ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on Innovation and Entrepreneurship Practice Education Path for College Students under the Perspective of New Media

Zhiqin Zhang[1,2,*]

[1] International Education College, Zheng Zhou University of Light Industry, Zhengzhou 450000. Henan. China
[2] New Era Ideological and Political Education Research Center, Zhengzhou University of Light Industry, Zhengzhou 450000, Henan, China

SUMMARY: *The study extracts the implied themes and mines the semantic attributes of the resources from the massive heterogeneous college innovation and entrepreneurship education resources through LDA topic model. And contextual dynamic clustering technology is added to instantly categorize the results and package the content-similar documents after each user retrieval. The model clustering identifies 6 major themes of innovative product service, team capability, marketing, resource funding, operation implementation and policy support. When the data volume reaches 70,000 documents, the F1 value of the fusion method in this paper is as high as 93.19%, while the traditional method is only 78.69%, leading by 18.43%. In terms of accurate retrieval, the overall checking accuracy for the six themes reaches 93.50%, meaning that more than 90% of the returned documents are truly relevant, while the traditional method's checking accuracy is only 74.10%, and nearly 30% of the results are invalid noise. Students using the smart repository in this paper visited the repository an average of 202 times in a semester, which is more than 2.5 times of the 79 times of the control group. And the quality of learning outputs of students in the experimental class is higher, and the business plans they wrote scored 87.7±7.96 and 88.5±6.92 on the dimensions of innovativeness and feasibility, which achieved 13 to 21 points of improvement compared with the control group. The path of integrating semantic understanding and dynamic organization proposed in the study can effectively crack the problem of retrieval accuracy of educational resources, and ultimately translate into the power of enhancing students' learning initiative and practical ability.*

KEYWORDS: *innovation and entrepreneurship education in colleges and universities; LDA topic modeling; context clustering; topic retrieval; repository*

1 Introduction

As an important driving force for future national development, the cultivation of college students' innovation ability is of great significance to the country, universities, and students themselves [1]. In the context of the new century, influenced by the new social environment, innovative development, and entrepreneurial education in colleges and universities has entered a new stage. However, the existing development routes and educational concepts of many colleges and universities can no longer fully keep pace with current social changes. The concept of innovation and entrepreneurship education in higher education should not remain at the level of curriculum design alone. To transform it into concrete educational practice and achieve the

*zhangzhiqin321321@163.com
https://doi.org/10.65102/is2026497
Published: 30 April, 2026

goal of cultivating innovation and entrepreneurship competence, it is necessary to build a more complete curriculum system through curriculum reform [2]. However, in the actual implementation of innovation and entrepreneurship courses, the course content taught by teachers is often derived from others' experiences. Teachers therefore need to answer students' new questions from the perspectives of idea generation, demand analysis, market analysis, product design, marketing, and publicity. At present, the imperfect curriculum system of innovation and entrepreneurship education in colleges and universities, together with the weak practicality of course content, has become one of the major problems facing innovation and entrepreneurship education in Chinese colleges and universities [3].

The quality of innovation and entrepreneurship courses is not high enough, and course teaching should be more closely combined with real development needs. Only in this way can colleges and universities cultivate innovative, application-oriented, and interdisciplinary talents. Innovation and entrepreneurship education in colleges and universities is an important branch of higher education. Therefore, it urgently needs to be reformed and upgraded [4, 5]. As far as college innovation and entrepreneurship education is concerned, it can not only help solve the employment difficulties faced by college graduates, but also stimulate young graduates' enthusiasm for innovation and entrepreneurship. It can further improve the overall creativity of society and promote the development of science and technology [6–8]. However, under the influence of traditional education models, current innovation and entrepreneurship education in Chinese colleges and universities still has a single teaching mode. Students' theoretical knowledge is disconnected from practical application, and teaching outcomes are not well aligned with social needs. This has led to the dual-creation education in colleges and universities being unable to fully meet the requirements of social productivity [9, 10].

To address these problems, innovation and entrepreneurship education in colleges and universities should follow the development requirements of the times and continue to reform and improve [11]. In the new era of educational development, innovation and entrepreneurship education urgently needs to be upgraded and further promoted toward deeper and more sustainable development. There is a natural connection between the concepts of innovation-driven development and innovation and entrepreneurship education. Both adhere to a people-oriented premise, follow human development laws, and focus on human development. At the same time, they create favorable conditions for promoting human development and enhancing human capabilities, so that human beings can fully understand the overall and free development of themselves and the comprehensive, coordinated development of society [12–15]. Therefore, in the new era, colleges and universities should regard innovation and entrepreneurship as a new development guide, so as to provide clearer direction and broader space for the implementation of innovation and entrepreneurship education in higher education [16, 17].

In the era of educational innovation and development, new media technology has brought new opportunities and challenges to innovation and entrepreneurship education in colleges and universities. The application of new media technology in optimizing innovation and entrepreneurship education has become an important means of improving students' entrepreneurial ability [18]. First, the popularization of the Internet has made students rely more on mobile devices and social media to obtain information and communicate with others. This change in information acquisition enables students to learn about the latest market trends, policy information, and technological development more conveniently in the process of innovation and entrepreneurship [19]. Secondly, the interactivity and participation of new media enhance students' innovation consciousness and entrepreneurial enthusiasm. By participating in online discussions, watching live courses and joining entrepreneurial communities, students can communicate and interact in real time with industry experts, entrepreneurial mentors and other entrepreneurs, and obtain opinions and suggestions from various parties [20, 21]. This kind of

interaction not only helps to broaden students' horizons and stimulate their innovative thinking, but also helps students to establish important human resources in the early stage of entrepreneurship, and enhance their entrepreneurial confidence and decision-making ability. Finally, the diversity and convenience of new media enable students to use various tools and platforms more flexibly in the process of entrepreneurship [22]. For example, by using online design tools, cloud storage services and project management software students can be more efficient in product design, team collaboration and project management [23]. Based on this, we provide feasible optimization solutions to further optimize the path of innovation and entrepreneurship education in colleges and universities, improve the quality of education and the practical ability of students, and ultimately achieve the goal of cultivating more high-quality skilled talents with the ability of “dual-creation”.

The research fuses topic modeling and context clustering to move resource retrieval from keyword matching to semantic understanding. And the implied semantic structure is mined and presented to the user. Firstly, a three-layer teaching resource base containing user layer, middle layer and data layer is built to establish the overall framework of data from storage to application. Then the LDA topic model is used to interpret the text in the repository, through the topic probability distribution θ and the topic lexical item probability distribution ϕ, accordingly to establish the topic index, which enables the system to expand the query according to the semantic relevance. At the same time, a dynamic clustering mechanism is introduced in context clustering. After user retrieval, the result set is quickly grouped, and documents with similar content are categorized into the same class, and concise category labels are automatically generated. This process will be applied to the previously extracted high-frequency words and semantic features, but also through the singular value decomposition and other technologies to compress the amount of information, improve processing speed, and ultimately form a clear and easy-to-understand clustering view, so that the user can grasp the whole picture of the retrieval results at a glance, and quickly locate the group of documents that he or she is really interested in.

2 Educational Resource Library Text Processing and Topic Modeling

2.1 Modeling of the Teaching Resource Library

2.1.1 User layer services

The innovative and entrepreneurial teaching resource library constructed in this paper for college students includes text library, picture library, audio library, video library, courseware library, preparatory resource library and user information library. The resource library model includes database server, middle layer application server and user layer browser, and its model is shown in Figure 1.

