each word in the document is expressed by word frequency only, which cannot express the semantic diversity; (2) synonymous words have the same contextual context but are independent of each other in the document matrix, and the semantic correlation between words is ignored, and thus, the document identity is also weakened, which affects the clustering effect.

LDA uses the relationship between words, topics, and texts to solve the problem of semantic mining in text clustering. The joint probability of the LDA model is shown in public in the following equation:

\[ p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta). \]

The core purpose of economics text clustering is to achieve intergovernmental and multigovernmental economics collaboration and to assist economic makers in improving the scientific and effective economics formulation based on the objective sense of documented documents generated by economics activities themselves [9]. The key to economics text clustering is to improve the accuracy of police word classification by introducing economics word lists and to improve the accuracy of clustering results by optimizing clustering algorithms to reduce operational errors. Therefore, this paper proposes an economics text clustering method based on the LDA topic model, in order to provide a reference for the formation of a new economics formulation, economics evaluation test, and two-way interaction generation mechanism.

2. Related Studies

A lot of research work has been carried out by domestic and foreign scholars on topic models and clustering methods, and literature studies have confirmed that text clustering results are directly related to topic extraction, mainly including three types of methods. Reference [9] proposed the LSA (latent semantic analysis) model, which is mainly used to explore the potential semantic association between text and words. Subsequently, Reference [10] improved LSA and proposed the PLSA (probabilistic latent semantic analysis) model, which is an extension of LSA and can effectively avoid complex computation by clustering texts with probabilistic models. However, the PLSA probabilistic model is not complete, and as the number of texts and words increases, the model becomes very large, and the computation becomes more complicated. Reference [11] proposed a clustering method based on the LDA (latent Dirichlet allocation) topic model, which determines the optimal number of topics by selecting the perplexity (Perplexity) [12] index to cluster the text, but the number of topics selected by this method is often too large, resulting in a high similarity between the extracted topics; however, the number of topics selected by this method is often too large, resulting in a high degree of similarity among the extracted topics, which affects the efficiency of economics formulation and economics evaluation testing.

LDA model combined with other methods for text clustering, such as the study in [13] proposed a density-based adaptive optimal LDA model selection method, proving that the model is optimal when the similarity between topics is minimal; Cagnin and Kännölä [14] proposed to use the Jensen-Shannon distance as the similarity measure of text based on LDA model and use hierarchical clustering method for clustering; the study in [15] used LDA and Word2Vec models to form the T-WV matrix of relevance and transformed the problem of selecting the number of topics in the traditional LDA model into the problem of evaluating the effect of clustering to calculate the optimal solution for the number of topic clusters; the study in [16] used the descriptors of the nearest topic in the cluster center as cluster labels for text clustering. However, the above studies generally have problems such as high model complexity, and the research method is limited to certain topics.

Based on the characteristics of economics texts and the problems in text clustering, this paper improves the accuracy of economics text clustering by introducing economics word lists and LDA model weighting algorithms.

To address the shortcomings of the traditional vector space model, researchers have proposed a number of methods to quantify the weight of feature words. Reference [17] proposed to construct a knowledge-based vector space model by combining similarity measures of average distance, position weighting, and edge counting to construct relationships among words in documents; the study in [18] proposed a new text representation model to reduce matrix dimensionality and computational complexity by implementing text feature word synonymy merging; the study in [19] represented documents into a three-layer model to reduce the high-dimensional word frequency information to low-dimensional information, so as to extract the semantic information of words in the document and then perform text clustering on this basis; the study in [20] determines the feature words by using similarity testing methods according to the composition structure and position of noun phrases; the study in [21] obtains the features of products in reviews by using association rules from customer reviews of
products through text mining and calculate the number of reviews with two types of emotions for each feature, so as to form a summary of product features. A domain term of information extraction is used, and some feature word pruning algorithms are added as an aid to represent the feature words [22].

