International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-14

https://doi.org/10.70917/ijcisim-2026-0150
Article

Study on the Correlation between Tang Dynasty Literary Creation and Socio-Political Background Based on Big Data Analysis Techniques

Chi Zhang *

College of Liberal Arts, Harbin Normal University, Harbin, Heilongjiang, 150080, China; zhangcuk@163.com

Abstract: Every major change in society has a profound impact on literary creation. The rapid development of big data technology provides new ideas and methods for literary research. In this paper, LDA topic model and K-Means++ text clustering algorithm are used to mine and divide the keywords respectively. The data processing processes such as word division, stop word deletion, and text feature word extraction are performed to analyze the relevance of the text data. The study is based on the database of "Tang Dynasty Literature Chronicle Map Platform" for text mining, and the keywords mined can be roughly divided into six types: ideology and culture, social class, political events, literary style, literary themes, and literary style. Combined with the perplexity index can LDAvis visualization analysis, determine the number of themes is 2 when the difference between the themes is large. Based on Apriori algorithm for association rule analysis, Tang Dynasty literary creation has a strong correlation with political events, social class, ideology and culture.

Keywords: LDA; K-Means++; TF-IDF; Apriori algorithm; Tang Dynasty literary creation

1. Introduction

We have now entered the era of big data. In this era, every aspect of people's lives is undergoing significant changes [1]. Big data is constantly influencing our lifestyles and ways of thinking, and the development of literature will inevitably be affected by it [2]. Of course, Chinese literary creation is also undergoing changes in the context of big data. Today, big data is driving the development of Chinese literary creation and has achieved certain successes in contemporary literary research. However, further improvement is still needed.

During the early years of the Tang Dynasty, the recovery of the social economy was relatively slow. The instability of the political situation and the low efficiency of agricultural production restricted the rapid development of the economy, and this background deeply influenced the content and form of literary creation [3-4]. Poetry, as the most important literary form of the Tang Dynasty, was closely related to the social and economic conditions [5]. Literature [6] employs a combination of qualitative and quantitative methods to analyze the history of Tang Dynasty literature in China, focusing on the role of exchange poetry in constructing the poetic subject. Through social network analysis and detailed interpretation, it reveals patterns and conclusions regarding literary activities, connections, and mobility. Literature [7] analyzes the geographical distribution and evolution of the Tang Dynasty poetry world through data, examining poets' places of origin and their activities, which to some extent also reflects the social and political changes of the Tang Dynasty.

During the early Tang Dynasty, the content of poetry reflected the social landscape transitioning from turmoil to stability, with poets depicting the scars of war and the post-recovery landscape in their works [8-9]. The poetic style of this period also underwent a noticeable transformation. Early Tang poetry predominantly employed five-character lines, inheriting the concise style of the late Sui Dynasty in form, with themes often focusing on descriptions of natural landscapes and expressions of personal emotions,

reflecting the societal process of transitioning from chaos to the reconstruction of order [10-11]. For example, reference [12] explores how contemporary Chinese poets engage with classical Tang poetry and argues that this engagement has made Tang poetry a vibrant, evolving classical tradition. At the same time, poetry creation gradually spread from the court and nobility to broader social strata, reflecting the lives and emotions of ordinary people, marking the initial trend toward the popularization of poetry [13]. Literature [14] examines changes in how poets in the Tang Dynasty (late 8th to 9th centuries) depicted painting in their literary works, analyzing new techniques for representing the illusory aspects of painting, the relationship between painters and their works, and the connection between painting and poetry. Poets of this period, such as the Four Great Poets of the Early Tang Dynasty, not only reflected their personal feelings about the changes of the times in their poetry, but also attempted to explore the future direction of social development through poetry [15].

The gradual stabilization of the economy and the restoration of social order provided more material for poetry creation. Poets began to focus on themes such as the recovery of agriculture and the reconstruction of social order, which were widely expressed in the poetry of the time [16]. This not only enriched the themes of poetry but also promoted stylistic diversity, laying the foundation for the prosperity of poetry during the High Tang period [17]. By reflecting the multifaceted nature of early Tang society, these works demonstrated innovation in both artistic expression and social function, providing an important perspective for understanding early Tang society.

The Tang Dynasty is a golden period in the history of literature, and the socio-political environment of this period is complex and changeable, so it is of great significance to explore the correlation between literary creation and socio-political background of the Tang Dynasty. In this paper, we first construct the LDA theme model and mine the keywords involving Tang Dynasty literature and socio-political background from the database of the Tang Dynasty Literature Chronology Map Platform. Then K-Means++ text clustering algorithm is applied to classify the keyword types so as to better classify the content of the text. After that, using the TF-IDF algorithm, the document-keyword weight matrix is established, and the number of topics is determined by combining the perplexity index can be LDAvis visualized and analyzed. Finally, based on Apriori algorithm, the potential connection between Tang Dynasty literary creation and socio-political background is mined.

2. Big Data Analytics

2.1. Text Mining

With the exponential growth of data information, data mining technology came into being, which refers to a set of tools and methods for refining patterns and summarizing knowledge from observed data to be used to guide work and practice [18].

The most important goal of text mining is to use Natural Language Processing (NLP) technology to convert text into computer-recognizable data, and then to find valuable information from the data, so as to discover or solve some practical problems in reality.