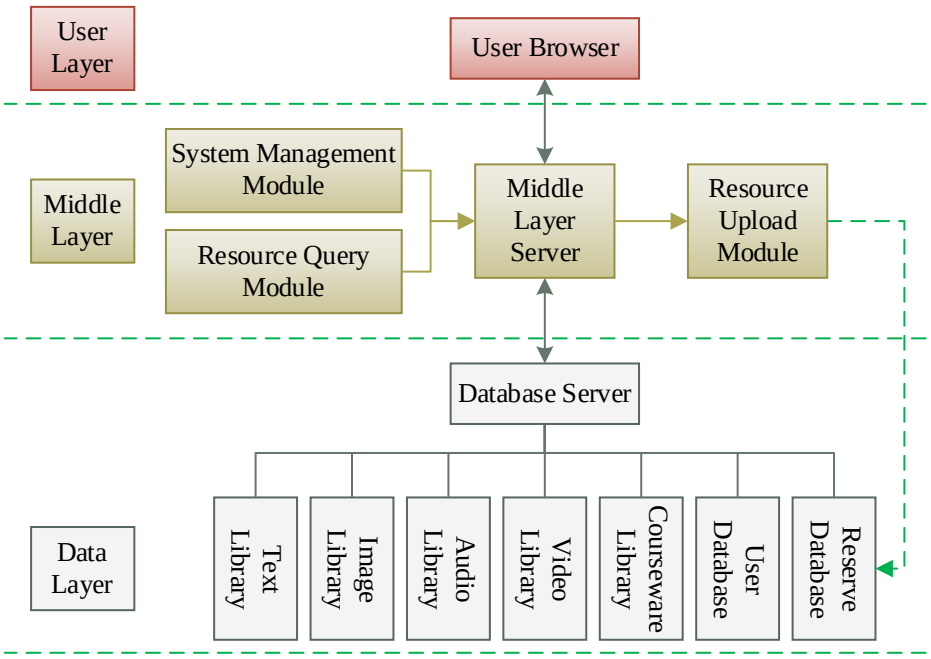


Figure 1: Teaching Resource Repository Model

The user layer is mainly responsible for the input of information requested by the learner and the output of data results, as well as the interaction with the middle-tier application server. Learners contact the server through the browser, query and download the required learning resources, and upload personal outstanding resources.

2.1.2 Intermediate layer services

The middle tier application server is directly connected to the database server. The modules stored in the middle tier are mainly the query module, the system management module and the resource upload module. The query module provides users with a multi-level search tool, which allows users to use 5-level or less than 5-level categorized query according to the difficulty level of the knowledge they wish to learn. The system management module is responsible for the management of user registration, login, and security operations; the resource upload module provides users with the function of uploading various types of resources into the library.

2.1.3 Data service layer

The data service layer is responsible for the daily management of the underlying database, including the management related to the entry, modification, deletion and attribute setting of resources. The database server stores and manages the application data and system data of the whole system. The database server is logically divided into teaching resources library, preparation library and user information library, in which teaching resources in the library include: text library, image library, audio library, video library and courseware library.

2.2 Text pre-processing

Educational resources from a wide range of sources, different forms, common document formats are word, pdf, ppt, etc.. In this paper, first of all, according to the different document formats to choose the appropriate document parser for its information processing, such as the use of PDFBox to operate pdf files, the use of POI access to have word, ppt documents and so on. As shown in Figure 1, the process of text preprocessing mainly includes three parts: text

breaks, Chinese word segmentation and stop word filtering. Among them, the text breaks as the basic link of natural language processing, to the end of the sentence punctuation as the basis for segmentation, the text will be divided into semantically complete sentence units; Chinese part of the lexical combination of user dictionaries, the sentence will be further segmented into finer-grained words; then, based on the deactivation dictionary filtering no actual semantic noise words, and ultimately get the more meaningful words for the retrieval. The text preprocessing flow of this paper is shown in Figure 2.

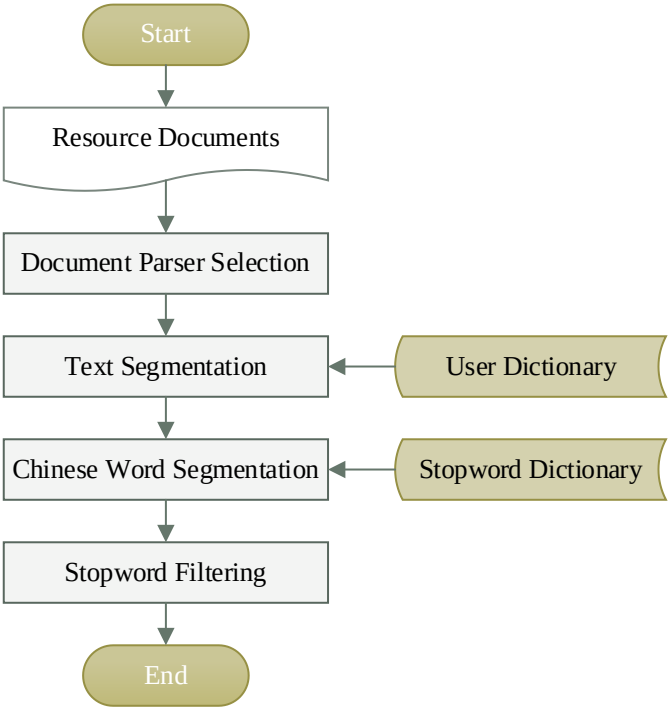


Figure 2: Text pre-processing procedure

2.3 LDA-based topic modeling and indexing

Based on the completion of text preprocessing, this section introduces the LDA topic model to identify document topics from a probabilistic generative perspective. By estimating the document-topic distribution and the topic-word distribution, the semantic structure of the text can be captured. On this basis, a hierarchical topic index is constructed for educational resources, which provides a feature representation for the subsequent clustering analysis.

2.3.1 LDA Subject Modeling

The LDA model is a multilayer generative Bayesian probabilistic model that includes a three-level structure consisting of words, topics, and documents. It treats a document as a mixture of multiple implicit topics, while each topic is represented as a probability distribution over word items. LDA assumes that words in a document and the co-occurrence relations among linked documents reflect potential topics. It builds the mapping relationship among word-document and document-topic structures. The LDA model is based on the bag-of-words (BOW) assumption, which assumes that the order of words within a document and the order of documents in a document collection can be ignored without affecting the model training results. In this way, the text can be converted into digital information that is easier to model.

The graphical representation of the LDA model is shown in Figure 3. Here, α and β are the *Dirichlet* prior parameters, K denotes the number of topics, M irepresents the

total number of documents, and N_m indicates the total number of words in document m, $W_{m,n}$ and $Z_{m,n}$ denote the nth word in document m and its topic, respectively, and ϕ_k represents the probability distribution of lexical items under topic k, while θ_m denotes the topic probability distribution of document m.

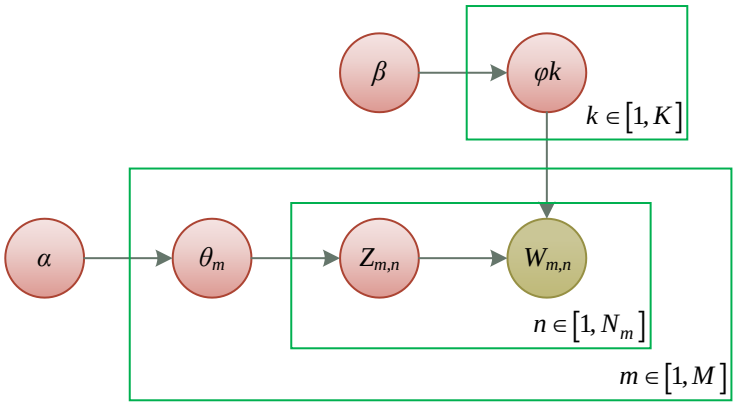


Figure 3: Graphical representation of the LDA model

The process of generating a document from the LDA model is described as follows

(1) Generate the topic probability distribution θ_m of document m by sampling from the *Dirichlet* distribution with parameter α.

(2) Generate the topic $Z_{m,n}$ for the nth word in document m by sampling from the topic probability distribution θ_m

(3) Generate the word-item probability distribution $\phi_{m,n}$ of the topic $Z_{m,n}$ by sampling from a *Dirichlet* distribution with parameter β

(4) Generate words $W_{m,n}$ by sampling from the word item probability distribution $\phi_{m,n}$

In the LDA model, the two most important sets of parameters are the topic probability distribution of the document θ and the lexical item probability distribution of the topic ϕ. Parameter estimation can be viewed as the inverse process of the document generation process: given a collection of documents, $W_{m,n}$ are the known variables, and α and β are the empirically given a priori parameters. The remaining variables $Z_{m,n}$, θ and ϕ are unknown implicit variables that need to be learned to be estimated based on the observed variables.

