Research and Applications of Foundation Models for Data Mining and Affective Computing (RAFDA)

Construction of Academic Innovation Chain Based on Multi-level Clustering of Field Literature

Wei Cheng; Tianshi Cong
Nanjing Agricultural University
(email: chengwei@stu.njau.edu.cn)

May 7, 2024

Introduction

Innovation stands as an indispensable requirement in scientific research, with the **extraction, measurement, and evaluation of innovative knowledge** emerging as prominent research areas within **academic evaluation**. It encompasses the discovery or creation of novel knowledge within the existing knowledge framework, characterized by both relevance and fundamental divergence from preexisting knowledge.

Various methods for analyzing **literature associations**, such as keyword cooccurrence networks, citation networks, and topic evolution networks, have been widely employed and yielded promising results. However, these methods encounter challenges in **visualizing potential connections between innovation points**, thereby limiting the comprehensive assessment of innovative contributions.

The **clustering algorithm**, a versatile and powerful tool, is widely applied across diverse domains for data analysis and pattern recognition. Notably, Pen et al. employ the Self-Organizing Map clustering algorithm for image segmentation. Additionally, clustering algorithms are invaluable for anomaly detection. Moreover, clustering methods play a pivotal role in social network analysis. These collective efforts underscore the efficacy of clustering algorithms across various applications.

Introduction

Text clustering, employing various clustering algorithms for analyzing text data, especially within the domains of **topic clustering and semantic clustering**, represents an effective strategy for categorizing similar texts. This approach provides a tangible means of organizing innovative concepts dispersed throughout diverse literature sources. As such, the principal aim of this study is to consolidate interconnected literature using text clustering algorithms. Additionally, our objective extends to establishing a chronological framework for innovation points by systematically extracting them from the literature in accordance with their order of publication.

This study introduces an innovative method for **constructing academic innovation chains through multi-level clustering**. We focus on Chinese journal literature in the knowledge element field as our experimental domain. Leveraging **tf-idf, LDA, and doc2vec** algorithms, we achieve feature representation of literature across three levels: **word frequency, topic, and semantic**. Subsequently, text clustering is performed using the **Kmeans** algorithm. To enhance the process, we integrate **a fusion rule method with the ALBERT pre-training model** to extract innovation points from the literature. Finally, we not only construct but also visualize academic innovation chains.

The primary contribution of this study lies in its ability to **observe innovation points within a progressive and dynamic chain structure**. This innovative approach holds significant implications for informing innovation assessment, knowledge organization, and related studies.

