A Recognition Method of Measuring Literature Topic Evolution Paths Based on K-means-NMF

Wenbo Cui*, Jinling Li**, Tao Zhang***, and Sibo Zhang****

School of Information Management, Heilongjiang University, harbin 150080, China

*<2212197@s.hlju.edu.cn> / **<2212201@s.hlju.edu.cn> /

<zhangtao@hlju.edu.cn> / * <1152579804@qq.com>

Wenbo Cui is a graduate student at Heilongjiang University. He received his Bachelor's Degree in Engineering Management from Xinyang Normal University. His research interests include user modeling, text mining, data analysis, and knowledge organization.

Jinling Li is a graduate student at Heilongjiang University. She received her Bachelor's Degree in Engineering Management from Shandong Normal University. Her research interests include information measurement and evaluation.

Tao Zhang is Associate Professor of School of Information Management, Heilongjiang University, CCF Senior Member. His research interests are text mining, natural language processing, data and algorithm security governance etc. He has presided over more than 10 important projects such as the National Social Science Fund, etc. He has published approximately 90 articles in refereed journals and conference proceedings including Symmetry, JACIII, IEIB etc. He is a reviewer of Library and Information Service, Journal of Library and Information Science in Agriculture, Science-Technology and Management etc.

Sibo Zhang is an undergraduate student at Heilongjiang University. His research interests include natural language processing and text mining.

Cui, Wenbo, Li Jinling, Tao Zhang and Sibo Zhang. 2023."A Recognition Method of Measuring Literature Topic Evolution Paths Based on K-means-NMF". *Knowledge Organization* 50, no. 4: 257-271. 27 references. DOI:10.5771/0943-7444-2023-4-257.

Abstract: In this study, we propose a recognition method of measuring literature topic evolution paths based on K-means-NMF in order to address problems such as the unobvious effect of topic clustering, high degree of mixing in clustering results, and unclear topic evolution paths that exist in the current research of topic evolution analysis. Firstly, we enhance the traditional NMF (Nonnegative Matrix Factorization) topic model by combining the K-means clustering algorithm with the NMF model to improve the accuracy of topic clustering and reduce the correlation among topics. Secondly, we perform the topic co-occurrence analysis based on the clustering results to identify important topic categories for recognizing critical evolution paths to solve the problem of multiple possible evolution paths in the experiment. Thirdly, we adopt the Word2Vec model to calculate topic word vectors in a semantic context to improve the accuracy of the correlation strength between topics at adjacent stages. Finally, we adopt the above method to conduct an empirical study using intelligent algorithms as an example. The experimental results show that this research method effectively identifies important topics and topic developments in the subject area, which can support scientific research and science and technology policy development.

Received: 21 Nov 2022; Revised: 16 May 2023; Accepted 01 June 2023

Keywords: K-means, nonnegative matrix factorization, topic models, recognition method

1.0 Introduction

Disciplinary fields undergo a process of dynamic change; namely, researchers must quickly grasp the hot spots of disciplinary development to find the breakthrough of innovation and promote disciplinary action. Analyzing the topic evolution of the disciplinary field is meant to reveal its law and improve the policy system in science and technology. From the structure perspective, topic evolution analysis belongs to microstructure, meaning that it usually represents the research direction of a particular field. From the content perspective, topics are collections of domain knowledge units, meaning that topics are usually implicit semantic structures (Jiang et al. 2022). Topic evolution in different fields contains different background factors of the times, such as the degree of technological development, the way data volume grows, and policy support. Indeed, it is necessary to obtain stage topics quickly and efficiently from massive information through topic extraction, then obtain the relations between topics at adjacent stages through topic evolution analysis. Overall, the evolution analysis of the disciplinary field has generally become a topic of concern in academia, and the innovation of the research methods has also received widespread attention from scholars.

Currently, LDA (Latent Dirichlet Allocation) and NMF (Nonnegative Matrix Factorization) topic models fusing temporal dimensions are typically represented in the development of topic evolution analysis (Lee and Sebastian 1999; Blei et al. 2003; Blei 2012). Specifically, potential topics in the initial document are extracted using LDA or NMF. Then, the extracted topics are vectorized by adopting the Word2Vec model. Finally, the temporal dimension is introduced to research the change process of topics over time. Although the above research method is gradually becoming a research paradigm, it still needs to address the following two problems: 1) Regarding topic extraction: LDA works well with long-text datasets while poorly with short-text datasets. In contrast, NMF works well with short-text datasets and can extract niche topics often underestimated in traditional LDA (Chen et al. 2019). However, NMF may have multiple possible decompositions, resulting in inefficient processing when the dataset is too large. Besides, the excessive number of topics due to multiple decompositions may lead to poorer fitting results, which are less robust than the LDA fits (O'Callaghan et al. 2015); 2) Regarding the evolution topic: there are multiple possible paths.

As a result, to solve these two problems, this study proposes a recognition method of measuring literature topic evolution paths based on K-means-NMF. To address problem number 1, we enhance the traditional NMF model by combining it with the K-means clustering algorithm to obtain the K-means-NMF model to improve the accuracy of topic clustering and reduce the correlation among topics. To address problem number 2, we perform the topic co-occurrence analysis based on the clustering results to identify important topic categories. Then the Word2Vec model is adopted to calculate topic word vectors in the semantic context, tracking the process of topic change in the time dimension. This method is applied to discovering topics in intelligent algorithms and identifying their evolution paths. Finally, we visualize and analyze the process of topic change to reveal the topics in different periods and topic evolution paths of the field, helping scholars understand the whole picture, evolution lineage, and research focus of their disciplinary field.

2.0 Related work

In this section, we reviewed two aspects of the methods of recognizing the topic evolution paths of literature. Firstly, we reviewed the definition of the topic model and introduced the related research of two important topic models, LDA and NMF. Afterward, we reviewed some research related to the development of methods used for topic evolution from machine learning to deep learning.

2.1 Topic extraction model

Text feature extraction based on topic models is the basis of topic evolution analysis. The topic model is a statistical model used to automatically compute and discover topics in a series of documents. Specifically, the topic model extracts a low-dimensional set of polynomial distribution from a collection of texts through parameter estimation to capture the relevant information between words and then collects, classifies, and reduces the dimension of the text to find the hidden semantic patterns in the set of unstructured documents.

At present, the research of topic models can be divided into two types: topic models based on matrix decomposition (e.g., Latent Semantic Indexing-LSI and NMF) and topic models based on probability (e.g., Probabilistic Latent Semantic Indexing-PLSI and LDA). Topic models are usually typified by LDA and NMF: LDA is a probabilistic model, while NMF is a matrix decomposition and multivariate analysis technique. Topic models can describe the content of information resources in an unsupervised manner using topics and topic words as units, facilitating the refinement of analysis units from documents to topic words. As a result, topic models are widely used in knowledge organization research (Macfarlane et al. 2020). Knowledge organization focuses on the semantic information in the text and emphasizes the associative relationships between knowledge (Zhang and Zhang 2022). Topic models can construct associative relationships between documents and feature words by mining the implicit topic features to facilitate knowledge

inference and are widely used in scholars' research of topic evolution.

Ambrosinod et al. (2018) applied LDA to the full text of economics articles stored in the JSTOR database to reveal the evolution of economics, illustrating the potential of LDA to study structural changes in economics in times of fragmentation. Dayeen et al. (2020) adopted LDA mining to analyze the literature on climate change in industrial ecology and study the evolution of major trends over time. Zhu et al. (2022) defined topic shift probability to analyze the evolution paths of topics based on LDA. Wang and Zhao (2018) proposed an improved dynamic NMF model with dynamic topic modeling of patent texts to study the dynamic evolution of technical topics. Mohotti and Nayak (2021) proposed a novel NMF model based on "cluster association awareness" to identify the dynamics of text clusters over time/domain. Jia et al. (2023) adopted the NMF model to cluster SCI papers in global carbon-neutral technologies to analyze the current status and trends of research in carbon neutrality. Therefore, topic models are the key method to study topic evolution.

2.2 Topic evolution analysis

The research of topic evolution is a method for developing a certain field. The research of topic extraction and path evolution on the content, semantics, and context of policies and papers in a certain field can reveal the development and changing trends of the field and thus infer its future direction.