According to the document generation model of LDA, the joint probability distribution of all variables is as follows

$$\begin{aligned} p\left(\overrightarrow{w_{m,n}}, \overrightarrow{z_{m,n}}, \overrightarrow{\theta_m}, \varphi \middle| \alpha, \beta\right) &= \prod_{n=1}^{N_m} p\left(w_{m,n} \middle| \varphi_{z_{m,n}}\right) \\ \cdot p\left(z_{m,n} \middle| \overrightarrow{\theta_m}\right) &\cdot p\left(\overrightarrow{\theta_m} \middle| \alpha\right) \cdot p\left(\varphi \middle| \beta\right) \end{aligned} \tag{1}$$

where the word $W_{m,n}$ is assigned to the probability distribution of the word class t, as shown in Equation (2):

$$p\left(w_{m,n}=t\left|\overrightarrow{\theta_m},\varphi\right.\right)=\sum_{k=1}^{K}p\left(w_{m,n}=t\left|\varphi_k\right.\right)\cdot p\left(z_{m,n}=k\left|\overrightarrow{\theta_m}\right.\right) \tag{2}$$

The likelihood function for the entire document collection is shown below:

$$p\left(W\left|\theta,\varphi\right.\right)=\prod_{m=1}^{M}p\left(\overrightarrow{w_m}\left|\overrightarrow{\theta_m},\varphi\right.\right)=\prod_{m=1}^{M}\prod_{n=1}^{N_m}p\left(w_{m,n}\left|\overrightarrow{\theta_m},\varphi\right.\right) \tag{3}$$

From the above equation, it can be seen that the likelihood function of the whole document collection is related to θ and ϕ. By maximizing the likelihood function, θ and ϕ can be found. The θ and ϕ obtained by using Gibbs parameter estimation method are shown in Eq. (4) and Eq. (5), where $\phi_{k,t}$ denotes the probability of the lexical item t in the topic k and $\theta_{m,k}$ denotes the probability of the topic k in the document m.

$$\varphi_{k,t}=\frac{n_k^t+\beta_t}{\sum_{t=1}^{V}n_k^t+\beta_t} \tag{4}$$

$$\theta_{m,k}=\frac{n_m^k+\alpha_k}{\sum_{k=1}^{K}n_m^k+\alpha_k} \tag{5}$$

where V is the total number of words in the document collection, n_k^t is the number of occurrences of the word item t in the topic k, and n_m^k is the number of occurrences of the topic k in the document m.

2.3.2 Subject indexing

The traditional full-text retrieval technology analyzes the relevance of resource documents to user queries by determining whether the keywords in the user query exist in the resource documents. This method cannot understand the implicit semantic connection between words, and cannot solve the problems of mismatch between the words used in the user query and the words used in the document, as well as the incomplete expression of the user query, which leads to unsatisfactory retrieval results.

According to the introduction in the previous subsection, the topic model can model the implicit topics of documents and automatically find out the semantic connections between words in the massive data. The topic probability distribution θ of a document and the lexical item probability distribution ϕ of a topic are the feature data of a collection of resource documents generated by the LDA modeling algorithm, and it can be understood in a mathematical sense that the whole collection of resource documents is a matrix compounded by m document topic scale vectors and k topic lexical item scale vectors: $D_{m,t}=\theta_{m,k}\times\varphi_{k,t}$. Meanwhile, it also has the following semantic meanings:

(1) Both θ and ϕ can reflect the associative relationship between words in a document collection, i.e., the probability of words under the same topic appearing at the same time is higher, and the probability of words between different topics appearing at the same time is smaller;

(2) Based on the assumption that the distribution of parent-child topics in the collection of resource documents should satisfy the inclusion relationship, the distribution of word items

calculated by θ and ϕ also reflects the parent-child characteristics of the topics from the statistical point of view, and the parent topic should be distributed in more documents, while the distribution of the child topic is more centralized.

Therefore, the concept of "topic index" is presented in this study, and the educational resource library's topic index is created using the LDA model's feature information, which shows the subjects' hierarchy and connection. In order to make up for the lack of user queries and improve search efficiency, the subject index of the educational resources can be used to expand the user queries semantically and provide additional subjects related to the user queries.

2.4 Context Clustering

After obtaining the text topic representation, the efficient clustering method based on context in this section realizes fast clustering and description generation of retrieval results through high-frequency unit extraction, vectorized representation with singular value decomposition dimensionality reduction to categorize semantically similar documents and present them to users.

2.4.1 Basic Strategies

The contextual clustering strategy proposed in the study is shown in Figure 4.

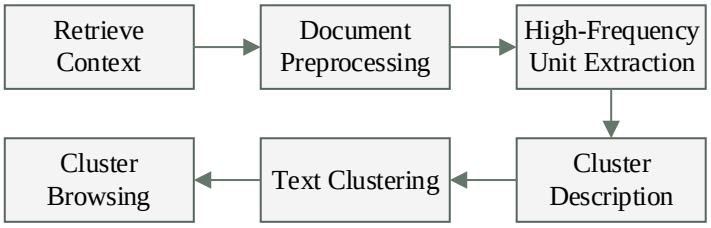


Figure 4: Context clustering strategy

(1) Document preprocessing

Retrieval results of the abstract as a clustering of document set object, the need for the necessary pre-processing:

1) Take the root word: Western need to take the root word, Chinese do not need.

2) Processing deactivated words: deactivated words are not involved in word frequency statistics, while deactivated words appearing in the middle of the clustered descriptions will not be removed, otherwise it will cause a bias in understanding. Processing can be done by comparing and labeling based on a list of deactivated words.

3) Segmentation Component: The Chinese and English segmentation components used in the text preprocessing process are used, and the components will automatically complete the segmentation according to the user's input character flow and his/her expectations.

(2) High Frequency Unit Extraction

The document description contains all or part of the query keywords, and the repeated use of words related to the topic is the collection of potential concepts selected for clustering, that is to say, often referred to as the document feature items. Such potential concepts express a single topic or can be related topics, and thus will be obtained in the next step by singular value decomposition of the matrix.

The criteria for word and phrase selection are:

1) The number of occurrences of the word exceeds a threshold so that relevance to the query is ensured.

2) Words are not across sentence and sentence boundaries so as to avoid breaking sentences to obtain potential concepts. The sentence boundaries used in this project are comma, semicolon, period, question mark, and exclamation point in both English and Chinese inputs.

3) Extraction of phrases formed by complete words. The beginning and end of words and phrases are not deactivated words, and deactivated words appearing in the middle will not be removed, otherwise it will affect the user's understanding.

The extraction of high-frequency units can be achieved by using a suffix array, aided by the Longest Common Prefix Array (LCPA), which is able to complete the recognition of excerpts of length N within $O(N)$ complexity.

(3) Clustering Description

Considering the word frequency and document frequency of the query word in the document, the TF-IDF formula is used to analyze the statistics of the obtained high-frequency units.

Abstract concepts are obtained by matrix singular value decomposition. After obtaining the clustering description, the abstract concept and high-frequency phrases appear in the same vector space, i.e., the column space of the original word-document matrix, and the cosine distance can be used to calculate the similarity between phrases or words and abstract concepts.

(4) Text Clustering

The classical vector space model is used to assign relevant documents to the classes where the clustering descriptions are located. The resulting cluster descriptions are used to re-query the input set of documents. Define the matrix Q, each column vector of Q represents a clustering description, and compute $Q' = Q^T A$, the matrix A is the item-document matrix of the original input documents. The element Q'_{ij} represents the subordination of the j th document to the i th cluster. If a document's Q'_{ij} exceeds the aggregation threshold, then it will become a member of the corresponding cluster. Documents that are not assigned to any cluster are treated as a class.

(5) Cluster Browsing

With the established clustering system, documents in the same class have greater similarity between them, while documents between classes have lesser similarity. The clustering results of the categories enable users to query the search results from a higher level of summarization, and to find relevant documents from the document classes that are similar to the subject.

2.4.2 Clustering description

Based on the established unit-document-frequency (UDF), that is, the item-document matrix (T-D) is established, where the rows represent the number of documents in the retrieval context, and the columns represent the associated high-frequency units of the documents. Here the document feature items (T) include not only words, but also items that characterize high-frequency units such as all for phrases.

