Scientometrics (2024) 129:4021–4042
https://doi.org/10.1007/s11192-024-05068-2

A deep learning-based method for predicting the emerging degree of research topics using emerging index

Zhenyu Yang[1] · **Wenyu Zhang**[1] · **Zhimin Wang**[2] · **Xiaoling Huang**[2]

Received: 31 July 2023 / Accepted: 21 May 2024 / Published online: 14 June 2024

Abstract

With the exponential growth of the volume of scientific literature, it is particularly important to grasp the research frontier. Predicting emerging research topics will help research institutions and scholars promptly discover promising research topics. However, previous studies mainly focused on identifying and detecting emerging research topics and lacked a method to efficiently represent and predict the emerging degree of research topics. Therefore, this study proposes a novel deep learning-based method to predict the emerging degree of research topics. First, a new indicator, the *emerging index*, is proposed based on the emerging attributes such as novelty, growth, and impact to quantitatively measure the emerging degree of research topics. Second, new features reflecting the emerging attributes of the research topics are extracted by constructing heterogeneous networks of bibliographic entities in the research domain. Finally, a deep learning-based time series model was employed to predict the future *emerging index* based on these new features. Data from the neoplasms and metabolism research domains in the PubMed Central database were used to validate the proposed method. The experimental results showed that the *emerging index* proposed effectively measures the emerging degree of the research topics. Furthermore, the deep learning-based model demonstrates superior performance to other models in predicting the *emerging index*, as evidenced by both error-based and rank-based metrics.

Keywords Emerging topics prediction · Heterogeneous networks · Deep learning · Emerging index

✉ Xiaoling Huang
huangxl@zufe.edu.cn

Zhenyu Yang
yangzhenyu1@zufe.edu.cn

Wenyu Zhang
wyzhang@e.ntu.edu.sg

Zhimin Wang
wzm@zufe.edu.cn

[1] School of Information Technology and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou 310018, China

[2] Library, Zhejiang University of Finance and Economics, Hangzhou 310018, China

Introduction

It is important for research institutions and scholars to identify emerging research topics in a timely manner, as it can provide a reference for finding prospective research directions and help the rational allocation of researchers and research funds (Behrouzi et al., 2020; Wang, 2018; Yang et al., 2022a). Several recent research projects, such as the European Research Council (ERC) supported emerging research areas and their coverage in 2009, and the Intelligence Advanced Research Projects Activity (IARPA) funded Foresight and Understanding from Scientific Exposition (FUSE) in 2011 are evidence of this growing attention (McKeown et al., 2016; Small et al., 2014).

Although there is currently no clear definition of the concept of emerging research topics, it has been observed that such attributes as novelty, growth, and impact are often seen as attributes that emerging research topics should have (Liang et al., 2021; Xu et al., 2019; Zhang et al., 2021). Studies on emerging research topics have mainly focused on the design of bibliometric indicators related to these attributes, and leveraged these indicators to identify and detect emerging research topics (Gao et al., 2021; Wang, 2018; Xu et al., 2021; Yang et al., 2022a).

Some studies treat the prediction problem of emerging research topics as a statistical learning problem, and train a predictive model using historical feature data as input and target indicators as output to achieve the prediction task of emerging research topics (Asooja et al., 2016; Kwon et al., 2019; McKeown et al., 2016). On one hand, for the choice of target indicators, these studies tend to choose the indicators that are relatively easy to obtain, such as the number of citations, term frequency and term frequency-inverse document frequency (TF-IDF). However, due to the multidimensional nature of emerging research topics, these intuitive indicators cannot fully reflect the emerging degree of research topics. On the other hand, for the choice of explanatory features, these studies usually consider the metadata of the research topics, such as the document frequency and term frequency. However, only considering the use of metadata as feature data does not reflect the emerging attributes of the research topics completely (Lu et al., 2021). In addition, it is difficult for traditional statistics-based prediction models to capture the complex relationships between the multivariate nonlinear data structures.

Therefore, to fill the above gap, this study proposes a deep learning-based prediction method of emerging research topics. Firstly, a new indicator, i.e., *emerging index*, is proposed to quantitatively measure the emerging degree of research topics, by considering the emerging attributes such as novelty, growth, and impact simultaneously. This new indicator enriches the existing assessment framework and equips academia and policymakers with a more effective tool for identifying and tracking emerging research topics. Secondly, the heterogeneous networks of bibliographic entities in a research domain are constructed to extract the new features that can reflect the emerging attributes of the research topics more precisely. This approach breaks through the limitation of relying solely on metadata as feature data, offering a new perspective for in-depth exploration of emerging research trends. Finally, the long short-term memory (LSTM) neural network, which is a deep learning-based time series model, is employed to represent the feature sequence and predict the future *emerging index* value of the candidate topics in the next two years. The data in the domains of neoplasms and metabolism from PubMed Central (PMC) database[1] were used

[1] https://www.ncbi.nlm.nih.gov/pmc/.

to conduct the in-depth comparative experiments, which verify the effectiveness and feasibility of the proposed method. In term of both error-based and rank-based metrics, the deep learning-based model achieves better performance in predicting the *emerging index* than other comparative models. This result highlights the advanced and practical nature of the proposed approach to the prediction of emerging research topics.

The remainder of the article is divided as follows: "Related work" section reviews the concept of emerging research topics, the works on identifying emerging research topics, and the works on predicting emerging research topics. "Methodology" section presents the proposed methodology for predicting the emerging degree of research topics. "Experimental setups and results" section illustrates the experimental setup and result analysis. Finally, "Conclusion" section concludes the research implications, research limitations and future work.

Related work

The concept of emerging research topics

Price (1965) was the first to develop the concept of research frontier and described it as a growth tip or epidermal layer, which emphasized novelty. Small and Griffith (1974) described research frontiers as highly interactive groupings of literature. Rotolo et al. (2015) evaluated previous work and defined an emerging technology as having five attributes: radical novelty, relatively rapid growth, coherence, prominent impact, and uncertainty and ambiguity. This concept has been widely recognized as a comprehensive definition of emerging technology, and thus survived to today. Wang (2018) gave a new definition of emerging topics based on the research of Rotolo et al. (2015), describing it as a novel and relatively rapidly developing research topic with some coherence and considerable scientific impact. Because the current study focuses on the specific research domain, the qualified coherence can be assured and need not be explored further. Therefore, based on Wang (2018), this study defines emerging research topics as those that are radically novel, developed quickly and have certain scientific impact. This definition employs three attributes to describe emerging topics: novelty, growth, and impact.

