A research topic evolution prediction approach based on multiplex-graph representation learning *

Yang Zheng[1], Kaiwen Shi[2], Yuhang Dong[1], XiaoGuang Wang[3] and Hongyu Wang[1,*]

[1] School of Management, Wuhan University of Technology, Wuhan, China
[2] School of Information Engineering, Zhongnan University of Economics and Law, Wuhan, China
[3] School of Information Management, Wuhan University, Wuhan, China

Abstract

The intensification of international technological innovation competition and the evolution of scientific research paradigms have led to a continuous expansion of scientific literature, making information analysis increasingly complex and diversified. To address the challenges of accurately assessing topic evolution within the context of vast literature big data, traditional methods of expert evaluation or visualization analysis based on scientific knowledge networks are inadequate. From the perspectives of artificial intelligence and big data, this paper proposes a universal method for automated and intelligent discrimination and prediction of research topic evolution hotness. This method involves integrating content and structural features of keywords to track the evolution of keyword frequency strength over time in research topic networks characterized by keywords. This study conducts a case analysis in the field of information science. The results demonstrate that the prediction of keyword strength is improved after integrating content and structural features, which has significant reference value for tasks such as future research topic evolution trend discrimination, research direction, and policy planning.

Keywords

topic evolution, keyword citation network, text mining, graph representation learning

1. Introduction

With the intensification of international technological innovation competition and the evolution of the fourth paradigm of scientific research driven by big data development, the growing volume of scientific literature, shifting scholarly interests, and the emergence of new research topics pose significant challenges to traditional methods of research topic analysis[1,2,3]. How to comprehensively and finely reveal the research topics and their characteristic keywords representing knowledge innovation within a vast array of scientific literature, track the evolution of research topics, and represent them on multiple knowledge networks that contain knowledge units and their complex interactions, so as to judge the future evolutionary trends of research topics, is a key direction for science and technology information construction and services, as well as a research focus in the fields of informetrics and scientometrics.

Current analyses of the evolution of scientific research topics largely unfold across three dimensions: content, structure, and strength[4,5]. Utilizing the powerful representation and feature learning capabilities of deep representation learning algorithms such as word embedding and graph embedding[6], it is possible to model the complex nonlinear relationships between entities represented by keywords, the smallest units of knowledge[7,8]. Tracking changes in topic strength, as indicated by

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 4th AI + Informetrics (EEKE-AII2024), April 22, 2024, Changchun, China
* Corresponding author.
zhengyang2002@whut.edu.cn (Y. Zheng); shikaiwen@stu.zuel.edu.cn (K. Shi); dongyuhang@whut.edu.cn (Y. Dong); wxguang@whu.edu.cn (X. Wang); hongyuwang@whut.edu.cn (H. Wang)

0000-0001-5635-1131 (Y. Zheng); 0000-0002-3563-982X (K. Shi); 0009-0005-4618-5906 (Y. Dong); 0000-0003-1284-7164 (X. Wang); 0000-0002-5063-9166 (H. Wang)

keyword frequency over time, can reflect the evolving trends of topics[9,10,11].

Therefore, this paper utilizes a multiplex-graph representation learning method combined with interactions in topic keyword content and structure to assess changes in topic strength, achieving prediction of topic evolution hotness. And select the field of "Information Science" for case analysis, aiming to address the following two scientific questions:

1. How to reveal the evolution process of topics at the micro level, thereby tracking the evolutionary trends of research topics?
2. How to effectively and comprehensively model and integrate the multi-dimensional features of research topic evolution representations on knowledge networks? After tracking these multi-dimensional evolutionary features, will the assessment of topic evolution trends become more accurate?

2. Methodology

This study aims to predict the strength of keyword frequency, using changes in keyword frequency strength to reflect variation in topic evolution hotness.

The research process is divided into 3 steps: Step 1 involves retrieving and cleaning the source data to obtain all the data needed for subsequent experiment. Step 2 involves obtaining content, structure, and strength representations of keywords. Step 3 utilizes deep learning models to integrate multi-dimensional data representations and conduct prediction of topic evolution hotness with the integrated experimental data. Figure 1 shows the detailed research process.