Vector space model (VSM) can be used for the occurrence of word frequency of each document, the basic idea is to represent the text as a vector: $(W1, W2, W3, \cdots, Wn)$, of which W_i is the weight of the i th feature item, you can generally choose the word, words or phrases, and here we choose to use STC to build the previous High-frequency units. If the number of occurrences of high-frequency units is not calculated, the initial vector representation is entirely in the form of 0, 1, i.e., if the word appears in the text, then the value of the text vector is 1, otherwise it is 0. The use of 0-1 cannot reflect the degree of the role of the word in the text, so the use of the word frequency of the method of 0, 1 is gradually replaced by a more accurate word frequency, the word frequency is divided into the absolute word frequency and relative

word frequency, the absolute word frequency, even though represent the text with the frequency of the word appearing in the text, relative word frequency is the normalized word frequency, here the normalization formula of lfc weighting method of TF-IDF is used to calculate W:

$$W\left(t,\vec{d}\right)=\frac{\left(1+\log_2 tf\left(t,\vec{d}\right)\right)\times\log_2\left(N/n_t\right)}{\sqrt{\sum_{t\in\vec{d}}\left[\left(1+\log_2 tf\left(t,\vec{d}\right)\right)\times\log_2\left(N/n_t\right)\right]^2}} \tag{6}$$

where $W\left(t,\vec{d}\right)$ is the weight of word t in text $\vec{d}$, while $tf\left(t,\vec{d}\right)$ is the word frequency of word t in text $\vec{d}$, N is the total number of training texts, n_t is the number of texts appearing t in the training text set, and the denominator is the corresponding normalization factor.

For the existing feature matrix, the singular value decomposition technique can be used for feature dimensionality reduction, which allows to approximate the existing matrix with a matrix of smaller dimensions.

$$W=T\times S\times D^T \tag{7}$$

Define the matrix W to represent the lexical item vector, the matrix S to represent the singular value matrix, k to represent its dimensionality reduction factor, the matrix T to represent the lexical item vector, and D to represent the document vector. The approximation matrix A_k in a sense maintains the intrinsic structure (underlying semantics) of the connection between the lexical items and documents reflected in the matrix A, but removes the effects due to word usage or polysemy of the language, etc.

Since the biggest advantage of singular value decomposition is that the elements on the diagonal of the singular values $\left(S_k\right)$ can be ranked according to their importance, the size of k can be set, and if one chooses to keep the largest k singular values and eliminate the remaining smaller singular values, the desired matrix can be obtained.

$$W'=T_kS_kD_k^T \tag{8}$$

where, if the largest k is retained, that one needs to retain the corresponding k columns in matrix T and matrix D, while deleting the rest of the columns of matrix T and D.

In this method, the dimension of the vector space is greatly reduced after processing by LSA theory, which can effectively improve the clustering speed of the document set. For the selection strategy of k, $\|Z_F\|$ can be used to compute the Frobenius paradigm to define a percentage coefficient t, which determines the correlation between the current approximation matrix and the original matrix:

$$\|Z\|_F=\sqrt{\left(\sum_{i=1}^{D}\sum_{j=1}^{W}\left|a_{ij}\right|^2\right)} \tag{9}$$

For the correlation t under k the following calculation can be used:

$$t = \frac{\|Z_k\|_F}{\|Z\|_F} \tag{10}$$

It can be seen that as k increases, t gets closer to 1 and its number of clusters k available for clustering increases, therefore, a correlation threshold t' is defined to determine the k value of correlation.

$$\min(t) \text{ And } t \geq t' \tag{11}$$

After obtaining the clustering description, the abstract concept and the high-frequency phrases appear in the same vector space, i.e., the column space of the original word-document matrix, so the original cosine distance can be used to compute the similarity between phrases or words and the abstract concept. The k value is obtained by the above method, and the W' matrix can be obtained by calculating according to the singular value formula. Based on the computed matrix W' and the current number of k, the available cluster descriptions $B = \{b_1, b_2, \cdots, b_k\}$ are obtained in order of Score score.

Define the score of each cluster as Score, the corresponding cluster description score as Cluster, and the number of documents present in the cluster as DocNum, then the cluster score can be used as the following formula:

$$Score = Cluster * DocName \tag{12}$$

where the corresponding clustering description score is the non-zero value with the largest absolute value occurring in the decomposition matrix U. DocNum is the number of occurrences of this word document, and since the recalculated $W' = T_k S_k D_k^T$ after decomposition will have some noisy data, the number of occurrences of DocNum is computed as the non-zero data in W.

$$DocNum(T) = count(Wt <> 0) \tag{13}$$

The product of the document weights $d_{W'}$ obtained from W' and the W of the unclassified documents is computed to give the final document weights.

3 Resource library topic modeling and clustering search test

Searching in the teaching resource base constructed above, setting the search formula as "SU=College+Innovation and Entrepreneurship", a total of 71,632 documents on innovation and entrepreneurship education for college students were obtained.

3.1 LDA Topic Model Performance Testing

Firstly, the performance of LDA topic model based on bag of words (BOW) is tested, and the study mainly examines its ability to model the topic of large-scale documents, taking the innovative entrepreneurship education documents of college students as a dataset, and randomly selecting 5,000, 30,000 and 50,000 documents in the dataset. Then choose different number of parallel computing nodes, run 5 times, take the average value, get the time used in LDA modeling for different sized datasets under the environment of multiple parallel

computing nodes, and take the running time under the single node as the benchmark, and take the ratio of the time used in LDA modeling under the environment of multiple nodes to the benchmark time as the acceleration ratio.

3.1.1 Computational time analysis

Figure 5 shows the time used for parallel computation of multiple nodes for LDA modeling with different sized corpus sets.

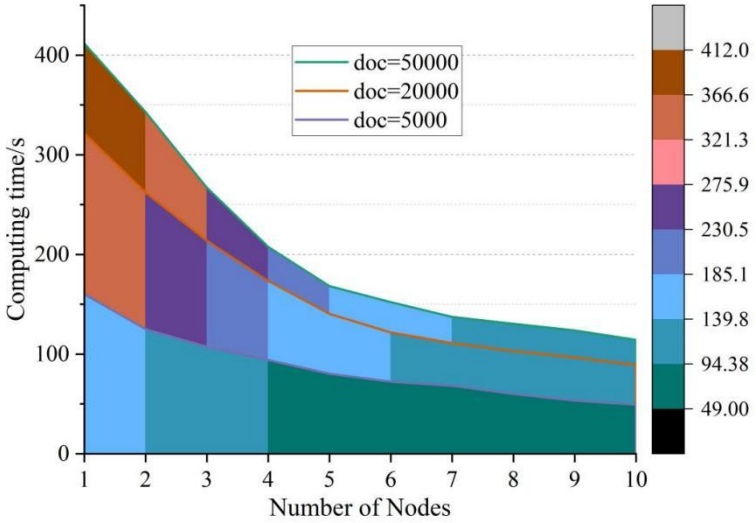


Figure 5: Time consumed for parallel computing of multiple nodes

It can be clearly seen that the computation time of the LDA model decreases significantly with the increase in the number of parallel computing nodes, which intuitively reflects the value of parallel computing. However, there is also a case of diminishing returns, for a set of 5000 documents, the computation time decreases from 160.16s for a single node to 93.94s for 4 nodes, with a larger decrease in efficiency, but the curve flattens out much more in the subsequent years, and when the number of nodes is increased from 60.13s for 8 to 49.23s for 10, the savings are only more than 10 seconds. The efficiency bottleneck occurs at about node = 5. For a large document set of 50,000 copies, the slope of the time descent curve is a bit larger, from 8 nodes to 10 nodes still saves about 16s, indicating that the model is able to eat more node arithmetic for large-scale tasks, and the parallelism advantage can be more fully released.

3.1.2 Acceleration ratio analysis

Based on the above computation time, it is derived that the speedup used for parallel computation of multiple nodes for LDA modeling under different sized corpus sets is shown in Fig. 6.