3. Related Technologies

3.1. Vector Space Model. Text clustering mainly consists of two steps: text representation and cluster analysis, where text representation is the process of converting unstructured or semistructured text data into a structured form that can be processed by computers. Usually, a collection of documents is represented by a matrix, where each row of the matrix represents a document, and each element of each row is the weight value of the words in the document, which is called a vector space model [23]. The weight value of an element in the model is the frequency of the word in the document. Given a set \( D = \{ \hat{d}_1, \hat{d}_2, \ldots, \hat{d}_n \} \) of \( n \) documents and a set \( W = \{ x_1, x_2, \ldots, x_m \} \) of \( m \) words appearing in all documents, the frequency of words appearing in the documents is represented by a two-dimensional \( m \times n \) matrix, denoted as

\[
X = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1n} \\
X_{21} & X_{22} & \cdots & X_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
X_{m1} & X_{m2} & \cdots & X_{mn}
\end{pmatrix}
\] (2)

3.2. Feature Word Selection and Weighting Calculation. After extracting feature words from Chinese text to characterize the text information and formalizing them by rules, the text is processed by the machine. Good features of text should have four types of characteristics: distinguishability, reliability, independence, and feature word conciseness. Commonly used feature selection methods are document frequency, information gain, mutual information, CHI statistics, etc. In this paper, we use the method of calculating word frequency in documents for feature selection [24]. Word frequency-inverse document frequency (TF-IDF) is a statistical mathematical method to determine the importance of a word or phrase in a document or document set. TF-IDF is calculated as

\[
W(t, d) = \frac{tf(t, d) \times \log (N/n_i + 0.01)}{\sqrt{\sum_{t \in d}[tf(t, d) \times \log (N/n_i + 0.01)]^2}},
\] (3)

where \( t \) denotes the weight of the feature item; \( n \) denotes the word frequency of the feature item; \( N \) denotes the number of all training documents; \( n_t \) denotes the number of documents in which the feature item \( t \) appears in the training set. The denominator is the normalization factor. If the TF-IDF feature value of a candidate keyword \( t \) is higher, the higher the probability that the candidate keyword will become a keyword of the document.

However, the TF-IDF algorithm has the following problems in practical applications [25]:

1. The simple structure of the IDF does not effectively reflect the importance of words and the distribution of feature words, making it unable to perform the function of adjusting the weight well
2. The positional relationship of words is not reflected in the TF-IDF algorithm
3. Word frequency is the number of times a word appears in a document, and feature selection by word frequency is to remove words with frequency less than a certain threshold and reduce the dimensionality of the feature space, which does not exactly match the actual situation, and sometimes, words with small frequency contain key information

Based on the above problems, this paper proposes that when using TF-IDF to calculate the weight value, the feature word vector is first spatially transformed by semantic clusters to obtain semantic information, and then, the weights of feature words are added with semantic information to obtain more accurate weight values.

4. Semantic Cluster-Based Text Representation

The text representation method based on semantic clusters, that is, using collocated word vectors to construct semantic clusters, finds the semantic information of document feature words, embeds them into document feature words, and then performs text clustering. The semantic clusters are a set of words with strong semantic relevance, and after the feature words are characterized by vectors, the clustering method can be used to cluster words with similar semantic features to form semantic clusters, and the central vector of the clustered classes is used to characterize the semantic clusters. The current clustering algorithms [26] mainly include the division method, hierarchical method, density algorithm, graph theory clustering method, grid algorithm, and model algorithm. In this paper, we use the hierarchical clustering method to extract the semantic clusters and get the semantic cluster centers, and the feature vectors of the text are spatially transformed based on the semantic cluster centers and then clustered by the K-means algorithm, and the clustering results have a high correct rate.

4.1. Building Semantic Clusters. The meaning of a word can be understood by the words that go with the word in context. When two words with high collocation frequency are similar, they may have a similar semantic relationship. According to this principle, suppose there are \( n \) words in the text data set that appear frequently with the combination of feature word \( i \). Define the vector of words with this feature word as

\[
T_i = (w_{i1}, w_{i2}, \ldots, w_{in}),
\] (4)

where \( w_{in} \) is the weight of the nth collocation of feature word \( i \). The collocation frequency is denoted by \( p(n, i) \).
where \( f_d(n,i) \) is the number of collocations between collocations \( n \) and feature word \( i \) in text \( d \); \( S_d \) is the number of sentences contained in text \( d \); \( D \) is the number of texts contained in the text dataset. The semantic similarity \( \text{Sim}(i, j) \) of the feature word vector is calculated as

\[
\text{Sim}(i, j) = \frac{\sum_{n=1}^{N} w_{nk} w_{kj}}{\sqrt{\sum_{n=1}^{N} w_{nk}^2} \sqrt{\sum_{n=1}^{N} w_{kj}^2}}.
\]

Hierarchical clustering divides clusters according to the data hierarchy, thus forming a tree in which each node of the tree is a divided cluster. In hierarchical clustering, clustering is performed layer by layer, and the method of partitioning the categories from top to bottom is called the splitting method; the method of aggregating small clusters from bottom to top is called coalescing method. The initial class method; the method of aggregating small clusters from the categories from top to bottom is called the splitting

the steps of the K-means algorithm clustering converge. The steps of the K-means algorithm clustering are as follows (Algorithms 1 and 2):