Identification of emerging research topics

The identification of emerging topics is mainly divided into three categories: methods based on words, methods based on citation network analysis, and hybrid methods. The methods based on words identify the emerging research topics by analyzing the text of publication. Liu et al. (2015) used natural language processing techniques to mine and analyze bibliographic entities such as topics and keywords in the literature to identify and explore emerging research topics. Weismayer and Pezenka (2017) used automatic longitudinal latent semantic analysis for terms to identify emerging research topics. Ma et al. (2021) used Latent Dirichlet Allocation (LDA) topic modeling and semantic Subject-Action-Object (SAO) analysis, and combined it with expert judgment and classification algorithms in machine learning to identify the emerging technical topics and potential development opportunities.

The methods based on citation network analysis can be further divided into direct citation network analysis (Waltman & Van Eck, 2012), co-citation network analysis (Small, 1973; Small & Griffith, 1974), and bibliographic coupling network analysis (Huang & Chang, 2014). Small et al. (2014) used both direct citation and co-citation analysis and identified emerging research topics using a difference function that combines novelty and growth. Wang (2018) used the direct citation model and proposed a set of rules to filter emerging research topics. Kwon et al. (2019) used the direct citation to verify that the degree of emerging technologies in a paper is proportional to its future impact. Hassan et al. (2018) used 64 dimensional indicators and combined citation analysis and deep learning to measure the importance of the research topic. In our previous work, Shi et al. (2018) combined direct citation network analysis and P-rank algorithm to estimate the interdisciplinarity in knowledge flow.

The hybrid methods combined the methods based on words and the methods based on citation network analysis networks to extract emerging research topics. Chen (2006) developed CiteSpace II combining citation network with text keywords to detect and visualize emerging trends in literature. Liu and Porter (2020) used a topic model to extract research topics from abstracts and titles, and developed a set of technology emergence indicators from multiple dimensions using citation networks to identify emerging research topics. Song et al. (2023) integrated bidirectional encoder representations from transformers (BERT) model and combined it with SAO semantic analysis to propose an emerging topic recognition framework.

Although the above studies have proposed methods that can identify emerging research topics from datasets, they cannot predict emerging research topics due to the delay in publication and citation. Therefore, the applicability of these studies is limited.

Predicting emerging research topics

Predicting research topics is a popular research field in information dissemination. Unlike topic identification, topic prediction requires indicators to quantify the attributes of research topics. Previous studies on research topics have focused on citation prediction or topic prediction based on the historical evolution of topics. For example, Xu et al. (2020) used the number of citations as an indicator of the scientific impact of research topics and leveraged bibliographic heterogeneous networks to anticipate the evolution of science impact, but the emerging degree could not be accurately represented. Similarly, McKeown et al. (2016) measured the impact of research topics as the exponentially weighted average of the numbers of relevant documents and predicted changes in impact using full-text characteristics and logistic regression models.

The keyword frequency has also been used as a measure of predicting emerging research topics. For example, to predict the emerging research trends, Asooja et al. (2016) used TF-IDF to quantify the annual relevance of each keyword, and utilized the regression-based methods to predict the distribution of keywords, but the emerging degree could not be comprehensively represented. Clearly, while the number of citations, the numbers of relevant documents and TF-IDF can all reflect the importance of research topics within a domain, they struggle to accurately represent the emerging degree of research topics due to their multidimensionality.

Multi-step methods can be used to improve the predictive accuracy of emerging research topics by separating and integrating emerging features. For example, to predict the emerging research trends, Xu et al. (2019) quantified the emerging attributes of candidate research topics, used the support vector regressor to predict each attribute separately, and then judged the emerging degree of research topics by analyzing whether the predicted

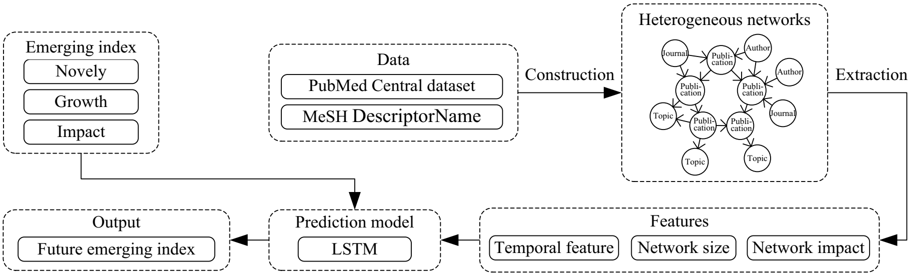

Fig. 1 Methodological framework for prediction of emerging research topics

value of each attribute was above the average level. Zhou et al. (2020) divided emerging technologies into seven dimensions and predicted emerging technologies from patent data by combining data enhancement and deep learning. Wei et al. (2022) used cover papers to mark emerging research topics, and used machine learning to judge whether the research topics would be emerging in the future from three types of features. Liang et al. (2021) separated the emerging attributes, applied two deep neural networks to predict the integrated indicator of growth and impact, and then used bibliometric methods to screen for novelty of research topics. These methods can more accurately predict emerging research topics, but they cannot measure the degree of emerging research topics. Therefore, the above methods cannot effectively reflect and predict the emerging degree of research topics.

Methodology

In order to improve the predictive efficiency of emerging research topics, this study proposes a deep learning-based method to predict the emerging degree of research topics. The overall framework is shown in Fig. 1. Firstly, a new indicator, i.e., *emerging index*, is proposed based on emerging attributes of novelty, growth, and impact, which can quantitatively measure the emerging degree of research topics. Then, considering that the traditional features such as metadata of documents cannot represent the research topics consistently and comprehensively, this study constructs the heterogeneous networks of bibliographic entities in a research domain to extract the new features such as temporal feature, network size, and network impact. Finally, a deep learning-based time series predictive model LSTM is employed to predict the future *emerging index* in the next two years by using the extracted features as the input. In addition, this study selects the PubMed Central (PMC) dataset as the data retrieval source, and the *DescriptorName* in the Medical Subject Headings (MeSH) thesaurus is used to denote the research topics of publications. MeSH thesaurus[2] is a hierarchically-organized vocabulary of biomedical and health-related information, where *DescriptorName* is the main heading of the MeSH thesaurus.

[2] https://www.nlm.nih.gov/mesh/meshhome.html.

Research topic selection

The research topics include research questions, methods, concepts, and techniques that are relevant to the discipline being studied by the researcher (Braam et al., 1991). Methods of extracting research topics using clustering (Wang, 2018), topic models (Xu et al., 2019), and natural language processing (Song et al., 2023) can be adapted to a variety of research topics. However, in the biomedical field, where research topics require uniform terminology to avoid ambiguity, these methods rely on human judgment and labeling. In order to reduce this reliance, this study uses *DescriptorName* in the MeSH thesaurus to represent the research topic of the publication.

In biomedicine and related fields, new concepts are constantly emerging, old concepts are constantly updated, and terminology and usage are modified accordingly. The MeSH thesaurus, as a very comprehensive and authoritative medical thesaurus, is able to update its thesaurus annually, including the introduction of new descriptors in place of old ones. Therefore, *DescriptorName* in the MeSH thesaurus can effectively represent the research topic.

