International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 7674-7684

https://doi.org/10.70517/ijhsa463658

Study on the Evolutionary Trend of the Styles of Modern and Contemporary Chinese Literary Works Based on Data Visualization Techniques

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

Abstract The stylistic evolution of modern and contemporary Chinese literary works carries the profound accumulation of history and culture, and the study of its development trend can reveal the social, cultural and ideological changes behind it. This paper analyzes the trend of style evolution of modern and contemporary Chinese literary works through text mining techniques. The research methodology includes using TF-IDF keyword extraction and LDA topic model analysis to process and analyze a large amount of literary work review data. The results show that "realism", "romanticism" and "modernism" are the most frequently occurring themes, and the styles of literary works have diversified over time. The results show that "realism", "romanticism" and "modernism" are the most frequent themes, and the styles of literary works have diversified over time. Specifically, "Literary Genres and Styles" accounted for the highest percentage of literature in the last ten years, at 19%, while "Creative Techniques and Narrative Techniques" and "Psychological and Humanistic Exploration" accounted for 16% and 13% respectively. In addition, through the LDA model analysis, it was found that the evolution of literary styles shows a gradual development from realism to romanticism and then to modernism. The study shows that with the development of the Internet and big data technology, the forms and techniques of literary creation are constantly innovated and show a diversified trend.

Index Terms data visualization, text mining, TF-IDF, LDA model, literary style, evolutionary trend

I. Introduction

In recent years, a number of representative works and courses on the interpretation of literary works have appeared in the field of the study of modern and contemporary Chinese literature, which have not only received the attention of professional researchers, but also the attention and praise of the public [1], [2]. These writings show several distinctive features, which are not only the development, integration and application of the resources of modern and contemporary Chinese literature, but also the new attempts and practices made to literary education [3].

In its special historical practice, modern and contemporary Chinese literature has given birth to the specific literary form of socialist literature with Chinese characteristics, which has become the center and mainstream of modern and contemporary literature, with obvious political, prescriptive and epochal character, a strong sense of social mission and noble aesthetic pursuit [4]-[6]. The authors seek to return to the social and epochal context of present-day China to fully understand the inevitability and complexity of expressing political ideals in present-day literature [7]. For this reason, it is only then that present contemporary literature establishes a new set of literary order from top to bottom, incorporating literature into institutional organization as a literary unit, and presenting literature as a planned mode of production through official formation and operation [8]-[10]. On the one hand, it reformed and cultivated the writers to realize the ideological unification and conceptual purification of the subjects of literary creation [11]. On the other hand, it guides and regulates the subject matter, form and style of creation, realizing a high degree of concentration and convergence in the process of literary production, thus achieving the educational and transformative function of literature on society [12], [13]. Based on this, a full understanding of the stylistic evolution of modern and contemporary Chinese literary works is a new response, exploration and practice of literary education and literary communication.

The style of literary works is the most personalized and contemporary character of literary creation, which not only reflects the artistic pursuit and value orientation of writers, but also embodies the social and cultural psychology of a specific historical stage. Since its development in the early 20th century, Chinese modern and contemporary literature has gone through a hundred years of changes, and the evolution of its style has not only inherited the traditional culture, but also absorbed the western literary trends, and even more uniquely interpreted the spirit of the

times. From the enlightenment and realism of the May Fourth New Literature Movement to the pioneering exploration of the modernists, from the grand narrative of socialist realism to the diversified development of new-era literature, the style of modern and contemporary Chinese literature has shown a complex and varied evolutionary trajectory. This evolution process not only witnesses the change of Chinese society, but also maps out the transmutation of national spirit, which has profound cultural meaning and research value. However, previous studies on the evolution of literary styles have mostly adopted qualitative descriptions and case studies, which lack systematicity and objectivity, making it difficult to fully grasp the overall trend and internal laws of style evolution. The traditional research method is limited to the subjective cognition of the researcher and limited samples, which often fails to accurately reflect the macroscopic pattern and microscopic details of style evolution. Moreover, literary style itself is multidimensional and ambiguous, and it is difficult to precisely define and quantitatively analyze it only by relying on traditional textual interpretation. Therefore, exploring new research methods and technical means to build a more scientific and comprehensive literary style research paradigm has become a necessary proposition for current literary research.

