Predicting Scholarly Impact Based on Novelty in Early Knowledge Structure: A Case Study in the Biomedical Field

Authors: Wu Zhixiang, Yihai Port, Yimeng Zhu, Wang Pei, Wang Hao, Wu Zhixiang

Date: 2024-10-21T19:01:16+00:00

Abstract

Purpose/Significance: This study leverages the innovative characteristics embedded in scholars' knowledge structures to design metrics for measuring the novelty of scholars' early-stage knowledge structures, thereby predicting their future impact and providing a new reference indicator for the early identification of academic talent. Method/Process: First, data for 57,927 scholars in the biomedical field were obtained from the PKG (PubMed Knowledge Graph) database, and scholars' knowledge structures were constructed using co-occurrence relationships of controlled subject terms. Second, six metrics were designed from two dimensions—knowledge topics and structural positions—to measure the novelty of scholars' early-stage knowledge structures. Subsequently, scholars were classified and labeled based on their later-stage impact, and machine learning models were trained. Finally, the classification performance of models with different variable combinations was experimentally evaluated, and the predictive performance of metrics based on knowledge structure novelty was analyzed. Results/Conclusion: The findings reveal that novelty metrics can effectively predict impact. In single-metric prediction, Topic Novelty (TN) demonstrates the best performance, while the four structural-level metrics all outperform Topic Combination Novelty (TCN). The integrated metrics achieve an average improvement of 2.7% in F1-score. This paper provides a perspective for predicting and understanding scholars' academic impact from content features, and the new metric design is valuable in addressing the limitations of existing predictive indicators.

Full Text

Preamble

Research on Predicting Scholar Influence Based on the Novelty of Early Knowledge Structure: A Case Study in the Biomedical Field

Wu Zhixiang[1], Yi Haigang[1], Zhu Yimeng[1], Wang Pei[1], Wang Hao[2]
[1]School of Economics and Management, Nanjing Tech University, Nanjing 211816, China
[2]School of Information Management, Nanjing University, Nanjing 211023, China

Abstract

[**Purpose/Significance**] This study leverages the innovative characteristics inherent in scholars' knowledge structures to design indicators that measure the novelty of early-career knowledge structures, thereby predicting scholars' future influence and providing new reference indicators for the early identification of academic talent. [**Methods/Process**] First, we obtained data on 57,927 biomedical scholars from the PubMed Knowledge Graph (PKG) database and constructed scholars' knowledge structures using co-occurrence relationships of controlled subject terms. Second, we designed six indicators from two perspectives—knowledge themes and structural positions—to measure the novelty of scholars' early knowledge structures. Next, scholars were classified and labeled based on their later influence, and machine learning models were trained. Finally, we evaluated the classification effectiveness under different variable combinations and analyzed the predictive performance of indicators based on knowledge structure novelty. [**Results/Conclusion**] The study found that novelty indicators can effectively predict influence. Among single-indicator predictions, Topic Novelty (TN) performed best, while the four structural-level indicators all outperformed Topic Combination Novelty (TCN). The F1-score of comprehensive indicators improved by an average of 2.7%. This paper provides a perspective for predicting and understanding scholars' academic influence from content characteristics. The newly designed indicators are valuable and can compensate for the shortcomings of existing prediction metrics.

Keywords: influence prediction, scholar knowledge structure, novelty, machine learning

This work is supported by the National Natural Science Foundation of China project "Research on Key Core Technologies-Experts Combination Model Based on Deep Learning and Semantic Association" (Grant No. 7190408) and the National Social Science Foundation of China project "Research on Knowledge Association and Value Mining of Documentary Resources of Chinese Meritorious Scientists" (Grant No. 23CTQ027).

Authors: Wu Zhixiang, Associate Professor, PhD, Master's Supervisor, E-

mail: cnwzx2012@njtech.edu.cn; Yi Haigang, Master's Student; Zhu Yimeng, Master's Student; Wang Pei, Master's Student; Wang Hao, Professor, PhD, Doctoral Supervisor.

1. Introduction

Predicting scholars' future influence based on their early academic performance and research capabilities is an important yet challenging research task that has attracted significant attention from both academia and research management practice [1-2]. A scholar's knowledge structure [3] constitutes an important foundation for their academic achievements and creativity, serving as a crucial carrier of their theoretical ideas and research methods, and is inextricably linked to their academic influence [4]. Highly influential scholars often tend to select novel topics for research during the early stages of their careers [5], making it promising to adopt indicators based on the novelty of scholars' early knowledge structures to predict their future influence.

Research on predicting scholar influence can help establish more comprehensive and objective academic evaluation systems and promote scientific development. Currently, influence measurement typically focuses on external indicators such as citation counts and journal impact factors [1]. The main problems with this approach are that some researchers one-sidedly pursue publication quantity, lack content innovation, and chase hot topics without addressing key scientific questions [7]. Meanwhile, highly novel research is often subject to greater uncertainty, and recognition of its influence (e.g., through citations) typically exhibits temporal lag [8]. For example, limited by peer reviewers' cognition, some major innovative achievements have experienced delayed recognition, becoming scientific "sleeping beauties" [9]. Therefore, predicting future influence based on early citations or H-index suffers from timeliness issues. As the main subjects of academic research, scholars' internal knowledge organization reflects the path of scientific progress. Consequently, indicators that reflect the innovative characteristics of scholars' knowledge structures should be incorporated into the indicator system for early evaluation and selection.

In view of this, this study designs novelty measurement indicators based on complex networks to predict scholars' future influence, compares their predictive performance with existing indicators across multiple dimensions, and explores more optimized combination prediction schemes. Unlike previous studies that start from external features of academic performance, this paper attempts to use the novelty presented in scholars' internal knowledge structures to predict future high-impact scholars, dynamically tracking scholars' academic development capabilities to compensate for the lag in academic recognition. Simultaneously, from the perspective of "breaking the five-only" policy, this study explores the significance of knowledge novelty—often overlooked by traditional quantitative evaluation indicators—for talent identification and selection.