Emerging index calculation

To quantitatively measure the emerging degree of research topics, this study proposes an *emerging index* based on emerging attributes of novelty, growth, and impact.

Firstly, the novelty indicator needs to be calculated. The novelty of research topics is mainly related to the length of time that topics were created. Therefore, this study uses the novelty indicator $Novelty_t$ defined by Huang et al. (2022) in Eq. (1):

$$Novelty_t = 2\left(1 - \frac{1}{1 + e^{\frac{-(t-t_0)}{\lambda}}}\right) \tag{1}$$

where t_0 represents the time when the research topic was first entered into the MeSH thesaurus. λ is a parameter that controls the decay rate of the novelty indicator, and set as 20 through empirical analysis following Huang et al. (2022).

Then, in order to measure the growth and impact of research topics, this study extends the popularity score (Liang et al., 2021) that measures rapid growth and impact through frequency. The original popularity score $PopularityScore_t$ defined by Liang et al. (2021) is shown in Eqs. (2) and (3):

$$Af_t = f_t + \alpha \cdot Af_{t-1} \tag{2}$$

$$PopularityScore_t = \ln(Af_t + 1) \cdot \frac{f_t+1}{f_{t-1}+1} \tag{3}$$

where f_t is the frequency of research topics in year t, which represents the impact of the research topics, and Af_t is the adjusted frequency of research topics after considering the historical impact. α is a decay factor, ranging from 0 to 1, that determines the extent to which historical impact should be taken into account at the current time. $f_t+1/f_{t-1}+1$ is the frequency ratio of research topics for two adjacent years, which represents the annual growth rate of research topics.

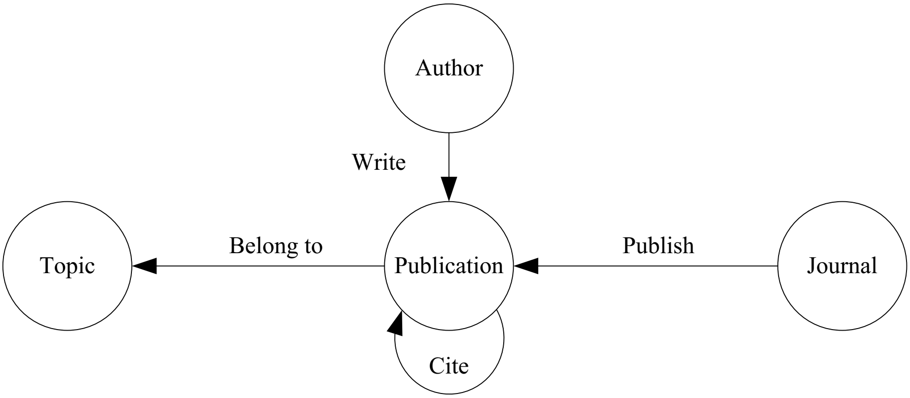

Fig. 2 Bibliographic heterogeneous network

However, random fluctuations such as database expansion or reduction may cause a sudden increase or decrease to the number of annual publications on a certain research topic, resulting in the unstable annual growth rate. Therefore, in order to calculate a more robust popularity score, this study employs the cumulated frequency ratio of the research topic instead of the frequency ratio of the research topic for the two adjacent years. The modified popularity score is calculated as shown in Eq. (4):

$$PopularityScore_t = \ln(Af_t + 1) \cdot \frac{1}{t - t_0} \sum_{t_0}^{t} \frac{f_i + 1}{f_{i-1} + 1} \tag{4}$$

where t_0 represents the time when the research topic was first entered into the MeSH thesaurus, and f_t is the frequency of research topics in year t. α is a decay factor, and set as 0.9 through empirical analysis following Liang et al. (2021).

Finally, the proposed *emerging index* of a research topic is calculated as the combination of novelty indicator and popularity score, as shown in Eq. (5):

$$EmergingIndex_t = Novelty_t \cdot PopularityScore_t \tag{5}$$

The logic of this equation is straightforward, because the higher level of novelty, growth and impact of the research topic mean the higher *emerging index*. Here, novelty is reflected in indicator $Novelty_t$, and growth and impact are reflected in indicator $PopularityScore_t$. This study empirically examines the proposed new *emerging index* in "Effectiveness analysis" section.

Network construction and feature extraction

Heterogeneous networks are particularly adept at capturing the intrinsic features of research topics, and the extraction of these features typically involves a constellation of interrelated entities (Sun et al., 2023). Therefore, to extract the features that better reflect the emerging attributes of the research topics, this study constructs bibliographic heterogeneous networks capable of accommodating diverse entity types and their interrelationships, as shown in Fig. 2. The bibliographic entities are *topic*, *publication*, *author* and *journal*, and the relationships include *belong to*, *cite*, *write* and *publish*. Inspired by Xie et al.

(2021), these bibliographic entities are chosen because they best reflect the characteristics of the research topic.

The selection of network features should be related to the attributes of the emerging research topic. Inspired by Lu et al. (2021), the following three categories of salient features are extracted because they best reflect the emerging degree of research topics: (1) Temporal feature, reflecting the novelty and timeliness of the research topic in a certain field, (2) Network size, reflecting the breadth and scale of bibliographic entities related to the research topics, and (3) Network impact, reflecting the depth and impact of bibliographic entities related to the research topics, as shown in Table 1.

Temporal feature

The temporal feature is part of the novelty. In this study, the temporal period from the creation time to the current time of research topic is used as the temporal feature. The temporal period is calculated as shown in Eq. (6):

$$t_{\text{period}} = t - t_0 \tag{6}$$

where t_0 is the time when the research topic was first entered into the MeSH thesaurus, and t represents the current time.

Network size

The change of network size at different period can reflect the growth and scale of research topics. Therefore, network size is selected to reflect the growth of research topics, and the number of bibliographic entities belonging to the research topics is used to express the size in this study. Let p_t denotes the number of publications belonging to a research topic in year t. Let a_t denotes the number of authors who write the publications belonging to a research topic in year t. Let j_t denotes the number of journals that publish the publications belonging to a research topic in year t.