Scholars mainly use the topic models to introduce the time dimension to study the change process of topics over time (Han 2020). Most studies vectorize the text by word vector models based on the time dimension and then explore the relationship between topics by calculating the similarity values, including growth, survive, shrink, split, merge, and dissolve of topics (İlhan and Öğüdücü 2015). Xi et al. (2021) researched the technical similarity visualization based on Word2Vec and LDA and conducted an empirical study in the field of NEDD (Nano Enabled Drug Delivery), proving that the model had good results in the analysis of technical similarity measurement. Xie et al. (2020) used BERT and LDA to analyze the similarity of monolingual and multilingual research topics evolution.

Therefore, accurately calculating the similarity between topics is a key step to measuring the evolution type of research topics and constructing the paths of topic evolution. The TF-IDF (Term Frequency-Inverse Document Frequency) method commonly used in previous studies is prone to fragmentation of word form and its meaning (Kim et al. 2016). With the application and development of deep learning, the topic model combined with word vector technology can fully explore the contextual context of words and inter-word relationships and has a strong semantic understanding of the topic. Thus, its modeling results are more interpretable. Typically represented by the Word2Vec model (Mikolov et al. 2013) and BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al. 2019), they can transform textual information from the unstructured form to vectorized form and generate semantically relevant word vectors, making up for the problem that topic models ignore potential semantic connections between topic words and lack semantic embeddings (Huang et al. 2022).

However, there is anisotropy in the vectors that BERT encodes for sentences because the vector values are affected by the word frequencies of the words in the sentences across all the training data, whose differences can lead to large distance biases. Therefore, the distance of word vectors is not a good representation of semantic relevance when using BERT (Li et al. 2020). In addition, prior knowledge for training models such as BERT is often not readily obtained from the detection tasks of knowledge structure for disciplinary.

In contrast, the Word2Vec model uses neural networks to learn word semantics and word-to-word contextual associations from large corpora in an unsupervised manner. As a result, each word is not limited to the meaning of the word itself but also maintains topic connections with related words. Therefore, in this study, Word2Vec is chosen to vectorize the text data in the experimental field, considering efficiency and scalability.

3.0 Methodology

To extract the key topics concerned in a certain disciplinary field and analyze the trend of topic evolution, this study proposes a recognition method of measuring literature topic evolution paths based on a K-means-NMF model. The topics are first extracted by the K-means-NMF model to explore the key problems in a certain disciplinary field. Then, the topic evolution in the disciplinary field is further investigated by adopting topic quantification based on the time dimension. The main research process includes four stages:

Stage 1: Data Collection and Pre-processing
Stage 2: Topic Extraction
Stage 3: Topic Evolution Relationship Analysis
Stage 4: Topic Evolution Path Analysis

The specific steps of each stage are shown in Figure 1.

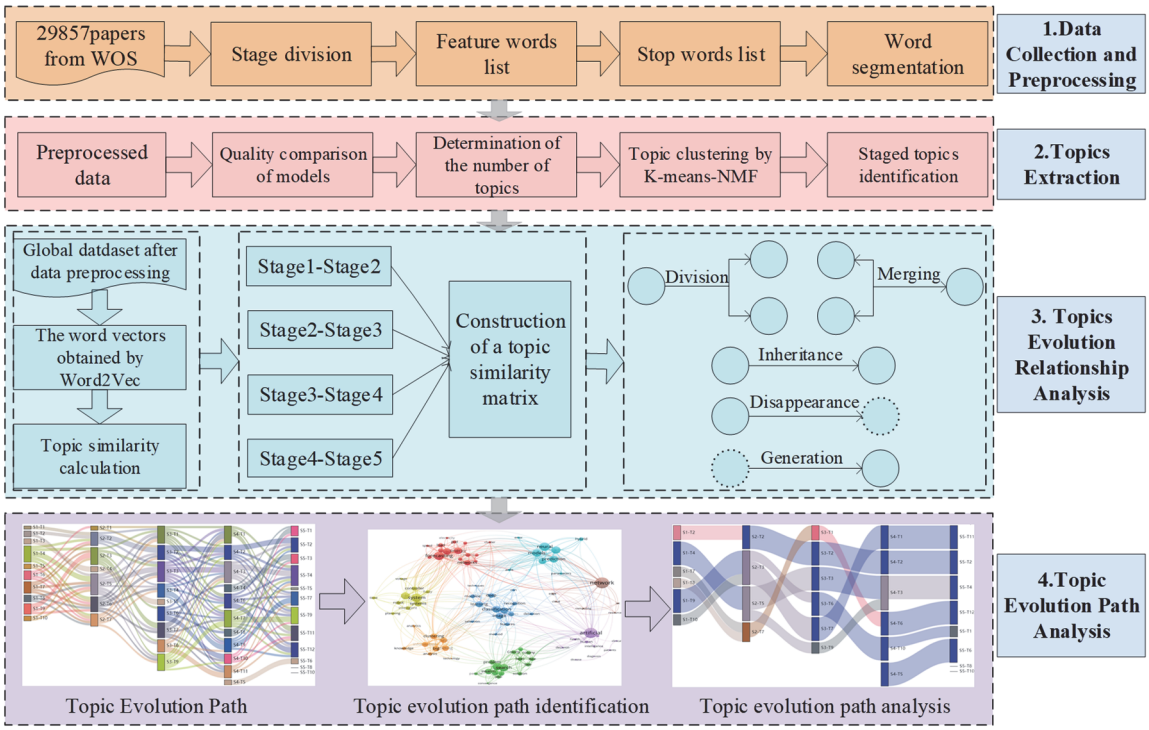


Figure 1. The overall research framework.

3.1 Data collection and pre-processing

Firstly, one needs to obtain papers in a certain disciplinary field from the Web of Science database, including their titles, keywords, abstracts, and publication dates, to form the text to be analyzed. After acquiring the data, one needs to pre-process the data, such as stage division, the construction of the feature word list and the stop word list, word segmentation, and stop word removal. Among them, the construction of word lists is the key to ensuring data quality and is the core step of data pre-processing, including the following two parts:

(1) Feature word list construction: Firstly, given the real dilemma that the target field is highly knowledgeable and possesses many specialized words, one needs to extract the keywords of each paper as the initial feature word list. Secondly, the domain-specific keywords are extracted from the text to be analyzed by the TF-IDF algorithm and then imported to the initial feature word list. Finally, one should de-emphasize and screen the combined feature word list.

(2) Stop word list construction: To improve the accuracy of topic recognition and prevent the interference of high-frequency invalid words, it is necessary to construct a target domain-oriented stop word list based on the generic stop word list. In specific cases, after the first topic clustering by K-means-NMF, one usually needs to check and extract the words that are not meaningful to the field, then import them into the generic stop word list. To ensure accuracy, one needs to iterate repeatedly the above process, usually at most five times, resulting in a final stop word list.

3.2 Topics extraction

3.2.1 Topics extraction by K-means-NMF

(1) NMF topic model

NMF is an unsupervised method for reducing the dimensionality of non-negative matrices to generate easily interpretable clusters of textual data. The core idea of the NMF model is to solve the topic extraction problem by transforming it into a constrained optimization problem. When dealing with textual data, it achieves topic extraction through the computation of matrix decomposition, where clusters are interpreted as topics, and each document is considered as an accumulative combination of multiple overlapping topics. Specifically, a non-negative matrix V of document-lexical items (consisting of n rows (lexical items) and m columns (documents)) is decomposed into the product of two non-negative matrices W and H, expressed as Equation 1:

$$V_{n\times m} \approx W_{n\times k} \times H_{k\times m}$$

Where k is the optimal number of topics that may be taken; w is the document-topic matrix whose size is related to k; H is $n\times k$ and H is the topic-word matrix, where the topic

words are obtained by calculating the distance between words through the semantic matrix.