2. Related Research

This section reviews relevant research on key concepts including scholar knowledge structure, scholar influence, and novelty, and summarizes the relationships between these concepts based on existing work to establish the logical foundation for this study.

2.1 Research on Scholar Knowledge Structure

Existing research has explored the concept of knowledge structure from both disciplinary and individual scholar perspectives. At the disciplinary level, knowledge structure is a collection of disciplinary knowledge elements and their interrelationships. Units representing knowledge elements, ordered by semantic granularity from small to large, include author keywords (AK), combined subject terms [10], individual papers [11], etc. For keywords or subject terms, document co-occurrence can be used to form topic networks; for papers, citation relationships can be used to form citation networks. Both types of networks are referred to as disciplinary knowledge structures. By designing complex network indicators, we can measure network attributes, discover structural characteristics, and observe their evolution over time. Changes in disciplinary knowledge structure may signify the publication of breakthrough achievements or the emergence of breakthrough topics. For example, Min et al. predicted breakthrough papers by constructing first-generation citation networks of focal papers and calculating network indicators [11], while Xu et al. identified breakthrough topics by constructing topic networks and observing sudden changes in network structural entropy [10].

Papers are not created out of thin air but are produced by numerous scholars. Therefore, it is appropriate to introduce the concept of knowledge structure from the disciplinary level to the scholar level. Scholar knowledge structure can be defined as "the composition and combination patterns of knowledge possessed by an individual" [3]. Construction methods for disciplinary knowledge structures (topic networks, citation networks) have also been introduced to the scholar level [12]. Scholar knowledge structure is also dynamic, resulting from scholars' selection and effective combination of appropriate knowledge elements. Brookes' fundamental equation of information science, $K[S] + \Delta I = K[S+\Delta S]$, describes the dynamic change process of individual knowledge structure, where $K[S]$ represents the initial knowledge structure, ΔI represents information, $K[S+\Delta S]$ represents the new knowledge structure formed after absorbing information ΔI, and ΔS represents the effect of information absorption [15]. Scholar knowledge behavior can be summarized as: absorbing new knowledge and reusing old knowledge [16]. The number and types of knowledge topics scholars choose reflect the breadth of their research vision, while the way they connect relationships reflects the depth of their research vision [3]. Therefore, scholar knowledge structure reflects scholars' academic foundation, academic resources, and academic capabilities to a certain extent [17-19], which ultimately manifest as academic influence. For example, Zeng et al. found that thematic concentra-

tion in scholars' early knowledge structures is conducive to generating greater influence, while later stages require appropriate diversification of research topics [4]. Unlike disciplinary-level knowledge structures, individual scholars' knowledge structures also implicitly contain tacit knowledge such as beliefs and ideas; however, considering measurement feasibility, this study limits it to the explicit knowledge level based on research topics.

2.2 Research on Scholar Influence Prediction

Scholar influence is a key indicator for measuring academic ability and contribution. Predicting scholars' future academic achievements and development has important application value in research management and resource allocation. Existing research typically predicts scholars' future influence from three aspects: academic performance, social resources, and personal characteristics: (1) Early academic performance includes the number of published papers, citation counts, h-index, and whether publications appear in top-tier journals. In 2012, Acuna et al. published in *Nature* demonstrating the feasibility of using polynomial regression based on the h-index [20] to predict scholars' influence over the next 5 and 10 years [21]. Additionally, publication journal diversity, journal impact factor, and top 10% journal indicators are commonly used to characterize scholars' academic performance [21]. (2) Scholars' social resources that can be quantified mainly include academic collaboration relationships, such as the number of collaborators, scholars' positions in collaboration networks, and whether they collaborate with top scholars [23], all of which are important factors for measuring future influence. (3) Furthermore, some personal characteristics such as gender, age, education level, and position changes are also important reference indicators [24]. In summary, most existing influence prediction research selects indicators from scholars' early academic performance external features.

Existing scholar influence prediction methods can be summarized as: (1) Statistical regression methods, based on bibliometric principles, treat influence as a continuous numerical dependent variable and fit mathematical formulas according to the correlation between independent and dependent variables to calculate future influence values [20]; (2) Machine learning methods, which extract features from scholars' early learning behaviors and predict future influence through supervised or unsupervised learning to achieve deep mining of large-scale data. Common research includes predicting outstanding scholars [25] (academic rising stars, academicians, Nobel laureates, etc.), treating whether one is an outstanding scholar as a classification label and converting the prediction into a binary classification problem, with applicable machine learning algorithms including KNN, SVM, XGBoost, Random Forest, etc.; (3) Social network methods, which construct social networks based on citation and collaboration relationships among scholars, predict scholars' future positions in the network through network evolution and node attribute characteristics, thereby predicting scholars' influence ranking [26].

2.3 Research on the Relationship Between Novelty and Influence

Although scholars' early academic performance (citations, h-index, etc.) has become mainstream indicators for predicting future influence, existing research also shows that there is a close relationship between academic achievement novelty and influence. Since academic research is not only an innovative activity that promotes scientific progress but also a social activity that promotes academic exchange, the relationship between academic achievement novelty and influence is not simply linear [27].

From the paper perspective, a paper's influence varies during dissemination and promotion depending on its novelty. Many studies have found a non-linear inverted U-shaped relationship between paper novelty and citation count [28-29]. This reflects that papers with low novelty often have outdated topics and lack research value, making it difficult to generate significant academic impact; meanwhile, papers with excessively high novelty may be too "novel" or "premature" to be widely understood and accepted, encountering resistance in dissemination [11]. Highly influential papers typically have unique innovative characteristics [30], balancing academic novelty and public recognition by connecting old and new knowledge. From the scholar perspective, scholars also seek balance between innovation and conservatism in research activities. Driven by the exploration-exploitation model (EEM), scholars will learn about and explore new topics based on their knowledge reserves to expand their knowledge structure and achieve academic innovation [13]. Huang et al. believe that successful scholars tend to research novel topics in their early careers and propose in their discussion section that "academic rising stars can be identified from scholars' early career topic selection behavior" [6].