Methodology: Overview of methodology

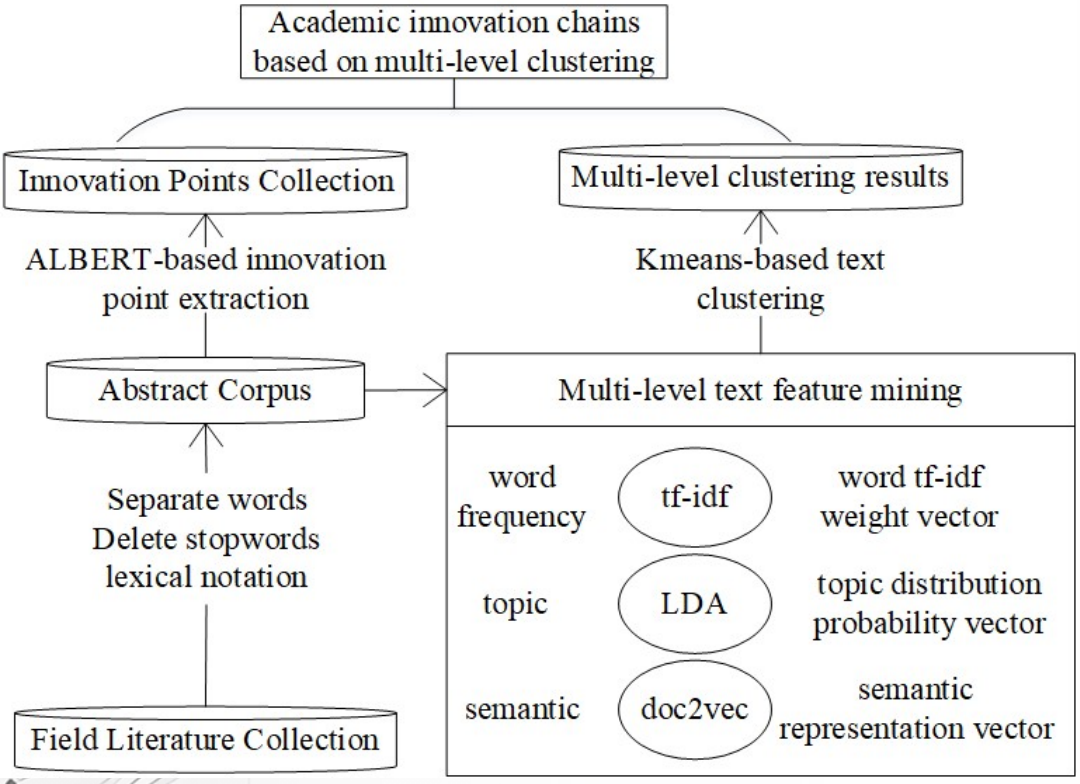


Fig. 1. Academic innovation chain construction method.

1. The title and abstract corpus is constructed by data preprocessing through separate words, delete stopwords and lexical labeling using the jieba library.
2. The textual feature vector representations of the field literature are mined in-depth from three levels based on the tf-idf, LDA and doc2vec algorithms.
3. Based on the Kmeans text clustering algorithm, the feature vector representations of the field literature are clustered at three levels.
4. Learning textual features of innovation points based on lightweight pre-trained model ALBERT to realize automatic extraction of innovation points.
5. The innovation points of the literature of the same type of clusters are chained and organized according to the chronological order of the publication of the literature, so as to realize the construction of the academic innovation chain.

Methodology: Multi-level text feature mining algorithms

Text feature (word frequency) mining algorithms: tf-idf

Tf-idf algorithm does not pay attention to the order and dependency of the words between the word, only from the point of view of the frequency of the word to determine a word's probability of occurrence. For any word in any document, its tf-idf weight is the product of tf (the number of occurrences of the word in the current document) and idf (the total number of documents divided by the number of words containing the word, the result is taken as a logarithm). Assuming that the collection contains N documents with a total of M non-repeating words, a N*M matrix can be constructed based on the tf-idf algorithm, which is a highly sparse matrix.

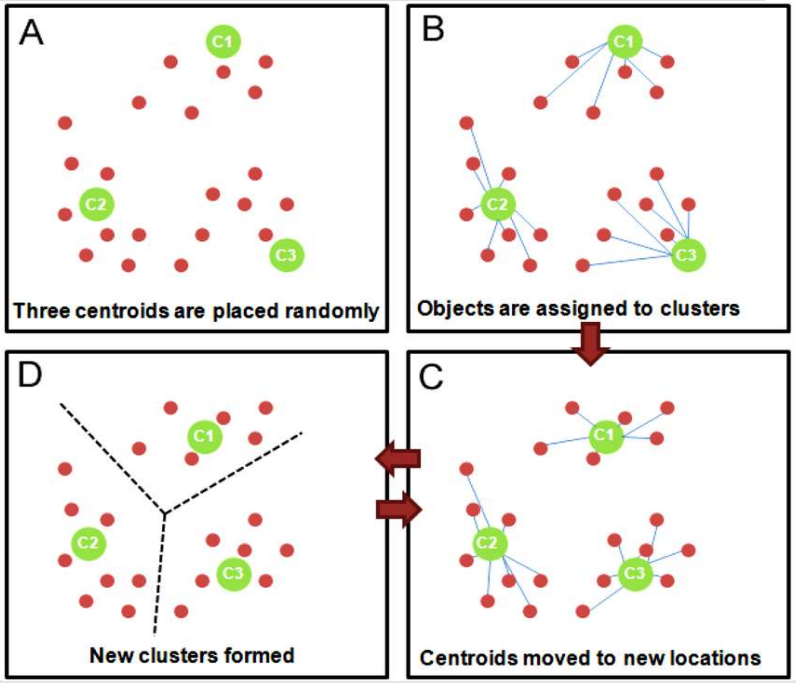
Text feature (topic) mining algorithms: LDA

LDA topic model is commonly used to infer the topic distribution of documents. LDA first needs to confirm the value of the number of topics k. The best k value is usually determined by the confusion value. Then the probability of distribution of words in k topics is obtained through the computation and training of the corpus. Finally, the topic distribution probabilities are calculated from the words contained in the documents. A k-dimensional document-topic distribution vector is constructed to characterize the topic content of the document, and the elements of the vector represent the probability that the document belongs to each topic.

Text feature (semantic) mining algorithms: doc2vec

Doc2vec integrates the dependency of a word with its preceding and following words as well as the order of occurrence of words. A certain part of the document is intercepted as a prediction word, and the word vectors of other words are used as input. The sentence vector D of the current document is taken as the feature vector input, and D is characterized as the semantic content of the document based on the inter-word semantic dependency. D is continuously updated during iterative training so that the model learns the missing content in the current context as well as the main idea content of the paragraph. Finally, vector W is generated for each word in the document, and vector D is generated for each document.