The rapid development of data visualization technology and text mining methods provides new ideas and tools for literary style research. Through the computational analysis of large-scale text corpus, the constituent elements and evolutionary laws of literary style can be revealed from multiple dimensions, such as word frequency statistics, semantic network, and theme model, so that the style research transcends the individual feelings and obtains more objective empirical support. Based on this, this study proposes a research framework for the evolution of Chinese modern and contemporary literary style based on data visualization and text mining. First, we collect research literature on the style of modern and contemporary Chinese literature through academic platforms such as Zhi.com, and construct a text database. Secondly, the text data are processed by Chinese word segmentation technology, and the keywords are extracted by TF-IDF algorithm to establish a vocabulary of style features. Once again, high-frequency keyword analysis, co-occurrence mapping and cluster analysis are applied to explore the constituent elements and intrinsic connections of literary style. Then, the LDA theme model is used to identify the main style types and analyze their percentage distribution and time evolution trend. Finally, based on the results of empirical analysis, the study summarizes the basic laws and characteristics of the evolution of styles in modern and contemporary Chinese literary works, providing new perspectives for literary creation and criticism. The study combines the qualitative interpretation of literary criticism with the quantitative analysis of data science, which maintains the humanistic concern of literary research and introduces the rigor of scientific method, and helps to establish a more systematic and objective paradigm for literary style research.

II. Data collection and text mining

II. A. Research Design Ideas and Processes

The research idea and process of this paper are briefly introduced here. Firstly, the research data sources and sample selection are explained and elaborated, and then online data mining and collection are carried out on TripAdvisor, Zhi.com and other websites, and the collected data are pre-processed to construct a text database for studying the evolutionary trend of the styles of modern and contemporary Chinese literary works. Then relevant text mining, analysis and processing were carried out through the application of text mining methods and software, such as Eight Catch Fish software, SPSSAU software and python technology, with the core techniques and analyses including TF-IDF [14] keyword extraction, and LDA topic model analysis (for discovering major themes and patterns in the reviews). The flowchart of the text mining research design is shown in Figure 1.

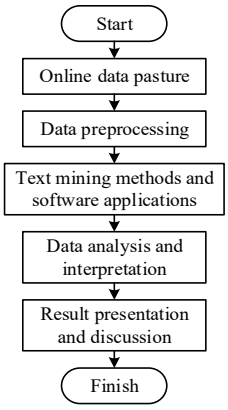

Figure 1: Flowchart of Text Mining Research Design

II. B.Data sources and data pre-processing

II. B. 1) Data sources

In order to more accurately explore the evolutionary trends of the styles of contemporary Chinese literary works, this study carefully examines the article collections used for thematic cluster mining. In the process of article collection selection, this study firstly circled the 15 journals with the highest total citations of published articles on entrepreneurship in Chinese and English respectively, and then filtered them through experts in the field, and finally comprehensively selected journals with greater influence in the academic world.

II. B. 2) Data pre-processing

Data preprocessing is a very critical step, in order to ensure the effectiveness and accuracy of the analysis, the collected comment data must be preprocessed The following is the specific data preprocessing process:

(1) Data cleaning: first of all, remove irrelevant information in the comments.

(2) Remove deactivated words: there are a large number of deactivated words in the online comments, in order to analyze the words with little significance, removing these words can focus on analyzing the words that are more meaningful.

(3) Segmentation: As the research object of this paper is Chinese, segmentation becomes especially important, and Chinese segmentation is a key part of text preprocessing. Sentences are broken down into words or phrases for subsequent analysis.

(4) Build a vocabulary list: according to the preprocessed text, build a vocabulary list when needed.

Text data preprocessing is a prerequisite for effective text mining. Through the above steps, the quality of data can be ensured, laying a foundation for subsequent text mining and analysis.

II. C.Methods for mining and analyzing text data

II. C. 1) Chinese Segmentation

Chinese word division, that is, the process of dividing a continuous sequence of words into separate words by following certain standardized principles. In English text, each word is separated by a space in the middle, and here lies the biggest difference between Chinese text and English text, which is generally composed of consecutive words and characters to form a sentence, and then divided by punctuation marks. The effect of word division will affect the text mining work afterwards. The common word separation methods are dictionary-based word separation method and statistical-based word separation method.

“Jieba Segmentation” is a Chinese segmentation tool that combines both dictionary-based and statistics-based methods. It uses both dictionary for exact matching and HMM (Hidden Markov Model) based model to solve the ambiguity problem in Chinese word segmentation. This combination makes jieba disambiguation able to handle both common disambiguation tasks efficiently and some special cases better, improving the accuracy and flexibility of disambiguation.

II. C. 2) Word frequency and frequency analysis

Word frequency usually refers to the number of times a word appears in a text, while frequency refers to the ratio of this word frequency relative to the total number of words in the text.

Frequency analysis in text mining usually refers to the statistical analysis of the number of occurrences of a word or phrase in a text in order to identify and evaluate key elements in the text. Commonly used methods include word frequency (TF) and inverse document frequency (IDF). Word frequency TF in text mining mainly refers to the calculation of the number or ratio of occurrences of each word in the text. And Inverse Document Frequency IDF is used to measure how much information words can provide, i.e. the general importance of words. Word frequency analysis is a basic but very important text analysis technique used to reveal features and trends in text content. This method can be used to identify and determine the most frequently occurring key words or phrases in a review text to understand the main concerns of a literary work's style.