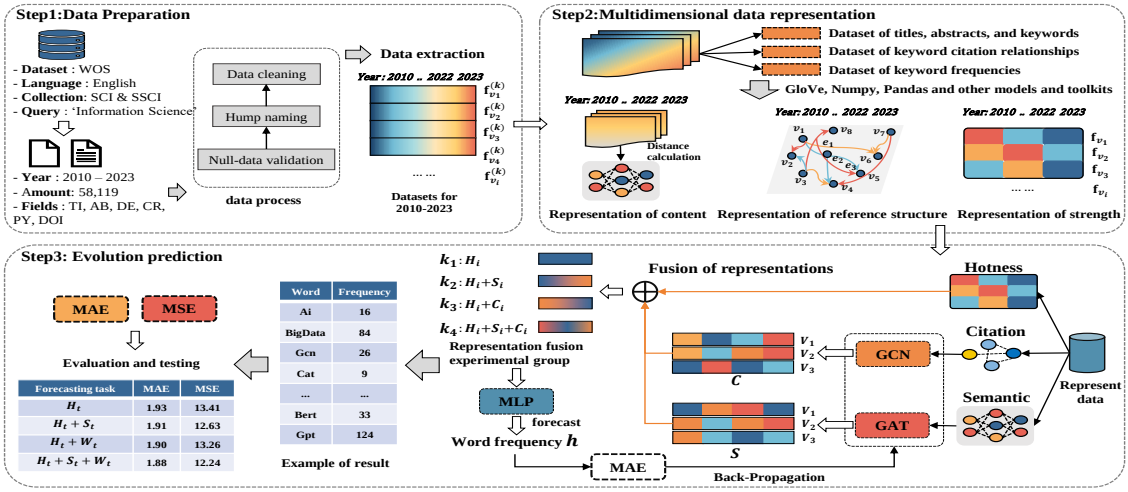

Figure 1: Research process.

2.1. Data preparation

Field-specific literature is selected from databases such as Web of Science and Scopus, with titles, keywords, abstracts, and references extracted as basic data. After cleaning and filtering the data, an original dataset U is constructed, which specifically includes the keyword citation relationship dataset C, the keyword frequency dataset K, and the integrated dataset N containing titles, abstracts, and keywords.

2.2. Multi-dimensional feature extraction

To prepare for multi-dimensional feature integration, this study will perform feature extraction on keyword data across three representational dimensions: content, structure, and strength of keywords.

（1）Content feature extraction

In the process of extracting keyword content features, this study opts to use the *GloVe* static word embedding method to capture the semantic relationships of keywords in their global context[12], facilitating the embedding of keywords, as it offers greater stability and requires less computational resources[13]. The principle is as shown in 2-1 and 2-2, where X_{ik} is the number of times word k appears in the context of word i, X_i is the total number of words appearing in the context of word i, and P_{ij} is the probability of word j appeare in the context of word i.

$$X_i = \sum_k X_{ik} \tag{2-1}$$

$$P_{ij} = P(j \mid i) = \frac{X_{ij}}{X_i} \tag{2-2}$$

The generated word vector is then used to calculate the cosine similarity between words using formula 2-3. This process results in obtaining the

semantic distance matrix $ES = ES_t = \left(es_{i,j}^t\right), i = j$, for keyword content feature extraction.

$$A = [a_1, a_2, \dots, a_n], \mathrm{B} = [b_1, b_2, \dots, b_n]$$
$$Distance = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (2\text{-}3)$$

（2）Structure feature extraction

Some scholars have proposed using a "keyword-citation-keyword" method to construct keyword citation networks[14], meaning when literature T_1 cites literature T_2, there exists a "Keyword-Cartesian product mapping" citation relationship between the keywords of the two literatures. The specific principle is shown in Figure 2. Based on this theory, this study constructs a keyword citation network with citation frequency as edge weight, resulting in a keyword citation matrix $EC = EC_t = \left(ec_{i,j}^t\right), i = j$.

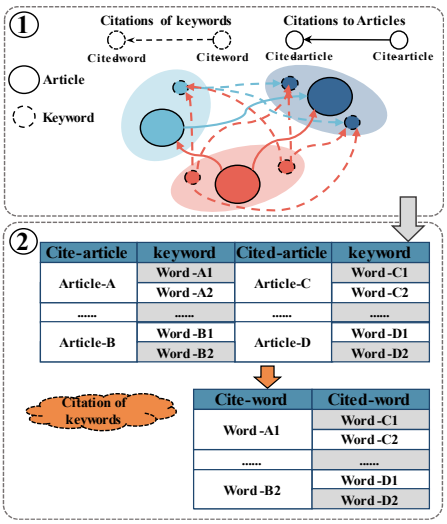

Figure2: Construction of keyword citation network

（3）Strength feature extraction

The size of keyword frequency reflects the strength of the keyword. In generating the keyword frequency matrix, this study chooses to use *Numpy* and *Pandas* packages to process the keyword frequencies k_w^t in dataset K, constructing a "frequency-year" frequency matrix $EH = EH_t = (eh_w^t)$. While extracting strength features, this also generates the strength representation H of keywords.