Methodology: Kmeans clustering algorithm



Kmeans is one of the most commonly used text clustering algorithms. The processing steps are as follows: first, based on the sample set S, randomly initialize k clustering centroids, which can be randomly selected from the sample set S or randomly generated. Next, traverse each sample x of the sample set S, calculate the space vector distance from x to the k centroids, and assign x to the class cluster y with the shortest distance. Then, the vector mean of all samples in each class cluster is calculated as the new clustering centroid. Finally, the above two steps are repeated until a certain number of iterations is reached or the clustering centroids are no longer changed.

Methodology: Innovation points extraction

Table 1. Rules and examples of innovation points extraction.

Trigger word	Example of extraction results
propose	An oil spill incident scenario model based on key scenario driver elements is proposed.
explore	The basic methods of structured processing of geological data texts are explored.
construct	Constructing an ontological knowledge base of chest paralysis evidence.
create	Creating a structured body of knowledge linked by concepts.
improve	Improving Bootstrapping methods.

1. Segment the abstract text into sentences.
2. For the clauses after clause splitting, the preliminary extraction of innovation points in the abstract is realized based on the trigger word rule. Some of the rules and examples are shown in Table 1.
3. Manually check and correct the results of the rule-based innovation extraction, and construct a "literature innovation points" combination and classiffcation corpus.
4. Adopt TensorFlow framework, create neural network model using Keras, and construct automatic extraction model of field innovation points based on lightweight pre-training model ALBERT to realize the generalized extraction of innovation points in field literature.

Empirical Research: Data collection and preprocessing

In this study, we take the field of **"knowledge element"** as an example to explore the feasibility of the academic innovation chain construction method. And **639 field literatures** are retained, spanning from 1981 to 2022.

The determination of k-value in Kmeans algorithm is very important work. The k-value interval is set to [2,50], and the values are traversed for multiple rounds of clustering, and **the best k-value is selected based on the following two ways**. (i) Silhouette Coefficient is used to evaluate the clustering effect and as a reference for selecting the optimal k-value. (ii) The T-SNE visualization dimensionality reduction algorithm is used to map the multi-dimensional text feature vectors into a two-dimensional space. The reasonableness of the clustering results is evaluated manually to select the best k-value.

Empirical Research: Multi-level clustering experiments

Word frequency. There are 5341 non-repeated words in the corpus. A 2D matrix of 639*5341 is obtained. In turn, Kmeans clustering is implemented based on sklearn library. The largest Silhouette Coefficient is 0.0207 (k takes the value of 50), and the smallest Silhouette Coefficient is 0.007 (k takes the value of 2), and the Silhouette Coefficient tends to 0. This indicates that there is a strong knowledge linkage within the field of knowledge elements, and when the dataset is classified into multiple class clusters, there is a certain amount of variability in the sample data between different class clusters compared to that in the same class clusters, but it is not particularly prominent, which objectively results in the effect of taking k value based on Silhouette Coefficient is not ideal. Therefore, it is manually judged that the overall clustering effect reaches the best when k is 4. The visualization of tf-idf+Kmeans clustering results based on T-SNE dimensionality reduction is shown in Figure 2. Number the individual clusters as T0, T1, T2, T3.

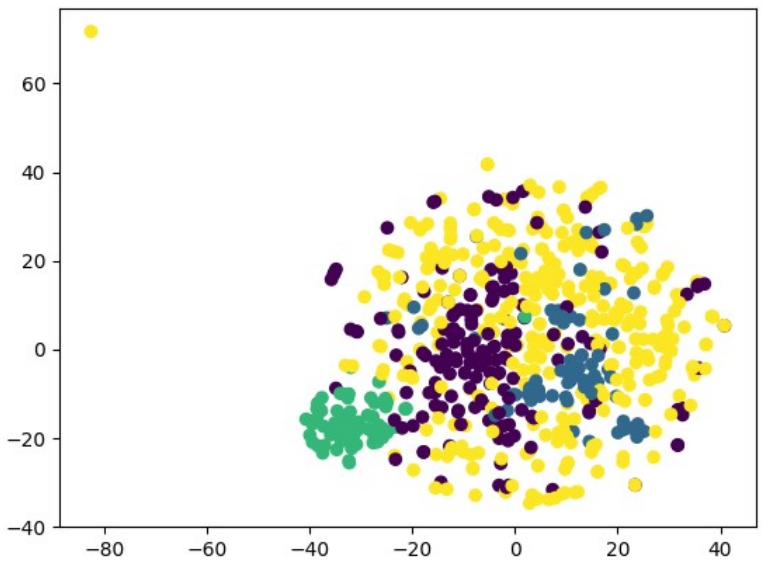


Fig. 2. Visualization of tf-idf+Kmeans clustering results based on T-SNE dimensionality reduction.