The basic formula for word frequency (TF) is:

$$TF = \frac{\text{The number of times a term appears in a document}}{\text{The total number of words in the document}} \tag{1}$$

II. C. 3) Keyword extraction

Keyword extraction is achieved through TF-IDF analysis, which is a common statistical method used in text mining to assess the importance of a word in a document. It is calculated by considering the frequency of occurrence of a word in a document and the frequency of its distribution across the entire collection of documents.TFIDF analysis takes into account both the word frequency (TF), which is the frequency of occurrence of a word in a document, and the inverse document frequency (IDF), which reflects the generalized importance of a word. Thus, TF-IDF is

both a frequency analysis and goes beyond simple frequency statistics by combining these two metrics to assess the importance of words in a set of documents.

In a study about the evolutionary trend of the style of modern and contemporary Chinese literary works, TF-IDF can identify keywords in the reviews. By analyzing these keywords, the evolutionary trend of the style of modern and contemporary Chinese literary works can be better understood. The specific principles and formulas are as follows:

(1) TF denotes the word frequency, i.e., the frequency or number of times a word appears in a document. In order for this algorithm to be tried on all texts, it is normalized in the process of calculation, that is, the number of occurrences is divided by the number of words in the whole document.

(2) IDF represents the Inverse Document Frequency, i.e. the general importance of a word in all documents. In this formula, 1 is added to the denominator in order to be able to avoid a denominator of 0. The larger the IDF value of a word, the greater the contribution of this word to the document, and more likely to become a keyword.

$$IDF = \log \frac{\text{Total number of documents}}{\text{The number of documents containing this word} +1} \tag{2}$$

(3) TF-IDF algorithm:

Combining TF and IDF, we can obtain TF-IDF values for evaluating the importance of words in a set of documents. The formula for TF-IDF is usually expressed as:

$$TF-IDF = TF \times IDF \tag{3}$$

Usually, in order to distinguish documents more clearly, the frequency of occurrence of the word will be measured, but if we consider this point alone, it is obviously also very unreasonable. Therefore, after comprehensive and practical consideration, the above two algorithms are used together to form a TF-IDF algorithm, which measures a word from two main aspects: word frequency and inverse document frequency. By combining these two aspects, the algorithm generates a higher weighted TFIDF, resulting in a more recognizable and representative word.

II. C. 4) LDA Thematic Model Analysis

The full name of the LDA model is Hidden Dirichlet Distribution [15]. In the present era of explosive growth of information, how to automatically extract the topic words in an article is a very important topic.

LDA topic model has many more obvious features: first of all, LDA topic model is a typical bag-of-words model, that is to say, the model assumes that an article is a set of words composed of a set of words, for the process of clustering the topic of the article, does not take into account the sequential order of the relationship between words and words, and only considers the count of different words in the article. At the same time, the LDA model considers that each word in a document is generated by one of the topics, so the topic of each document can be given in the form of a probability distribution. In addition to this, the LDA model is a typical unsupervised learning algorithm, with the advantage that there is no need to manually label the training set during training, all that is needed is the document set and a specified number K of topics. Finally, the topics obtained from the clustering of the LDA model can all be described by finding some words, which is more intuitive.

For a collection of articles, where the number of articles is D and the number of all possible topics in the collection of articles is K, the LDA model considers that each article consists of a mixture of probability distributions of K topics, and at the same time, each topic is composed of a mixture of probability distributions of different words. Therefore for a collection of articles to be clustered using the LDA algorithm, for any article d in the document repository D, the distribution θ_d of its topics is calculated a priori as:

$$\theta_d = Dirichlet(\vec{\alpha}) \tag{4}$$

where $\vec{\alpha}$ is a vector of dimension K and is a hyperparameter of the Delikere distribution.

In addition to the fact that the topic of an article is considered to obey a Delikeray distribution, the prior distribution of the words in the topic is likewise considered to obey a Delikeray distribution. And for any one of the K topics k, the distribution of words β_k is calculated a priori as:

$$\beta_k = Dirichlet(\vec{\eta}) \tag{5}$$

where $\vec{\eta}$ is a vector of the same dimension as the size of the word list and is a hyperparameter of the Dirichlet distribution.