Currently, methods for measuring paper novelty are mainly based on topic novelty or citation diffusion effects. Calculation methods based on topic novelty can be obtained through topic age [6] or relative topic frequency. Calculating topic novelty requires a background corpus with sufficient time span and complete database. Meanwhile, paper subject terms are also knowledge elements of scholars' knowledge structures, and scholar knowledge structure novelty can also be calculated based on topic age.

2.4 Research Review

Existing research has achieved rich results in scholar knowledge structure, scholar influence prediction, and novelty, laying a solid foundation for this study. Summarizing existing research reveals: (1) Scholar knowledge structure can be constructed from knowledge topics and their interrelationships, and knowledge structure can reflect scholars' research capabilities, which are ultimately manifested as influence. (2) Predicting scholars' future influence through early academic performance (such as citation counts, h-index) is the current mainstream approach, but it has defects such as citation lag and insufficient historical data for young scholars. More importantly, most existing

research has overlooked the possibility of predicting scholars' future influence based on their early knowledge structure. (3) There is an inverted U-shaped relationship between academic achievement novelty and influence, and the underlying logic of academic achievement novelty is determined by scholars' selection of knowledge topics and establishment of inter-topic relationships, which is precisely the connotation of scholar knowledge structure. Therefore, there is also a close relationship between reasonable scholar knowledge structure and academic influence. The novelty characteristics of knowledge structure can be used to predict scholar influence; moreover, prediction from the perspective of scholars' research content features has the characteristic of immediacy, can compensate for the lag of traditional citation-based indicators, and can complement traditional indicators to jointly improve influence prediction effectiveness.

In summary, the internal logic of this paper is: constructing scholar knowledge structure through knowledge topics and their relationships, designing new indicators that fuse "complex network indicators + novelty" to measure scholar knowledge structure novelty, and using this as a basis to predict scholar influence. Through extensive experiments and comparisons, we verify the effectiveness of our indicators and analyze their potential value.

3.1 Research Design

To verify the capability and value of scholar knowledge structure novelty indicators in predicting scholar influence, this paper centers on indicator design, uses large-scale domain data, and conducts machine learning-based prediction tasks to verify indicator rationality. From an empirical process perspective, the research framework is designed as shown in Figure 1 [Figure 1: see original paper].

Figure 1 Research framework of this paper
1. Data source and processing
2. Indicator design and calculation
3. Experimental design and evaluation
- Large-scale biomedical scholar database
- Screen appropriate scholar groups
- Extract scholars' controlled subject terms
- Construct scholar knowledge structure based on term co-occurrence
- Topic-based indicators
- Network structure-based indicators
- Measure knowledge structure novelty under different indicators
- Machine learning classification prediction
- Algorithms: Logistic Regression, Back Propagation Neural Network, Random Forest
- Prediction evaluation (P/R/F1)
- Result analysis and discussion

Specifically: (1) This paper adopts a large-scale, long-time-span complete biomedical scholar database as the data source, screens scholars from the database according to certain criteria as research subjects, and classifies and labels scholars based on specific prediction tasks. (2) For screened scholars, we construct each scholar's knowledge structure based on subject terms annotated for their published articles in the database. We discuss and design novelty indicators from two levels—theme/theme combinations and their positions in the knowledge structure—and measure scholars' knowledge structures. (3) We transform scholar influence prediction into a binary classification task based on machine learning algorithms, examine the predictive performance of scholar knowledge structure novelty over long time periods, and finally evaluate and analyze prediction results to provide understanding of the relationship between novelty and scholar influence.

Main experimental steps: (1) Calculate scholar knowledge structure novelty indicators and influence indicators for each time window in scholars' early careers using 5-year, 10-year, 15-year, and 20-year periods. (2) Rank scholars based on cumulative citation counts in each time window, labeling the top 10% (following Bornmann's definition of highly cited papers [22,32]) as outstanding scholars for pre-classification. (3) Design two sets of time-span predictions according to scholar growth patterns: using the first 5 years of academic performance to predict influence classification in the 15th year, and using the first 10 years to predict influence classification in the 20th year.

3.2 Scholar Knowledge Structure and Measurement Indicators

Following the description in study [3], this paper defines scholar knowledge structure as "the composition and combination patterns of knowledge possessed by an individual." Similar to study [10], we use topics as basic knowledge elements and present knowledge combination patterns through document co-occurrence of topics (topic combinations). Since this paper takes the biomedical field as an example, we can obtain controlled subject terms annotated for each scholar's paper. Subject terms cover the paper's subfield and research content, and the topic network contains the knowledge and inter-topic relationships that scholars absorbed and created early in their careers, which is defined as the scholar knowledge structure in this paper. Compared with author keywords, controlled subject terms are more standardized, better solving synonym problems and thus reducing noise in scholar knowledge structures. Additionally, since the number of controlled subject terms is far less than the numerous author keywords, it is more convenient and accurate for large-scale calculation of term novelty.

The specific construction process is shown in Figure 2 [Figure 2: see original paper].

Figure 2 [Figure 2: see original paper] Workflow for constructing scholar's knowledge structure

Based on the generated knowledge structure, this paper designed six novelty indicators. Among them, topic novelty and topic combination novelty are commonly used indicators in existing research [6]; while indicators that fuse novelty connotation from complex network structure perspectives are newly designed in this paper, considering that the positions of topics and topic combinations in the knowledge structure contain rich and non-negligible innovative behaviors. The symbols and definitions used for related concepts are explained in Table 1 .