In Figure 2, the T3 category (yellow nodes) has the largest number of samples, which causes some interference to the correct categorization of the samples of the other clusters, and there is a problem of unclear boundaries with other clusters as well as mutual confusion. Therefore, its clustering results have some limitations. In contrast, the other three class clusters have clearer boundaries with each other.

Empirical Research: Multi-level clustering experiments

Topic. The LDA topic model is invoked using the gensim library. The number of passes through the corpus during training is 10, and the prior alpha of the document-topic distribution and the prior of the topic-word distribution are both set to auto. After manual judgment, the integers in the interval [5,25] are used as the candidate topic numbers for iterative experiments. Perplexity is calculated in each round of experiments, and the inflection point of perplexity is used to set the optimal number of topics, and the topic model achieves the best effect when the number of topics is 15. The corpus is transformed into a 2D matrix of 639*15. During the iterative clustering process, the difference in Silhouette Coefficient in each round of experiments is small, and when k is 14, the overall clustering effect is judged manually to be optimal. The visualization of LDA+Kmeans clustering results based on T-SNE dimensionality reduction is shown in Figure 3. Number the individual clusters as L0, L1, ..., L12, L13.

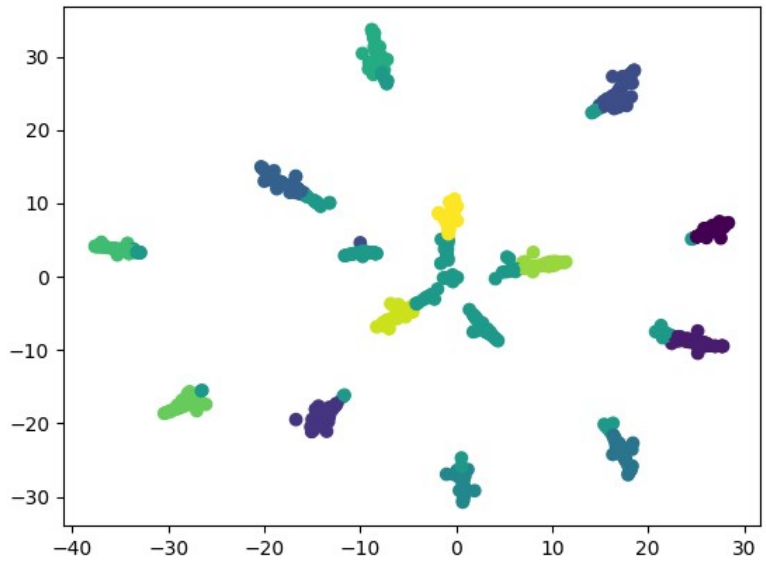


Fig. 3. Visualization of LDA+Kmeans clustering results based on T-SNE dimensionality reduction.

In Figure 3, the 15 topics modeled based on LDA form 15 "clusters" with obvious differentiation in the spatial vector distribution, except for L7 (dark green nodes), which has the largest number of samples and its samples are distributed in multiple "communities". The samples in the other clusters are distributed in relatively fixed "communities", and the differences in topic are more obvious.

Empirical Research: Multi-level clustering experiments

Semantic. The doc2vec sentence vector semantic training model is invoked using the gensim library. The sentence vector dimension is defined as 100 dimensions, and the window value is set to 3.The minimum word frequency is set to 1, and the number of iterations is set to 100. A word vector representation of 5,341 words is obtained, as well as a 2D matrix of 639*100. In the iterative clustering process, the difference between the Silhouette Coefficient in each round of experiments is still small, and when k is 7, the overall clustering effect is judged manually to be optimal. The visualization of doc2vec+Kmeans clustering results based on T-SNE dimensionality reduction is shown in Figure 4. Number the individual clusters as D0, D1, ..., D5, D6.