(2) K-means-NMF model

K-means-NMF is obtained by combining the K-means clustering algorithm with the NMF model to improve the traditional NMF model. Although the NMF model can handle short text datasets well, NMF may have multiple possible outcomes during multiple iterations of decomposition, resulting in low processing efficiency when the dataset is too large, and the excessive number of topics may lead to poor and not robust fitting results. In addition, since the NMF model is initialized randomly, it is greatly affected by the initial value, and different initial values may lead to different convergence results. Based on the advantages that the K-means clustering algorithm has better scalability and good interpretability, to improve the accuracy of NMF topic clustering and reduce the correlation between topics, as well as to make the initial values of the NMF model more informative, this requires the combination of K-means clustering algorithm and NMF mode. Firstly, the vector of all words in the text data is trained using the Word2Vec model; Secondly, the K-means clustering algorithm is used to cluster and analyze all words in the overall text to obtain the clustering space of words. Then, the clustered words are reconstituted into sentences based on the clustering space, and all the sentences are formed into matrix *L*, which is normalized to eliminate the effect of category magnitudes. Finally, the normalized matrix *L* and matrix *V,* which is the original input of the NMF model, are fused into matrix *V′*, which is then decomposed and iterated to identify text topics. This process can enrich the original features of the traditional NMF model, reduce the number of iterations and obtain better-quality experimental results.

3.2.2 Topic model evaluation and topic number determination

Perplexity and Topic coherence are evaluation metrics that can characterize the probability distribution or the prediction effectiveness of a model and can be used to measure how good a topic model is and a key basis for determining the number of topics in a model. However, it was found that Perplexity is ineffective in recognizing topic models. When the number of topics is too large, the model is prone to overfitting, resulting in a large difference or even a negative correlation with the topic results judged by experts (Chang et al. 2009), while Topic coherence takes more into account the topic context, such as the co-occurrence of words, and can compensate for the shortcomings of Perplexity to a greater extent (Greene and Cross 2017). Based on this, Topic coherence is more suitable for assessing model quality and determining the optimal number of topics for a document, as shown in Equation 2:

$$\mathrm{Coh}(\mathrm{T}) = \frac{1}{\mathrm{k}}\sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{1}{\binom{\mathrm{t}}{2}} \sum_{\mathrm{j}=1}^{\mathrm{t}-1} \sum_{\mathrm{z}=\mathrm{j}+1}^{\mathrm{t}} \cos(\mathrm{w2v}(\mathrm{w}_{\mathrm{i,j}}), \mathrm{w2v}(\mathrm{w}_{\mathrm{i,z}}))$$

Where, k represents the possible number of topics to be taken, which is set according to the test text; $\mathrm{T} = [\mathrm{t}_1, \mathrm{t}_2, \cdots, \mathrm{t}_\mathrm{k}]$; $\mathrm{t}_\mathrm{i} = [\mathrm{w}_{\mathrm{i},1}, \mathrm{w}_{\mathrm{i},2}, \cdots, \mathrm{w}_{\mathrm{i},\mathrm{t}}]$; t_i represents the set of the top t words most relevant to the i-th topic. *w2v(w)* represents the word vector of the word *w*. *cos* represents the cosine similarity between the word vectors. In Equation 2, each *k* value corresponds to a *Coh(T)* value, and when *Coh(T)* is the maximum value, its *k* value represents the optimal number of topics.

3.3 Topic evolution relationship analysis

3.3.1 Topic similarity calculation

Measuring the similarity between topic word vectors can assess the degree of relevance between topics and thus determine the relationship between topics. Word2Vec yields two classical language models: Skip-Gram predicts nearby words through the central word, while CBOW (Continuous Bag of Words) predicts the current value through the context. Firstly, one needs to use the Word2vec model to train the word vector model on the pre-processed text, which uses the Skip-Gram model based on the current word to predict the context. Secondly, one needs to determine the word vector corresponding to each topic. Finally, the similarity values between adjacent stage topics are calculated based on the word vector using cosine similarity, and the inter-stage topic similarity matrix is constructed, as shown in Equation 3:

$$\mathrm{Similar}(\mathrm{T}_\mathrm{i}, \mathrm{T}_\mathrm{j}) = \cos(\mathrm{T}_\mathrm{i}, \mathrm{T}_\mathrm{j})$$

3.3.2 Topic evolution type determination

After calculating the inter-topic similarity, the degree of correlation between topics is judged by setting a threshold gradient to further track the type of topic evolution under different time stages, as shown in Equation 4:

$$\sigma \le \mathrm{Similar}(\mathrm{T}_{\mathrm{t}-1/\mathrm{t}+1}, \mathrm{T}_\mathrm{t}) \le 1$$

Where, $\sigma \in [0,1]$, σ is the threshold value determined by previous research and experimental results. According to whether there is a correlation between topics and the presentation of topics in different time stages, the types of topic evolution are classified into five types: division, merging, inheritance, disappearance, and generation, and their determination conditions are shown in Table 1.

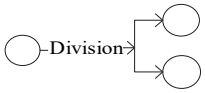
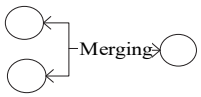
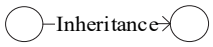
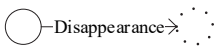
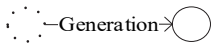
Topic Types	Description of determination criteria
	When $\sigma \le Similar(T_{t-i}, T_t)$, and a topic at the previous stage can be divided into two or more topics at the current stage.
	When $\sigma \le Similar(T_{t-i}, T_t)$, and two or more topics at the previous can be merged into one topic at the current stage.
	When $\sigma \le Similar(T_{t-i}, T_t)$, it means that a topic at the previous stage is continuous with a topic at the current stage.
	When $\sigma > Similar(T_{t-i}, T_t)$, and the topic at the previous stage doesn't exist at the current stage.
	When $\sigma > Similar(T_{t-i}, T_t)$, and the topic at the current stage doesn't exist at the previous stage.

Table 1. The criteria for topic evolution type determination

3.4 Topic evolution path analysis

Considering that there are many different potential topic evolution paths, in order to recognize the important evolution paths, it requires the performance of a co-occurrence analysis to identify the core topics on the clustering results of "topic-topic word" at each stage by the K-means-NMF model. Specifically, the topics with high relevance and influence are identified by calculating the intensity of topic co-occurrence, which is used as the basis for selecting important evolution paths and presented in a visualization map, as shown in Equation 5. At the same time, to show the evolution between topics at adjacent stages and construct a clear topic evolution path at the whole stage, this needs to adopt the Sankey visualization in the pyecharts package of Python language to implement. In Sankey, the element blocks represent objects, and the connecting lines indicate the direction and connection of the energy flow generated by the objects. This feature can visually represent the updates and additions generated by the topic content over time, as shown in Equation 5:

$$CI(T_{i,j}) = \frac{S_{ij}^2}{S_i S_j}$$

Where, $CI(T_{i,j})$ is the intensity of topic co-occurrence; S_i and S_j are the frequency of topic i and topic j occurring in the topic cluster; S_{ij} is the frequency of the same topic i and topic j occurring together in the topic cluster.

4.0 Empirical Research

With intelligent algorithms becoming the new production factor and tool, marked by ChatGPT and Artificial Intelligence Generated Content (AIGC) generic big models, they are gradually penetrating and influencing various disciplinary fields. Therefore, a comprehensive grasp of the knowledge system and development frontier in intelligent algorithms is of certain practical significance for each disciplinary field to find the entry point of research and discipline development. In this study, we take the papers on intelligent algorithms as an example to conduct an empirical study on the recognition method of measuring literature topic evolution paths based on K-means-NMF.

4.1 Data collection and pre-processing

Literature is an important information carrier for government staff and scientific researchers to keep abreast of developments in the field and an important reference for sci-

entific decision-making and effective implementation by institutions at all levels. Based on this, we selected the papers in intelligent algorithms from 2013-2022 as the data and used two years as a stage to explore the content and laws of the topic evolution in the field of intelligent algorithms for ten years. We used the core collection database in WOS (Web of Science) as the source of literature data. In the search process, we used ("intelligence algorithm" OR "artificial intelligence algorithm" OR "AI artificial algorithm") as the search condition and set the "Document Types" to "Article" and the "Languages" to "English." After the above search, we screened the papers according to the titles and abstracts, eliminating the papers with irrelevant and non-academic topics. We finally screened 29857 papers and performed data pre-processing, such as data cleaning, word separation, and removal of stop words for the screened papers. As shown in Table 2, the papers on intelligent algorithms show an increasing trend yearly, reflecting the hotness and importance of research in intelligent algorithms.