2.2. LDA Subject Modeling

LDA is an unsupervised learning document topic generation model, also known as a three-level Bayesian probabilistic model containing words, topics, and documents. Documents are probability distributions of topics, while topics are probability distributions of words.The purpose of LDA is to identify the hidden topic information in all documents and to turn the document-word matrix into a topic-word matrix and a document-topic matrix [19]. The LDA model can be visualized by the following probabilistic formula:

$$P(word \mid document) = P(word \mid topic) \times P(topic \mid document) \tag{1}$$

2.2.1. Principles of Generation

The LDA generation model is shown in Figure 1 with the following three assumptions:

(a) There are a total of D documents, a total of K topics, and the number of all words in the dictionary is V.

(b) The prior distribution of document topics is the Dirichlet distribution, i.e., the topic distribution θ_d of any document d satisfies $\theta_d = Dirichlet(\vec{\alpha})$, α is a K-dimensional vector that is the prior parameter of the Dirichlet distribution.

(c) The prior distribution of words in a topic is a Dirichlet distribution, i.e., the topic distribution β_k

for any topic k satisfies $\beta_k = Dirichlet(\vec{\eta})$, and η is the V-dimensional vector that is the prior parameter of the Dirichlet distribution.

For the n th word in any document d, the distribution of topic numbers $Z_{d,n}$ can be drawn from the topic distribution θ_d:

$$Z_{d,n} = multi(\theta_d) \tag{2}$$

And for that topic number, the probability distribution of the word $W_{d,n}$ is obtained:

$$W_{d,n} = multi(\beta_{Z_{d,n}}) \tag{3}$$

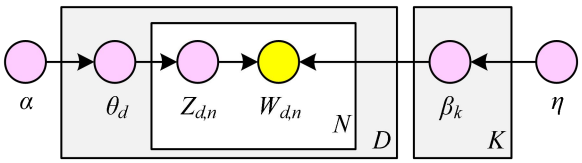


Figure 1. LDA generation model.

The next thing to be solved is the distribution of topics $Z_{d,n}$ for each document and the distribution of words $W_{d,n}$ for each topic, which are generally sampled using Gibbs' algorithm.

2.2.2. Gibbs Sampling

As can be seen from the previous section, α and η are known a priori parameters, and the goal is to obtain the document topic distribution θ_d and the topic word distribution β_k. To perform Gibbs sampling, it is first necessary to obtain the conditional probability distribution of each feature dimension of the corresponding distribution. Currently, the word vectors w are known, while the distribution of topic z is unknown. It is necessary to find the joint distribution $p(w,z)$ of w, z, and then the conditional probability distribution $p(z_i = k \mid w, z_{-i})$ of the topic feature z_i corresponding to a word w_i, where z_{-i} denotes the distribution of topics after removing the words with subscript i. After obtaining the conditional probability distribution, Gibbs sampling can be performed, and its algorithm flow is as follows:

(1) Choose a suitable number of topics K, hyperparameter vectors α, η.

(2) Assign a random topic number z to each word of each document in D.

(3) Rescan D and for each word, update the topic number of the word using Gibbs sampling and update the number of the word in D.

(4) Repeat (3) until Gibbs sampling converges.

(5) Count the topics of each word of each document in D to get the document topic distribution θ_d, and count the topics of all words in D to get the word distribution of each topic of LDA β_k.

2.3. Text Clustering

Text clustering is an automatic process of dividing the set of documents into clusters in the presence of unknown document categories, such that documents within clusters have high similarity but documents in different clusters have high dissimilarity. Similarity and dissimilarity are usually measured by the distance between documents.

2.3.1. K-Means++ Clustering

K-Means++ is an improvement of K-Means clustering algorithm [20]. To establish the K-Means++ cluster analysis model, the main steps are as follows.

(1) Determine N samples and M features of the corresponding samples, and construct the matrix of $N \times M$.

(2) Select the best number of clusters K, which can be determined by contour coefficient method, elbow method or according to the actual situation.

(3) Randomly select K samples far away from each other as the initial clustering centers.

(4) Enter the loop, calculate the distance between each sample point and the clustering center, assign each sample point to the nearest clustering center, generate K clusters, and take the average of all sample points in each cluster as the new clustering center.

(5) When the position of the clustering center no longer changes, stop the iteration and the clustering is complete.

(6) Plot the visualization of the K clusters and analyze.

2.3.2. Distance Metrics

The distance from each sample point to the cluster center in the K-Means++ text clustering model can be measured using the Euclidean distance d, expressed as:

$$d(X,\mu)=\sqrt{\sum_{i=1}^{n}\left(X_i-\mu_i\right)^2} \tag{4}$$

Where X denotes a sample point in a cluster, μ denotes the cluster center of the cluster, n denotes the number of features, and i denotes each of the features that make up the sample point X. X_i denotes the i th feature of the sample point X. μ_i denotes the i th feature of the clustering center of the cluster.

2.4. Analysis of Association Rules

Association rules are used to find, mine and describe the correlations that exist between data. The related concepts of association rules are as follows:

(a) Items and item sets: if $I=\{i_1,i_2,\ldots,i_m\}$ is a set of m distinct items, then each $i_k\left(k=1,2,\ldots,m\right)$ in it is an item and I is the the set of terms. In this paper, each job posting can be regarded as a term set, and the workplace is a term.

(b) Transaction and Transaction Set: Transaction T is a subset of item set I, and the whole of a transaction is called a transaction set.

(c) Association rule: i.e. $A\Rightarrow B$, where both A and B belong to the itemset I and A does not intersect with B.