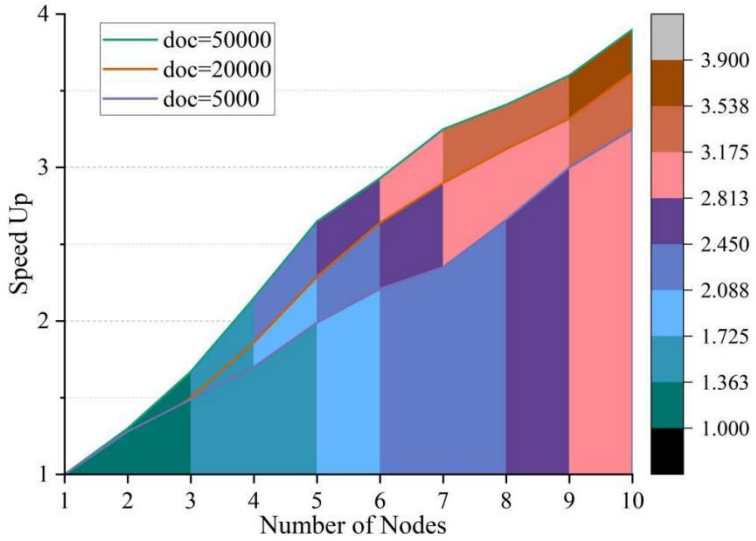


Figure 6: The speedup of time used for parallel computing of multiple nodes

Figure 6 provides a more standardized view of parallel computing scalability by profiling the acceleration ratio at different nodes. Ideally, the acceleration ratio for 10 nodes should be 10, but the actual data is much lower than that due to the unavoidable extra overhead in distributed computing. The larger the data size, the more efficient the acceleration. This is consistent with the previous time-consuming slope analysis, at 10 nodes, the acceleration ratio of 5000 documents is 3.25, while 50000 documents reaches 3.90. The study fully confirms the superiority of the LDA model for large-scale data.

3.2 LDA-based clustering analysis of innovation and entrepreneurship themes in colleges and universities

For the above 71,632 documents on innovation and entrepreneurship education for college students obtained from the resource library, the text preprocessing method in section 2.2 was used to clean the text and input it into the LDA topic model.

3.2.1 Clustering Visualization of Innovation and Entrepreneurship Themes in Colleges and Universities

Before modeling, the optimal number of themes is first established. The consistency value corresponding to each theme is calculated by Python, and it is concluded that K=6 is in the highest area of the fold, which indicates that the clustering effect between the themes is better, and it can reflect the clustering effect of theme clustering of innovation and entrepreneurship research in colleges and universities. The six theme clusters obtained based on LDA model contextual clustering are shown in Figure 7.

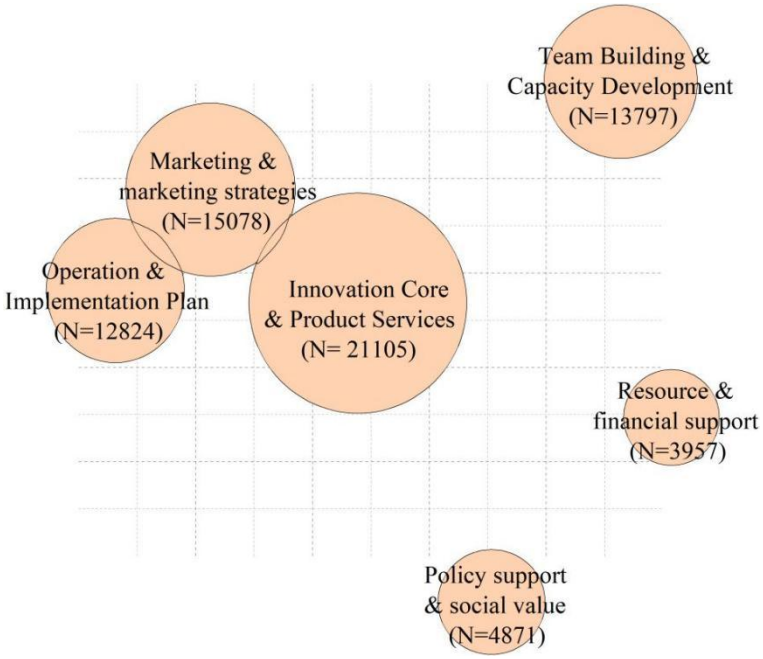
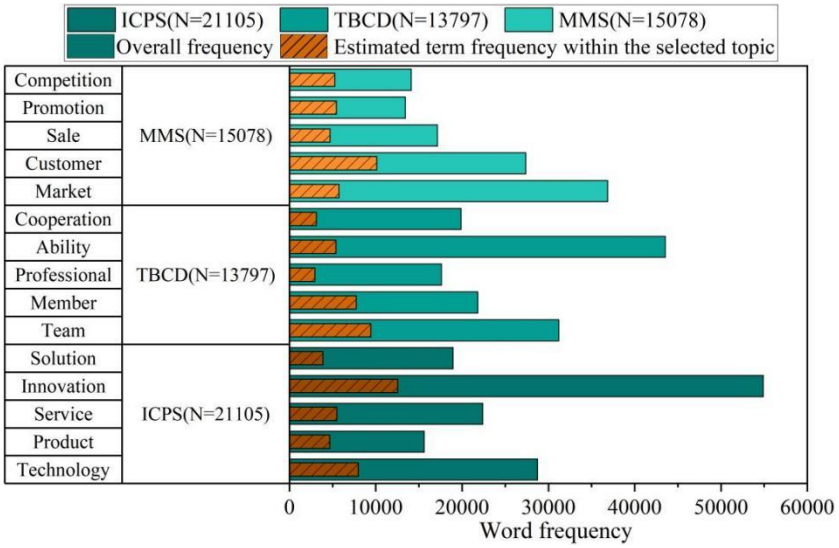


Figure 7: Cluster results of the innovation and entrepreneurship themes in universities

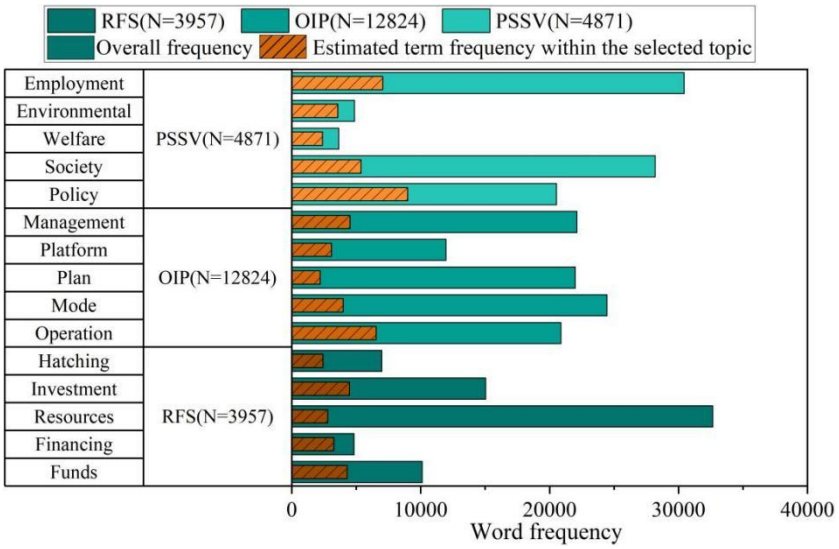
Clustering visualization shows that "Innovation Core and Product Services" is undoubtedly the most central and contains the most text content, N=21105. It can be found that there is a certain overlap between "Market and Marketing Strategy" (N=15078) and the central cluster. This is because the innovation core of product services must be deeply combined with user needs and market feedback. At the same time, there is also a certain overlap between operation and marketing, which includes 12,824 documents. The sample sizes of the other three theme clusters, "Team and Capacity Building", "Resources and Financial Support", and "Policy Support and Social Value", are 13,797, 3,957, and 4,871 respectively, and they are relatively far apart from each other, indicating good classification. From the perspective of overall layout, the six themes still form relatively clear groups. Only some themes have inevitable semantic proximity. The LDA topic model that integrates the context clustering strategy has an excellent clustering effect on the text in the field of innovation and entrepreneurship in colleges and universities.

3.2.2 Analysis of high-frequency words

The five high-frequency words with the highest frequency of occurrence within each theme were also selected to analyze their overall word frequencies and the estimated word frequencies within the selected themes in 71,632 documents on innovation and entrepreneurship education for college students extracted from the repository. The word frequency analysis is visualized in Figure 8.