4.2 Topic extraction

4.2.1 Topic model quality evaluation and topic number determination

(1) Quality comparison of three topic models

To verify the effectiveness of the K-means-NMF model proposed in this paper, we compared it with two traditional topic models, LDA and NMF, in the case of the dataset of intelligent algorithms as the experimental object. In the comparison experiments, as shown in Figure 2, we obtained the different scores of topic consistency of different models at different stages and with different numbers of topics to compare the effects of topic extraction of different models. The higher the score of topic consistency, the better the effect of topic extraction. As shown in Figure 2, K-means-NMF, the topic extraction model used in this study, outperforms the other two models.

Stages	2013-2014	2015-2016	2017-2018	2019-2020	2021-2022
Number of Literature	1368	1632	2845	7809	16923

Table 2. Description of stage division

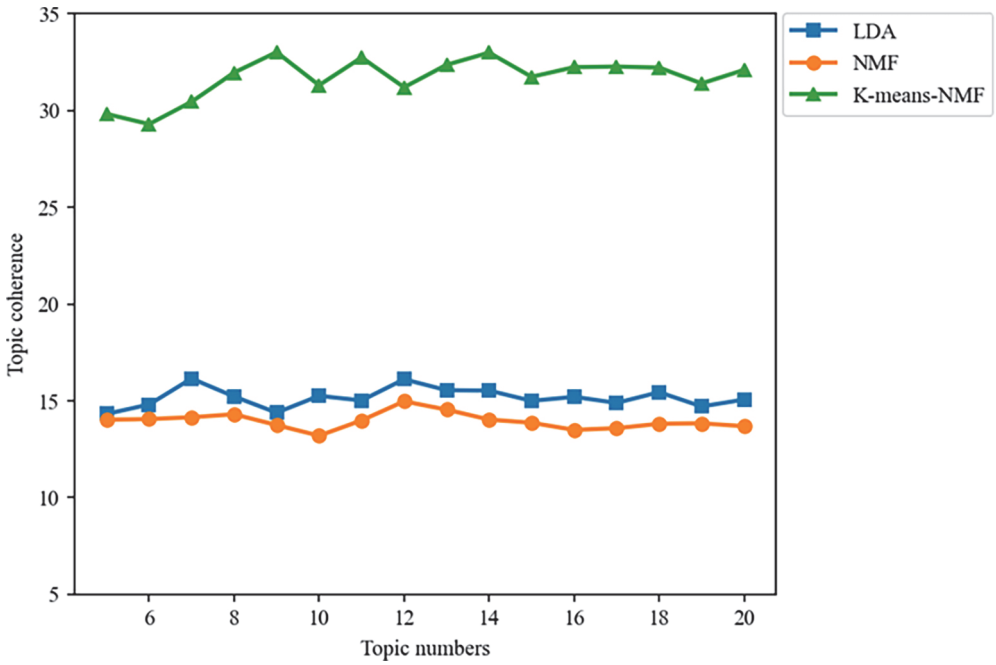


Figure 2. Comparison curves of topic consistency for different models

(2) Determination of the number of topics

The number of topics, often referred to as the k-value, is the most important hyperparameter in literature topic modeling, directly impacting the effect of extracting potential topics and needing to be set manually. Therefore, the optimal number of topics needs to be determined before modeling. According to the experimental tests, it was determined that $k \in [5,20]$ in this study. Figure 3 gives the scores of topic coherence of the K-means-NMF model with different numbers of topics set in each stage. It needs to select the number of topics with high scores through comparative analysis. Finally, we selected the number of topics for the five stages of the model as 10, 7, 9, 11, and 12.

4.2.2 Topic extraction analysis

Based on determining the optimal number of topics at each stage, we performed model training and stage topic clustering by K-means-NMF on the preprocessed texts with the help of Python. In this, we set the number of iterations to 200 and chose default values for the rest of the parameters. Then we showed the top 10 topic words with higher weight for each topic meaning, as shown in Table 3, where each topic word is sorted by weight from largest to smallest.

4.3 Topic evolution relationship analysis

4.3.1 Topic similarity calculation

Topic similarity calculation is the premise and basis of the evolution of topic content. In our experiments, we first used the functions with the Word2Vec library in the gensim package of the Python language to train the word vector model, where the window size was set to 5 and the vector dimension was 150. Secondly, we made the similarity calculations of topics at adjacent stages to form a heat map of the topic similarity matrix between adjacent stages, as shown in Figure 4-Figure 7. Finally, we set a threshold, named σ, to determine the types of topic evolution between adjacent stages, whose selection was obtained based on continuous iterative testing in the experiment. According to the experimental results, when σ=0.36, the interpretability was strong and the type of topic evolution could be clearly shown.

4.3.2 Topic evolution type determination

(1) Topics of division, merging, and inheritance types

The topics of division, merging, and inheritance types, usually with large proportions, strong evolutionary ability, and great influence, are the main type of topic evolution and the focus in intelligent algorithms. Among the three types of topic evolution, inheritance-type topics are usually included in division-type and merging-type topics. In this study, the division-type topic was exemplified by the topic

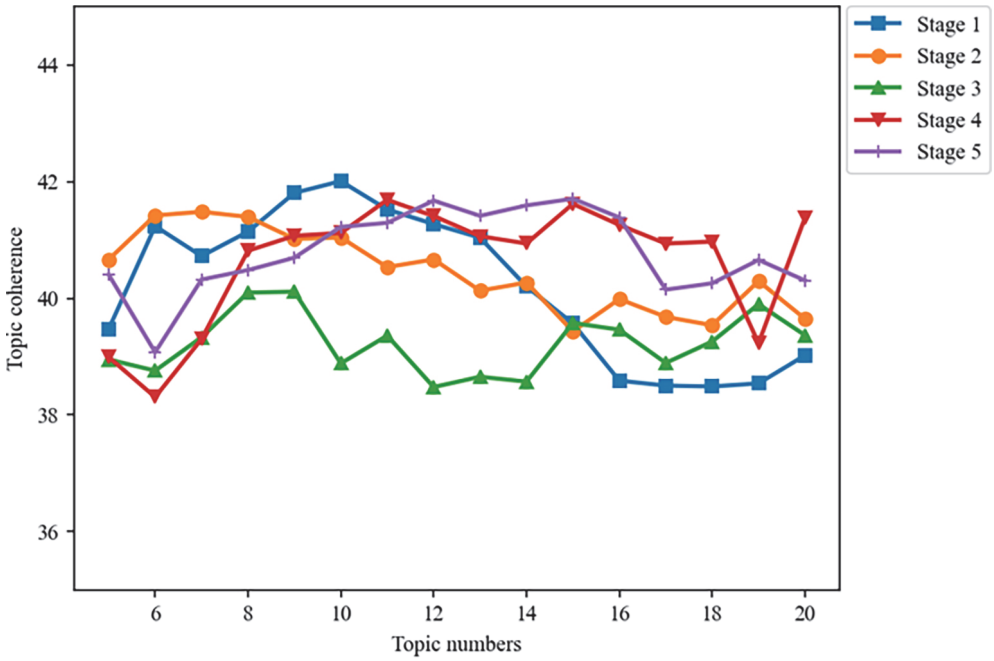


Figure 3. Curves of topic consistency at different stages

Stages	Topics-Topic words（some of them）
Stage 1	T1 (search problem algorithm); T2 (particle swarm optimization); T3 (artificial bee colony); T4 (model neural ANN); T5 (power wind energy); T6 (routing ad hoc); T7 (learning system systems); T8 (fuzzy system control); T9 (data clustering mining); T10 (classification image feature);
Stage 2	T1 (planning systems decision); T2 (optimization swarm algorithm); T3 (data learning classification); T4 (artificial bee colony); T5 (model forecasting models); T6 (network sensor wireless); T7 (control system power);
Stage 3	T1 (system control systems); T2 (optimization algorithm swarm); T3 (model models ANN（Artificial Neural Network）); T4 (energy routing network); T5 (artificial bee colony); T6 (image feature classification); T7 (data big clustering mining); T8 (forecasting wind power); T9 (learning deep machine);
Stage 4	T1 (data big mining); T2 (optimization algorithm swarm); T3 (model models ANN); T4 (edge computing cloud); T5 (ai artificial clinical); T6 (control system robot); T7 (learning machine deep); T8 (detection fault intrusion); T9 (energy consumption power); T10 (image images classification); T11 (forecasting wind power);
Stage 5	T1 (detection image images); T2 (optimization algorithm swarm); T3 (edge iot computing); T4 (model models prediction); T5 (teaching students english); T6 (ai artificial intelligence); T7 (energy power consumption); T8 (covid chest disease); T9 (learning machine ml); T10 (cancer patients breast); T11 (data big information); T12 (control system robot);