Network impact

Network impact refers to the scientific impact of the network to which the research topic belongs, and also reflects the impact of research topics. The number of citations to publications belonging to the research topics is used as the measure in this study. Let c_t denotes the total number of citations for a research topic in year t. The total numbers of publications written by authors and published by journals also reflect the impact of research topics. Let ai_t denotes the total number of publications written by all authors for research topic in year t. Let ji_t denotes the total number of publications published by all journals for research topic in year t. These metrics are calculated as shown in Eqs. (7) and (8):

$$ai_t = \sum_{a \in A_t} ap_{a,t} \tag{7}$$

$$ji_t = \sum_{j \in J_t} jp_{j,t} \tag{8}$$

Table 1 Summary of the features extracted in this study

Category	Sub-indicator	Indicator definition
Temporal feature	t_{period}	The temporal period from the creation time to the current time of research topic
	p_t	The number of publications belonging to a research topic in year *t*
Network size	a_t	The number of authors who write the publications belonging to a research topic in year *t*
	j_t	The number of journals that publish the publications belonging to a research topic in year *t*
	c_t	The total number of citations for a research topic in year *t*
Network impact	ai_t	The total number of publications written by all authors for research topic in year *t*
	ji_t	The total number of publications published by all journals for research topic in year *t*

where A_t is the set of authors for a research topic in year t, and $ap_{a,t}$ is the total number of publications written by author a in year t. J_t is the set of journals for a research topic in year t, $jp_{j,t}$ is the total number of publications published by journal j in year t.

Predictive models

The prediction task of the emerging degree of research topics is to use historical data features related to the research topics to fit and infer the future *emerging index* in the next two years. These historical data features constitute a multivariate non-stationary time series, and traditional models have limitations in dealing with them. Different from the traditional models, deep learning models can better adapt to multivariate nonstationary time series data due to their powerful expressive power and adaptability. As a kind of deep learning model, LSTM can effectively fit the sequence characteristics of time series data, and has obvious advantages in avoiding gradient disappearance explosion (Lu et al., 2021). Therefore, the LSTM is chosen in this study as the predictive model.

In this study, $X = \{X_1, X_2, ..., X_n\}$ is defined as the set of feature time series for n research topics, and a feature time series X_i is a $t \times m$ matrix with m features at t time steps. Similarly, $Y = \{Y_1, Y_2, ..., Y_n\}$ is defined as the set of target time series for n research topics, and a target time series Y_i is a $t' \times m'$ matrix with m' features at t' time steps. The target time series in this study is a single output using the *emerging index* as the target value.

Experimental setups and results

Data description

Both datasets used in this study were drawn from the PubMed Central (PMC) database, which houses biomedical and life science journal literature from the National Library of Medicine of the National Institutes of Health, with a unique identifier PubMed ID set for a large number of publications. The first dataset consists of 423,665 publications in the field of neoplasms, and the second dataset consists of 139,405 publications in the field of metabolism. The publication years of both datasets were taken from 1969 to 2020. After removing a small number of duplicate and missing publications, the first dataset totals 420,086 publications, 1,166,069 authors and 3755 journals, and the second dataset totals 139,399 publications, 540,124 authors, and 3041 journals. These two fields were chosen because they are not only important areas of biomedical research with a profound impact on human health and disease, but also attract the interest of many interdisciplinary researchers, which facilitates knowledge crossover and innovation.

This study uses *DescriptorName* in the MeSH thesaurus to denote the research topics of publications, and the domain experts had assigned the appropriate *DescriptorName* to the publications based on the research content when they were uploaded to PubMed. *DateCreated* is the time when *DescriptorName* is created in MeSH thesaurus, and this study takes *DateCreated* as the initial year of a research topic. The initial time of research topics was selected from 2000 to 2020. After removing those without publication match, 6926 research topics were obtained in the first dataset, and 6688 research topics were obtained in the second dataset. In addition, in order to facilitate the experimental verification, the raw data used in this study have been uploaded to figshare.com (https://doi.org/10.6084/m9.figshare.25493512).

Table 2 Bibliographic nodes and relationships of heterogeneous networks

Dataset	Node	Number	Relationship	Number
Neoplasms	Topic	6926	Belong to	6,356,486
	Publication	420,086	Cite	3,425,238
	Author	1,166,069	Write	3,736,385
	Journal	3755	Publish	420,086
Metabolism	Topic	6688	Belong to	1,857,069
	Publication	139,399	Cite	1,236,279
	Author	540,124	Write	913,880
	Journal	3041	Publish	139,399

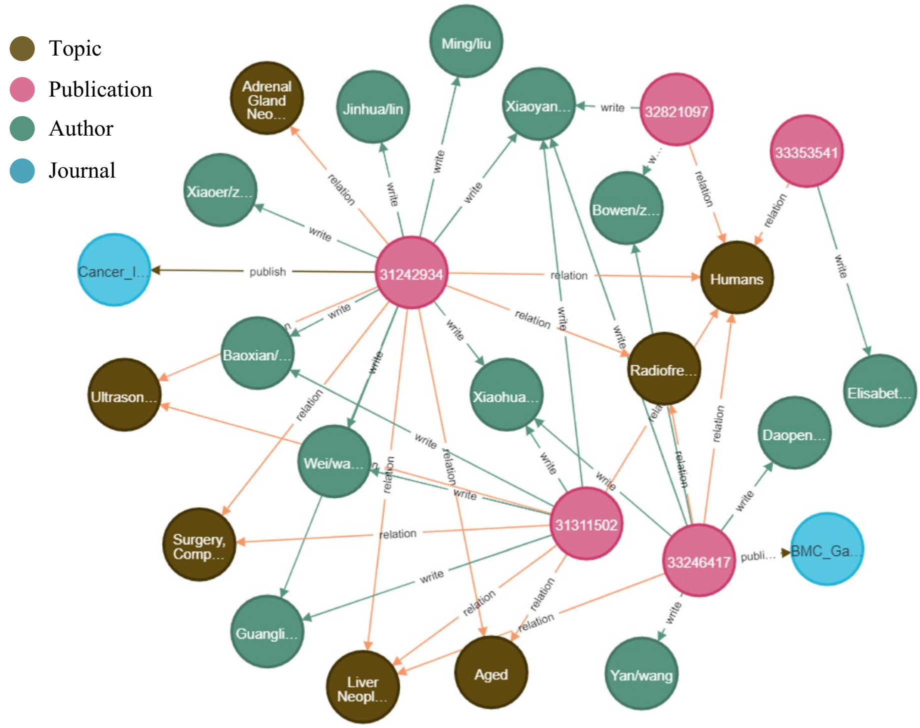

Fig. 3 Portion of heterogeneous networks

The construction of heterogeneous networks

This study used Neo4j[3] to construct and visualize the heterogeneous networks. Table 2 shows the nodes and relationship information of neoplasms heterogeneous networks and metabolism heterogeneous networks. Fig. 3 shows a portion of heterogeneous networks.

[3] https://neo4j.com/.

