NovSRC: A Novelty-Oriented Scientific Collaborators Recommendation Model

Xiuxiu Li[1], Mingyang Wang[2*], Chaoran Wang[3], Yujia Fu[4], Xianjie Wang[5]
College of Computer and Control Engineering, Northeast Forestry University, Harbin, China[1, 2, 3, 4]
Harbin Institute of Technology, Harbin, China[5]

***Abstract*—Collaborator recommendation is a crucial topic in research management. This paper proposes a Novelty-Oriented Scientific Research Collaborator recommendation model (NovSRC). By recommending collaborators under the guidance of novel indicators, NovSRC aims to broaden scholars' research perspectives and facilitate the progress of research innovation. NovSRC utilizes heterogeneous academic networks composed of different academic entities and their relationships to learn vector representations of scholars and quantify their novelty metrics. A weighted academic collaboration network was constructed by measuring the novelty collaboration strength (NCS) among scholars under the novelty index, and based on this network, the final vector representation of scholars under the guidance of novelty characteristics was learned. By calculating the similarity between scholar vectors, NovSRC generates a Top-N recommendation list with a focus on novelty. The experimental results indicate that NovSRC achieved the best recommendation performance. Compared with the baseline models, the recommendation precision of NovSRC has improved by 6.9%, the F1 value has increased by 17.3%, and the novelty collaboration strength among scholars has increased by 3.3%. The analysis of the recommended list shows that compared to the target scholars, scholars recommended by the NovSRC model exhibit a wider distribution of research interests, which confirms that novelty has become a key benchmark factor for scholars seeking collaborators.**

Keywords*—Scientific collaborator recommendation; novelty; heterogeneous academic collaboration network; network representation learning

I. INTRODUCTION

Nowadays, scientific research is developing towards the direction of synthesis and diversification of disciplines. It is also increasingly difficult for scholars to discover new knowledge and propose new theories, which makes academic cooperation a new trend to break through scientific research problems. Academic cooperation can remove geographical restrictions and promote complementary advantages for scholars. However, in the face of academic big data and information overload, researchers often find it difficult to effectively select collaborators who match their research interests and can bring novel insights. How to help scientific researchers quickly and efficiently find their interested collaborators in massive data has always been a bottleneck that restricts the effectiveness of academic cooperation recommendations.

Existing collaborator recommendation methods focus on the similarity of scholars' research interests to achieve high similarity recommendation results, aiming to recommend collaborators closest to the target scholar's research interests. However, this strategy is difficult to bring more sparks of innovative thinking to the target scholars. A new recommendation strategy is needed to help them expand their innovative perspectives and improve their research level. The introduction of novelty is the key to solving this problem, as it can enrich academic cooperation models and meet the diverse cooperation needs of scholars.

In this paper, a new research collaborator recommendation model NovSRC is proposed which considers the novelty characters of collaborators. By examining the similarity and diversity of research interests among scholars, as well as the differences in academic influence among scholars, this model establishes an indicator system to measure the intensity of novelty cooperation among scholars. Under the guidance of this indicator system, the model learns the novelty representation vector of scholars and generates a recommendation list of collaborators based on this. The main contributions of the NovSRC model are as follows:

1) NovSRC model quantifies the intensity of collaboration between scholars in terms of the orientation of novelty. Based on a heterogeneous academic network composed of heterogeneous academic entities and their relationships, NovSRC quantifies the similarity and diversity of research interests between scholars, as well as the differences in academic influence between scholars. Based on these three indicators, NovSRC calculates the strength of novelty cooperation between scholars.

2) NovSRC model achieves novelty-oriented representation vectors of scholars. Based on a collaborative network with the novelty cooperation strength as the edge weight, NovSRC designed a random walk process guided by the edge weight, and finally learned the novelty orientation representation vectors of scholars.

3) NovSRC model obtains a list of collaborator recommendations based on novelty orientation. Based on the novelty scholar vectors, NovSRC calculates the similarity between scholar vectors and generates the novelty-oriented scholar recommendation list. Experimental results show that compared with the baseline models, the NovSRC model can achieve more accurate recommendation results.

*Corresponding Author.

II. Related Work

In collaborator recommendation research, researchers mainly recommend potential collaborators to target scholars from the perspective of similarity.