(d) Support: the probability that the terms contained in A and B occur together in the transaction set:

$$\sup port(A\Rightarrow B)=P(A\cup B) \tag{5}$$

(e) Confidence: the probability of simultaneous occurrence of B in a transaction if the transaction contains A:

$$\begin{aligned} confidence\left(A\Rightarrow B\right)&=\frac{\sup port\ (A\Rightarrow B)}{\sup port(A)}\\ &=\frac{P(A\cup B)}{P(A)}=P(B\mid A)\end{aligned} \tag{6}$$

(f) Enhancement: The ratio of confidence level to support level, which represents the extent to which the presence of A affects B, with an enhancement level of more than 1 to be of practical significance, indicating that A has a positive influence on B:

$$lift(A\Rightarrow B)=\frac{confidence(A\Rightarrow B)}{\sup port(A)} \tag{7}$$

(g) Frequent itemsets: If the support of itemset I is greater than or equal to the minimum support threshold, then I is a frequent itemset.

(h) Strong association rules: association rules that are greater than or equal to the minimum support and minimum confidence, i.e. strong association rules to be mined.

When mining association rules, it is necessary to find out all frequent itemsets before generating

strong association rules mining process is shown below:

(a) Mining frequent itemsets

Apriori algorithm is a typical association rule algorithm for mining frequent itemsets, and the most crucial steps are joining and cropping [21]. Joining refers to the use of $k-1$ frequent itemsets to obtain $k-$ candidate itemsets by joining, and only itemsets differing by one item can be joined, such as $\{A,C\}$ and $\{C,B\}$ are joined to become $\{A,C,B\}$. Pruning is done based on a property: for a $k-$ term set, if one of its subsets is not frequent, then it cannot be frequent itself, based on this property, candidate sets can be pruned by judging the prior property.

(b) Generating association rules

After connecting and pruning, i.e., all the frequent itemsets are found, on the basis of which strong association rules can be generated. The steps are as follows:

(a) For each frequent itemset I, generate all non-empty subsets of I.

b) For each non-empty subset x of I, compute $confidence\left(x \Rightarrow (I-x)\right)$, if $confidence\left(x \Rightarrow (I-x)\right)$ is greater than the minimum confidence threshold, then the rule $x \Rightarrow (I-x)$ holds.

3. Data-Processing Methods

Most of the data in real-world scenarios cannot be used directly for data mining because they tend to be incomplete, cluttered, high-latitude, missing, and very noisy. Using these data directly to perform data mining operations usually produces very poor results. Data preprocessing techniques refer to a series of standardized operations on the collected data for the purpose of improving the quality of data mining, thus greatly improving the effectiveness of data mining.

3.1. Text Segmentation

The first step in preprocessing is word splitting, which is simply the process of breaking down a long, whole paragraph of text into separate words and restoring it to the initial form of the document. A complete text consists of a combination of words and strings. Because a single word can only convey very little information, while the document usually consists of a large number of words, and some words in Chinese have many different meanings, and one meaning corresponds to a dimension, then if such words are used as the standard of measurement in the process of feature selection, then it will lead to the problem of dimensionality catastrophe accordingly. Although phrases provide more information than words, a phrase often consists of more than one character, and a phrase often occurs less often, so it is a bit unrealistic to use phrases to replace text, so in the process of selecting the participle, the word is usually chosen as the feature item.

3.2. Stop Word Deletion

Stop words don't have any practical use value for the text, so they are usually deactivated and deleted. Stop words are usually some, interjections can also be some mood prepositions, adverbs, mood particles and other common "of", "to", "ah", etc., so it can be seen that they do not actually exist in the sentence, and can only be used to express the corresponding meaning and exclamation in the specific context. Such words will inevitably appear in many articles, and if these words appear more, the less information the article itself can convey. Therefore, it is necessary to delete these stop words in the article to ensure that the article has enough valid information. In addition, punctuation marks need to be stripped of key objects. After the word segmentation is completed, these words are deleted to remove miscellaneous information and improve the reliability of the analysis. The definition of the stop synonym dictionary contains words such as "some, I, again, once, before, in one word". According to different requirements, words can be added or deleted from the stop thesaurus table to achieve the expected word segmentation effect.

3.3. Text Feature Word Extraction

In the process of representing the text, if the generated disambiguation, that is, the molecule generated after the text processing is used for the representation, then it is obvious that it will make the dimension of the text vectors too large, which is very complicated for the actual operation, and also affects the clustering operation. So in order to reduce the amount of operation, it is necessary to use the method of compression processing for the operation process of text words to reduce the dimension of text features.

In this paper, document frequency algorithm is used to extract text feature words. Document frequency (DF) refers to the frequency of a feature word in a document set, that is, in how many documents such a word occurs. For a feature word, if it has a high document frequency, then we understand that the word is less capable of recognizing the document in which it is found, and therefore the weight of the word selected as a feature term should be reduced. It can be found that this method, in operation, is simpler and consumes less time accordingly. However, this method also has the problem that if there is still a corresponding feature item in only one text category, it will be easily filtered out, resulting in a document frequency of 0.

3.4. Text Representation Model

3.4.1. Vector Space Model

Although the text has been pre-processed, but it is also essentially belongs to the category of text, the computer in the process of processing text objects, the text needs to be converted first. In the research process of this paper, first of all, through the establishment of a suitable model, the text is manipulated so as to obtain a variety of different structured data models, the commonly used models are bag-of-words model, word vector model, vector space model, topic model and so on. This paper analyzes and finally applies the vector space model (VSM), which can also be expressed as bag-of-words model.