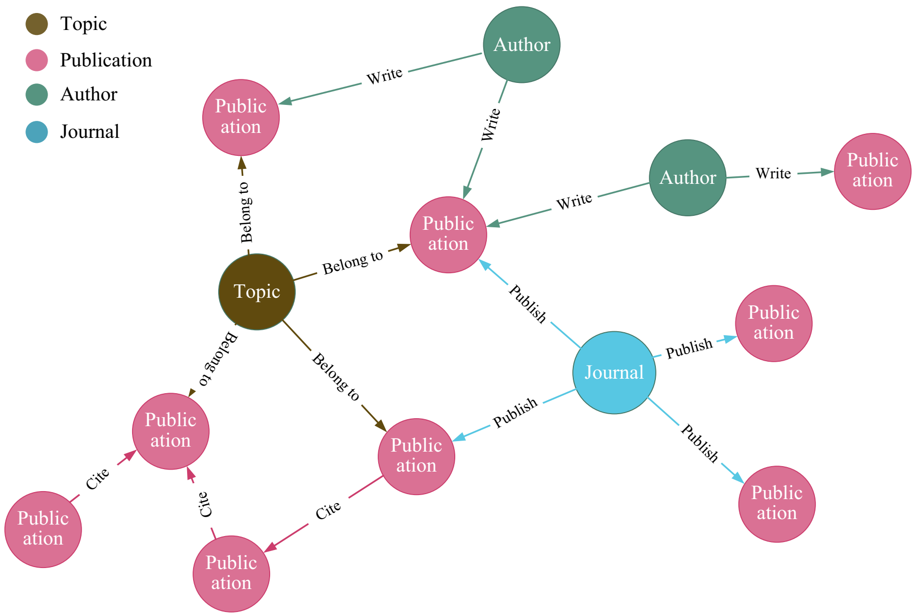

Fig. 4 An illustrative example for the feature network of a topic in a given year

Table 3 The distribution of emerging index values of research topics in 2020

Range	Neoplasms	Metabolism
0–1	2420	2884
1–2	1591	1855
2–3	1490	1006
3–4	848	529
4-	577	414

Then, the heterogeneous networks constructed can be used to extract the network features of research topics. As shown in Fig. 4, the values of network size including p_t, a_t, and j_t in this example are 4 (the number of publications belonging to a certain topic), 2 (the number of authors), and 1 (the number of journals), respectively. The values of network impact including c_t, ai_t, and ji_t in this example are 2 (the number of citations), 4 (the number of write relationships), and 4 (the number of publish relationships), respectively.

Effectiveness analysis

To verify the effectiveness of the proposed *emerging index*, the research topics in 2020 were selected as the investigation specimens. Table 3 shows the distribution of the values of *emerging index* for research topics in the two datasets in 2020, and the *emerging index* values of most research topics are concentrated in the ranges of 0–1 and 1–2. The 169 emerging research topics of neoplasms and the 149 emerging research topics of metabolism in 2020 were selected as a standard set by using the method of Liang et al. (2021) due to two main reasons. First, the proposed method is the extension to Liang et al.'s method.

Second, not only has Liang et al.'s method been published in high-quality journals, but it has been cited approximately 32 times in Google Scholar in less than two years (as of July 8, 2023). In addition, this study selected two emerging research topic identification methods as benchmarks, i.e., the method by Tu and Seng (2012), and the method by Yang et al. (2022a).

The top *N* research topics for *emerging index* will be considered emerging research topics. As shown in Table 4, the values of *N* were set to 0.5%, 1%, 2% and 3% respectively. Precision and recall were calculated using $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$, respectively, where *TP* is the number of positive samples in the prediction results, *FP* is the number of negative samples that were identified as positive samples, and *FN* is the number of negative samples in the prediction results. $F1$ scores is the harmonic average of prediction and recall, calculated as $F1 = 2P \times R/(P + R)$. The results are shown in Table 4.

As shown in Table 4, by comparing the precision, recall, and $F1$ on both datasets, the proposed method outperforms both benchmark methods for each value of *N*. For the proposed method, the precision is high (>0.7) for *N*-values at 0.5% and 1%, indicating that the topics ranked at the top of the *emerging index* are emerging. With the increase of *N*-value, the recall value also increases. Especially the *N*-value at 3% will achieve the recall value at 0.746 and 0.859, respectively, which means that the research topics ranked at the top 3% of the emerging research index encompass most of the research topics in the standard set. At the same time, the $F1$ scores increase with the balance of precision and recall, reaching 0.669 and 0.732 for *N*-value at 3%, respectively, which is a qualified score and means that the proposed *emerging index* is able to select emerging research topics to the top of the ranking. Therefore, the proposed *emerging index* is an effective indicator that ranks the topics with high emerging degrees at the top of the ranking.

Evaluation metrics

To compare the performance of LSTM with other prediction methods, this study selected four common machine learning models as baselines, i.e., the k-nearest neighbor (KNN) (Yan et al., 2012), linear regression (LR) (Abramo et al., 2019), support vector regression (SVR) (Chakraborty et al., 2014), light gradient boosting machine (LightGBM) (Liang et al., 2021), and graph convolutional network (GCN) (Yang et al., 2022b), which have achieved good performance in previous topic prediction studies.

To evaluate the proposed predictive model, two error-based metrics and one ranking-based metric are adopted.

The two error-based metrics mean absolute error (MAE) and root mean squared error (RMSE) are defined as shown in Eqs. (9) and (10):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

where y_i represents the actual value of *emerging index*, $\hat{y}_i$ represents the predicted value of *emerging index*, and n represents the number of samples.

Table 4 Performance on different N-values

Dataset	N-value (%)	Number	Tu and Seng (2012)			Yang et al. (2022a)			The proposed method		
			Precision	Recall	$F1$	Precision	Recall	$F1$	Precision	Recall	$F1$
Neoplasms	0.5	35	0.743	0.154	0.255	0.829	0.172	0.284	0.886	0.183	0.303
	1	69	0.580	0.237	0.336	0.710	0.290	0.412	0.797	0.325	0.462
	2	139	0.511	0.420	0.461	0.590	0.485	0.532	0.706	0.568	0.630
	3	208	0.438	0.538	0.483	0.457	0.562	0.504	0.606	0.746	0.669
Metabolism	0.5	33	0.727	0.161	0.264	0.818	0.181	0.296	0.909	0.201	0.329
	1	67	0.672	0.302	0.417	0.657	0.295	0.407	0.836	0.376	0.519
	2	134	0.575	0.517	0.544	0.604	0.544	0.572	0.761	0.685	0.721
	3	201	0.502	0.678	0.577	0.517	0.698	0.594	0.637	0.859	0.732

Table 5 Parameters of the LSTM

Parameters	Neoplasms	Metabolism
Number of units in each layer	256, 128, and 128	512, 512, and 512
Activation function	ReLU	ReLU
Initial learning rate	0.0001	0.0001
Optimizer	Adam	Adam
Batch size	256	256
Epochs	100	100