Each semantic cluster obtained after hierarchical clustering is represented by a cluster center vector \( C_i \). The cluster center vector is calculated as

\[
C_i = \frac{T_1 + T_2 + \ldots + T_M}{|T_1| + |T_2| + \ldots + |T_M|} = (W'_1, W'_2, \ldots, W'_K),
\]

where \( M \) is the number of cluster samples in the semantic cluster; \( T_i \) denotes the vector of cluster samples; \( K \) is the dimension of the collocation vector. The central vector is the collocation vector of all cluster samples in the synthetic semantic cluster and then normalized.

4.2. Spatial Transformation of Text Vectors. Since each feature word belongs to a certain semantic cluster, the similarity between the collocation vector of the feature word and the central vector of the semantic cluster, that is, the cosine of the angle between the collocation vector and the central vector, reflects the similarity between the feature word and the central concept of the semantic cluster. The closer this similarity is to 1, the higher the degree of conformity; the closer it is to 0, the lower the degree of conformity. Therefore, the projection can be used to project the weights of feature words belonging to the same semantic cluster onto the semantic cluster center vector.

The spatial transformation is to transform the text vector on the feature word space to the semantic cluster space and embed semantic information on the feature words. Suppose the spatial vector of feature words is \( D = (t_1, t_2, \ldots, t_K) \), where \( t_K \) is the weight of feature word \( k \) in the text vector. The semantic information vector \( D' = (\omega_1', \omega_2', \ldots, \omega_K') \) of the embedded semantic information is obtained by spatial transformation, where \( \omega_k' \) is the weight of feature word \( k \) of the embedded semantic information, which is calculated by the similarity between feature word \( k \) and the semantic cluster center vector \( C_i \) to which it belongs

\[
\omega_k' = t_k \text{Sim}(C_i, t_k) = t_k \frac{C_i t_k}{|C_i||t_k|}.
\]

For each component of the document vector \( D = (t_1, t_2, \ldots, t_K) \), the spatial vector of the document combined with the semantic information vector is

\[
D'' = (t_1 + \omega_1', t_2 + \omega_2', \ldots, t_K + \omega_K').
\]

Although the dimensionality of the text vector is not reduced after the spatial transformation, the semantic information is embedded through the weights of the feature words so that the semantic correlation between the feature words is introduced into the text vector, which reduces the information loss compared to the spatial vector of the feature words directly replaced by the center vector on the space of the semantic clusters.

The set of 3 semantic clusters and the centers of the corresponding clusters are input to the machine for storage as vectors; REPEAT: each member of the semantic cluster is projected onto the center of its corresponding cluster according to the spatial transformation formula to obtain the new vector value; UNTIL: all members of the \( n \) semantic clusters obtain the new vector value.

4.3. Algorithm Design. The new vector of the test data set \( D \) is extracted, and the clustering analysis is performed based on this combined with the \( K \)-means clustering algorithm. First, we input \( n \) initial centers of mass, calculate the dissimilarity of each sample point in the test set \( D \) to the cluster center, divide them into the clusters with the lowest dissimilarity, and then recluster the centers of each of the \( k \) clusters according to the clustering results, which is calculated by taking the average of the respective dimensions of all the elements in the clusters, and then reclustering all the elements in the test set \( D \) according to the new centers until the clustering results converge. The steps of the \( K \)-means algorithm clustering are as follows.

5. Research Methodology

Economics text clustering mainly includes text preprocessing, determining the number of topics, determining the optimal topics, optimal clustering, etc. Figure 1 shows the economics text clustering process.

5.1. Data Preprocessing. The economics text has the characteristics of rigorous and standardized, so it is necessary to import the economics word list in the preprocessing process, which mainly includes the following steps:
(1) Collect economics text-related corpus
(2) Manual screening of economics texts
(3) Importing the economics word list
(4) Word separation of the economics text corpus, which is as accurate as possible
(5) To remove deactivated words, punctuate, and annotate the economics text
(6) Forming the economics text to be processed

The selection of economics text corpus and data preprocessing is one of the most important aspects of text clustering, and the accuracy and efficiency of the text clustering results are closely related to this process. So, each step from the selection of economics text corpus, word separation, and deactivation must ensure the best results, so as to ensure the accuracy of the experimental results to the greatest extent possible. In this way, the accuracy of the experimental results can be maximized.