In Vector Space Model, the document is represented by vectors, so the feature terms and weights constitute the Vector Space Model, which can process the text data in a simpler and faster way.

The vector space model needs to define a dictionary, which is a set of feature words of the text sample. The dictionary can be imported by defining itself according to the demand, or it can be generated in the sample set. So in the process of representing the text, you can use the dictionary, the length of the dictionary as a benchmark, so that the initialization of the vector operation, you can get the dictionary in the corresponding feature word position, for all the words in the text traversal operation. If the word to be found appears in some process, a certain value needs to be entered into the corresponding position, and accordingly a certain value can be assigned to the corresponding position and the corresponding feature weight. Word frequency-inverse document frequency weights are two operations commonly used in the process of feature word assignment methods.

3.4.2. TF-IDF Weights

TF-IDF essentially belongs to a weighting algorithm, which can be used for information retrieval as well as text mining processing, and the importance of words can be measured effectively. If a word has a higher probability of appearing in a document, the more important it is accordingly. Nevertheless, its importance decreases with its frequency of occurrence in the corpus or its frequency of representation in the document set, i.e., the frequency of document occurrence.

Word frequency in TF-IDF is represented by TF and inverse document frequency is represented by IDF.

Word frequency is the frequency or number of times a particular word appears in a particular document:

$$tf_{i,j} = \frac{t_{i,j}}{\sum_k t_{k,j}} \tag{8}$$

The $tf_{i,j}$ in Eq. (8) denotes the magnitude of the total frequency of occurrence of the feature word i in the document j.

The inverse document frequency also provides a measure of how much a word contributes or is important to the whole document. This metric measures the feature word in terms of its general degree of importance. Its exact formula is as follows:

$$idf_j = \log \frac{|D|}{\left|\left\{ j : t_i \in d_j \right\}\right|} \tag{9}$$

Using “Tang Poetry” as an example, if the total number of documents is 1000, and the number of documents in which “Tang Poetry” appears is 100, the corresponding IDF result is $230\left(\log \frac{1000}{100} \right)$:

$$tf-idf_j = tf_{i,j} \times idf_j = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{\left|\left\{ j : t_i \in d_j \right\}\right|} \tag{10}$$

4. Analysis of the Relevance of Tang Dynasty Literary Creation to the Socio-Political Context

This chapter obtains the basic information of the poets, the time and place of creation and other data from the database of the Tang Dynasty Literature Chronological Map Platform. The text preprocessing is carried out according to the method proposed above, after which the co-occurrence matrix is constructed to determine the theme of the text, and finally the association rules between Tang Dynasty literary creation and socio-political background are mined based on the Apriori algorithm.

4.1. Co-Word Analysis

4.1.1. Co-Occurrence Matrix

The jieba method is used to segment sentences, and the frequency of key words is counted according to the word segmentation results. Some of the keywords extracted from the Tang Dynasty Literary Chronicle Map Platform are shown in Table 1. It can be seen that the literature of the Tang Dynasty mainly revolves around "poetry", among which the more prominent types of poetry include "Biansai" poems, "palace resentment" poems, "landscape" poems, etc. In addition, it also includes other keywords with high frequency such as "prose, graceful, bold, imperial examination, Anshi Rebellion, legend, Xuanwumen change, Confucianism, Buddhism and Taoism, scholars, Shu landlords, and ordinary people".

Table 1. Some of the keywords extracted.

N	Key words	Word frequency	N	Key words	Word frequency
1	Poetry	1152	9	An Lushan Rebellion	241
2	Edge plug	963	10	Legend	217
3	Palace	852	11	The Xuanwu Gate Incident	193
4	Landscape	655	12	Confucianism	152
5	Prose	496	13	Buddhist thought	106
6	Euphemism	411	14	Sclerics	99
7	Trove	362	15	Landowner	86
8	Imperial examination	355	16	Common people	73

The partial keyword co-occurrence matrix is shown in Figure 2. According to the contextual co-occurrence relationship, the co-occurrence matrix is generated, in which the data in the upper or lower triangular cell is the number of times two keywords appear in the same document at the same time. For example, the co-occurrence frequency of “poetry” and “border fortress” is 263, which means that these two keywords appear 263 times in the same document. The more times the two keywords appear together in the co-occurrence matrix, the closer the connection between the two keywords.