Table 1 Definitions of symbols used in this paper

Symbol	Definition
A	All scholars in the biomedical field
α	Individual scholar, α A
nα	Cumulative number of articles published by scholar α
nα,i	The i-th article of scholar α
mα,i	Number of subject terms in article nα,i
Xα,i	Set of subject terms in article nα,i
Xα,i,k	The k-th subject term in article nα,i
tα,i	Publication year of article nα,i
txα,i,k	Year when subject term Xα,i,k was first introduced into the research field

(1) Topic Novelty (TN)

The main elements of a scholar's knowledge network are topics and topic combinations. What topics scholars choose to build their knowledge structure is a key factor determining their output, and conducting research using highly novel knowledge is the source of innovation. Therefore, we must first determine whether topic units in a scholar's knowledge structure are novel. Tu and Seng proposed an algorithm to quantify the novelty index of subject terms based on time [31], which uses the reciprocal of the time span between the year a term was first adopted in a specific field and the year it is currently being used to measure the term's novelty when used. Building on this, Huang et al. introduced a temperature parameter λ to control the decay rate of the Sigmoid curve [6], with an appropriate λ value (20 in the paper) effectively distinguishing the novelty of different subject terms at specific times. The calculation is as follows:

$$TN(x_{\alpha,i,k},t)=\frac{2}{(1+e^{-\lambda(t-t_{x_{\alpha,i,k}})})},\quad t\in[t_{x_{\alpha,i,k}},t_{\alpha,i}]\quad \text{Formula (1)}$$

Where $TN(x_{\alpha,i,k},t)$ is the novelty of the k-th subject term in article $n_{\alpha,i}$ published by scholar α in year t, and $t-t_{x_{\alpha,i,k}}$ is the time span between the year the subject term was first introduced into the field and the year the scholar

used it. The novelty value obtained from this formula ranges between 0 and 1, with larger values indicating higher novelty of the subject term, reaching 1 at $t = t_{x_{\alpha,i,k}}$ and approaching 0 at $t = +\infty$.

This paper uses the first recorded year of a subject term as the reference time for its first appearance in the field. Scholar α's Topic Novelty (TN) in year t is the average of novelty indices of all topics in the scholar's knowledge structure in year t, as shown below:

$$TN(\alpha,t) = \frac{1}{\sum_{i=1}^{n_\alpha} m_{\alpha,i}} \sum_{i=1}^{n_\alpha} \sum_{k=1}^{m_{\alpha,i}} TN(x_{\alpha,i,k},t), \quad t \in [t_{x_{\alpha,i,k}}, t_{\alpha,i}] \quad \text{Formula (2)}$$

(2) Topic Combination Novelty (TCN)

The arrangement and combination of topics by scholars are also important aspects of knowledge structure shaping and key pathways to achieving innovation. The more novel, larger-span, and less common the topic combination, the more likely it contains knowledge with higher innovative value and can lead to breakthrough progress [33]. Similarly, Topic Combination Novelty is derived from the time-decay algorithm using the reciprocal of the time span between when the topic combination was first adopted by the field and when the scholar used it [6]. A scholar's TCN in year t is the average novelty of all topic combinations in papers published that year, as shown in Formulas (3) and (4):

$$TCN(x_{\alpha,i,j}, x_{\alpha,i,k}, t) = \frac{2}{(1 + e^{-\lambda(t - t_{x_{\alpha,i,j}, x_{\alpha,i,k}})})}, \quad t \in [t_{x_{\alpha,i,j}, x_{\alpha,i,k}}, t_{\alpha,i}] \quad \text{Formula (3)}$$

$$TCN(\alpha,t) = \frac{1}{\sum_{i=1}^{n_\alpha} \binom{m_{\alpha,i}}{2}} \sum_{i=1}^{n_\alpha} \sum_{1 \le j < k \le m_{\alpha,i}} TCN(x_{\alpha,i,j}, x_{\alpha,i,k}, t), \quad t \in [t_{x_{\alpha,i,k}}, t_{\alpha,i}] \quad \text{Formula (4)}$$

(3) Novelty of Topic Degree Center (NTDC)

Beyond considering the novelty of topics/topic combinations, the position of topics in the knowledge structure is also important. When topics with higher novelty occupy core positions in a scholar's knowledge structure, it indicates that the scholar has conducted extensive research centered around these topics.

Using complex network principles, in a scholar's knowledge network, the topic node with the highest degree centrality is the one that has combined with the most other topics, representing the scholar's high-frequency use and multi-dimensional research of this topic knowledge. From the perspective of novelty maximization, ideally, the node with the highest novelty should be the node with the highest degree centrality, with other nodes ranked accordingly. Based

on this, this study designed an indicator to measure the gap between actual network conditions and the ideal state regarding node degree centrality—the closer to the ideal state, the higher the knowledge structure novelty. The specific calculation of NTDC is as follows.

NTDCE represents the ideal state novelty of topic degree center, where n is the number of topic nodes in the network. N_i is the novelty value of the topic ranked i-th (in descending order), and C_i is the degree centrality value of the topic node ranked i-th (in descending order). The product of corresponding N_i and C_i is calculated and averaged. The formula is:

$$NTDCE = \frac{1}{n}\sum_{i=1}^{n} N_i \times C_i \quad \text{Formula (5)}$$

In a real scholar knowledge network, we calculate the average of the product of each topic node's actual novelty N_i' and actual degree centrality C_i' to obtain the actual novelty of topic degree center, NTDCR:

$$NTDCR = \frac{1}{n}\sum_{i=1}^{n} N_i' \times C_i' \quad \text{Formula (6)}$$

By dividing the ideal knowledge unit degree center novelty by the actual situation, the ratio is used to measure the novelty degree of the scholar's knowledge network. NTDC values range between 0 and 1, with larger values indicating higher knowledge structure novelty:

$$NTDC = \frac{NTDCR}{NTDCE} \quad \text{Formula (7)}$$

(4) Novelty of Topic Combination Degree Centrality (NTCDC)
Topic combinations exist as edges in scholar knowledge networks. Although complex network theory lacks formulas for edge centrality, from the practical meaning of this study, we use the average degree centrality of the two nodes in a topic combination as the degree centrality of that topic combination. Using the same calculation method as NTDC, we can obtain NTCDC. Since it considers two topics simultaneously, NTCDC more prominently emphasizes the importance of novel topics' positions in the knowledge structure compared to NTDC.