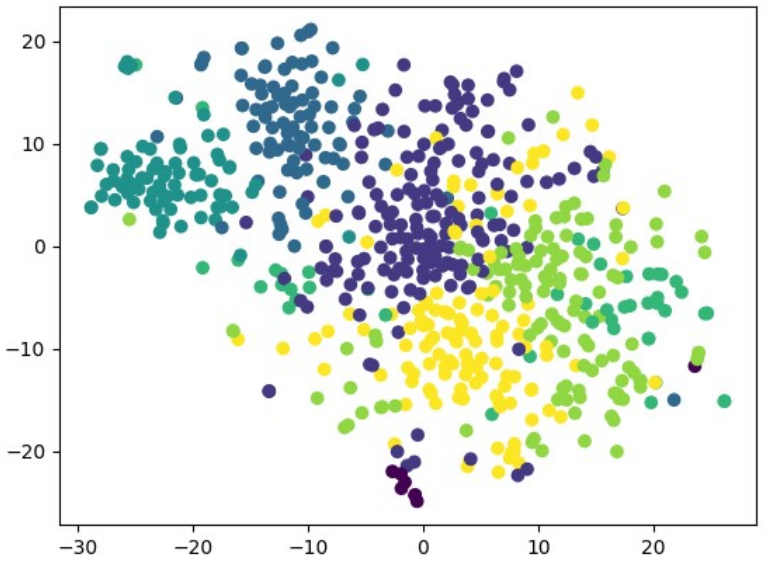


Fig. 4. Visualization of doc2vec+Kmeans clustering results based on T-SNE dimensionality reduction.

In Figure 4, the semantic-based clustering results are relatively the most obvious differentiation, and the main content of the literature within each category cluster is more unified. It can also be found that the distribution of samples of the same category is relatively centralized and clear, except for a certain amount of confusing interference at the boundaries of each cluster.

Empirical Research: Innovation points extraction experiments

The innovation points classification corpus contains 1982 positive samples for innovation points clauses and 4288 negative samples for non-innovation points clauses. The training set and test set are divided according to 9:1. The pretraining model is Google's open-source ALBERT large Chinese pre-training model, **Albert-large-zh**. The training parameters are adjusted as follows. The batch size is 128; the maximum text length is 128; the epoch is 10; the learning-rate is 0.00001. The performance of the innovation points auto-extraction model on the test set is evaluated using the performance measures commonly used in text categorization: **precision (P), recall (R), and the F1-score (F1)**, and the results of the calculations are retained in two decimal places. The evaluation results are shown in Table 2.

Table 2. The evaluation results of automatically innovation points extracted models.

Classification	Micro-P	Micro-R	Micro-F
innovation points	95.73%	79.29%	86.74%
non-innovation points	91.13%	98.36%	94.61%

The overall classiffcation accuracy of the auto-extraction model on the test set with 198 positive samples and 428 negative samples is **92.33%**, which achieves good results. The precision of automatic extraction of innovation points is higher, while the recall is **not yet 80%**. It shows that the extracted innovation points have high confidence, but it leads to about one-fifth omission. Although there are certain defects, the model can assist the automatic extraction of field innovation points.

Empirical Research: Constraction of academic innovation chain

Comparing the clusters to which different innovation points belong at different levels, there are 583 literatures with the same clusters as at least one literature whose **innovation points belong to the same clusters at the three levels of word frequency, topic and semantic**. They can be considered to be highly correlated with each other with unified multi-level features, and the combination of the three clusters at the three levels is called the **public cluster combination**. There are **119** such public cluster combinations such as "T3-L7-D3".

Taking the **public cluster combination "T3-L2-D4"** as an example, its academic innovation chain is visualized in Figure 5, in which the innovation points are basically the proposed special object-oriented knowledge element representation, extraction and association methods. It reflects the high correlation between a series of innovation points and also provides a visualization method for the comparative assessment of the differences between innovation points.