5.2. Determine the Number of Topics. The number of topics is determined by calculating the number of topics based on the base data formed by the LDA topic model. It is a difficult problem to determine the number of topics in the LDA model, and traditionally the number of topics is set empirically and estimated according to the size of the data volume, which is less reasonable. In this paper, the optimal number of topics is determined by calculating the weighted values of the maximum average distribution probability of text-topic and the average probability of topic-word similarity [15].

Parameter settings: \( d \) represents a text, \( n \) represents the number of texts, \( z \) represents a topic, \( w \) represents a topic word, \( k \) represents the number of topics, let \( d_n \) be a text in the text set \( D \), as shown in equation (10), \( z_k \) is a topic in the topic set \( Z \), \( E \) represents the maximum average distribution probability of topics and texts, \( T \) represents the average similarity between topics, and \( G \) represents the weighted similarity between topics.
(i) Set the number of topics as $k$, and get the initial model.

(ii) The distribution probability of topics and texts and the distribution probability of words on topics are calculated by the formula.

(iii) This paper proposes to obtain the maximum average distribution probability of topics and texts by using the maximum average method, as shown in equation (4), and obtains the $E$ value.

$$E = \frac{1}{n} \sum_n \max[p(z|d_n)] : i = 1, \ldots, k.$$  \hspace{1cm} (10)

(iv) The average similarity between topics is calculated using the cosine similarity theorem [16]. $T$

$$T = \frac{2}{k(k-1)} \sum_{k=(k-1)/2} \sum_{j=2}^k \sum_{i=1}^k \sim(z_{1}, z_{2}), \sim(z_{1}, z_{3}), \ldots, \sim(z_{k-1}, z_{k}).$$  \hspace{1cm} (11)

(v) The values of equations (4) and (6) are weighted to form the weighted $G$ value, as shown in the following equation:

$$G = \frac{E}{T}.$$  \hspace{1cm} (12)

(vi) Adjust $k$ value and retrain the economics text.

(vii) Repeat step (ii) to obtain the optimal $k$ value when the $G$ value is maximum.

5.3. Text Clustering Evaluation. For assessing the effect of economics text clustering, Purity value and $F$ value are used as evaluation indicators [18].

The calculation method of Purity value is relatively simple but highly applicable, and the specific formula is as follows:

$$\text{Purity}(\Omega, D) = \frac{1}{N} \sum_k \max_{\omega_k \cap d_j},$$  \hspace{1cm} (13)

where $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_k\}$ is the set of clusters, $\omega_k$ denotes the set of $k$th cluster, and $D = \{d_1, d_2, d_3, \ldots, d_l\}$ is the set of texts, which represents the $j$th text. This method only calculates the number of correctly clustered texts to the total texts, and the Purity value is between 0 and 1, the higher the value means the better the clustering effect, and the worse the opposite.

6. Experimental Data and Analysis of Results

6.1. Selection of Experimental Tools and Experimental Procedure. The purpose of this experiment is to effectively cluster these economics texts by the LDA topic model.

(1) Importing economics word lists, which are formed by collecting and organizing economics keywords in related fields.

(2) Using ICTCLAS [27] to classify the sampled texts, the accuracy of which has an important impact on the clustering results.

(3) Performing operations such as deactivation, lexicality, and punctuation removal.

(4) Forming a sample of economics text after word separation.

(5) Importing the processed economics text documents into the LDA model.

In the LDA model, $\alpha$ is set to 50/$k$, $\beta$ is set to 0.01, the number of topics $k$ is trained at fixed intervals from 2–50, the number of samples is set to 1,000, the number of iterations is set to 1,000, and the keywords are set to 20 [21]. After generating the base data by the LDA model, the experimental results data were programmed in Java language according to the formula [28].

6.2. Analysis of Experimental Results. As shown in Figure 2, the horizontal coordinates represent $k$ values, and the vertical coordinates represent $G$ values and $T$ values. $K$ is repeated from 2 to 50, $G = 62,047$ is the largest value when $k = 4$, and then, the $G$ value gradually decreases to convergence as the $k$ value gradually increases, so it is confirmed that the optimal number of topics for this sample is 4, and the optimal number of topics is directly related to the preinitialization of the economics text corpus [29].