(5) Novelty of Topic Unit Betweenness Centrality (NTBC)
Using the same calculation method, but replacing degree centrality with betweenness centrality of topics in the network, we obtain NTBC. Betweenness centrality focuses on nodes' control over information flow in the network. Nodes with high betweenness centrality may not have high degree centrality but play a potential indirect role in network connectivity. Therefore, NTBC measures

whether novel topics serve to connect other topics. Higher NTBC values indicate that scholars are better at using highly novel topics to connect existing old topics in their knowledge structure.

(6) Novelty of Topic Combination Betweenness Centrality (NTCBC) Similarly, using the average betweenness centrality of the two subject term nodes in a topic combination in the knowledge network as the betweenness centrality of that topic combination, we can obtain NTCBC. Since it considers the betweenness positions of two subject terms, NTCBC further highlights scholars' behavior of using highly novel topics to connect old topics, building upon NTBC.

In total, six novelty indicators are obtained. To compare with traditional methods, we include three external features in the indicator system: scholars' early citation counts, journal impact factors, and h-index. These three classic indicators are selected because they measure scholars' academic performance from the perspective of influence. To avoid excessive interference factors, collaboration indicators are not included in the prediction analysis. The names, abbreviations, and descriptions of the nine indicators involved in the experiments are summarized in Table 2 .

Table 2 The indicators of novelty for scholar's knowledge structure

Indicator Type	Name	Code	Description
Knowledge Structure Novelty Indicators	Topic Novelty	TN	Temporal advancement degree of topics [6,31]
	Topic Combination Novelty	TCN	Temporal advancement degree of topic combinations [6]
	Novelty of Topic Degree Center	NTDC	Matching degree between topic novelty and degree centrality
	Novelty of Topic Combination Degree Centrality	NTCDC	Matching degree between topic combination novelty and degree centrality
	Novelty of Topic Unit Betweenness Centrality	NTBC	Matching degree between topic novelty and betweenness centrality

Indicator Type	Name	Code	Description
	Novelty of Topic Combination Betweenness Centrality	NTCBC	Matching degree between topic combination novelty and betweenness centrality
Influence Indicators	Citation Count	C	Citation count of scholar's publications
	H-index	H	Scholar's h-index
	Journal Impact Factor	IF	Impact factor of journals where scholar's papers are published

3.3 Prediction Models and Evaluation Metrics

The focus of this paper is to explore the predictive capability of scholar knowledge structure novelty indicators, not to pursue more powerful prediction algorithms. Therefore, three classic algorithms are selected for experiments: Logistic Regression (LR), Back Propagation Neural Network (BP), and Random Forest (RF). LR is a generalized linear model suitable for linearly separable problems; BP is a classic algorithm for handling non-linear problems; RF is a classic ensemble algorithm. These three algorithms have distinct characteristics and are used to test the effectiveness of prediction tasks.

Since this paper transforms prediction into a binary classification machine learning task—judging whether a scholar will enter the high-influence category in a future period based on early knowledge structure novelty indicators—we adopt the conventional F1-score to evaluate model performance. F1-score is the harmonic mean of Precision and Recall, with higher values indicating better prediction performance.

4.1 Data Sources and Processing

In the post-pandemic era, human health has become a global priority, making biomedicine an increasingly hot topic in academia. Meanwhile, due to the rich knowledge concepts in the biomedical field, some biomedical literature retrieval databases have become relatively mature. The PubMed Knowledge Graph (PKG) is a large-scale knowledge base built on PubMed, jointly developed by Professor Ying Ding, Professor Jian Xu, and their teams [34]. It represents literature, concepts, entities, and their relationships from PubMed in graph form, integrating vast amounts of biomedical knowledge. The database uses a hybrid method for author name disambiguation, claiming an F1-value of

98.09% for author disambiguation, indicating high data quality. Additionally, literature in the database is annotated with subject terms from the Medical Subject Headings (MeSH) compiled by the U.S. National Library of Medicine (NLM). MeSH terms can summarize detailed topic concepts within papers and have high knowledge representation capability.

Therefore, this paper statistically obtained all 29,576 MeSH major subject terms appearing up to 2020 (the last complete data year in PKG) and recorded the publication year of the first article indexed with each MeSH term, marking it as the initial year when the knowledge concept entered the biomedical community. For scholar selection, we screened scholars whose academic careers began in 2000 or later, lasted at least 15 years, and had publication counts between 5 and 300, initially obtaining data on 72,156 scholars. After further cleaning and organization, we obtained data on 57,927 scholars with academic careers of 20 years or more, including their published papers, MeSH subject terms, and citation relationships between papers. Through large-scale computation, we obtained data on paper citations and journal impact factors; simultaneously, we constructed scholar knowledge structures and calculated novelty indicators.

4.2.1 Stratified Scholar Indicator Differences

Scholars were classified based on cumulative citation counts in the first 5 years. We compared the early academic performance of high-influence scholars (top 10%) and ordinary scholars (bottom 90%) across various indicators, as shown in Table 3 . The data in Table 3 shows that high-influence scholars have higher early-career citation counts, h-index values, and journal impact factor indices than ordinary scholars. Additionally, high-influence scholars have higher values in Topic Novelty (TN), Topic Combination Novelty (TCN), and Novelty of Topic Combination Betweenness Centrality (NTCBC); while ordinary scholars have higher values in Novelty of Topic Degree Center (NTDC), Novelty of Topic Combination Degree Centrality (NTCDC), and Novelty of Topic Unit Betweenness Centrality (NTBC).