On the basis of the prior distributions, the procedure to compute the posterior distributions of the topic distribution parameter θ_d and the word distribution parameter β_k based on the likelihoods observed in the collection of

articles is as follows: firstly, using the bag-of-words model, statistically obtain the word distribution of each article in the collection of articles, i.e., for article d, the kth topic word of the article is $n_d^{(k)}$, which is calculated by the formula:

Combined with the prior distribution of article topics, the posterior distribution of the parameters θ_d based on the Dirichlet function can be obtained by the method of Bayesian inference as:

$$Dirichlet(\theta_d \mid \vec{\alpha} + \vec{n}_d) \tag{6}$$

In the same way, for between words and topics, the number $n_k^{(V)}$ of the V th word in the kth topic can be expressed as:

$$\vec{n}_k = (n_k^{(1)}, n_k^{(2)}, ...n_k^{(V)}) \tag{7}$$

Thus the posterior distribution of β_k is:

$$Dirichlet(\beta_k \mid \tilde{\eta} + \tilde{n}_k) \tag{8}$$

Since the article topic distribution and word distribution are independent, the topic distribution of each article and the word distribution in each topic can be solved using the Gibbs sampling method, and the topic distribution θ_d and the word distribution β_k that converge to a smooth one can be obtained using the Gibbs sampling method.

When facing an unknown or complex distribution, a very common method is to use MCMC (Markov Chain Monte Carlo) method to sample the distribution, and the purpose of sampling is to get the samples of this distribution, through which the specific structure of the distribution is clarified. So MCMC itself solves the problem of distributions that cannot be sampled or understood directly, so it is not sampling a known distribution. And Gibbs sampling is an improvement strategy of MCMC method. In the LDA model, the posterior probability cannot be obtained directly, and the distribution can be sampled by Gibbs sampling to get the model structure.

Therefore, in the process of solving the LDA model, the core methods used are MCMC (Markov Chain Monte Carlo) and Gibbs sampling, and the process is:

(1) Determine the appropriate number of topics K based on the available set of articles. Randomly initialize the hyperparameters $\vec{\alpha}$ and $\vec{\eta}$ that generate the Dirichlet distribution.

(2) Iterate over all the words in the word list obtained from the article collection, and for each word w, randomly initialize a topic number z.

(3) Iterate through all the words in the article collection, and for each word w, update the topic number to which w belongs according to the Gibbs sampling formula.

(4) Repeat step 4 until convergence of the model occurs.

(5) When the model converges, count the topic-word co-occurrence frequency matrix in the article collection, i.e., get the final converged LDA model.

Usually, during the LDA training process, the results of n iterations after convergence by Gibbs sampling are averaged for parameter estimation to improve the quality and stability of the model.

After the converged LDA model is obtained by Gibbs sampling, for a new article, the algorithmic flow of the LDA model to get the topic distribution of the article is:

(1) Randomly initialize, traverse all the words w in the new article, and randomly assign a topic number z to each word w.

(2) Re-traversing all the words in the article and resampling each word w to get its topic according to the Gibbs sampling formula.

(3) Repeat step 2 until Gibbs sampling converges.

(4) Statistical distribution of topics in the article, i.e., get the topic distribution result of the new article.

Since the LDA model needs to set the number of target topics artificially, it needs to determine the number of topics of the articles input to LDA in a certain way, and the input of different number of topics may have a large impact on the results of LDA training.

For the problem of the number of topics in the LDA model, there are usually two main types of methods: the first type is to refer to industry research or experience. The second category is determined by the authors of the paper themselves.

Since an LDA model is essentially a probability distribution model, some work has considered the use of a perplexity metric that measures the goodness of a probability distribution to determine the degree to which each model perplexity metric is used to measure how well a probability distribution or probability model predicts a sample,

and it can also be used to compare two probability distributions or probability models. For the LDA topic probability model, essentially two probability distributions are generated unsupervised from the full document collection, namely the article-topic probability distribution and the topic-word probability distribution. Therefore, by calculating the perplexity degree obtained under the condition of setting different numbers of topics, the advantages and disadvantages of different numbers of topics can be well measured. The formula for calculating the perplexity degree is:

$$perplexity(D_m \mid M) = \exp\left|-\frac{\sum_{i=1}^{M}\log(P(d_i))}{\sum_{i=1}^{M}N_i}\right| \tag{9}$$

II. C. 5) Semantic network analysis

Semantic network analysis (SNA) [16] is Semantic network analysis is a text analysis technique.

In semantic networks, each node represents a keyword, and the co-occurrence relationship between keywords and keywords describes the internal compositional relationship and its structure in a certain field, and can also be used to reveal the dynamics and development trends of a discipline.

Briefly, in semantic networks, each node represents a concept or entity, while edges represent semantic relationships between nodes, such as synonymous, contextual, and associative relationships. By analyzing the semantic network, information such as the degree of correlation between the concepts of the main participants, events, places and other entities in the text, the distance between the concepts, as well as the set structure of the concepts, can be revealed, so as to gain a deeper understanding of the organization of the language or the knowledge, the semantic correlation, and so on.