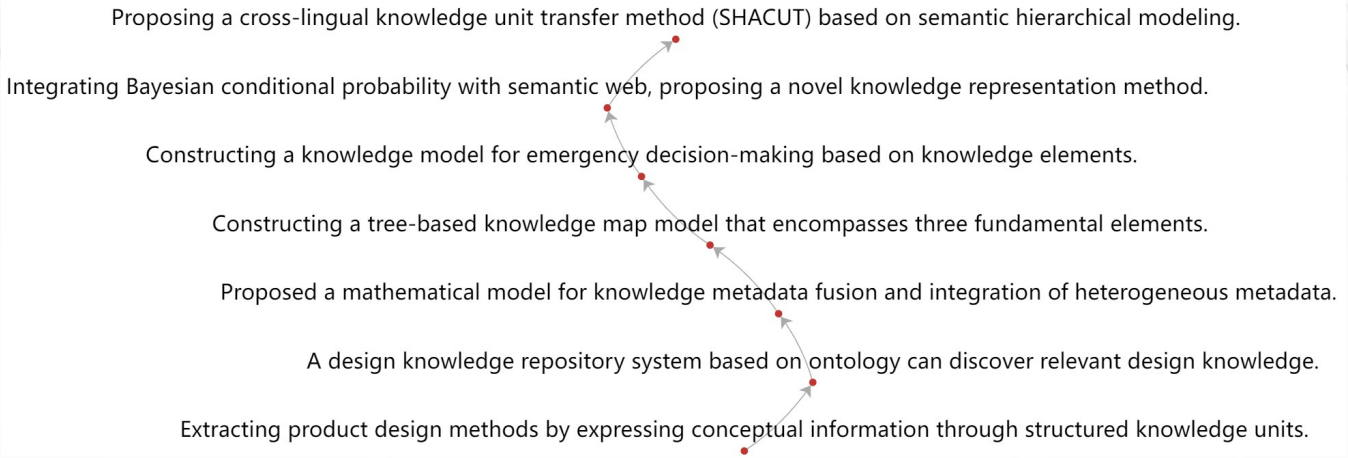


Fig. 5. Academic innovation chain based on "T3-L2-D4" public cluster combination.

Empirical Research: Constraction of academic innovation chain

Taking the innovation points of the literature **"Research on the Extraction Rules of Knowledge Element Contents in the South China Sea Documents of the Republic of China Period"** as an example, its localized academic innovation chain examples on **three feature levels** are shown in Figure 6, where the green connecting line indicates the correlation based on the word frequency feature, the red connecting line indicates the correlation based on the topic feature, and the blue connecting line indicates the correlation based on the semantic feature. In Figure 6, it shows other related innovation points. Such as the knowledge element entity and relationship extraction method based on word frequency features, the knowledge network fusion method based on topic features, and the knowledge graph construction model and integrated construction method based on semantic features. It reveals the potential correlation between the studies at different levels, and also intuitively reflects the essential differences between the related studies, highlight the different innovation points.

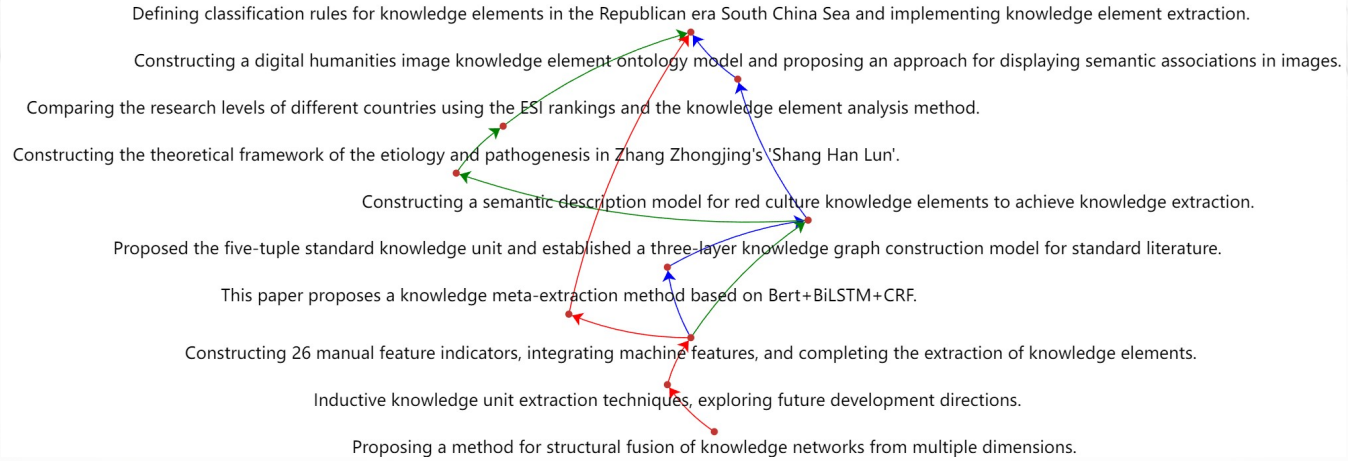


Fig. 6. Example of academic innovation chain based on multi-level clustering.

Empirical Research: Comparison experiment

References provide borrowing for new literature, and promote or facilitate the generation of new literature. Therefore, **citation networks** are often used for **literature association analysis**. As a comparative experiment, the citation networks of 639 literature are constructed as shown in Figure 7.

In Figure 7, the citation network can be used to quickly discover high-impact papers in the field, as well as to discover the linkages between the literature. It is able to better show the literature associations globally. Citation networks are also relatively easy to build. However, it is difficult to visualize the linkages between the literature at the content level, especially at the innovation points. And it is difficult to reflect the degree of differences in literature associations. Therefore, academic innovation chain can be used as a complement to citation networks to assist in field innovation knowledge discovery and assessment.