2.3. Model construction and prediction

Graph Attention Network(GAT)based on the attention mechanism, can effectively capture complex semantic dependencies between keywords, while Graph Convolutional Network(GCN) can efficiently process the strcture information of the graph structure itself by aggregating the features of neighboring nodes. Therefore, this study chooses to use GAT and GCN graph neural network models to capture the relationships between nodes in graph-structured data from content and structure perspectives, respectively, and employs an Multilayer Perceptron(MLP) regression model to integrate multi-dimensional features of keywords for strength prediction.

This study constructs an ablation experiment group, as shown in Table 1, for predicting the hotness of topic evolution. And details of the model settings are shown in Figure 1. It's worth noting that set an initial identity matrix allows the neural network to gradually adjust and optimize feature representations during the learning process. Therefore, after obtaining the content matrix ES and the structure matrix EC, GAT and GCN are used to perform convolution operations on these two matrices on a predefined 50-dimensional identity matrix. After obtaining content representation S and structure representation C of keywords, the three representation data are concatenated directly for integration, and prediction is made based on MLP[15].

Table 1
Deep Learning model group setting

Tasks	Model	Composition
$Group_1$	$Model_1$	MLP
$Group_2$	$Model_2$	$GAT + MLP$
$Group_3$	$Model_3$	$GCN + MLP$
$Group_4$	$Model_4$	$GAT + GCN + MLP$

After obtaining the prediction results, the study chooses to use two metrics, Mean Square Error (MSE) and Mean Absolute Error (MAE), to measure the predictive capability of the model[16,17]. The specific formulas are as follows.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (2\text{-}4)$$
$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (2\text{-}5)$$

Above, y_i represents the ith element of y, and n is the number of elements.

3. Experiment

The detailed process of data acquisition can be found in the appendix under $B. Data\ Resources$.

3.1. Experiment Preparation

When extracting features from experimental data, this study choose to work with four sets of yearly data:

2017-2019, 2018-2020, 2019-2021, and 2020-2022. The first three sets are used as the training groups, and the last set as the test group. Specifically, the data from 2019 to 2022 serve as the basis for operations (all operations on yearly data will follow this four-year standard). Taking 2019 as an example, semantic content matrix ES, citation structure matrix EC, and frequency strength matrix EH are constructed for that year's keyword data.

3.2. Model training and prediction

The objective is to determine the optimal parameters for various model groups in order to ensure accurate predictions. This study has set four hyperparameters: learning rate (1e-2, 1e-3, 1e-4), number of training epochs (10, 50, 100, 200), hidden layer dimensions (10, 30, 50), and stopping steps (5, 10, 20), using the training data from 2019 to 2021 to train models across four experimental groups, and evaluating the final MAE and MSE results on the training set to determine the most suitable hyperparameters for each model. The optimal parameter settings for different models are shown in Table 2.

Table 2
Optimal setting of model parameters

Model	LearningRate	Epoch	EarlyStop	HiddenDim
$Model_1$	0.01	100	20	50
$Model_2$	0.001	200	20	50
$Model_3$	0.01	200	20	50
$Model_4$	0.01	100	20	50

Utilizing the model groups above and employing the test data of 2022 to conduct the prediction of topic evolution hotness for 2023.

3.3. Results and Discusstion

The prediction results are evaluated using two indicators: MAE and MSE, with the evaluation results listed in Table 3. From the table, it can be observed that using only the strength representation H for predicting topic hotness, the MAE and MSE between the predicted and actual values are 1.93484 and 13.41658, respectively. However, after integrating the content representation S or structure representation C, the values of MAE and MSE both decrease. The best result for predicting topic hotness are achieved by integrating all three types of representations, resulting in the lowest values of MAE and MSE.

The results indicate that predicting the evolution hotness of research topics by integrating multi-dimensional features such as content and structure through multiplex-graph representation learning is more accurate than traditional prediction methods.