(a)Theme of ICPS, TBCD and MMS



(b)Theme of ICPS, TBCD and MMS

Figure 8: Frequency Analysis Visualization of 6 Theme

The high-frequency words of each theme perform differently in the overall and specific fields. "Innovation", "employment" and "ability" are the three words that appear most frequently in the innovation and entrepreneurship documents, far ahead of other words. It is one of the five most frequently used words in defining themes and also a universal core concept that runs through multiple fields. When focusing on their respective themes, some high-frequency words also exhibit specific characteristics of the themes. For instance, the word "incubation" appears only 3,086 times in total, but it is mentioned 1,693 times in the "Resources and Financial Support" theme document. Similar terms include "financing" and "environmental protection", which have a strong presence in their respective themes and are key words that define topics in their fields. This distribution ingeniously confirms the quality of topic clustering based on the LDA context clustering model in this paper. Exclusive words anchor the topic characteristics, while general words connect the semantic networks between different topics, jointly forming a knowledge graph.

3.2.3 Thematic probability distributions

Based on the above analysis, taking the word "technology" as an example, it is a high-frequency term in "innovation core and product services", appearing 58,051 times in this topic, which is 27.81% of the total frequency of 208,742. In other words, it has a 27.81% probability of belonging to this topic. Similarly, the probability of each word belonging to each topic is shown in Figure 9.

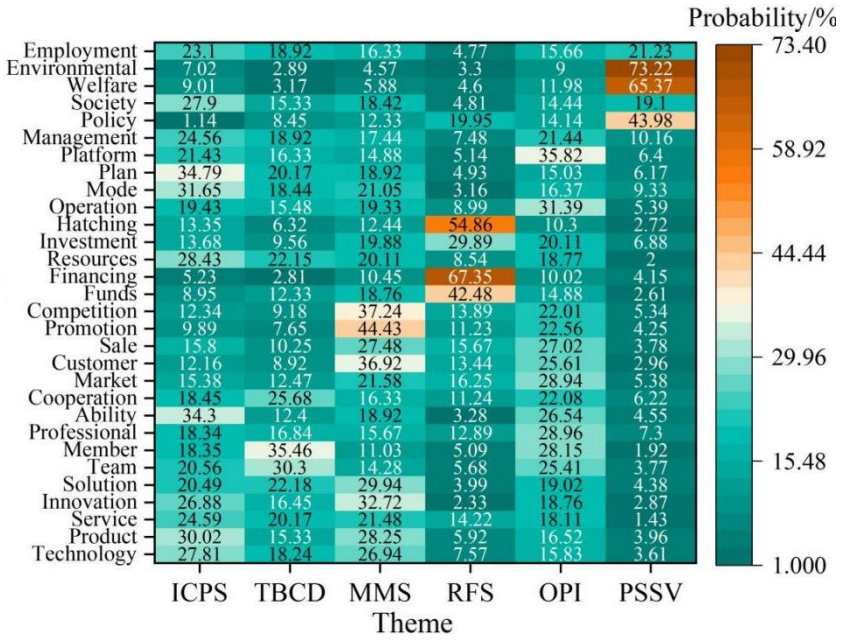


Figure 9: The probability distribution of different types of themes

Generally speaking, the frequency of high-frequency words in each region belonging to the same topic is still the highest. By observing the probability heat map, it can be found that the color mapping extends in a 5-gradient from the lower left to the upper right. At the same time, it can be found that "innovation core and Product services" has a strong affinity for each term. Take the term "resources" as an example. It should be highly related to the theme of financial support, but its affinity probability in this theme is only 8.54%, while its affinity probability for "innovation core and product Services" is 28.43%. This is because of its large sample size. The absolute number of times this term appears in this document reaches 37,719, thus giving it an advantage in probability calculations. This is also the reason why the probability of the terms belonging to the small sample themes "Resources and Financial Support" and "Policy Support and Social Value" is relatively low.

However, as mentioned in the above word frequency analysis, specific words such as "environmental protection", "public welfare", "financing" and "incubation" can still maintain a very high attribution probability in small sample themes, indicating that their semantic specificity is very strong. It can resist the attraction of sample gravity and be anchored in its own field. The probability value of vocabulary is not only related to the strength of the true semantic association with the topic, but also affected by the statistical cardinality brought by the sample size of the topic.

3.3 Comparative Experimental Analysis of Clustering and Retrieval Performance

3.3.1 Comparative analysis of clustering effects

In order to further explore the superiority of the clustering effect of the LDA model that introduces contextual clustering, comparative experiments are conducted with the traditional clustering methods, and the accuracy, recall and F1 are used as the evaluation indexes. The text categorization performance for different specifications of university students' innovation and entrepreneurship education documents is shown in Fig. 10.

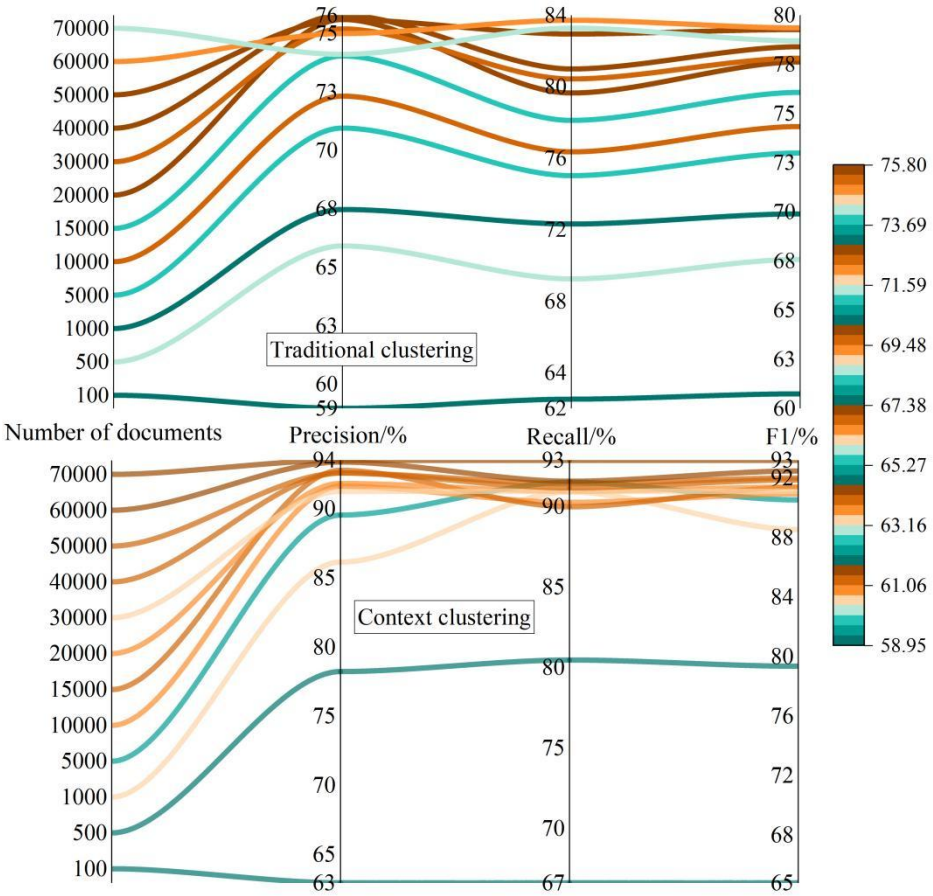


Figure 10: Performance of text classification for different specifications

When the data volume is only 100 documents, both methods perform poorly, with F1 of 64.79% and 60.74%, respectively. As the document size expands, the advantage of studying LDA topic model + context clustering over traditional clustering methods becomes more and more obvious. Its F1 value is raised from 88.53% at 1,000 documents to 93.19% at 70,000 documents, showing a strong large-scale effect. On the other hand, the F1 value of traditional clustering method is weak after the document volume exceeds 5000, hovering around 78%. Although its search rate can exceed 80%, the lower search rate still limits its development. The bottleneck of the traditional method is a direct reflection of the inability to understand the implicit semantic links mentioned in Section 2.3. Thanks to the deep subject index constructed by the LDA model in this paper, the semantic understanding problem is solved, while the contextual clustering strategy refines the retrieval results to keep the degree of completeness and accuracy.