The ranking-based metric, i.e., normalized discounted cumulative gain (NDCG@k), gauges the quality of ranking based on the order of the top *k* results (Järvelin & Kekäläinen, 2000). In the field of information retrieval, this metric is frequently employed by search engines and recommendation systems to evaluate their effectiveness. The NDCG@k is calculated as shown in Eqs. (11)–(13):

$$DCG = \sum_{i=1}^{|pred|} \frac{2^{y_i} - 1}{\log_2(i + 1)} \tag{11}$$

$$IDCG = \sum_{i=1}^{|true|} \frac{2^{y_i} - 1}{\log_2(i + 1)} \tag{12}$$

$$NDCG = \frac{DCG}{IDCG} \tag{13}$$

where y_i represents the true *emerging index* value of a research topic and *i* is its position number, *pred@k* denotes the list of the top *k* research topics sorted by predicted value of the *emerging index*, and *true@k* denotes the list of the top *k* research topics sorted by true value of the *emerging index*. NDCG@k ranges from zero to one, and measures the predictive power of the model in terms of the sequential quality of the top k research topics. The higher NDCG@k means the superior performance. In this study, *k* was set to 20, that is, NDCG@20 was utilized to evaluate the ranking quality of the top 20 research topics.

Model setups

To fully utilize the historical data of each research topic, a sliding window with a fixed step size of 3 is set, and the target time is set to 2, that is, the target value after two years is predicted with three years of historical data, which is similar to Lu et al. (2021). This study employs a time-based validation approach, using data prior to 2015 for training, data from 2015 to 2017 for validation, and data from 2018 to 2020 for testing. Finally, 29,032 training pairs, 16,003 validation pairs, and 18,177 test pairs were obtained in the field of metabolism, and 30,101 training pairs, 15,946 validation pairs, and 17,845 test pairs were obtained in the field of metabolism, containing input values and target values.

To improve the performance of predictive model as much as possible, the structure of the deep learning-based model is fine tuned. The number of units in each layer, activation function, optimizer, and other parameter settings are shown in Table 5.

Table 6 Overall evaluation results of different models

Models	Neoplasms			Metabolism		
	MAE	RMSE	NDCG@20	MAE	RMSE	NDCG@20
KNN	0.394	0.514	0.727	0.312	0.615	0.716
LR	0.375	0.593	0.834	0.315	0.633	0.821
SVR	0.362	0.532	0.704	0.375	0.790	0.768
LightGBM	0.367	0.492	0.841	0.308	0.591	0.786
GCN	0.374	0.469	0.793	0.309	0.522	0.745
LSTM	**0.299**	**0.412**	**0.898**	**0.275**	**0.496**	**0.901**

The best results are marked in bold

Prediction results

The prediction performance of the baselines and LSTM models are illustrated in Table 6. Table 6 reveals that, in terms of MAE, RMSE and NDCG@20 metrics on the test sets for both datasets, LSTM significantly outperforms the baseline models. In the neoplasms dataset, the MAE of LSTM is 24.11% lower than that of KNN, 20.27% lower than that of LR, 17.40% lower than that of SVR, 18.53% lower than that of LightGBM, and 20.05% lower than that of GCN. The RMSE of LSTM is reduced by 19.84% compared to that of KNN, 30.52% compared to that of LR, 22.56% compared to that of SVR, 16.26% compared to that of LightGBM, and 12.15% compared to that of GCN. The NDCG@20 of LSTM is improved by 23.52% compared to that of KNN, 7.67% compared to LR, 27.56% compared to SVR, 6.78% compared to LightGBM, and 13.24% compared to GCN. And in the metabolism dataset, the MAE of LSTM is 11.86% lower than that of KNN, 12.70% lower than that of LR, 26.67% lower than that of SVR, 10.71% lower than that of LightGBM, and 11.00% lower than that of GCN. The RMSE of LSTM is reduced by 19.35% compared to that of KNN, 21.64% compared to that of LR, 37.22% compared to that of SVR, 16.07% compared to that of LightGBM, and 4.98% compared to that of GCN. The NDCG@20 of LSTM is improved by 25.84% compared to that of KNN, 9.74% compared to LR, 17.32% compared to SVR, 14.63% compared to LightGBM, and 20.94% compared to GCN. This indicates that LSTM can not only achieve the better results in terms of prediction error, but also be more confident in predicting the *emerging index* for the top ranked research topics.

Feature importance

In order to compare the relative importance of the input features given in this study, one of seven features is removed each time, and the remaining features are used to train the model (i.e., “leave-one-out model”), then the RMSE of the test set is calculated. The deep learning-based model is adopted to compare the feature importance, and the results are shown in Table 7.

Firstly, Table 7 shows that when the temporal feature t_{period} is removed, the increase of RMSE is the largest on both datasets, which indicates that the temporal feature can greatly improve the prediction performance. Secondly, the second most salient features to improve the prediction performance of the model is the network impact. Finally, the network size

Table 7 The increase in RMSE between each 'leave-one-out' model and the full features model

Category of the removed feature	Removed feature	Neoplasms Increase in the RMSE	Metabolism Increase in the RMSE
Temporal feature	t_{period}	0.0160	0.0154
	p_t	0.0053	0.0036
Network size	a_t	0.0041	0.0028
	j_t	0.0015	0.0011
	ai_t	0.0065	0.0076
Network impact	ji_t	0.0079	0.0083
	c_t	0.0034	0.0023

Table 8 Top 20 emerging research topics in neoplasms

Rank	Research topic	Rank	Research topic
1	Epithelial–mesenchymal transition	11	Gene ontology
2	Gastrointestinal microbiome	12	Drug liberation
3	Organs at risk	13	Robotic surgical procedures
4	Protein interaction maps	14	Extracellular vesicles
5	Machine learning	15	Xenograft model antitumor assays
6	Cell proliferation	16	Kaplan–Meier estimate
7	Theranostic nanomedicine	17	MicroRNAs
8	Protective factors	18	Gene regulatory networks
9	Early detection of cancer	19	Radiation dose hypofractionation
10	Tumor microenvironment	20	Genetic association studies

is the least helpful to improve the prediction performance of the model, where the RMSE increase for the two datasets after removing j_t is only 0.0015 and 0.0011, respectively, which is the smallest value among all RMSE values.

Nominated emerging research topics

Using the datasets from 2000 to 2020, this study predicts the *emerging index* of research topics for 2022, and selects the top 20 topics as the nominated emerging research topics. Tables 8 and 9 list the top 20 emerging research topics in the fields of neoplasms and metabolism predicted using LSTM, respectively.