Table 3. Training results of the K-means-NMF model

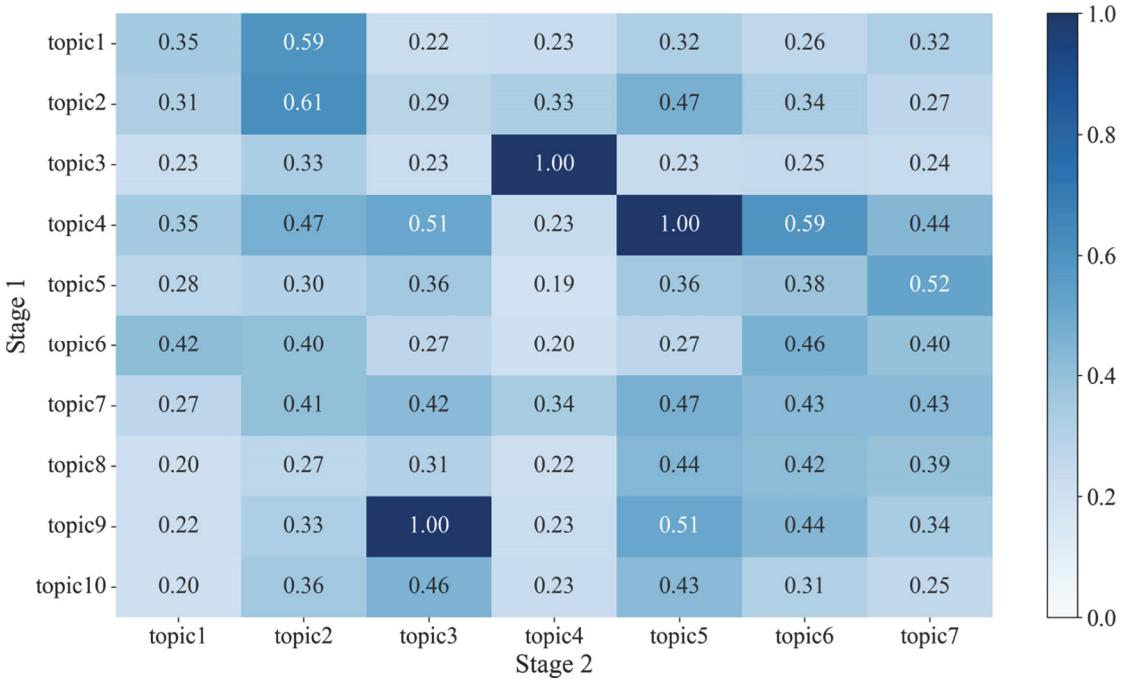


Figure 4. Similarity matrix between Stage 1 and Stage 2

S1-T7, divided into topics S2-T2, S2-T3, S2-T5, S2-T6, and S2-T7. Taking the specific meanings of each topic into the above analysis, we concluded that machine learning systems (S1-T7) were fundamental research in the field of intelligent algorithms, based on which they were divided into optimization algorithms (S2-T2), classification learning (S2-T3), model prediction (S2-T5), network sensors (S2-T6), control systems (S2-T7) and other topics related to the development and application of intelligent algorithms; The merging-type topic was exemplified by the topic S4-T1, which was merged from topics S3-T1, S3-T3, S3-T6, S3-T7, S3-T8, and S3-T9. Taking the specific meanings of each topic into the above analysis, we concluded that the topics of control systems (S3-T1), artificial neural networks (S3-T3), image feature classification (S3-T6), big data clustering analysis (S3-T7), temperature prediction (S3-T8), and deep learning (S3-T9) were merged into new research topics such as big data mining and information analytics.