Since influence indicators (C, H, IF) and novelty indicators (TN~NTCBC) have different dimensions (novelty indicators range from 0 to 1), they are not directly comparable. After conducting t-tests, we found significant differences in means and standard deviations for all indicators ($P < 0.001$), demonstrating that the indicators designed and selected in this paper are conducive to prediction. Second, due to the inherent correlation between early and future influence, high-influence scholars' early indicators are significantly higher than those of ordinary scholars. However, novelty indicators show inconsistent patterns: high-influence scholars exhibit higher novelty at the topic and topic combination levels in early stages, but the opposite is true for structural-level novelty (the reasons for which are discussed in the next subsection's correlation analysis). This inconsistency enhances indicator diversity and provides more options for predicting and deeply understanding scholar knowledge structure characteristics.

Table 3 Comparison of indicators between ordinary scholars and high academic impact scholars

Indicator	High-Influence Scholars (Top 10%)	Ordinary Scholars (Bottom 90%)
C	Higher	Lower
H	Higher	Lower
IF	Higher	Lower
TN	Higher	Lower
TCN	Higher	Lower
NTDC	Lower	Higher
NTCDC	Lower	Higher
NTBC	Lower	Higher
NTCBC	Higher	Lower

Note: All differences in means and standard deviations of indicators in the table are statistically significant ($P < 0.001$).

4.2.2 Correlation Between Novelty and Influence Indicators

To further understand the relationship between scholar knowledge structure novelty and academic influence, we conducted correlation analysis on the nine scholar indicators involved in the experiments. Using scholars' knowledge structure data from the first 5 years of their careers, we measured the strength of correlation between these nine indicators using Pearson correlation coefficients. We also measured the correlation between the nine indicators and Y (cumulative citation count in the 15th year). The results are shown in Table 4 .

Table 4 Correlation analysis of indicators

	C	H	IF	TN	TCN	NTDC	NTCDC	NTBC	NTCBC
C	1								
H	0.89	1							
IF	0.65	0.62	1						
TN	0.23	0.21	0.18	1					
TCN	0.15	0.14	0.12	0.67	1				
NTDC	-0.18	-0.16	-0.14	-0.05	-0.08	1			
NTCDC	-0.16	-0.15	-0.13	-0.06	-0.09	0.89	1		
NTBC	-0.12	-0.11	-0.10	-0.04	-0.07	0.78	0.75	1	
NTCBC	0.13	0.12	0.11	0.31	0.28	-0.15	-0.13	-0.11	1

Note: All correlations are statistically significant ($P < 0.001$).

The results show interesting correlations between knowledge structure novelty indicators and academic influence indicators. TN and TCN have significant positive correlations with citation count, h-index, and journal impact factor; while

network structure-based novelty indicators show mixed results—only NTCBC is positively correlated with academic influence, while the other three indicators are negatively correlated. This aligns with the findings in Table 3.

Analyzing the meaning behind the indicators, scholars with higher topic or topic combination novelty (i.e., more novel research topics) may achieve higher academic influence. However, scholars with higher network structure novelty (i.e., novel topics occupying central positions in the knowledge structure) show decreased academic influence, contrary to the original design expectation but bringing new insights for understanding scholar knowledge behavior. A possible explanation is that when nodes in overly central positions of the knowledge structure are too novel, it indicates that the scholar has made them their primary research direction. When such topics lack support from other mature, traditional topics, it suggests the research is too niche or specialized, making the output more difficult for other scholars to understand and accept. Interestingly, higher NTCBC often indicates that scholars have integrated old and new knowledge, making the knowledge structure novel yet reasonable. Compared to other network structure-based novelty indicators, NTCBC better reflects scholars' comprehensive ability to connect and apply knowledge, and therefore has a positive relationship with academic influence.

In summary, there are significant correlations between knowledge structure novelty and academic influence indicators, with differences across various indicators, supporting subsequent prediction experiments using different indicator combinations.

4.3 Prediction Results Analysis

We designed four groups of prediction experiments: single-indicator prediction, knowledge structure novelty indicator prediction, early academic influence indicator prediction, and comprehensive indicator prediction. Experiments divided data into training and test sets at an 8:2 ratio and used 5-fold cross-validation to ensure prediction reliability. The F1-values of the four groups of models under different prediction time windows are shown in the table.

Table 5 Evaluation results of four groups of experiments (F1 Score)

Indicator Type	Input Variables	5$→15|10→$20	
Single Novelty	TN	0.723	0.738
	TCN	0.654	0.671
	NTDC	0.712	0.728
	NTCDC	0.705	0.719
	NTBC	0.698	0.714
	NTCBC	0.718	0.732
Influence	C+H+IF	0.835	0.842
All Novelty	TN+TCN+NTDC+NTCDC+NTBC+NTCBC	0.756	0.768

Indicator Type	Input Variables	5$→15|10→$20	
Comprehensive	All 9 indicators	0.874	0.878

4.3.1 Single-Indicator Prediction Performance Analysis

The evaluation results in Table 5 show that academic influence-related indicators generally achieve higher prediction accuracy than scholar knowledge structure novelty indicators. Scholars with high early influence are more likely to maintain high influence characteristics in the next period. Although classification is based on citation counts, scholars' h-index shows better prediction performance than citation counts themselves, with average accuracy above 82%, confirming the scientific validity of the h-index [20].

Among knowledge structure novelty indicators, topic indicators generally outperform topic combination indicators in prediction effectiveness (Figure 3 [Figure 3: see original paper]). The best-performing indicator is TN, achieving the highest results across different time periods and models. A deeper understanding reveals that the novelty of topics used by scholars in early stages greatly relates to later academic influence, but topic combinations are too complex in principle to be suitable as single indicators for scholar classification prediction. Additionally, the four network structure-based indicators (NTDC, NTCDC, NTBC, and NTCBC) also outperform TCN, with maximum differences reaching 10.9% (10$→$20, BP, NTDC vs. TCN). This is an interesting finding. In current novelty research, topic combination novelty is the most commonly used indicator besides topic novelty. Therefore, this discovery provides inspiration for future research: network structure-level novelty indicators deserve adoption.