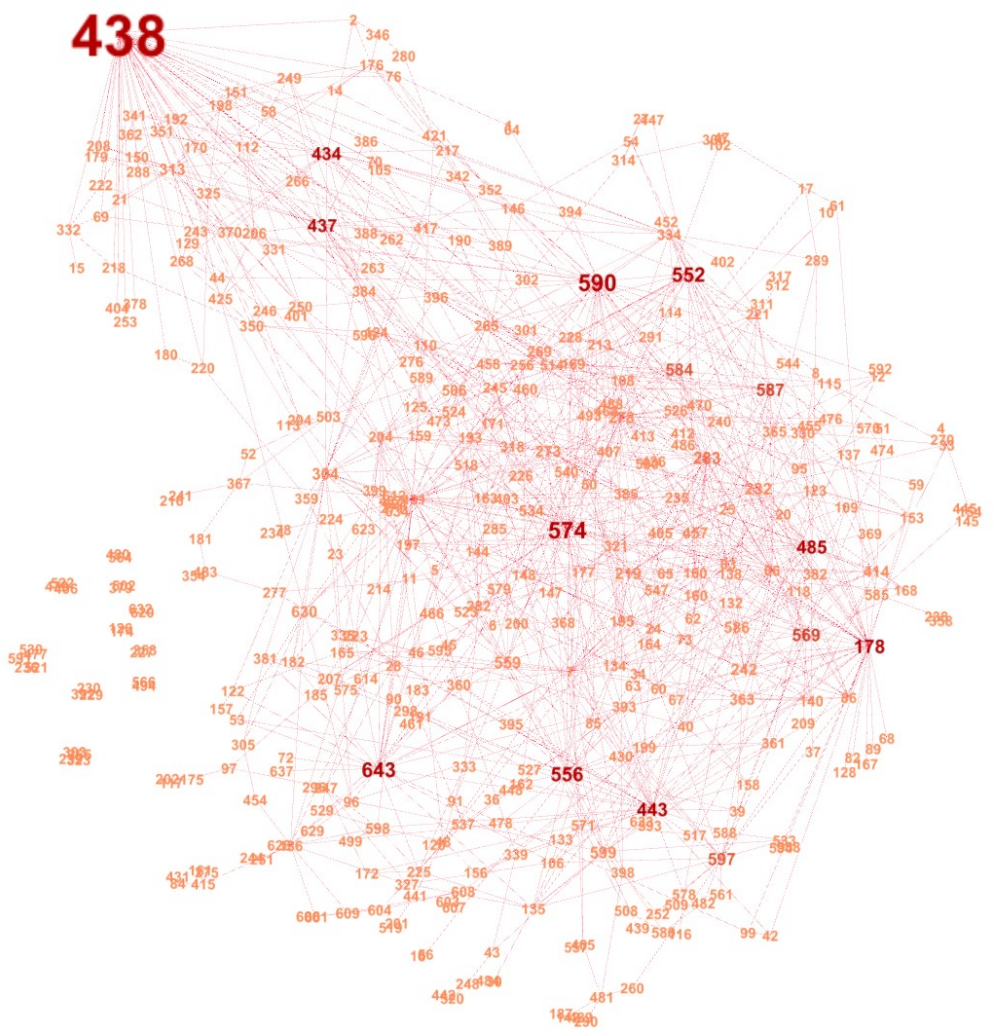


Fig. 7. Literature association network based on citation relationships.

Discussion

Academic innovation chains linearly correlate otherwise isolated innovation points. It can either organize highly relevant innovation points in an orderly manner based on the potential correlation of the unity of textual features at multiple levels, or organize relevant innovation points from the perspective of the correlation of textual features at different levels. **The academic innovation chain clearly and intuitively shows the evolution path and characteristic trend of innovation in chronological order, providing a visualization tool for the comparative analysis of the relevance and difference between innovation points in the field.** It can help scholars to grasp the historical research trend and inspire their innovation selection and discovery, and it can also assist experts to carry out innovation evaluation more efficiently and objectively. **In addition, when a batch of new field literature are published, the academic innovation chain based on multi-level clustering can be quickly updated.** It can not only detect the most cutting-edge innovations in the field in time, but also assess the innovation value based on the force of innovation points on the updating of the academic innovation chain.