3.3.2 Comparative analysis of search performance

It is known that the LDA model based on contextual clustering has 93.50% and 92.87% search accuracy and completeness for 70,000 large-specification documents, while only 74.10% and 83.27% under the traditional clustering method, and now focuses on the specific retrieval performance for the 71,632 articles of the six major topics, comparing the number of retrieved documents of the two methods for each topic. The results of the retrieval performance comparison test are shown in Figure 11.

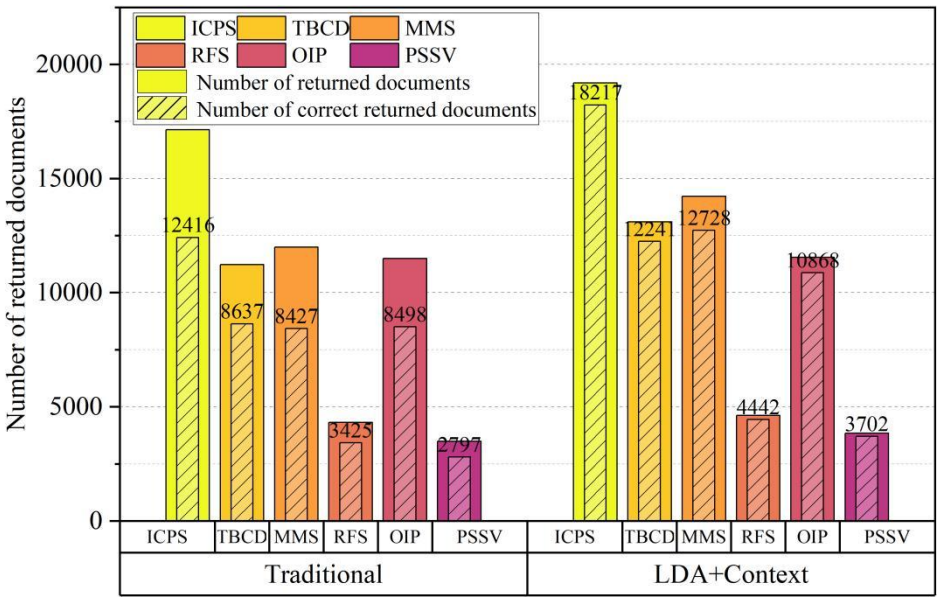


Figure 11: Search performance comparison test results

As the results of clustering analysis in section 3.3.1, the traditional method and this paper's method return a total of 59,646 and 66,524 documents for 71,632 documents, with a total search rate of 92.37% and 83.27%, respectively. The traditional clustering method for each topic retrieval returned documents are mixed with 684-6419 irrelevant information, the accuracy is only 74.1%, which indicates that relying only on the surface feature matching can not cross the gap between words and semantics, resulting in mixed results and low accuracy. And this paper introduces the context clustering method to effectively solve this pain point, which returns accurate retrieval of documents for all topics. Taking the theme of "innovation core and product services" as an example, the number of returned documents is 19,178, with a search rate of 90.87%, of which the number of accurate documents is 18,217, with an accuracy of 94.99%. For the small sample of "policy support and social value" theme retrieval accuracy is as high as 96.16%, only 148 irrelevant documents exist in 3702 documents. Once again, the effectiveness of combining deep semantic understanding and contextual clustering strategy is fully confirmed.

4 Assessment of the effectiveness of the application of the repository

4.1 Experimental setup

As a way of validating the success of the teaching resource library proposed in this paper which combines LDA topic indexing with context clustering strategies, an empirical study was

conducted for verifying the effectiveness of its use in promoting innovation and entrepreneurship among college students. A total of 135 students in two parallel classes of business administration major students from the batch of 2024 at a certain university were chosen as respondents. In class (1), there were 67 respondents who were included in the control group where conventional learning methods were employed, while in class (2), there were 68 respondents who formed the experimental group wherein the teaching resource library that employs the combination of "LDA topic model + context clustering" was employed.

The experiment lasted for 16 weeks, during which the two groups of students used the same core textbook, the same instructor, and the same class hours, and completed the same theoretical teaching and homework assignments in the course of "Fundamentals of Innovation and Entrepreneurship". The only independent variable was the resource retrieval support system they used.

In order to measure the effect of resource library application in a multi-dimensional way, this paper does not use a single subjective questionnaire, but evaluates it through both the total number of times students accessed the resource library and the quality of their business plans.

4.2 Comparative analysis of experimental results

4.2.1 Total number of student visits to the resource library

After a semester of teaching practice, the statistical results were obtained by counting the total number of times students in each class accessed the repository during 16 weeks as shown in Figure 12.

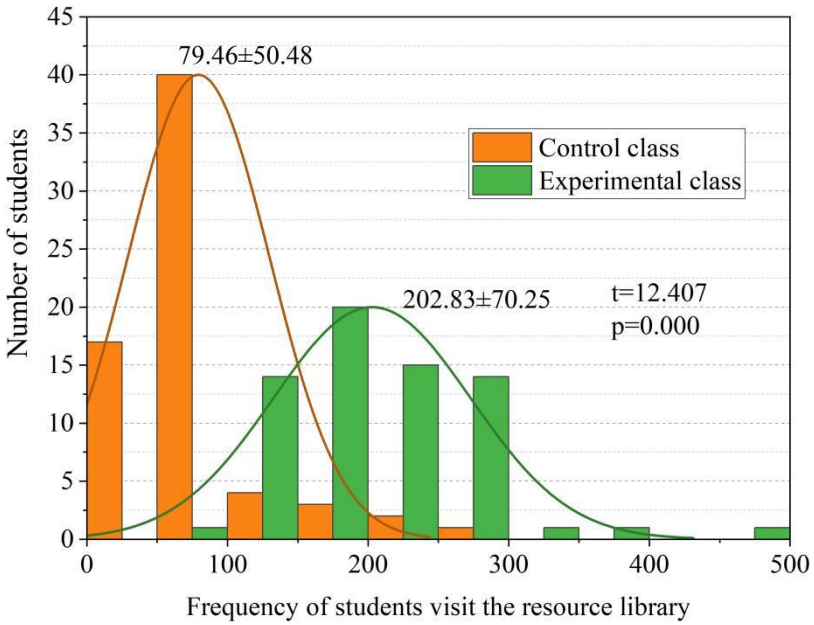


Figure 12: The total number of visits to the resource library by students in each class

The total number of times students in the experimental class, who used the resource library with integrated intelligent search function, accessed the resources averaged 202.83±70.25, which was much higher than that of the control group which was 79.46±50.48, and the difference between the two reached a statistically very high level of significance, p=0.000. Looking closely at the distribution of the number of times students in the two classes accessed the resources, the experimental group of students mostly accessed the resource library in the range of 150-300, and visited the resource library on average 1-2 times per day, while 59.7%

(40/67) of the students in the control group visited the resource library only between 50-100 times. This shows that the “LDA topic model + context clustering” fusion retrieval system constructed in this paper significantly improves the attractiveness and stickiness of the repository to students. In traditional teaching methods, students do not actively search the repository.

4.2.2 Quality of business plans

Students from each class were divided into 10 groups, each consisting of 6 to 7 people. A total of 10 business plans were produced. Three teachers scored the completion quality of the business plans for the final group assignments on a 100-point scale in the four dimensions of "innovation, feasibility, logic, and financial planning". The evaluation results of the business plans are shown in Figure 13.