Figure 3 Comparison of prediction results using single novelty indicators (two group experiments)
(a) Predicting the 15th-year impact based on novelty indicators from the previous 5 years
(b) Predicting the 20th-year impact based on novelty indicators from the previous 10 years

Combining the numerical values in Table 5 with the visual comparisons in Figures 3(a) and 3(b), we can see that different indicators and algorithms show varying performance across different prediction windows (5$→15, 10→$20). However, overall prediction effectiveness remains consistent, demonstrating the robustness of novelty indicators in prediction tasks.

4.3.2 Comprehensive Indicator Prediction Performance Analysis

Using all six knowledge structure novelty indicators as variables simultaneously, the average prediction accuracy across all time periods and models reached over 75%, higher than any single indicator included. This demonstrates the scientific validity of the knowledge structure novelty indicator system design: measuring scholar knowledge structure novelty requires comprehensive evaluation from

multiple perspectives. Additionally, although academic influence indicators (C+H+IF) achieve higher overall accuracy than knowledge structure novelty indicators (6-indicator combination), prediction accuracy further improves when knowledge structure novelty indicators are added to form comprehensive indicators (all 9 indicators) (Figure 4 [Figure 4: see original paper]), with an average improvement of 2.7% across different time windows and algorithms. Specifically, in the 5$→15$prediction, the RF algorithm's F1-value improved from 84.3→$20 prediction, the comprehensive indicator achieved the maximum F1-value of 87.8% across all experiments.

It is worth noting that there is a natural autocorrelation between scholars' early and future influence in both connotation and numerical value, exhibiting cumulative effects. However, the relationship between novelty indicators and future influence is hidden, without numerical necessity. Achieving good prediction results using novelty indicators demonstrates their effectiveness for scholar influence prediction in the biomedical field. Similar to Figure 3, prediction tasks across different time windows show stable performance.

Figure 4 Comparison of prediction results using different combined indicators (two group experiments)
(a) Predicting the 15th-year impact based on novelty indicators from the previous 5 years
(b) Predicting the 20th-year impact based on novelty indicators from the previous 10 years

4.3.3 Model Comparison Analysis

Overall, the accuracy differences among the three classification algorithms are not substantial. In single-indicator predictions, BP performs best, as it excels at identifying complex features and has stronger fitting capabilities for non-linear indicators. In comprehensive indicator predictions with multi-variable input, Random Forest shows superior performance, reflecting its ability to integrate the respective advantages of multiple indicators.

5. Discussion and Conclusion

Using the PKG database, this study takes 57,927 scholars in the biomedical field as research subjects, designs a set of novelty indicators based on scholar knowledge structure characteristics, applies machine learning algorithms to predict scholars' future influence categories, and compares them with traditional influence-based external indicators through different indicator combination schemes to verify the predictive performance of knowledge structure novelty indicators. The study found that although knowledge structure novelty indicators' predictive performance is not as high as traditional influence indicators (h-index, etc.), the comprehensive indicator system incorporating knowledge structure novelty indicators improved prediction accuracy by an average of 2.7%. The 5-year prediction for the 15th year reached up to 87.4%, and the

10-year prediction for the 20th year reached up to 87.8%, demonstrating that novelty indicators can compensate for the shortcomings of relying solely on external indicators.

Compared with previous research, this paper provides a content-based perspective from scholars' knowledge structures, deeply 挖掘 the academic value inherent in knowledge structures, and demonstrates the effectiveness and applicability of scholar knowledge structure novelty indicators. Under the "breaking the five-only" policy orientation for sci-tech evaluation, the academic community needs to explore establishing a more comprehensive, fair, and transparent academic evaluation system. The identification and prediction of outstanding scholars also need to start from the underlying logic of scholar innovation. Different from traditional academic achievement indicators, scholar knowledge structure novelty focuses on measuring scholars' comprehensive qualities and deep-level knowledge contributions, not relying on the "quantity" of academic achievements but paying more attention to the innovative potential contained within, providing new thinking for scholar influence prediction work. Meanwhile, this study also provides inspiration for innovative design of novelty evaluation indicators: indicators based on scholar knowledge structure characteristics (such as NTDC, NTCBC) have better predictive capability than topic combination novelty indicators.

This study has several limitations: First, the experiments are conducted in the biomedical field, and whether the predictive performance of the designed knowledge structure novelty indicators adapts to other fields remains to be verified. However, the indicators themselves are not domain-restricted. Since the PubMed database underlying PKG annotates papers with controlled subject terms, reducing the cost of knowledge structure construction; when studying other disciplines (e.g., humanities and social sciences lacking controlled subject term annotations), one can follow the approach in literature [10] to extract topics through natural language processing techniques, construct knowledge networks, and calculate knowledge structure novelty indicators under long time-span datasets to conduct prediction tasks. Second, this study uses traditional binary classification models and long-period experiments across two time segments to validate indicator performance, leaving room for improvement in parameter adjustment granularity and indicator combination richness. Future work will incorporate factors such as scholar stratification ratios (e.g., identifying top 2%, top 5% scholars as outstanding) and influence indicator autocorrelation into the research scope to further refine experimental designs. Additionally, investigating the predictive capability of early knowledge structure novelty for future disruptive achievements [29] is also an interesting direction.

References

[1] XIA W, LI T, LI C. A review of scientific impact prediction: tasks, features and methods[J]. Scientometrics, 2023, 128(1): 543-585.

[2] MUSTAFA G, RAUF A, AFZAL M T. GK index: bridging Gf and K indices for comprehensive author evaluation[J]. Knowledge and information systems, 2024: 1-36.

[3] LIU P, SHI X, XING Y Y. Modeling and representation of tacit knowledge infrastructure of outstanding scholars[J]. Journal of the China society for scientific and technical information, 2023, 42(08): 915-925.