III. Co-word analysis

III. A. High-frequency keyword analysis

Keywords are important information in articles and journals. Based on the high-frequency keywords reflected in a research field, the focus and hotspot of research in this field can be grasped. Through the statistics of text data in the Knowledge Network database, keywords are extracted and meaningless and invalid keywords are eliminated, and finally 4,500 keywords are obtained from 2,000 documents. In the paper, the high-frequency word threshold is estimated by the Price formula, and the high-frequency word threshold M is calculated to be about 15, i.e., when the frequency of keywords appearing in the text is more than or equal to 15, it is determined to be a high-frequency keyword, and there are a total of 30, and the keywords are shown in Table 1.

Table 1: Key words

Serial number	Key words	Frequency	Serial number	Key words	Frequency
1	Reality	245	16	Poetics	111
2	Romanticism	235	17	Critically	95
3	Modernism	221	18	Warmth	90
4	Postmodernism	213	19	Cold and steep	81
5	Magic realism	200	20	Massiness	73
6	Avant-garde	189	21	Lightness	62
7	Realism	175	22	Deep	55
8	Lyricism	165	23	Freshness	43
9	Satirical	158	24	Ornate	39
10	Sense of humor	152	25	Simplicity	25
11	Absurdity	145	26	Complexity	22
12	Exquisite	142	27	Multivision	20
13	Grand narrative	130	28	Fragmentation	18
14	Onlooo	125	29	Symbolism	16
15	Narrative	120	30	Regionality	16

In the process of studying the evolution trend of the style of modern and contemporary Chinese literary works, the frequency of keywords such as "realism", "romanticism", "modernism", "postmodernism", "magical realism", "avant-garde", "realism", "lyricism", "irony" and "sense of humor" is relatively high, which also shows that scholars

are relatively concentrated in the hot spots and directions of the evolution trend of modern and contemporary Chinese literary works. In addition, other key words such as "lightness", "depth", "freshness", "gorgeousness", "simplicity" and "complexity" have also appeared, indicating that writers have begun to shift from the reality and idealism of literary works to pay attention to the language style of literary works.

III. B. Keyword co-occurrence mapping construction

In order to deeply excavate the co-occurrence situation within the keywords of the evolution trend of the styles of Chinese modern and contemporary literary works, the paper uses Python to obtain the keywords from the literature, extracts the keyword frequency, and defines the high-frequency keywords using the Price formula, constructs the high-frequency keyword matrix by using Python, and finally inputs the obtained co-occurrence matrix into the Gephi software, and plots the corresponding keyword co-occurrence map as shown in Fig. 2. This reflects the research hotspots and the basic situation of the stylistic evolution trend of Chinese modern and contemporary literary works when keywords are used as the research corpus.

It can be seen that there are obviously two related network structures. One is centered on realism, reflecting the close connection between realism and romanticism, modernism, postmodernism, magical realism, and so on. It shows the continuity of the evolution of literary style, from realism to romanticism to modernism, literary style has not been developed in a fragmented way, indicating that the evolution of literary style is not an isolated jump, but an orderly advance based on the results of the predecessors, and the subsequent style of literary work is produced after the understanding and absorption of the established style of literary work.

The other center is warmth, which is more closely related to the relationship between lightness, depth and freshness. It reflects the richness and complexity of emotions and thoughts in Chinese literary works, as well as the deep heritage and unique aesthetic concepts of Chinese culture. It can be seen that if the keywords are used as the corpus to study the literature on the trend of style evolution of Chinese modern and contemporary literary works, the number of effective themes extracted can be effectively integrated across time and space, and become the treasures of the cultural treasury of Chinese literary works.

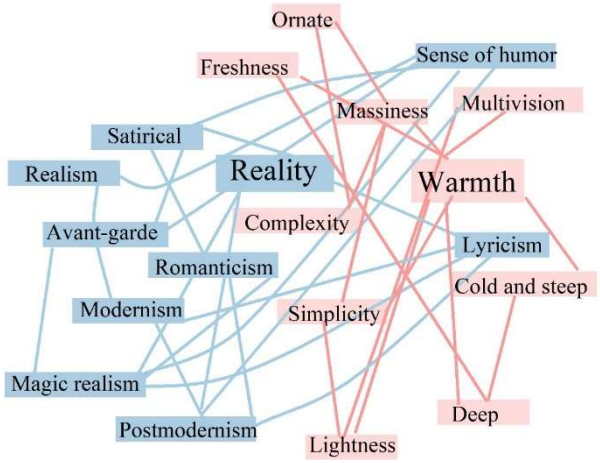

Figure 2: High frequency keywords

III. C. Keyword clustering analysis

Figure 3 is the keyword clustering analysis made by using the modular algorithm in gephi, the keyword clustering can react to the concentration degree of the trend of the evolution of the style of Chinese modern and contemporary literary works, and there is a node slightly larger in each color classification, which is equivalent to the theme after clustering. There are three colors in the Knowledge Network Keyword Clustering Knowledge Graph of the Evolutionary Trends of the Styles of Modern and Contemporary Chinese Literary Works, which represent the themes after clustering. After in-depth analysis and organization of the clustered information, the 3 keywords with the highest tag value are obtained.