Conclusion

This study explores the academic innovation chain construction method based on multi-level clustering of field literature. Combining the text feature mining algorithms of **tf-idf, LDA, doc2vec and the Kmeans text clustering algorithm**, the field literature is clustered from the three levels of **word frequency, topic and semantic**, respectively. Then the academic innovation chain under the time dimension is constructed on the basis of innovation points extraction by **rule method and ALBERT pre-training model**. The feasibility of the method is verified by taking the knowledge element field as an example. The potential value of academic innovation chain for innovation discovery and innovation assessment is explored, which provides a new perspective for related research. **This study also has the following limitations.** More algorithms are not introduced in the clustering process for comparative experiments to achieve better clustering results. In the clustering process, only the title and abstract text are used, and the **full text of literature**, which contains richer semantic information, is not used. This will be the direction of improvement for subsequent research. Further, in the subsequent experiments we will use **larger scale data** for more comprehensive innovative knowledge mining and organization to build a practical experimental platform. We will also optimize the construction method and display form of the academic innovation chain through quantitative evaluation and qualitative evaluation.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3, 993–1022 (2003)
2. Boyack, K.W., Klavans, R.: Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for information Science and Technology 61, 2389–2404 (2010)
3. Cui, J., Wang, Z., Ho, S.B., Cambria, E.: Survey on sentiment analysis: evolution of research methods and topics. Artificial Intelligence Review 56, 8469–8510 (2023)
4. Curiskis, S.A., Drake, B., Osborn, T.R., Kennedy, P.J.: An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. Information Processing Management 57, 102034 (2020)
5. Ghosal, T., Edithal, V., Ekbal, A., Bhattacharyya, P., Chivukula, S.S.S.K., Tsatsaronis, G.: Is your document novel? let attention guide you. an attention-based model for document-level novelty detection. Information Processing Management 27, 427–454 (2021)
6. Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., Bhattacharyya, P.: Novelty detection: A perspective from natural language processing. Computational Linguistics 48, 77–117 (2022)
7. Grames, E.M., Stillman, A.N., Tingley, M.W., Elphick, C.S.: An automated approach to identifying search terms for systematic reviews using keyword cooccurrence networks. Methods in Ecology and Evolution 10, 1645–1654 (2022)
8. Huang, J., Gong, S., Zhu, X.: Deep semantic clustering by partition confidence maximisation (2020)
9. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimedia Tools and Applications 78, 15169–15211 (2019)
10. Kim, D., Seo, D., Cho, S., Kang, P.: Multi-co-training for document classiffcation using various document representations: Tf–idf, lda, and doc2vec. Information sciences 477, 15–29 (2019)
11. Kim, S., Park, H., Lee, J.: Word2vec-based latent semantic analysis (w2v-lsa) for topic modeling: A study on blockchain technology trend analysis. Expert Systems with Applications 152, 113401 (2020)
12. Kleminski, R., Kazienko, P., Kajdanowicz, T.: Analysis of direct citation, cocitation and bibliographic coupling in scientiffc topic identiffcation. Journal of Information Science 48, 349–373 (2022)
13. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. International Journal 1, 90–95 (2013)
14. Li, J., Izakian, H., Pedrycz, W., Jamal, I.: Clustering-based anomaly detection in multivariate time series data. Applied Soft Computing 100, 106919 (2021)
15. Onan, A.: Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. IEEE Access 7, 145614–145633 (2019)
16. Pen, H., Wang, Q., Wang, Z.: Boundary precedence image inpainting method based on self-organizing maps. Knowledge-Based Systems 216, 106722 (2021)
17. Shahapure, K.R., Nicholas, C.: Cluster quality analysis using silhouette score. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). pp. 747–748. IEEE (2020)
18. Steinley, D.: K-means clustering: a half-century synthesis. British Journal of Mathematical and Statistical Psychology 59, 1–34 (2006)
19. Wang, J., Ma, X., Zhao, Y., Zhao, J., Heydari, M.: Impact of scientiffc and technological innovation policies on innovation efffciency of high-technology industrial parks–a dual analysis with linear regression and qca. International Journal of Innovation Studies 6, 169–182 (2022)
20. Wang, X., Wang, H., Huang, H.: Evolutionary exploration and comparative analysis of the research topic networks in information disciplines. Scientometrics 126, 4991–5017 (2021)
21. Wang, Z., Tong, V.J.C., Xin, X., Chin, H.C.: Anomaly detection through enhanced sentiment analysis on social media data. In: 2014 IEEE 6th international conference on cloud computing technology and science. pp. 917–922. IEEE (2014)
22. Wei, X., Shen, L.: A research review of the academic paper innovativeness. Documentation,Information Knowledge 39, 68–79 (2022)

Thanks For Listening

Construction of Academic Innovation Chain Based on Multi-level Clustering of Field Literature

Wei Cheng; Tianshi Cong
Nanjing Agricultural University
(email: chengwei@stu.njau.edu.cn)

May 7, 2024