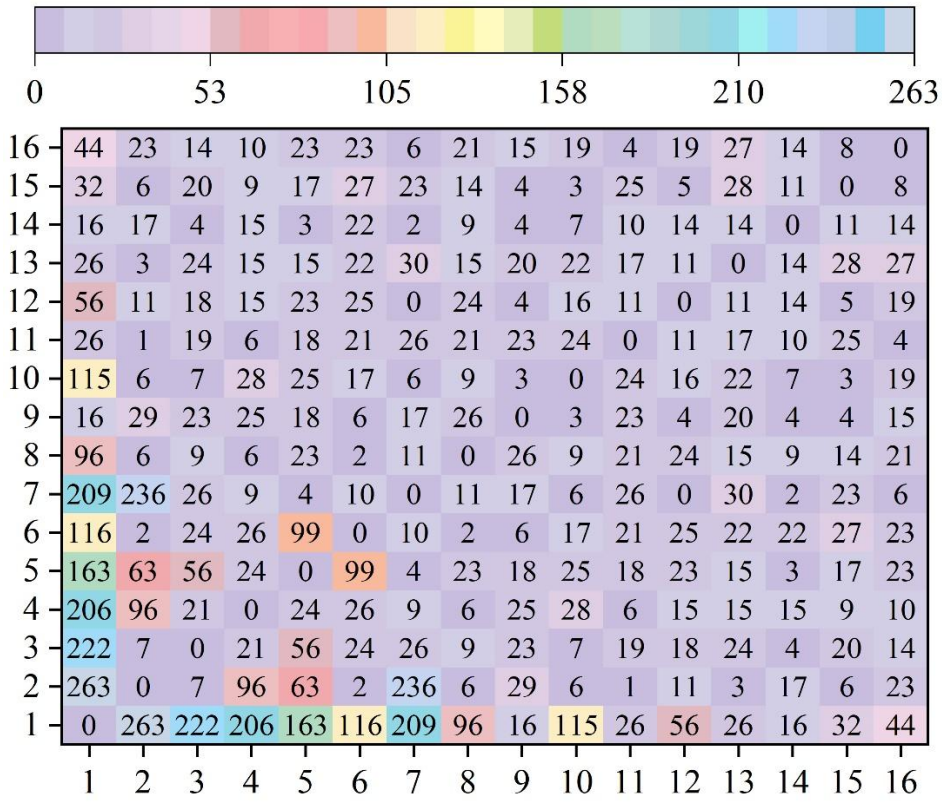


Figure 2. Some key words common matrix.

4.1.2. Cluster analysis

In this paper, K-Means++ clustering is used, and the co-occurrence matrix obtained is imported into SPSS software, which can initially determine the degree of association between each keyword.

The clustering results are shown in Figure 3. The above keywords can be clustered into six types: ideology and culture, social class, political events, literary style, literary themes, and literary genres.

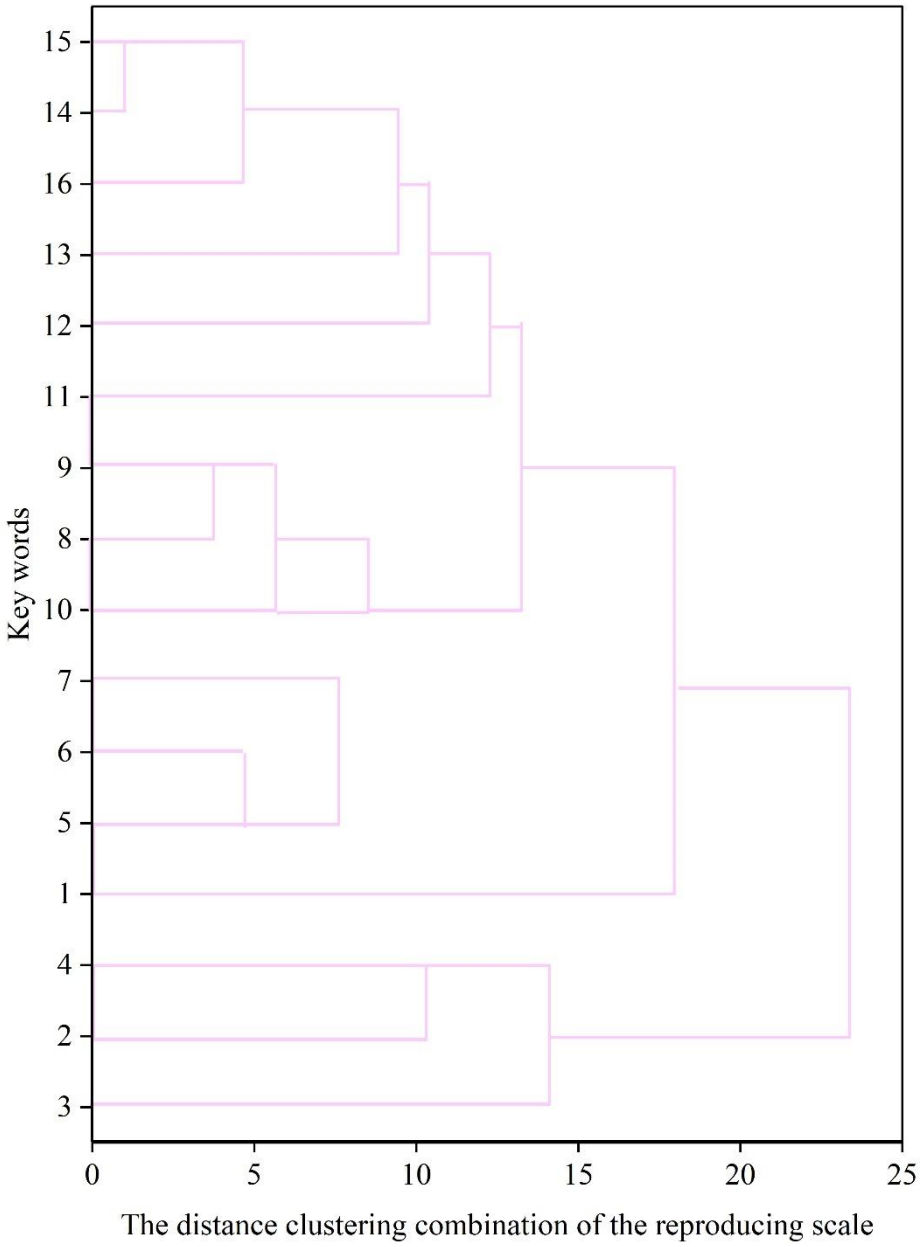


Figure 3. Cluster result.