The number of nodes for a keyword indicates the number of keywords included in the text set. In terms of the number of nodes included in the graph, “realism” has the highest number of nodes, 15, which indicates that it is closely related to many keywords and has been studied in various fields. The thickness of the lines in the graph indicates the degree of association. It can be seen that “realism” is closely related to postmodernism, which is in line with the trend of the evolution of the style of real Chinese literary works. To a certain extent, postmodernism

has inherited modernism's challenge to traditional concepts, and further expanded the exploration of language, culture and meaning.

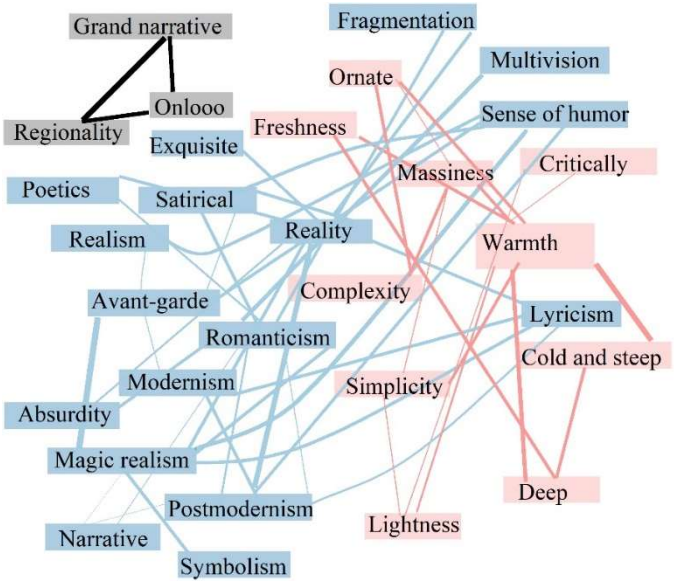

Figure 3: Knowledge map of gateway key word clustering

IV. LDA Thematic Modeling Analysis Results

IV. A. Determination of the number of subjects

The final result of the LDA topic model is greatly influenced by the number of topics K. The LDA topic model can be used to determine the optimal number of topics. The main indicators for evaluating the goodness of LDA topic models are perplexity indicator, log-likelihood indicator and U-mass indicator, etc. In this paper, the popular perplexity indicator is used as the basis for determining the optimal number of topics K. The perplexity indicator, which is also called the perplexity metric, which calculates the inverse of the geometric mean of the probability of occurrence of all words in the test set. It is a commonly used criterion for evaluating the quality of clustering, which decreases monotonically with the number of topics, and the lower the perplexity, the better the model. In the paper, the gensim open source package is utilized, the hyperparameters alpha and eta are set to auto-optimization, the number of iterations iteration=100, and the number of traversal of the document set is set to 20 times in batch LDA mode, the model is built under different number of topics K and the respective perplexity values are calculated, and the results are shown in Fig. 4. It can be seen that the perplexity value gradually decreases with the increase of the number of topics, leveling off near K=15 and reaching the global minimum. Therefore, the number of topics K is set to 15.

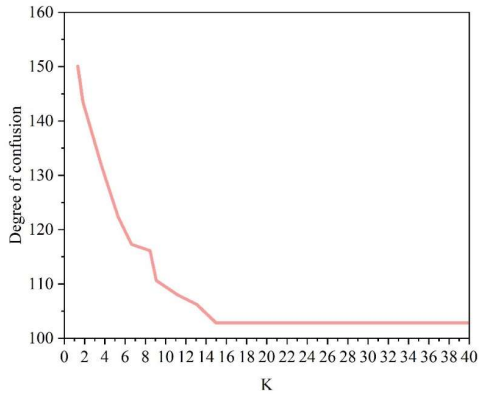

Figure 4: Searching optimal number of topics K with perplexity

IV. B. Theme identification

Using the GenSim open source package, the model was established under the condition that the number of topics was 15, and the topic identification was carried out. The top 10 words with the highest probability are selected for each topic as the high-frequency representative words of the topic. If only the confusion is used to determine the number of topics K, it will often lead to the semantic ambiguity of some topics. Therefore, this paper further examines the high-frequency representative words of each theme, and finds that the five words "delicate", "complex", "fresh", "gorgeous" and "simple" appear in most of the themes, and these words have no practical significance for distinguishing each theme. Therefore, without considering these words, the meaning of the remaining words under each topic is synthesized and the content of the corresponding summary is reviewed to label the topic. Finally, 10 valid themes were identified, and the results are shown in Table 2. The 10 themes are literary genre and style, creative technique and narrative technique, emotion and thematic expression, criticism and reflection, narrative structure and content, language and style, culture and regional characteristics, psychological and human exploration, artistic pursuit and innovation, and depth of emotion and ideology.