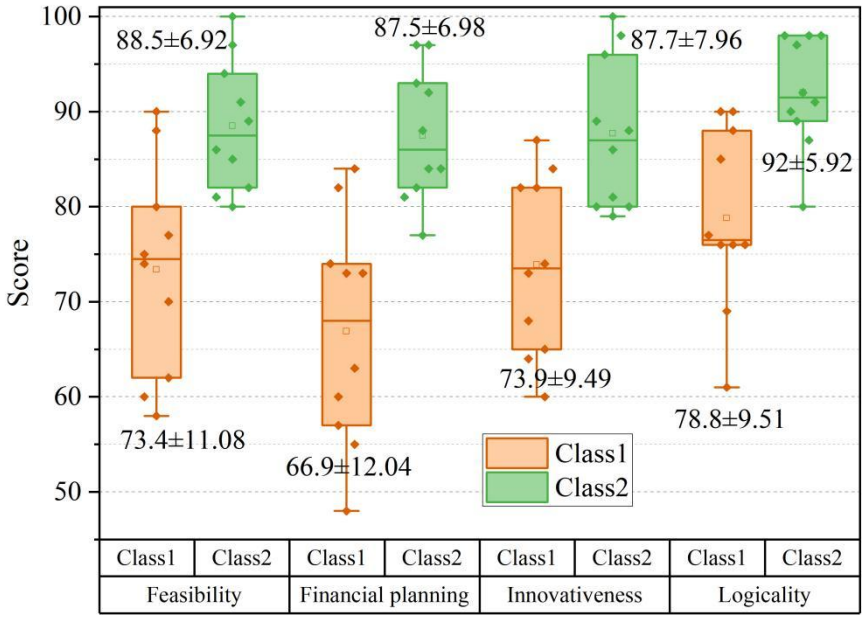


Figure 13: Comprehensive degree assessment of the business plan

The experimental group using the intelligent retrieval system scored higher than the control group in all dimensions of business plan quality, in which the teacher of the two groups of students to develop a business plan on the “financial planning” dimension of the largest difference in scores, the experimental class of four groups were more than 90 high scores, the overall score of 87.5 ± 6.98, while the control class The average score of the control class was only 66.9±12.04, with only two groups scoring above 80 and only three groups scoring below 60. Because this area involves a lot of terminology and structured knowledge, the resource library under the traditional program cannot cope with it well, while the LDA thematic model in this paper is able to understand the semantics of the theme of “resources and funds” behind each word, and the contextual clustering further clusters the dispersed resources. This enables the students to obtain high-quality information in this vertical field more efficiently, which significantly improves the professionalism of writing the financial part of the proposal.

In addition, the business plans of students in the experimental group scored about 14, 15 and 13 points in terms of “innovation”, “feasibility” and “logic” respectively. The advantages of this paper are summarized as follows Once again, this paper confirms the application value of the intelligent semantic retrieval system.

5 Conclusion

The study focuses on the core issue of how to make the retrieval of innovative and entrepreneurial educational resources for college students more efficient, and outlines the efficacy profile and application value of the fusion semantic understanding and dynamic organization method in this paper, from the construction of the method to the experimental validation, and then to the teaching application.

The model works best for the node parallel operation of 50000 samples, taking only 147.2 s. Meanwhile, no matter processing 5000 or 70,000 documents, the F1 value of the fusion method is always stable above 90%, with little fluctuation. In the retrieval task of subdivided topics, it performs especially well on “resources and financial support” and “policy support and social value”, which are highly specialized and easy to be confused by traditional methods, with the accuracy rates of 95.97% and 96.16%, respectively.

Meanwhile, students in the experimental group who used the smart resource library visited it more frequently and their learning trajectories were more in-depth. When writing a business plan, they scored 87.5±6.98 on financial planning, and the system was able to accurately map their fuzzy needs to the topic of “resources and funding” and present systematic knowledge clusters, thus lowering the threshold of specialization and assisting students in constructing a more complete knowledge framework.

Funding

1. Henan Provincial Higher Education Research Project. Project Name: Research on the Construction of a Digital-Intelligent Fusion Model for the Hierarchical Cultivation of Undergraduate Abilities Based on Big Data of Learning Behaviors (2025SXHLX237)

2. Humanities and Social Sciences Project of the Henan Provincial Department of Education. Project Name: Research on the Practice of University Culture Construction Led by Xi Jinping's Cultural Thought (2025 - ZZJH - 219)

References

[1] Selznick, B. S., & Mayhew, M. J. (2018). Measuring undergraduates’ innovation capacities. Research in Higher Education, 59(6), 744-764.

[2] Chen, X., Xiong, L., & Sun, Q. (2017, June). Research and practice on innovation and entrepreneurship education system in vocational colleges. In 2017 International Conference on Management, Education and Social Science (ICMESS 2017) (pp. 642-645). Atlantis Press.

[3] Wang, Y., & Ma, Y. (2022). Innovation and entrepreneurship education in Chinese universities: Developments and challenges. Chinese Education & Society, 55(4-5), 225-232.

[4] Wu, Z., Liu, H. Y., & Deng, X. Y. (2024). Teaching Practices for the Cultivation of “AI+ X” Composite Talents in Higher Education: Challenges and Strategies. Education Science and Management, 2(3), 156-175.

[5] Wang, X. (2024). Research on the influencing factors and upgrading paths for innovation and entrepreneurship education in universities under the background of sustainable

development goals: a QCA empirical study on new engineering of Chinese and foreign universities. International Journal of Sustainability in Higher Education, 25(7), 1426-1452.

[6] Liang, Q. (2023). Innovation and entrepreneurship education development on employment anxiety of college students. CNS Spectrums, 28(S2), S127-S128.

[7] Kirkley, W. W. (2017). Cultivating entrepreneurial behaviour: entrepreneurship education in secondary schools. Asia Pacific Journal of Innovation and Entrepreneurship, 11(1), 17-37.

[8] Lei, J., Hock, O. Y., & Asif, M. K. (2020). The influence of entrepreneurship education on innovation capability among Chinese undergraduate students in COVID-19 pandemic era: a framework of analysis. Solid State Technology, 63(6), 2279-2296.

[9] Gowda, R. S., & Suma, V. (2017, February). A comparative analysis of traditional education system vs. e-Learning. In 2017 International conference on innovative mechanisms for industry applications (ICIMIA) (pp. 567-571). IEEE.

[10] Ding, Y. Y. (2017). The constraints of innovation and entrepreneurship education for university students. Journal of Interdisciplinary Mathematics, 20(6-7), 1431-1434.

[11] Zhu, H. B., Zhang, K., & Ogbodo, U. S. (2017). Review on innovation and entrepreneurship education in Chinese universities during 2010-2015. Eurasia Journal of Mathematics, Science and Technology Education, 13(8), 5939-5948.

[12] Čapienė, A., & Ragauskaitė, A. (2017). Entrepreneurship education at university: Innovative models and current trends. Research for rural development, 2(23), 284-291.

[13] Zhang, M. (2025). The Systematization of Economic Principles Curriculum with Ideological and Political Education Based on the New Development Philosophy. Journal of Sociology and Education, 1(9).

[14] Xing, Z. (2025). Research on Talent Cultivation Strategies for Business Administration Majors in Applied Colleges and Universities Based on New Development Concepts. Evaluation of Educational Research, 2(8).

[15] Frolov, Y. V., & Bosenko, T. M. (2020). Training of personnel for the development of innovative entrepreneurship. Academy of Entrepreneurship Journal, 26(1), 1-6.

[16] Moica, S., Socaciu, T., & Rădulescu, E. (2012). Model innovation system for economical development using entrepreneurship education. Procedia Economics and Finance, 3, 521-526.

[17] Akhmetshin, E. M., Mueller, J. E., Chikunov, S. O., Fedchenko, E. A., & Pronskaya, O. N. (2019). Innovative technologies in entrepreneurship education: The case of European and Asian countries. Journal of Entrepreneurship Education, 22(1), 1-15.

[18] Aderogba, A. A. (2022). Entrepreneurship education and new media in Nigeria. NIU Journal of Social Sciences, 8(1), 133-139.

[19] Ye, J. (2024). Exploring Pathways for Mobile Interaction Technologies to Foster Innovation in Entrepreneurial Education Models. International Journal of Interactive Mobile Technologies, 18(10).

[20] Krämer, N. C. (2017). The immersive power of social interaction: Using new media and technology to foster learning by means of social immersion. In Virtual, augmented, and mixed realities in education (pp. 55-70). Singapore: Springer Singapore.

[21] Chen, L., Ifenthaler, D., & Yau, J. Y. K. (2021). Online and blended entrepreneurship education: a systematic review of applied educational technologies. Entrepreneurship education, 4(2), 191-232.

[22] Linzalone, R., Schiuma, G., & Ammirato, S. (2020). Connecting universities with entrepreneurship through digital learning platform: functional requirements and education-based knowledge exchange activities. International Journal of Entrepreneurial Behavior & Research, 26(7), 1525-1545.

[23] Marion, T. J., & Fixson, S. K. (2021). The transformation of the innovation process: How digital tools are changing work, collaboration, and organizations in new product development. Journal of Product Innovation Management, 38(1), 192-215.