[4] ZENG A, SHEN Z, ZHOU J, et al. Increasing trend of scientists to switch between topics[J]. Nature communications, 2019, 10(1): 3439.

[5] YU H, MARSCHKE G, ROSS M B, et al. Publish or perish: selective attrition as a unifying explanation for patterns in innovation over the career[J]. Journal of human resources, 2023, 58(4):1307-1346.

[6] HUANG S, LU W, BU Y, et al. Revisiting the exploration-exploitation behavior of scholars' research topic selection: Evidence from a large-scale bibliographic database[J]. Information processing & management, 2022, 59(6): 103110.

[7] PARK M, LEAHEY E, FUNK R J. Papers and patents are becoming less disruptive over time[J]. Nature, 2023, 613(7942):138-144.

[8] LIANG X K. Novelty, conventionality, and scientific impact of papers in library and information science in China: evidence from papers in CSSCI (2000-2019) [J]. Library and information service, 2022, 66(20): 148-161.

[9] QIN C L, ZHANG C Z. Problems and responses strategy of peer review in the context of big data[J]. Information studies: theory & application, 2021, 44(4): 99-112.

[10] XU H, LUO R, WINNINK J, et al. A methodology for identifying breakthrough topics using structural entropy[J]. Information processing & management, 2022, 59(2): 102862.

[11] MIN C, BU Y, SUN J. Predicting scientific breakthroughs based on knowledge structure variations[J]. Technological forecasting and social change, 2021, 164: 120502.

[12] ZHANG T, TAN F, YU C, et al. Understanding relationship between topic selection and academic performance of scientific teams based on entity popularity trend[J]. Aslib journal of information management, 2023, 75(3): 561-588.

[13] JIA T, WANG D, SZYMANSKI B K. Quantifying patterns of research-interest evolution[J]. Nature human behaviour, 2017, 1(4): 0078.

[14] TENG G Q, HE D F, PENG J, et al. Structure and order: the evolution of thought on structuralism in knowledge organization[J]. Information studies: theory & application, 2015, 38(04): 6-10.

[15] BROOKES B C. The foundations of information science. Part I. Philosophical aspects[J]. Journal of information science, 1980, 2(3-4): 125-133.

[16] WU Z X, HE C, ZHAO K R. Research on the knowledge accumulation behavior of scholars based on the measurement of subject headings distribution[J]. Information studies: theory & application, 2022, 45(07): 140-147.

[17] QIU P F, SUN J J, MIN C. Review of and thoughts on mentor-ship in scientific research[J]. Library and information, 2018, (05): 50-55+118.

[18] CHEN G, ZHAO Y X. A network evolution model for domain knowledge driven by multiple factors: Following suit, conservatism, and innovation[J]. Journal of the China society for scientific and technical information, 2020, 039(001):1-11.

[19] ZHOU C Y, HE Y J, LIU L F. A method for predicting rising stars by combining co-authors' diversity and influence[J]. Information studies: theory & application, 2020, 43(02): 78-83+71.

[20] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the national academy of sciences, 2005, 102(46): 16569-16572.

[21] ACUNA D E, ALLESINA S, KORDING K P. Predicting scientific success[J]. Nature, 2012, 489(7415): 201-202.

[22] LINDAHL J. Predicting research excellence at the individual level: The importance of publication rate, top journal publications, and top 10% publications in the case of early career mathematicians[J]. Journal of informetrics, 2018, 12(2): 518-533.

[23] PANAGOPOULOS G, TSATSARONIS G, VARLAMIS I. Detecting rising stars in dynamic collaborative networks[J]. Journal of informetrics, 2017, 11(1): 198-222.

[24] LINDAHL J, COLLIANDER C, DANELL R. Early career performance and its correlation with gender and publication output during doctoral education[J]. Scientometrics, 2020, 122(1): 309-330.

[25] NIE Y, ZHU Y, LIN Q, et al. Academic rising star prediction via scholar's evaluation model and machine learning techniques[J]. Scientometrics, 2019, 120(2): 461-476.

[26] DAUD A, SONG M, HAYAT M K, et al. Finding rising stars in bibliometric networks[J]. Scientometrics, 2020, 124: 633-661.

[27] DONG K, CHEN X P, WU J C. Research on the correlation between creativity and citation impact of scientific research paper: measurement from semantic perspective[J]. Information studies: theory & application, 2023, 46(10): 24-31.

[28] YAN Y, TIAN S, ZHANG J. The impact of a paper's new combinations and new components on its citation[J]. Scientometrics, 2020, 122: 895-913.

[29] RUAN X, AO W, LYU D, et al. Effect of the topic-combination novelty on the disruption and impact of scientific articles: evidence from PubMed[J].

Journal of information science, 2023: 1-15.

[30] HE J J, MIN C. Research on the features and influencing factors of papers with continuous citation growth[J]. Library and information service, 2022, 66(07): 110-119.

[31] TU Y N, SENG J L. Indices of novelty for emerging topic detection[J]. Information processing & management, 2012, 48(2): 303-325.

[32] BORNMANN L. How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature[J]. Research evaluation, 2014, 23(2): 166-173.

[33] LU Q, QIN Y Y, CHEN J. Identification of interdisciplinary "technology-topic" innovation combinations: take artificial intelligence technology driving the innovation in LIS as an Example[J]. Library and information service, 2024, 68(02): 50-61.

[34] XU J, KIM S, SONG M, et al. Building a PubMed knowledge graph[J]. Scientific data, 2020, 7(1): 205.

Author Contributions

Wu Zhixiang: Designed research framework, wrote and revised the paper;
Yi Haigang: Data collection and computation, prediction experiments;
Zhu Yimeng: Indicator design, paper writing;
Wang Pei: Data collection and computation;
Wang Hao: Proposed research ideas, finalized the paper.

Note: Figure translations are in progress. See original paper for figures.