Table 2: Effective theme recognition results

Topic number	Theme label
1	Literature genre and style
2	Creative techniques and narrative skills
3	Emotional and thematic expression
4	Criticism and reflection
5	Narrative structure and content
6	Language and style
7	Cultural and regional characteristics
8	Psychological and human exploration
9	Art pursuit and innovation
10	Emotional and ideological depth

IV. C. Thematic Share Analysis

The LDA topic model calculates a topic probability distribution for each literature summary. The topic probability distribution determines that the meaning of the topic to which a document belongs is consistent with the description of the corresponding topic identified. Therefore, the topic probability distribution is calculated for all the literature abstracts in the document set, and then the number of literature under each topic is summarized to obtain the distribution of the proportion of the topic of the entire document set, and the results are shown in Fig. 5, in which p is the proportion of the number of literature under each topic, and i is the topic number.

The proportion of “Literary Genres and Styles” is 19%. This is because literary genres and styles are the cornerstone of Chinese literary works, not only the core features of literary works, but also the diversity and complexity, which are able to dominate other themes and are of great significance for historical, cultural and academic research. Secondly, the two themes of “creative techniques and narrative skills” (16%) and “exploration of psychology and human nature” (13%) are the embodiment of artistic innovation in modern and contemporary Chinese literature, and therefore the literary community is also very concerned about these two themes.

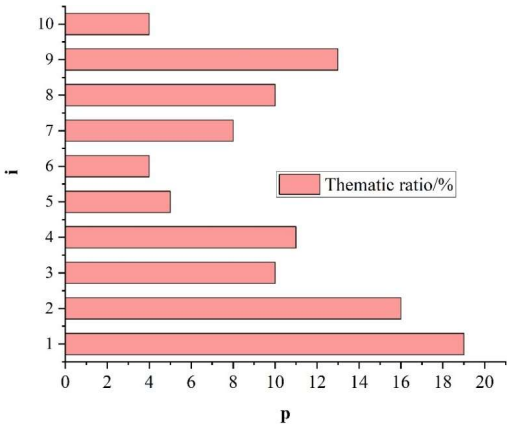

Figure 5: Overall topic proportions for whole corpus

IV. D. Trend analysis of the evolution of the theme

Figure 6 shows the trend of the number of literatures of the 10 themes with the year, in which N is the number of literatures per year for each theme and T is the year of the publication of the literature. The pattern of rising or falling number of literature for the themes can reflect the research trend of the themes. The identification of fast-growing and changing research themes can provide valuable information for the study of hot issues in the field of the evolutionary trend of the styles of modern and contemporary Chinese literary works.The Mann-Kendall nonparametric trend test is used to predict the trend change of each theme. The results show that under the condition of two-sided test significance level of P=0.05, one theme shows a significant upward trend, which is “creative techniques and narrative skills”. This indicates that with the rapid development of the Internet and big data and other technologies, the medium and form of literary creation have undergone profound changes, and the rise of network literature has made creative techniques and narrative skills a hot topic of research, and scholars are concerned about how to create literary expressions in the new media environment.

The number of literature on “exploration of psychology and human nature” is relatively small and the trend of decline is obvious. This is because with the rapid changes in society and the diversification of cultural consumption, the delicate portrayal of psychology and human nature in literature has been gradually replaced by a wider range of socio-cultural issues, which has led to a gradual decline in the attention paid to this theme.