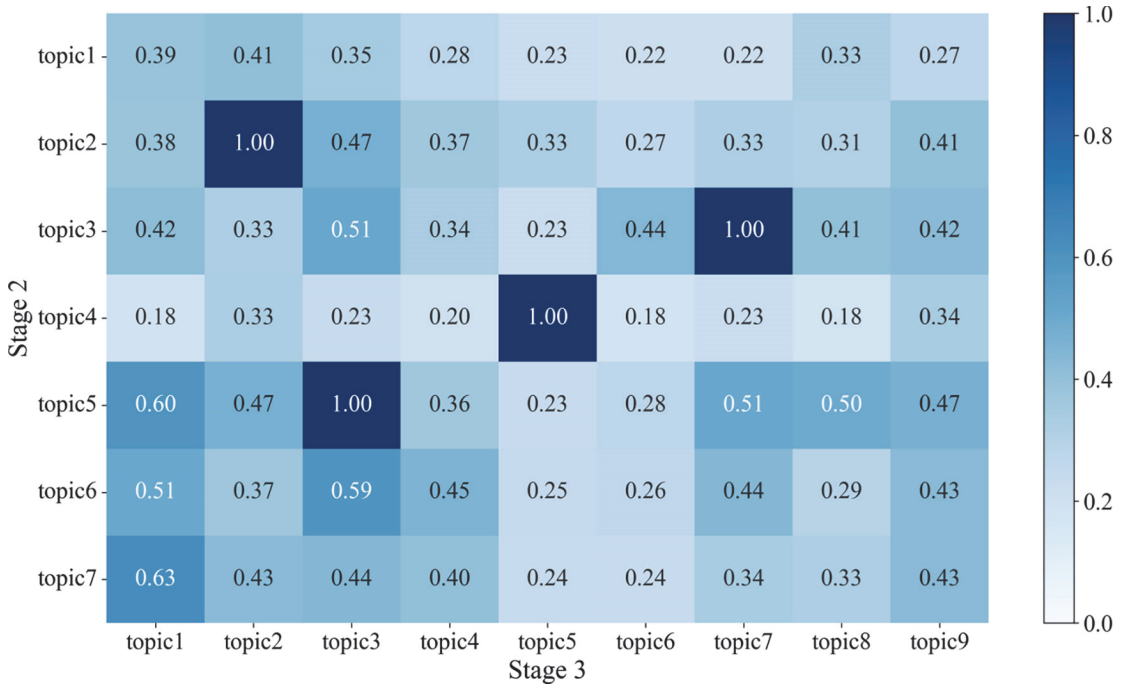


Figure 5. Similarity matrix between Stage 2 and Stage 3

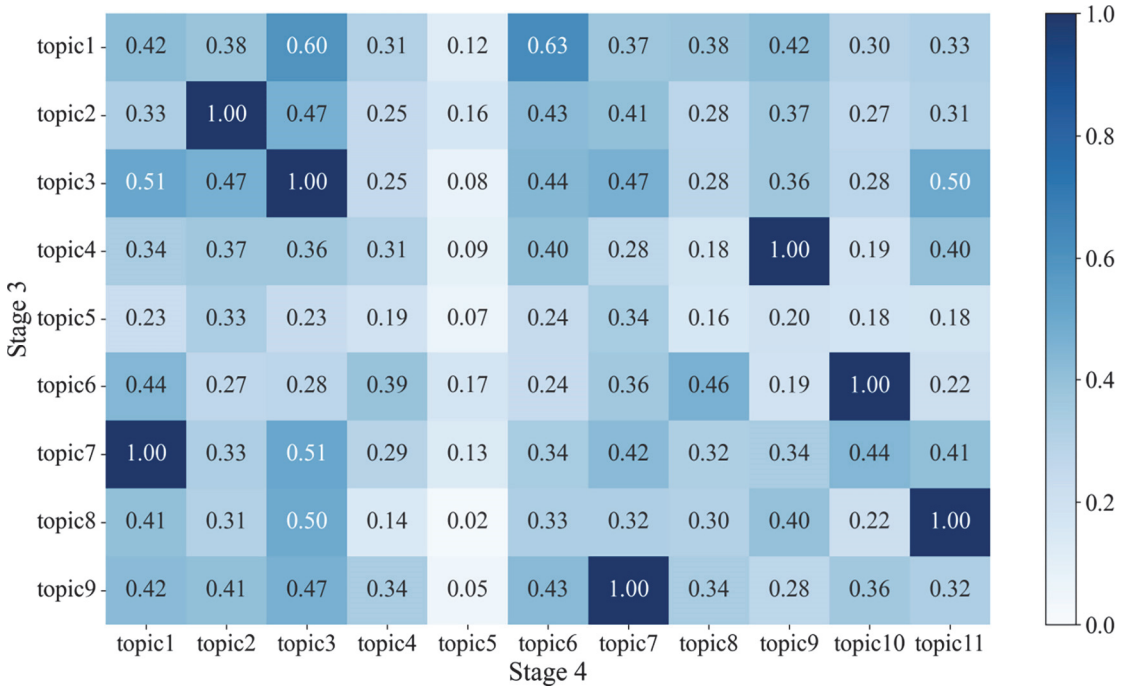


Figure 6. Similarity matrix between Stage 3 and Stage 4

(2) Topics of disappearance and generation types

The topics of disappearance and generation types, usually with small proportions and weak evolutionary ability, reflect the turnover of research topics in intelligent algorithms. Because the popularity of disappearance-type topics tends to decline or even disappear, identifying such topics can provide a reference for researchers. In this study, the division-type topic was exemplified by the topic S3-T5, which was focused on the research of artificial bee colony algorithms. The artificial bee colony algorithm is a novel global

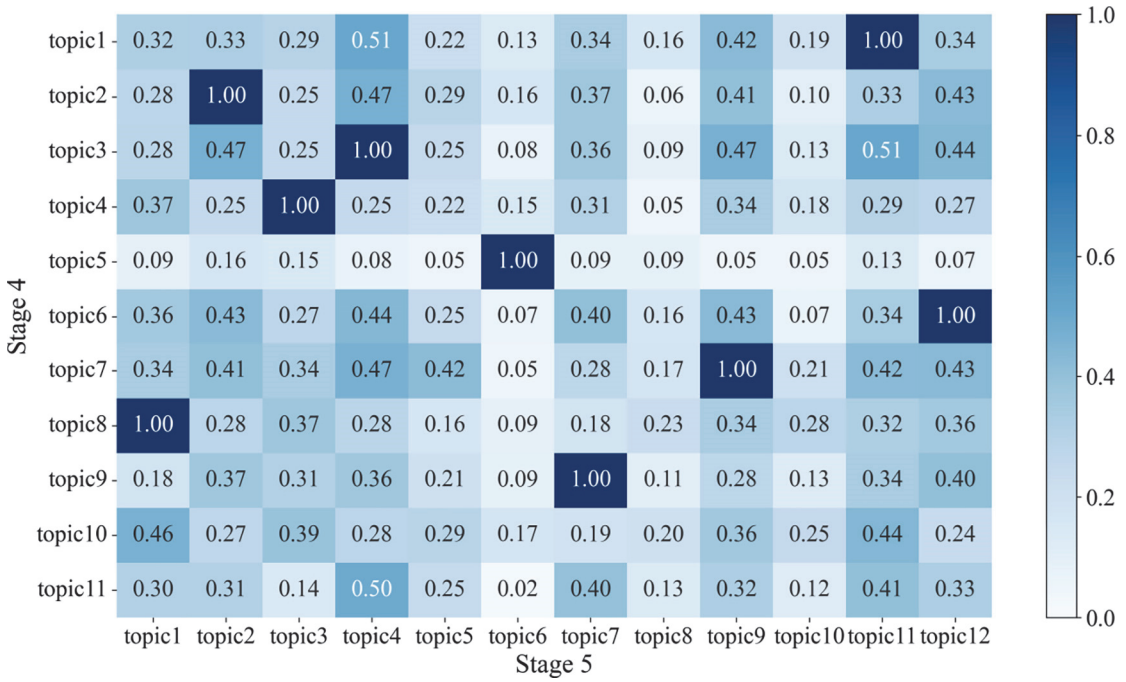


Figure 7. Similarity matrix between Stage 4 and Stage 5

optimization algorithm based on swarm intelligence proposed in 2005, whose research fervor has gradually decreased since 2018. Emerging-type topics are novel and belong to emerging hot topics, reflecting the trend of research contents in the field of intelligent algorithms with strong evolutionary ability. In this study, such topics included S4-T5, S5-T8, and S5-T10, which were mainly related to AI clinical care (S4-T5), COVID-19 (S5-T8), and cancer treatment (S5-T10). These topics mostly occurred at stage 4 and stage 5 and are the smart medical and health issues that people are more concerned about and studied in the current big data era.

4.4 Topic evolution paths analysis

As scientific research has progressed, there are links between the topics of each discipline, both at the same stage and at different stages. Also, there are changing relationships between research topics at adjacent stages as technology has evolved. We drew a Sankey diagram of the topic evolution in the field of intelligent algorithms according to the similarity between topics, as shown in Figure 8. The Sankey diagram visualizes the topic evolution paths in intelligent algorithms at five stages over the last ten years, where nodes represent topics and lines represent the association of topics. The horizontal perspective of the Sankey diagram can visually show the types of topic evolution on each path, and the vertical perspective can clearly show the distribution and proportion of topics at each stage. By focusing on horizontal analysis and supplementing it with vertical analysis, we analyzed the changes in topic contents and topic evolutionary ability in intelligent algorithms.

4.4.1 Topic evolution path recognition

In order to select the important topic evolution paths in intelligent algorithms, we used Equation 5 to analyze the co-occurrence of topic clustering results at each stage from the perspective of the overall evolution. From the co-occurrence analysis, we obtained a co-occurrence matrix of size 90*90, which was converted to Pajek format using Ucinet software and then imported into VOSviewer software to draw a co-occurrence map of topic words. In the results, shown in Figure 9, the color represents the category, and the size of the circle represents the degree of influence of the topic, meaning the larger the circle, the greater the influence of the topic word.

In the process of co-occurrence analysis, we selected topic words with great influence to characterize the topics and finally identified eight categories. However, among them, “forecasting” and “prediction” belonged to the applications of “network,” so these three categories were combined into one called “network.” In summary, we finally identified six categories: neural network, classification, clustering system and system control, algorithm optimization, and intelligent healthcare. These six categories, with high topic relevance and impact, usually have the characteristics of sustainability and growth and are the focus in the field of intelligent algorithms.