Some words with a high probability of distribution were selected from the theme mining results. In Theme 1, the words related to tourism, city, and leisure were classified; in Theme 2, the words related to information security, big data, and data security were classified; in Theme 3, the words related to tradition, crafts, and cultural relics were classified; in Theme 4, the words related to sports and sports industry were classified. We can see that there are some characteristic words that do not match the theme, such as “attention” and
“process” in Theme 1, “plan” in Theme 3, “development” and “development” in Theme 4. This is related to the model parameters and the preinitialization of the economics text corpus [30].

Figure 3 shows the clustering of topics and texts when \( k = 4 \). The maximum probability of distribution of texts on topics is over 90%, which is the probability relationship between sample 30 and topic 4, indicating that the topics and texts fit very well, and all texts under this cluster can be indexed quickly by finding the set of texts with high fit. The minimum probability of 0.3589 is the probable relationship between sample 38 and topic 1, and the distribution probability of 0.3435 with topical 4 on this text means that this text contains two topics with high probability of occurrence. In Figure 3, 10 texts belong to topic 1, 16 texts belong to Topic 2, 7 texts belong to Topic 3, and 17 texts belong to topic 4. If the set of texts tested is large, this method can also be used to reduce the dimensionality of large-scale texts, and the model can be repeated for the reduced results to form refined clusters again.

The Purity value and \( F \)-score of the LDA topic model were calculated as the evaluation index of clustering, and if the text and the topic match well, the clustering was judged to be correct. The \( G \) value is the largest when the weighted text clustering result is \( k = 4 \), which is consistent with the initial classification number. The Purity value, \( P \), \( R \) and \( F \) values were calculated for \( k = 4 \). The final Purity value is 0.88, and the \( F \) value is 0.83. The clustering results are shown in Figure 4.

The clustering was performed using the method of calculating the \( T \)-value similarity in the LDA model, and the \( T \)-value was maximum when \( k = 5 \). The clustering results were compared with the manual clustering results, and there were deviations from the initial 4 classifications. The clustering was performed by the perplexity index, but the value of this method was too large and deviated from the original classification. Therefore, based on the calculation of the Purity value and the \( F \) value and comparison analysis with other LDA clustering methods, it can be confirmed that this method is reasonable and effective.

After reacquiring the feature word vectors of the documents, the transformed feature word vectors were used to determine the number of clusters and the initial cluster centers before performing K-means clustering, and the clustering results are shown in Figure 5. As seen in Figure 5, the 100 vocational education documents can be clearly classified into 3 classes by the method of K-means clustering after the spatial transformation of semantic clusters. While K-means clustering is performed on the experimental data without spatial transformation, the clustering results are shown in Figure 6.

Instead of evaluating the clustering results separately for each category, we directly count the accuracy and recall of the clustering results as a whole and calculate the \( F1 \) value. The accuracy rate is the percentage of relevant documents that are correctly clustered; the recall rate, also known as the complete rate, is the percentage of relevant documents that are correctly clustered among all relevant documents; the \( F1 \) value is the weighted summed average of the accuracy rate and recall rate.
7. Conclusions

The purpose of this paper is to propose a new economics text clustering method through the data processing algorithm in the weighted LDA model, so that it can better serve the economics interpretation and analysis. The LDA model can reduce the dimensionality of the economics text set, especially for the text with a large amount of sampling data, and then set the model after the reduction of dimensionality, which can greatly reduce the computation and improve the execution efficiency and can achieve fine. This can greatly reduce the computation and improve the implementation efficiency and can achieve fine clustering.

This paper proposes a method for clustering economics texts based on the weighted values of the probability of maximum average distribution of text-topic and the probability of average similarity of topic words and finally verifies that the method is reasonable and effective through data analysis and evaluation index values. It is hoped that more researchers will optimize, validate, and improve this method with large-scale economics text corpus in the future.

Accurate clustering of economics texts can help to improve the efficiency of police collaboration. Economics synergy is an important element in the process of economics formulation, which ensures the internal consistency of a consultation mechanism, information exchange, and sharing mechanism to make cross-regional economics comparisons feasible, and economics synergy research can reduce the error rate of economics formulation. Accurate clustering of economics texts provides guidance and scientific prediction for economics formulation and provides technical support for economics formulation. Analysis of clustered texts can make economics formulation more scientific and can also consider the effect of existing economics implementation, avoiding the phenomenon of missing and unobjective policies. Economics text clustering can help improve the efficiency of economics synergy, promote the formation of a two-way interaction mechanism of economics documents, and provide a reference for the reverse evaluation of economics documents already made.

Data Availability

The dataset used in this paper is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

References