4.2. Mining results and analysis

4.2.1. Keyword weighting matrix

Using the TF-IDF algorithm, the document-keyword weight matrix is established, and part of the keyword weight matrix is shown in Table 2. The weights in the matrix represent the degree of correlation between the keywords in the Tang literature and the socio-political context, and the higher the weight, the higher the degree of correlation. For example, the weight of "frontier" is high, reaching 0.83, which indicates that the creation of Tang Dynasty literature included a large number of frontier poems, which was precisely connected with the military expansion and frontier defense policy of the Tang Dynasty. For example, the frequent border wars during the reign of Emperor Xuanzong of the Tang Dynasty led to the emergence of a large number of frontier poems, which allowed people to express their views on the wars through poetry, including praising the soldiers and exposing the cruelty of the wars, which were related to the political and military decisions of the time.

Table 2. Keyword weight matrix.

Key words	Poetry	Edge plug	Palace	Landscape	Prose	Euphemism	Trove	……
1	0	0.83	0	0	0.01	0	0	
2	0.63	0	0	0	0	0	0	
3	0	0	0.55	0	0.32	0	0	
4	0	0.11	0	0	0.61	0	0	
5	0	0.59	0	0	0.35	0	0	……
6	0	0	0	0.22	0	0	0	
7	0	0	0	0	0	0.21	0	
	0.11	0	0	0	0	0	0.63	
……	……							

4.2.2. Thematic Results and Analysis

Customizing the number of topics is an important feature of LDA topic model, however, considering the number of texts is too large, it is difficult to determine the number of topics autonomously by human beings to achieve the optimal effect, and there is a certain degree of subjectivity. Therefore, this paper introduces the confusion degree indicator to determine the number of topics, the smaller the confusion degree indicator indicates that the better the modeling ability, the number of topics is optimal, the formula is as follows:

$$perplexity = \exp\left\{-\frac{\sum_m \log\left(P\left(w_m\right)\right)}{\sum_m N_m}\right\} \tag{11}$$

Where: $perplexity$ denotes the perplexity degree, m denotes the m th document, $P\left(w_m\right)$ denotes the probability of each keyword in the m th document, and N_m denotes the total number of lexical items in the m th document. The confusion metric curve is shown in Figure 4. Where the horizontal coordinate is the number of topics and the vertical coordinate is the perplexity, the figure shows that the perplexity metric curve is at its minimum when the number of topics is 2, followed by the perplexity metric value when the number of topics is 3.

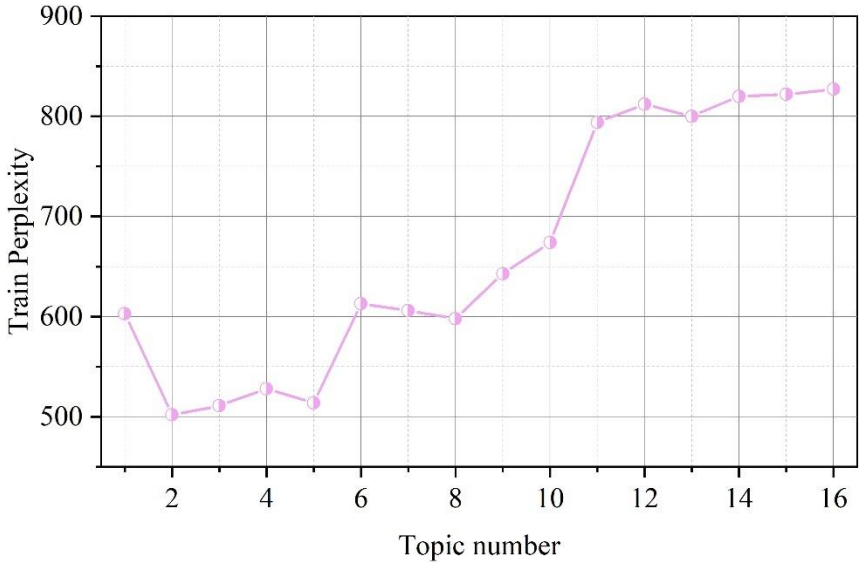
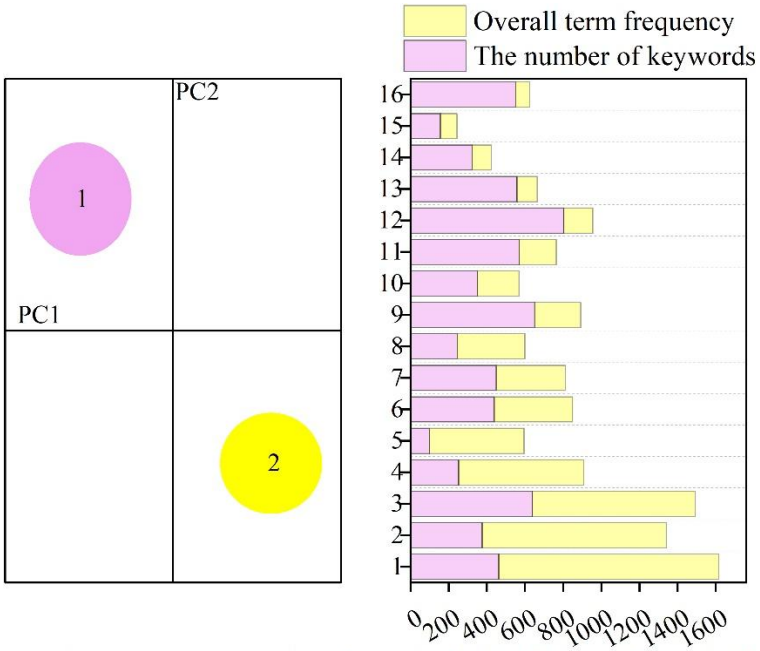


Figure 4. Index curve.