Table 3
Evaluation of research topic evolution hotness prediction results in 2023

Forecasting task	MAE	MSE
H_t	1.93484	13.41658
$H_t + C_t$	1.91702	12.63112
$H_t + S_t$	1.90118	13.26141
$H_t + S_t + C_t$	1.88017	12.24382

4. Conclusion

This study proposes a novel approach based on multiplex-graph representation learning to predict the evolution of research topics. And the contributions are follows: First, in feature modeling, GCN and GAT graph neural network models are used to perform convolution operations on content and structure features on unit matrices of specified dimensions, adaptively aligning data across different dimensions and time windows to ensure comparability. Second, this study integrates semantic content features, citation structure features, and frequency strength features of keywords for research topic hotness prediction, showcasing the interaction between knowledge structures and cognitive structures from a multidimensional perspective, offering a deeper insight into predicting research topic evolution hotness. Third, after integrating content and structure features, a domain case analysis is conducted, and the result indicates that combining these two types of features indeed makes the prediction of research topic evolution hotness more accurate.

Owing to the desire to directly validate whether integrating multiple representations of topic evolution enhances the accuracy of topic evolution analysis, this paper choose to predict the future frequency of topic keywords, which has certain limitations. Subsequent tasks such as research topic trend discrimination, research direction, and policy planning can be developed based on the effective analysis results of this study.

Acknowledgements

This work was funded by the National Natural Science Fund of China (No. 71874129), the Open-end Fund of Information Engineering Lab of ISTIC and the Independent Innovation Foundation of Wuhan University of Technology (No. 233103002).

References

[1] Zhu, Hengmin, et al. "Evolution analysis of online topics based on 'word-topic'coupling network." Scientometrics 127.7 (2022): 3767-3792.

[2] Hu, Kai, et al. "Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis." Information Processing & Management 56.4 (2019): 1185-1203.

[3] Huo, Chaoguang, Shutian Ma, and Xiaozhong Liu. "Hotness prediction of scientific topics based on a bibliographic knowledge graph." Information Processing & Management 59.4 (2022): 102980.

[4] Z. Liu, X. Wang, and R. Bai. "Research on Visualization Analysis Method of Discipline Topics Evolution from the Perspective of Multi Dimensions:A Case Study of the Big Data in the Field of Library and Information Science in China." Journal of Library Science in China 42.6 (2016): 67-84. (in Chinese)

[5] K. Cui. The Research and lmplementation of Topic Evolution Based on LDA. Diss. National University of Defense Technology 2010. (in Chinese)

[6] Zhou, Yuan, et al. "A deep learning framework to early identify emerging technologies in large-scale outlier patents: An empirical study of CNC machine tool." Scientometrics 126 (2021): 969-994.

[7] Şenel, Lütfi Kerem, et al. "Learning interpretable word embeddings via bidirectional alignment of dimensions with semantic concepts." Information Processing & Management 59.3 (2022): 102925.

[8] Shi, Bin, et al. "RelaGraph: Improving embedding on small-scale sparse knowledge graphs by neighborhood relations." Information Processing & Management 60.5 (2023): 103447.

[9] Raamkumar, Aravind Sesagiri, Schubert Foo, and Natalie Pang. "Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems." Information Processing & Management 53.3 (2017): 577-594.

[10] Yoon, Young Seog, et al. "Exploring the dynamic knowledge structure of studies on the Internet of things: Keyword analysis." ETRI Journal 40.6 (2018): 745-758.

[11] Ohniwa, Ryosuke L., and Aiko Hibino. "Generating process of emerging topics in the life sciences." Scientometrics 121.3 (2019): 1549-1561.

[12] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

[13] Wang, Yuxuan, et al. "From static to dynamic word representations: a survey." International Journal of Machine Learning and Cybernetics 11 (2020): 1611-1630.

[14] Q. Chen, J. Wang, and W. Lu. "Discovering Domain Vocabularies Based on Citation Co-word Network" Data Analysis and Knowledge Discovery 3.6 (2019): 57-65. (in Chinese)

[15] Liu, Weijia, et al. "Category-universal witness discovery with attention mechanism in social network." Information Processing & Management 59.4 (2022): 102947.

[16] Yan, Yuwei, et al. "Data mining of customer choice behavior in internet of things within relationship network." International Journal of Information Management 50 (2020): 566-574.

[17] Gandhudi, Manoranjan, et al. "Causal aware parameterized quantum stochastic gradient descent for analyzing marketing advertisements and sales forecasting." Information Processing & Management 60.5 (2023): 103473.

A. Online Resources

The resources of this article can be downloaded at https://github.com/Hipkevin/EEKE-hotness.

B. Data resources

This study uses "Information Science" as a case study topic, selecting the SCI and SSCI core databases in WOS. Conducting literature searches in the "Information Science & Library Science" field with the search query "Document Types: Article or Review Article; Languages: English," ultimately selecting literature from 2010 to 2023, totaling 58,119 articles, as experimental data.