As shown in Table 8, "Eearly detection of cancer" ranked the ninth is predicted as one of the most important research topics. In the recent scientific discourse, a review article titled *Early detection of cancer* was published in the esteemed journal *Science* in 2022, which delineates the prevailing challenges and extant research endeavors within the early detection of cancer (Crosby et al., 2022). The technologies and theories explicated in this review are closely related to several other highly-ranked research topics, including "Machine learning" ranked the fifth, "Therapeutic nanomedicine" ranked the seventh, "Tumor microenvironment" ranked the tenth, "Kaplan–Meier estimate" ranked the sixteenth and "Genetic association study" ranked the twentieth. In addition, "Epithelial–mesenchymal

Table 9 Top 20 emerging research topics in metabolism

Rank	Research topic	Rank	Research topic
1	Tumor hypoxia	11	Machine learning
2	Protein interaction maps	12	Cellular reprogramming
3	Metabolic networks and pathways	13	Cardiorespiratory fitness
4	Metabolic engineering	14	Molecular sequence annotation
5	Transcriptome	15	Deep learning
6	MCF-7 cells	16	Gene regulatory networks
7	Metabolic flux analysis	17	Protein aggregates
8	Gene knockdown techniques	18	CRISPR-cas systems
9	Molecular targeted therapy	19	Wnt signaling pathway
10	Endoplasmic reticulum stress	20	Tandem mass spectrometry

transition" ranked the first has been the most dynamic research topic in the neoplasms research domain in the most recent practice. Similarly, "Protein interaction maps" ranked the fourth and "Gene regulatory networks" ranked the eighteenth have also been widely used in the study of epithelial–mesenchymal transition, and combined with "machine learning" ranked the fifth to predict the evolution of cancer in the most recent practice (Foo et al., 2022; Pillai et al., 2023; Sherman et al., 2021).

As shown in Table 9, "Metabolic engineering" ranked the fourth and "Metabolic flux analysis" ranked the seventh are predicted to be the two most important research topics. In the recent practice, *Nature Reviews Bioengineering* published a review article titled *Engineered autonomous dynamic regulation of metabolic flux* in 2023, which proposed the role of metabolic engineering and metabolic flux in editing cell metabolism (Ream & Prather, 2023). The technologies and theories mentioned in this article relate to "Metabolic networks and pathways" ranked the third, "Transcriptome" ranked the fifth, "Cellular reprogramming" ranked the twelfth, and "Gene regulatory networks" ranked the sixteenth. In addition, "Machine learning" ranked the eleventh and "Deep learning" ranked the fifteenth have been used to optimize metabolic engineering and signaled pathway discovery (Helmy et al., 2020).

In addition, six medical experts (three in the field of neoplasms and three in the field of metabolism) were consulted through both online and offline methods to evaluate each of the top 20 emerging research topics predicted by the study. The six medical experts agreed that more than half of the research topics were emerging research topics for year 2022. Therefore, through the above qualitative analysis, it is believed that the prediction method of emerging research topics proposed in this study can nominate emerging research topics.

Conclusion

Theoretical implications

Theoretically, this study provides valuable idea and method on measurement and prediction of emerging research topics by combining a deep learning-based model with heterogeneous networks. For the measurement of emerging research topics, this study proposes an *emerging index* that considers combination of three attributes: novelty, growth and impact.

It can measure the emerging degree of research topics and serve as a target indicator of predictive models. This study also explores the factors that affect the emerging degree of research topics, and put forward new features such as temporal feature, network size, and network impact, which can be extracted by constructing a heterogeneous network of bibliographic entities in a research domain. The experimental results show that temporal feature and network size play the more important role in predicting the emerging degree of research tops. In addition, the experimental results also show that the deep learning-based model is superior to other machine learning models in predicting the *emerging index* in term of both error-based and rank-based metrics, and has better generalization.

Practical implications

Practically, this study provides policy makers and researchers with a method that can quickly and effectively predict the emerging research topics to support policy making and grant allocation. The proposed *emerging index* can quantitatively measure the emerging degree of research topics in the research domain, and can reflect the development prospect of a research topic (the research topic with an *emerging index* greater than the average level has a better development prospect). The topic representation method based on MeSH thesaurus can ensure the professionalism and authority of the extracted research topics. In addition, the proposed *emerging index* as the target output and the extracted temporal feature, network size and network impact as input features can be applied to a variety of machine learning models, enabling subsequent researchers to improve the prediction accuracy of emerging research topics.

Limitations and future work

Admittedly, this study has some limitations. Firstly, the datasets selected in this study is domain-specific and does not consider the situation of interdisciplinarity. The PMC database used in this study only covers the publications in the field of biomedical and life sciences, and the research topics of each publication is denoted by the *DescriptorName* in MeSH thesaurus and assigned by the domain experts according to the content of the publication. To explore the more databases covering more comprehensive research fields, annotation through domain experts is too time-consuming and semantic-based methods such as topic models, co-word clustering or natural language processing could be considered in the future. Secondly, the heterogeneous network of bibliographic entities constructed in this study only focuses on the few entities including research topic, publication, author and journal and the relationships between them. As a text analysis method, content analysis can not only extract features such as common research methods and arguments, but also identify emotional tendencies and evolution trends in publications. Future research may conduct content analysis of publications to capture more features such as emotional tendencies that permeate the literature and the shifting patterns that indicate the progression of topics over time, to more comprehensively reflect the emerging features of the research topics. Thirdly, this study focuses only on novelty, growth and impact for the measurement of emerging research topics, but there are still some other intrinsic attributes such as coherence and uncertainty that may affect the *emerging index* and are worth further exploring. Finally, to improve the performance and generalization ability of predictive models, future research may consider the integration of embedding mechanisms, attention mechanisms, and semantic analysis into the deep learning architecture. This includes integrating pre-trained word embedding models to capture richer semantic information, employing attention mechanisms to improve the model's focus on key information, and incorporating semantic analysis techniques to enhance understanding of text context.

Supporting information

The raw data used in this study have been uploaded to figshare.com (https://doi.org/10.6084/m9.figshare.25493512), including the publications collected from PMC database and the MeSH data from MeSH thesaurus.

Acknowledgements This work has been supported by Zhejiang Natural Science Foundation of China (No. LZ20E050001), and Zhejiang Province Key R&D Program of China (Nos. 2022C03166, 2024C01034).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. *Journal of Informetrics, 13*(1), 32–49.

Asooja, K., Bordea, G., Vulcu, G., & Buitelaar, P. (2016). Forecasting emerging trends from scientific literature. In *Proceedings of the 10th international conference on language resources and evaluation*, May 23–28, Portorož, Slovenia (pp. 417–420).

Behrouzi, S., Sarmoor, Z. S., Hajsadeghi, K., & Kavousi, K. (2020). Predicting scientific research trends based on link prediction in keyword networks. *Journal of Informetrics, 14*(4), 101079.

Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233–251.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2014). Towards a stratified learning approach to predict future citation counts. In *Proceedings of the 2014 IEEE/ACM joint conference on digital libraries (JCDL)*, September 8–12, London, United Kingdom (pp. 351–360).

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., et al. (2022). Early detection of cancer. *Science, 375*(6586), eaay040.