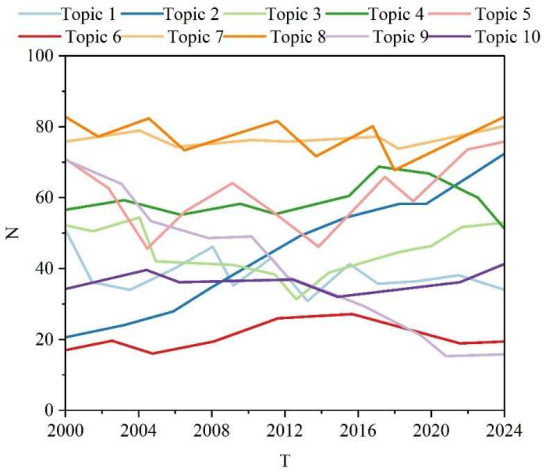

Figure 6: Research topic trends during 2000—2024

V. Conclusion

Data-driven literary style research provides a new perspective for revealing the evolution of modern and contemporary Chinese literature. By mining and analyzing 4,500 keywords, it is found that the evolution of Chinese modern and contemporary literary styles is characterized by diversification and stages. The analysis of high-frequency keywords shows that among the 30 major keywords, realism (245 times), romanticism (235 times) and modernism (221 times) constitute the main vein of literary style evolution, reflecting the development trajectory of Chinese literature from realism to romanticism to modernism. The keyword co-occurrence mapping reveals the continuity of the evolution of literary styles, and that literary genres do not develop in a ruptured manner, but rather influence each other and intermingle in a symbiotic manner.

The LDA theme model analysis identified 10 valid themes, of which literary genres and styles accounted for 19%, taking the first place, followed by creative techniques and narrative skills (16%). The analysis of theme evolution trend shows that creative techniques and narrative skills show a significant upward trend, reflecting the impact of the rise of Internet literature on creative forms; while the theme of psychological and human nature exploration declines significantly, accounting for only 13% of the total, indicating that the focus of literature has shifted from the individual's inner heart to socio-cultural issues. Semantic network analysis further confirms that there is a close connection between realism and postmodernism, and that postmodern literature challenges tradition while continuing and developing established styles. These findings not only validate the existing knowledge of literary history, but also provide a more objective and detailed picture of stylistic evolution through data analysis, enriching the methodological and epistemological basis of literary criticism.

References

[1] Lindsay, A. A. (2019). An Overview of Developments and Trends in Modern and Contemporary Chinese Literature. Emergence, 62.

[2] Ning, W., & Ross, C. (2016). Contemporary Chinese Fiction and world literature. MFS Modern Fiction Studies, 62(4), 579-589.
[3] Zhang, X., & Lin, J. (2021). Between modern and postmodern: Contemporary Chinese poetry from the outside in. Journal of Modern Literature, 44(2), 34-48.
[4] Wang, J., Li, H., & Bryant, S. (2017). The Distant Person in the Near Scene: Kubin's Study of Modern and Contemporary Chinese Literature. Chinese Literature Today, 6(2), 110-117.
[5] Chen, X., & Kubát, M. (2022). Rural versus urban fiction in contemporary Chinese literature-Quantitative approach case study. Digital Scholarship in the Humanities, 37(3), 681-692.
[6] Hladíková, K. (2021). In the name of stability: Literary censorship and self-censorship in contemporary China. In The Routledge Handbook of Chinese Studies. Taylor & Francis.
[7] Shen, Z. H. A. N. G. (2020). Traditional Genealogy and the Study of Modern and Contemporary Chinese Literature in the New Century. Journal of Southwest University Social Science Edition, 46(2), 133-142.
[8] Jiang, H. (2022). Modern and contemporary literature courses in colleges and universities using the teaching mode of deep learning. Mobile Information Systems, 2022(1), 3517022.
[9] Møller-Olsen, A. (2017). Fictional dictionaries: Power and philosophy of language in contemporary Chinese fiction. Modern Chinese Literature and Culture, 29(2), 66-108.
[10] Zhang, J., & Huang, Y. (2019). The Not So Soft Power of Chinese Literary Theory and Criticism: A Review of Literature and Literary Criticism in Contemporary China by Zhang Jiong. Style, 53(2), 257-264.
[11] Manqoush, R. A., & Al-Wadhaf, Y. H. (2021). Stylistics as a literary approach: A historical & critical analysis. Journal DOI, 7(1).
[12] Eder, M. (2017). Visualization in stylometry: cluster analysis using networks. Digital Scholarship in the Humanities, 32(1), 50-64.
[13] Widyahening, E. T., & Wardhani, N. E. (2016). Literary works and character education. International journal of language and literature, 4(1), 176-180.
[14] Seonghyun Park,Seungmin Oh & Woncheol Park. (2025). Automated Classification Model for Elementary Mathematics Diagnostic Assessment Data Based on TF-IDF and XGBoost. Applied Sciences,15(7),3764-3764.
[15] Qamar Muneer & Muhammad Asif Khan. (2025). Role of YouTube in creating awareness of sustainable transportation: A Latent Dirichlet Allocation approach. Sustainable Futures,9,100607-100607.
[16] Jun Dong,Haixiang Shang & Hang Yin. (2025). Semantic network analysis of spatial gene sequence in Dabaodao neighbourhood. Journal of Asian Architecture and Building Engineering,24(3),1999-2016.