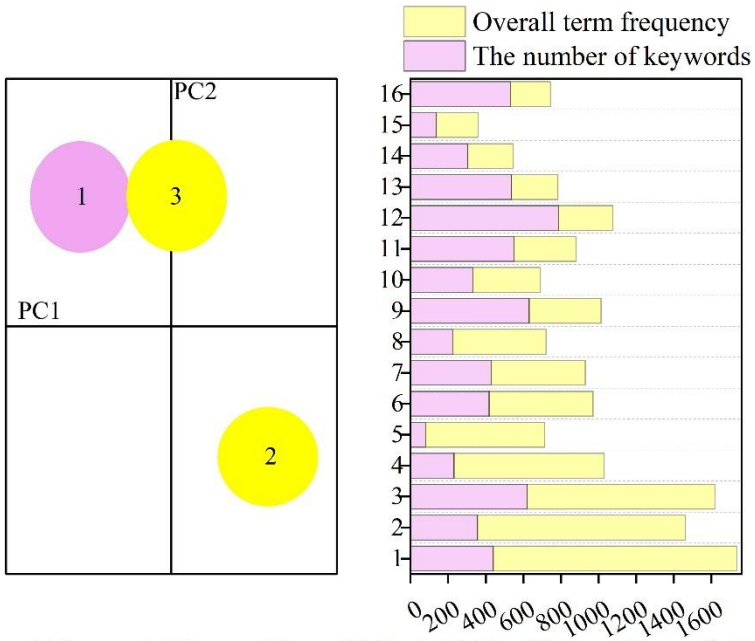
LDAvis displays the core technology topics in a dynamic and interactive graph, where the size of the

circle can represent the probability of the topic in the document, and the distance between two circles represents the similarity of the topics. The corresponding topic models when the number of topics is 2 and when the number of topics is 3 are visualized by LDAvis respectively, and the plotting results are shown in Fig. 5 and Fig. 6, (a) and (b) represent the distance map between topics and the most relevant keywords of topics 1 and 2 respectively. When the number of topics is 2, the two circles are similar in area and basically uniformly distributed at the two ends of the coordinate system, and the distance is relatively far away, so the similarity of the topics is the lowest, therefore, the highest degree of differentiation between the topics when the number of topics is indicates that the LDA topic model is more effective. On the contrary, when the number of themes is 3, the distance between theme 1 and theme 3 is very close and partially overlapped, indicating that these two themes are more similar, so the degree of differentiation is not good. In summary, combining the consistency score, the results of theme model visualization and the actual situation, it is considered that the theme model with 2 themes is selected to be the most effective.



(a) Intertopic Distance Map (b) Topic-16 Most Relevant Terms for Topic 1

Figure 5. The topic number is a two-time theme model visualization.



(a) Intertopic Distance Map (b) Topic-16 Most Relevant Terms for Topic 2

Figure 6. The topic number is a three-time theme model visualization.

4.3. Association Rule Mining

In this paper, the support degree is set to 0.002 and the confidence degree is set to 0.3, and the partial output of the association rule obtained is shown in Table 3. According to the principle of association rules, the support degree indicates the probability that the two feature item transactions before and after appear simultaneously in the total feature item transaction set. Confidence can be understood as the probability of the occurrence of the following word after the occurrence of the preceding word. A boost greater than 1 indicates a valid strong association rule.

Table 3. The correlation rule is partial output.

Rules	Support	Confidence	Lift
Edge plug=>An Lushan Rebellion	0.41%	82.0%	49.324
Palace=>The Xuanwu Gate Incident	0.32%	93.7%	10.473
Landscape=>Prose	0.31%	72.2%	29.112
Euphemism=>Prose	0.54%	98.7%	11.114
Confucianism=>Ideological culture	0.59%	91.5%	10.266
Buddhist thought=>Ideological culture	0.59%	92.3%	10.767
Sclerics=>Social class	0.42%	65.7%	16.048
Landowner=>Social class	0.66%	87.8%	15.966
Common people=>Social class	0.35%	75.5%	59.679

Analyzing the association rules generated above, it can be seen that:

Rule 1: If keywords such as "Biansai and Palace Resentment" appear in a literary work, it is very likely that the work was created during or after the "Anshi Rebellion and the Xuanwumen Change".

Rule 2: If keywords such as "scholars, Shu landlords, and common people" appear in a literary work, the work may be related to "social class".

Rule 3: If keywords such as "Confucianism, Buddhism and Taoism" appear in a literary work, the work may be related to the "ideology and culture" of the Tang Dynasty.

5. Conclusion

This study is based on the database of Tang Dynasty Literature Chronological Map Platform to reveal the close connection between Tang Dynasty literary creation and socio-political background.