Foo, J., Basanta, D., Rockne, R. C., Strelez, C., Shah, C., et al. (2022). Roadmap on plasticity and epigenetics in cancer. *Physical Biology, 19*(3), 031501.

Gao, Q., Liang, Z. T., Wang, P., Hou, J. R., Chen, X. X., & Liu, M. M. (2021). Potential index: Revealing the future impact of research topics based on current knowledge networks. *Journal of Informetrics, 15*(3), 101165.

Hassan, S. U., Imran, M., Iqbal, S., Aljohani, N. R., & Nawaz, R. (2018). Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics, 117*, 1645–1662.

Helmy, M., Smith, D., & Selvarajoo, K. (2020). Systems biology approaches integrated with artificial intelligence for optimized metabolic engineering. *Metabolic Engineering Communications, 11*, e00149.

Huang, M. H., & Chang, C. P. (2014). Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics, 98*(3), 1721–1744.

Huang, S. Z., Lu, W., Bu, Y., & Huang, Y. (2022). Revisiting the exploration-exploitation behavior of scholars' research topic selection: Evidence from a large-scale bibliographic database. *Information Processing and Management, 59*(6), 103110.

Järvelin, K, & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, July 24–28, Athens, Greece (pp. 41–48).

Kwon, S., Liu, X. Y., Porter, A. L., & Youtie, J. (2019). Research addressing emerging technological ideas has greater scientific impact. *Research Policy, 48*(9), 103834.
Liang, Z. T., Mao, J., Lu, K., Ba, Z. C., & Li, G. (2021). Combining deep neural network and bibliometric indicator for emerging research topic prediction. *Information Processing & Management, 58*(5), 102611.
Liu, X. Y., & Porter, A. L. (2020). A 3-dimensional analysis for evaluating technology emergence indicators. *Scientometrics, 124*(1), 27–55.
Liu, Z. G., Yin, Y. M., Liu, W. D., & Dunford, M. (2015). Visualizing the intellectual structure and evolution of innovation systems research: A bibliometric analysis. *Scientometrics, 103*(1), 135–158.
Lu, W., Huang, S. Z., Yang, J. Q., Bu, Y., Cheng, Q. K., & Huang, Y. (2021). Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management, 58*(4), 102594.
Ma, T. T., Zhou, X., Liu, J., Lou, Z. K., Hua, Z. T., & Wang, R. T. (2021). Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. *Technological Forecasting and Social Change*, *173*, 121159.
McKeown, K., Daume, H., III., Chaturvedi, S., Paparrizos, J., Thadani, K., et al. (2016). Predicting the impact of scientific concepts using full-text features. *Journal of the Association for Information Science and Technology, 67*(11), 2684–2696.
Pillai, M., Hojel, E., Jolly, M. K., & Goyal, Y. (2023). Unraveling non-genetic heterogeneity in cancer with dynamical models and computational tools. *Nature Computational Science, 3*, 301–313.
Price, D. J. D. S. (1965). Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science, 149*(3683), 510–515.
Ream, M., & Prather, K. L. (2023). Engineered autonomous dynamic regulation of metabolic flux. *Nature Reviews Bioengineering*. https://doi.org/10.1038/s44222-023-00140-7
Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology? *Research Policy, 44*(10), 1827–1843.
Sherman, J., Verstandig, G., Rowe, J. W., & Brumer, Y. (2021). Application of machine learning to large in vitro databases to identify drug-cancer cell interactions: Azithromycin and KLK6 mutation status. *Oncogene, 40*(21), 3766–3770.
Shi, S. S., Zhang, W. Y., Zhang, S., & Chen, J. (2018). Does prestige dimension influence the interdisciplinary performance of scientific entities in knowledge flow? Evidence from the e-government field. *Scientometrics, 117*, 1237–1264.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269.
Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450–1467.
Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies, 4*(1), 17–40.
Song, B. W., Luan, C. J., & Liang, D. N. (2023). Identification of emerging technology topics (ETTs) using BERT-based model and sematic analysis: A perspective of multiple-field characteristics of patented inventions (MFCOPIs). *Scientometrics, 128*(11), 5883–5904.
Sun, Y. H., Jia, M. Y., Liu, C., & Shao, M. L. (2023). Heterogeneous network representation learning based on role feature extraction. *Pattern Recognition, 144*, 109870.
Tu, Y. N., & Seng, J. L. (2012). Indices of novelty for emerging topic detection. *Information Processing & Management, 48*(2), 303–325.
Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392.
Wang, Q. (2018). A bibliometric model for identifying emerging research topics. *Journal of the Association for Information Science and Technology, 69*(2), 290–304.
Wei, W. J., Liu, H. X., & Sun, Z. L. (2022). Cover papers of top journals are reliable source for emerging topics detection: A machine learning based prediction framework. *Scientometrics, 127*(8), 4315–4333.
Weismayer, C., & Pezenka, I. (2017). Identifying emerging research fields: A longitudinal latent semantic keyword analysis. *Scientometrics, 113*(3), 1757–1785.
Xie, Q., Zhang, X. Y., & Song, M. (2021). A network embedding-based scholar assessment indicator considering four facets: Research topic, author credit allocation, field-normalized journal impact, and published time. *Journal of Informetrics, 15*(4), 101201.
Xu, H. Y., Winnink, J., Yue, Z. H., Zhang, H. L., & Pang, H. S. (2021). Multidimensional Scientometric indicators for the detection of emerging research topics. *Technological Forecasting and Social Change, 163*, 120490.

Xu, S., Hao, L. Y., An, X., Pang, H. S., & Li, T. (2020). Review on emerging research topics with key-route main path analysis. *Scientometrics, 122*(1), 607–624.
Xu, S., Hao, L. Y., An, X., Yang, G. C., & Wang, F. F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13*(4), 100983.
Yan, R., Huang, C. R., Tang, J., Zhang, Y., & Li, X. M. (2012). To better stand on the shoulder of giants. In *Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries*, June 10–14, Washington DC, United States (pp. 51–60).
Yang, J. Q., Lu, W., Hu, J. M., & Huang, S. Z. (2022a). A novel emerging topic detection method: A knowledge ecology perspective. *Information Processing and Management, 59*(2), 102843.
Yang, M., Jiang, C. J., Yan, X. F., Ai, T. H., Cao, M. J., & Chen, W. Y. (2022b). Detecting interchanges in road networks using a graph convolutional network approach. *International Journal of Geographical Information Science, 36*(6), 1119–1139.
Zhang, Y., Wu, M. J., Miao, W., Huang, L., & Lu, J. (2021). Bi-layer network analytics: A methodology for characterizing emerging general-purpose technologies. *Journal of Informetrics, 15*(4), 101202.
Zhou, Y., Dong, F., Liu, Y. F., Li, Z. F., Du, J. F., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics, 123*(1), 1–29.