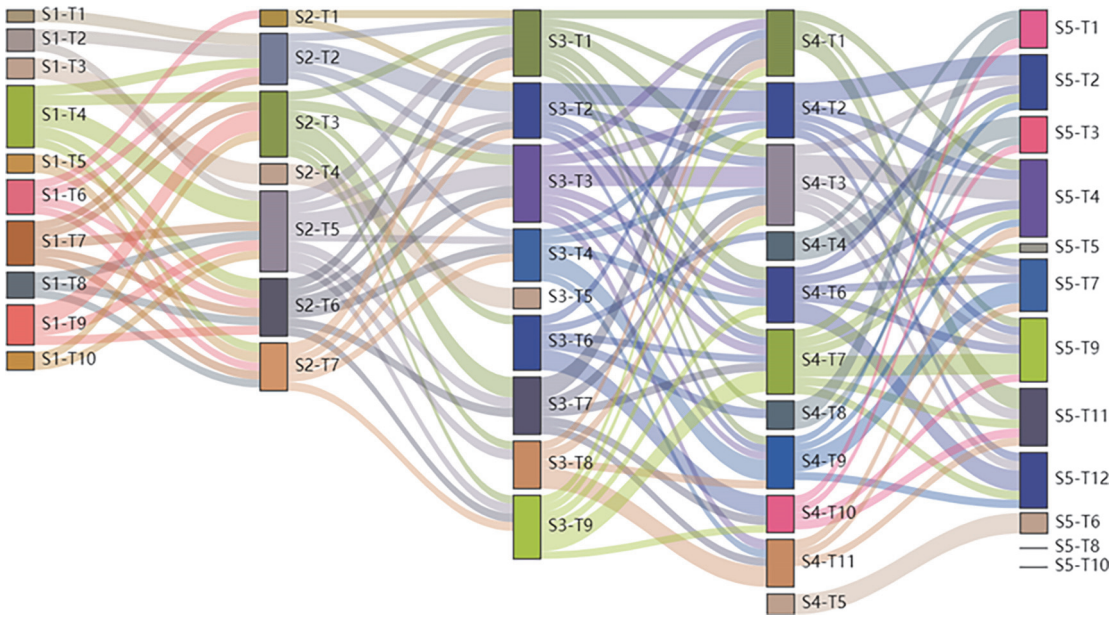


Figure 8. Topic Evolution Path

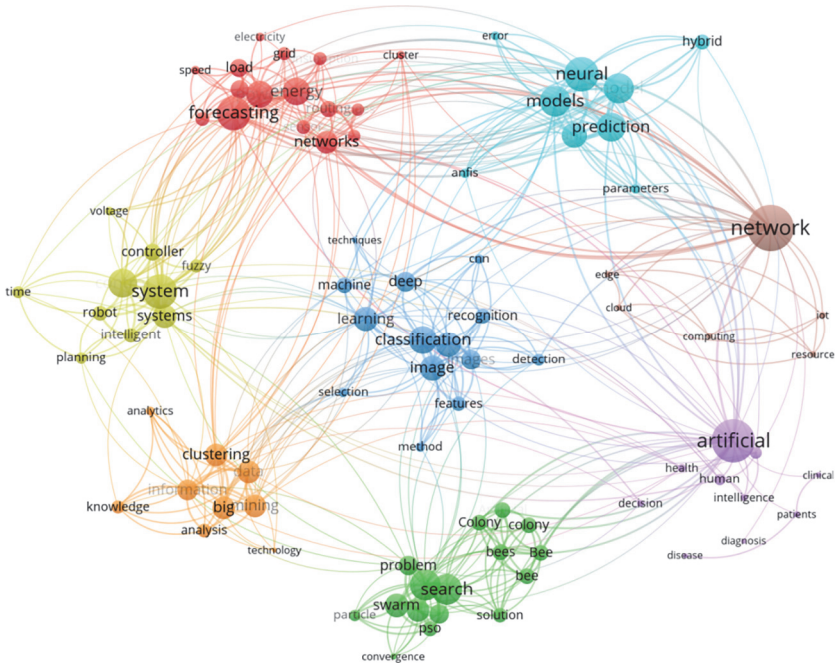


Figure 9. Topic co-occurrence mapping

4.4.2 Topic evolution path analysis

Taking the six categories identified from the topic co-occurrence analysis as the main reference, combined with the evolution types, we selected paths with higher importance to analyze. Finally, we formed five important evolution paths, as shown in Figure 10.

(1) The first path is S1-T4(model neural ANN)→S2-T5(model forecasting models)→S3-T3(model models ANN) & S3-T9(learning deep machine)→S4-T3(model models ANN) & S4-T9(energy consumption power)→S5-T4(model models prediction). This evolution process is about neural networks and is an inheritance-type evolution path, reflecting the heat and importance of neural network

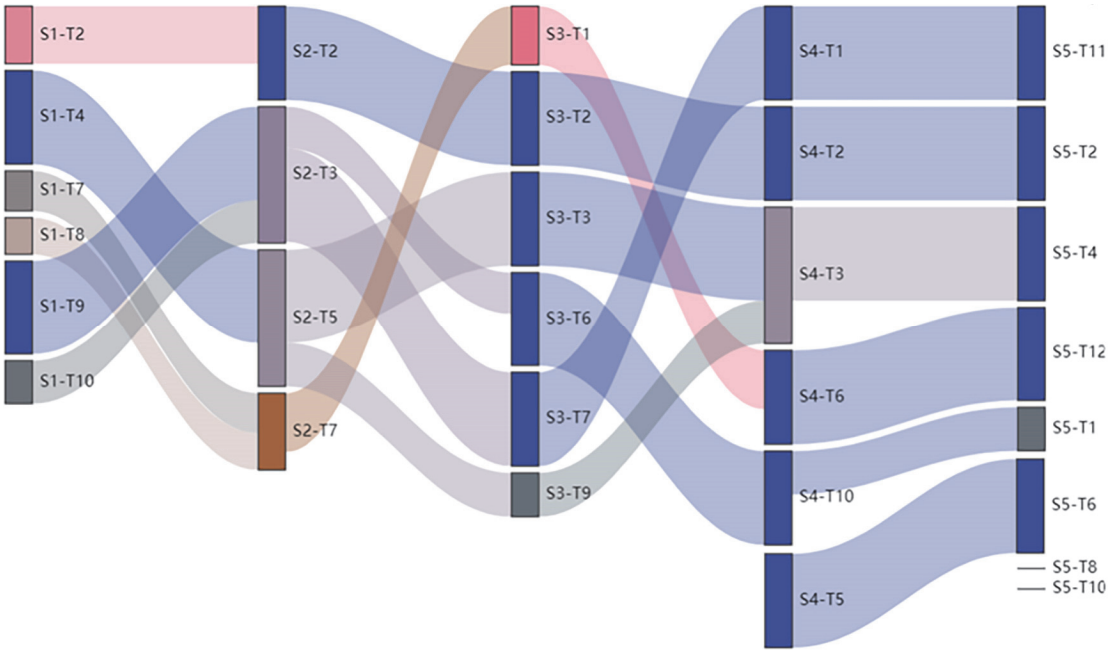


Figure 10. Path evolution diagram of higher topic importance

development. The main algorithms are ANN (Artificial Neural Network), CNN (Convolutional Neural Network), and deep learning networks, mainly for prediction and recognition.

(2) The second path is S1-T7(learning system systems) & S1-T8(fuzzy system control)→S2-T7(control system power)→S3-T1(system control systems)→S4-T6(control system robot)→S5-T12(control system robot). This evolution process, belonging to both merging-type and inheritance-type evolution paths, is about systems and system control. A system can be defined as a set of interconnected factors that interact and influence each other in a specific pattern of behavior over time. Therefore, not only specific systems such as ANIS (Adaptive Neuro-Fuzzy Inference System) but also process control theory are usually integrated into the research to ensure the proper operation of the system.

(3) The third path is S1-T9(data clustering mining) & S1-T10(classification image feature)→S2-T3(data learning classification)→S3-T6(image feature classification) & S3-T7(data big clustering mining)→S4-T1(data big mining) & S4-T10(image images classification)→S5-T1(detection image images) & S5-T11(data big information). This evolution process, belonging to both merging-type and division-type evolution paths, is about classification and clustering. Among them, classification is mainly studied by images, while clustering is mainly studied by big data. With the massive growth of various heterogeneous data, image classification and big data mining have become hot research directions for scholars.

(4) The fourth path is S1-T2(particle swarm optimization)→S2-T2(optimization swarm algorithm)→S3-T2(optimization algorithm swarm)→S4-T2(optimization algorithm swarm)→S5-T2(optimization algorithm swarm). This evolution process is about particle swarm optimization algorithms and is an inheritance-type evolution path, reflecting the heat and importance of the research of particle swarm optimization algorithms. Particle swarm optimization algorithm has the advantages of fast convergence, few parameters, and simple and easy implementation. It converges to the optimal solution faster than a genetic algorithm for high-dimensional optimization problems. However, there is also the problem of falling into local optimal solutions, which becomes a hot direction of scholars' attention.

(5) The fifth path is S4-T5(ai artificial clinical)→S5-T6(ai artificial intelligence) & S5-T8(covid chest disease) & S5-T10(cancer patients breast). This evolution process, belonging to both inheritance-type and generation-type evolution paths, is about smart healthcare, reflecting the heat and importance of smart healthcare research at present. This evolution path focuses on adopting intelligent algorithms or artificial intelligence for paramedicine and medical decision-making under the background of COVID-19 outbreaks.

In the three stages of 2013-2014, 2015-2016, and 2017-2018, the number of research topics in intelligent algorithms is small, and most need more attention. Thus, the evolutionary ability between topics is moderate, and the evolution types are mainly inheritance-type, division-type,

and merging-type. As scientific research has continued to progress, the number of topics has increased, and topics' evolutionary ability and influence have gradually increased in 2019-2020. At the same time, the merging and division between topics have become obvious, and the cross-research phenomenon among multiple topics gradually enriches the evolution paths. Notably, new topics have been generated in 2021-2022, which inject new vitality into scientific research in intelligent algorithms. Few research topics that have matured at different stages belong to disappearance-type topics. Through an empirical analysis of the field of intelligent algorithms, the results of this study are found to be in good agreement with the development and evolutionary process of intelligent algorithms proposed by scholars Heaton (2015) and Mahesh (2020), further validating the feasibility of the method we proposed for recognizing the evolution paths of literature topics based on K-means-NMF.