(1) The extracted keywords are clustered using the K-Means++ algorithm, and a total of six types of ideology and culture, social class, political events, literary style, literary themes, and literary genres are classified.

(2) Using the TF-IDF algorithm, the document-keyword weight matrix is established, in which the weight of "Border Plugs" is higher, reaching 0.83, indicating that the creation of Tang Dynasty literature contains a large number of border plugs poems.

(3) The TF-IDF model is used to extract the themes of the article, and the text is divided into two theme types: literary creation and socio-political background. Among them, the literary works have a close connection with political events, social classes, and ideology and culture.

Future research can further expand the data sources, optimize the analysis model, and dig deeper into more details and laws between literary creation and socio-political background, so as to provide comprehensive support for literary research.

References

1. Cuiñas, A. G. (2023). Literature Seen Through Big Data and Artificial Intelligence: Key Concepts and Critical Challenges. Humanidades Digitales y Big Data en Iberoamérica Digital Humanities and Big Data in Ibero-America, 25.
2. Zhang, L., & Luo, X. (2021, May). The creation of literature communication science based on big data of internet of things and the study of ancient chinese literature communication. In Journal of Physics: Conference Series (Vol. 1915, No. 2, p. 022086). IOP Publishing.
3. Zorkina, M. (2023). Daoist Immortals as a Poetic Image in the Tang Dynasty: A Corpus Study. Digital Humanities and Religions in Asia: An Introduction, 3, 177.
4. Xia, C. (2021). Poetry and emotion in classical Chinese literature. In The Routledge handbook of Chinese studies (pp. 289-303). Routledge.
5. Ding, F. (2022). The Making of Classics: Li Bai and Du Fu's Poems in Anthologies of Tang Poetry between the Tang and the Ming Dynasties. Journal of chinese humanities, 8(2), 163-188.
6. Mazanec, T. J. (2018). Networks of exchange poetry in late medieval china: Notes toward a dynamic history of tang literature. Journal of Chinese Literature and Culture, 5(2), 322-359.
7. Zhaopeng, W., & Junjun, Q. (2018). Geographic distribution and change in Tang poetry: data analysis from the "Chronological Map of Tang-Song Literature". Journal of Chinese Literature and Culture, 5(2), 360-374.
8. Liu, S. (2022). Poetry under Imperial Order and Seven-Character Metrical Poetry in the Early Tang Dynasty. Theoretical Studies in Literature and Art, 42(3), 174-184.
9. Wu, S. (2010). The development of poetry helped by ancient postal service in the tang dynasty. Frontiers of Literary Studies in China, 4, 553-577.
10. Chen, L., & Chaetnalao, A. (2023). Analysis on "The Spirit of the Prosperous Tang Dynasty" and "Picturesense" of Frontier Fortress Poetry in the Prosperous Tang Dynasty. Journal of Community Development Research (Humanities and Social Sciences), 16(3), 1-13.
11. Xu, J. (2024). The Historical Conceptualization of the Three Changing Trends in Tang-Dynasty Prose and the Evolution of Literature in Different Dynasties. Theoretical Studies in Literature and Art, 43(5), 31-42.
12. Klein, L. (2021). What Does Tang Poetry Mean to Contemporary Chinese Writers? Li Bai and the Canonicity of Tang Poetry in Liu Liduo, Ha Jin, Yi Sha, and Xi Chuan. Prism: Theory and Modern Chinese Literature, 18(1), 138-169.
13. Ning, L. (2023). Literature in the Late Years of the Tang Dynasty and the Five Dynasties. In Concise Reader of Chinese Literature History (pp. 237-259). Singapore: Springer Nature Singapore.
14. Egan, R. (2021). Poems on Painting from the High Tang to Later Tang Periods. Early Medieval China, 27(1), 19-44.
15. Cui, F. (2014, August). A study on Universal Values of Literati in Tang Dynasty–Take Idylls as an Example. In 2014 2nd International Conference on Education Technology and Information System (ICETIS 2014) (pp. 92-95). Atlantis Press.
16. Zorkina, M. (2018). Describing Objects in Tang Dynasty Poetic Language: A Study Based on Word Embeddings. Journal of Chinese Literature and Culture, 5(2), 250-275.
17. Tian, T. (2024). The relationship between Tang-Song poetry and Zen Buddhism thought. Trans/Form/Ação, 47(4), e0240064.

18. Lucia Trapanese,Francesca Petrocchi Jasinski,Giovanna Bifulco,Nicola Pasquino,Umberto Bernabucci & Angela Salzano. (2024). Buffalo welfare: a literature review from 1992 to 2023 with a text mining and topic analysis approach. Italian Journal of Animal Science,23(1),570-584.
19. Ozcan Ozyurt,Hakan Özköse & Ahmet Ayaz. (2024). Evaluating the latest trends of Industry 4.0 based on LDA topic model. The Journal of Supercomputing,80(13),19003-19030.
20. Youdong Yuan,Ping Yang,Hanbing Jiang & Tiange Shi. (2024). A Multi-Robot Task Allocation Method Based on the Synergy of the K-Means++ Algorithm and the Particle Swarm Algorithm. Biomimetics,9(11),694-694.
21. KeyiShen,YeTian,BisongHu,JinLuo,ShuhuaQi,SongliChen & HuiLin. (2024). Association rule mining of air quality through an improved Apriori algorithm: A case study in 244 Chinese cities. Transactions in GIS,28(4),726-745.