5.0 Conclusion

This study proposed a recognition method of measuring literature topic evolution paths based on K-means-NMF to meet scientific research needs and policy-making in science and technology. The innovation points of the proposed method mainly include topic extraction, the selection of critical topic paths, and their analysis results.

Firstly, we enhanced the traditional NMF model by combining it with the K-means clustering algorithm to solve the problems such as the unobvious effect in topic clustering and the high degree of mixing in clustering results caused by the situation that the traditional LDA and NMF were confined to the long or short texts. Secondly, we performed the topic co-occurrence analysis based on the clustering results to identify important topic categories for recognizing critical evolution paths to solve the problem of multiple possible evolution paths in the experiment. Thirdly, the Word2vec model was adopted to calculate topic word vectors in a semantic context to improve the accuracy of topic association strength between adjacent stages and then realize the analysis of topic evolution paths at different stages in the field's life cycle. Finally, to verify the validity of the proposed model, we conducted an empirical study that adopted the model to the field of intelligent algorithms.

The results of this paper show that, firstly, the K-means-NMF model outperforms LDA and NMF in topic extraction for text. Secondly, the co-occurrence analysis of clustered topics effectively recognizes critical evolution paths. Thirdly, Word2vec can better measure the relationships and evolution laws between topics in adjacent stages. Since the proposed method in this study is an unsupervised approach, its scientific validity and effectiveness just only be verified using empirical studies and comparison of the clustering effect of standard topic models such as LDA and NMF. For this, we will further explore how to perform supervised training based on accurate clustering in the subsequent research.

Acknowledgements

This work was Supported by National Social Science Foundation of China "Research on Risk Governance Path of Intelligence Analysis Algorithm in Data Intelligence Environment (Grant No. 22BTQ064).

References

Ambrosino, Angela, Mario Cedrini, John B Davis, Stefano Fiori, Marco Guerzoni and Massimiliano Nuccio. 2018. "What topic Modelling Could Reveal About the Evolution of Economics." *Journal of Economic Methodology* 25, no. 4:329-348. https://doi.org/10.1080/1350178X.2018.1529215

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Machine Learning Research* 3: 993-1022. https://dl.acm.org/doi/10.5555/944919.944937

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55, no.4:77-84. https://doi.org/10.1145/2133806.2133826

Chen, Yong, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. 2019. "Experimental explorations on short text topic mining between LDA and NMF based Schemes." *Knowledge-Based Systems* 163, no. 1:1-13 https://doi.org/10.1016/j.knosys.2018.08.011

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang and David M. Blei 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Proceedings of the 22nd International Conference on Neural Information Processing Systems* (NIPS'09), Red Hook, NY, USA, Curran Associates Inc, 288-296. https://dl.acm.org/doi/10.5555/2984093.2984126

Dayeen, Fazle Rabbi, Abhinav S. Sharma, and Sybil Derrible. 2020. "A Text Mining Analysis of the Climate Change Literature in Industrial Ecology." *Journal of Industrial Ecology* 24, no.2: 276-284. https://doi.org/10.1111/jiec.12998

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Minneapolis, Minnesota. Association for Computational Linguistics, 4171-4186. https://doi.org/10.48550/arXiv.1810.04805

Greene, Derekand and James P. Cross. 2017. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *Political Analysis* 25, no.1:77-94. https://doi.org/10.1017/pan.2016.7

Han, Xiaoyao. 2020. "Evolution of Research Topics in LIS Between 1996 and 2019: an Analysis Based on Latent Dirichlet Allocation Topic Model." *Scientometrics* 125: 2561-2595. https://doi.org/10.1007/s11192-020-03721-0

Heaton, Jeff. 2015. "Artificial Intelligence for Humans, Volume 3: Deep Learning and Neural Networks. Chesterfield: Heaton Research, Inc.

Huang, Lu, Xiang Chen, Yi Zhang, Changtian Wang, Xiaoli Cao and Jiarun Liu. 2022. "Identification of Topic Evolution: Network Analytics With Piecewise Linear Representation and Word Embedding." *Scientometrics* 127:5353-5383. https://doi.org/10.1007/s11192-022-04273-1

lhan, Nagehan and Şule Gündüz Öğüdücü. 2015. "Predicting Community Evolution Based on Time Series Modeling." In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Paris, France, IEEE, edited by Jian Pei, Fabrizio Silvestri and Jie Tang, 1509-1516. https://doi.org/10.1145/2808797.2808913

Jia Jia, Lu Qing, and Yang Fan. 2023. "Trend Analysis of Global Carbon Neutrality Related Field Research: Based on Machine Learning and Bibliometrics." *Science Focus* 18, no. 1:27-44. https://doi.org/10.15978/j.cnki.1673-5668.202301006

Jiang, Lei, Tao Zhang, and Taihua Huang. 2022. "Empirical Research of Hot Topic Recognition and its Evolution Path Method for Scientific and Technological Literature." *Adv. Comput. Intell. Intell. Inform* 26, no. 3:299-308. http://dx.doi.org/10.20965/jaciii.2022.p0299

Kim, Ha Jin, Jeong Yoo Kyung and Min Song. 2016. "Content- and Proximity-based Author Co-citation Analysis Using Citation Sentences." *Journal of Informetrics* 10, no. 4:954–966. https://doi.org/10.1016/j.joi.2016.07.007

Lee, Daniel D. and Seung H. Sebastian. 1999. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401:788-791. https://doi.org/10.1038/44565

Li, Bohan, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, 9119-9130. http://dx.doi.org/10.18653/v1/2020.emnlp-main.733

Macfarlane, Andrew, Sondess Missaoui, and Sylwia Frankowska-Takhari. 2020. "On Machine Learning and Knowledge Organization in Multimedia Information Retrieval." K*nowledge Organization* 47, no.1:45-55. https://doi.org/10.5771/0943-7444-2020-1-45.

Mahesh, Batta. 2020. "Machine Learning Algorithms-A Review." *International Journal of Science and Research* 9:381-386. http://dx.doi.org/10.21275/ART20203995

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems* 26: 3111-3119. https://dl.acm.org/doi/10.5555/2999792.2999959

Mohotti, Wathsala Anupama and Richi Nayak. 2021. "Discovering cluster evolution patterns with the Cluster Association-aware matrix factorization." *Knowledge and Information Systems* 63:1397-1428. https://doi.org/10.1007/s10115-021-01561-9

O'Callaghan, Derek, Greene Derek, Carthy Joe, and Cunningham. 2015. "An Analysis of the Coherence of Descriptors in Topic Modeling." *Expert Systems with Applications (ESWA)* 42, no.13:5645-5657. https://doi.org/10.1016/j.eswa.2015.02.055

Wang, Yuanyuan and Yajuan Zhao. 2018. "Evolution Analysis of Technological Topic: An Approach Based on Non-negative Matrix Factorization." *Library and Information Service* 62, no. 10:94-105. https://www.cnki.net/kcms/doi/10.13266/j.issn.0252-3116.2018.10.013.html

Xi, Xiaowen, Ying Guo, Xinna Song, and Jin Wang. 2021. "Research on the Technical Similarity Visualization Based on Word2vec and LDA Topic Model." *Journal of the China Society for Scientific and Technical Information* 40, no. 9:974-983. https://doi.org/10.3772/j.issn.1000-0135.2021.09.007

Xie, Qing, Xinyuan Zhang, Ying Ding, and Min Song. 2020. "Monolingual and Multilingual Topic Analysis Using LDA and BERT Embeddings." *Informetrics* 14, no. 3:1751-1577. https://doi.org/10.1016/j.joi.2020.101055

Zhang Dongxin and Min Zhang. 2022. "A Review on Application Studies of LDA Topic Models in Library and Information Science Field." *Documentation, Information & Knowledge* 39, no. 6:143-157. https://doi.org/10.13366/j.dik.2022.06.143

Zhu, Hengmin, Li Qian, Wang Qin, Jing Wei, and Chao Shen. 2022. "Evolution Analysis of Online Topics Based on 'Word-topic' Coupling Network." *Scientometrics* 127:3767-3792. https://doi.org/10.1007/s11192-022-04439-x