As a popular method, the similarity-based research collaboration recommendation system builds scholars' interest profiles and constructs their "portraits" by extracting the research topics or keywords of their published papers, and accordingly recommends collaborators with similar research interests [1]. Chen et al. [2] constructed a heterogeneous network of institutions and collaborator networks and based on this, a random walk algorithm method was used to recommend academic collaborators. Zhang et al. [3] proposed a research collaboration recommendation method that integrates network representation learning and author topic models, and combines author structural similarity and author topic similarity to generate a recommendation list. Pradhan et al. [4] designed DRACoR, a multi-level fusion-based model for collaborator recommendation, which integrated the deep learning-boosted collaborator recommendation model and meta-path aggregated random walk based collaborator recommendation model, to generate a list of collaborators to recommend. Hu et al. [5] proposed a collaborator recommendation model CRISI that integrates the author's cooperation strength and research interests on the attribute graph. The quality of the recommended nodes is improved by double-weighting the structure and attributes and using the node replacement method. Kumara et al. [6] used Google Scholar archives to construct collaborative networks by extracting co-authors, similarities in areas of interest, citation rates, and multiple papers co-authored between scholars. Du et al. [7] utilized the Node2vec representation learning method to capture information from nodes in the research network, and integrated the institutional cooperation preferences among authors and the similarity in academic levels to obtain recommendation results. Du et al. [8] proposed an academic collaborator recommendation model ACR-ANE based on attribute network embedding. This model makes full use of the network topology and multi-type scholar attributes to enhance scholar embedding, and employs a deep auto-encoder to encode the structure of the academic collaboration network and attributes of scholars into low-dimensional representation vectors for collaborative recommendation. Jagadishwari et al. [9] used a collaborative filtering method to help identify collaborators based on the research interests and the papers published by the researchers. Liu et al. [10] proposed a heterogeneous network embedding recommendation model HNERec. This method uses four meta-path random walks of topic relationship, authorship, citation relationship, and venue relationship to traverse the heterogeneous network randomly, and utilizes the skip-gram model to embed the nodes, and finally generates a recommendation list based on the similarity between the corresponding node vectors.

However, considering similarity alone makes it difficult to broaden the research perspectives, and over time, it may reduce scholars' satisfaction with the collaboration recommendation system [11]. In recent years, researchers have gradually integrated novelty indicators into recommender systems [12]. By introducing novelty indicators, the recommendation results are no longer limited to high similarity, improving the innovation of the recommendation results, and providing surprise choices for target users. Zhang et al. [13] proposed a serendipity-oriented next point-of-interest recommendation model, SNPR, and designed a transformer-based neural network to capture the complex interdependencies of POIs in a user's clicking sequence by weighing relevance and unexpectedness. Ziarani et al. [14] proposed a deep neural network approach for a serendipity-oriented recommendation system, using unexpectedness and relevance parameters to compose focus shift points to generate novelty recommendations by integrating Convolutional Neural Networks and Particle Swarm Optimization algorithm. However, most of these studies are based on product recommendation systems, and only a few studies have introduced them into the research collaborators recommender systems. Gao et al. [15] proposed a community outlier detection algorithm to identify abnormal academic conferences and scholars with more research topics in the academic community. Xu et al. [16] proposed the Seren2vec network representation learning algorithm to provide serendipitous scientific collaborators by generating accidental bias vectors of scholar nodes. Ding [17] proposed a paper recommendation algorithm based on novelty and influence, which improved the traditional citation network graph by combining the novelty and impact of a paper, and used a restarted random wandering algorithm to make recommendations.

In summary, collaborator recommendations based on similarity can improve the relevance of recommendations and ensure that the research directions of the recommenders and the target scholar are highly consistent. However, relying solely on similarity to generate collaborators is difficult to effectively expand the research perspective of the target scholar. In the field of academic collaboration, researchers hope to collaborate with scholars with different research perspectives to obtain relevant but different ideas or knowledge. Therefore, introducing novelty elements into recommendation systems will help meet the needs of researchers.

III. Methodology

Fig. 1 shows the overall framework of the NovSRC model. The NovSRC model consists of four modules: Initial encoding module, Novelty indicator calculation module, Novelty-oriented encoding module, and Collaborator recommendation module. These modules are used for encoding the initial vectors of scholars, quantifying and calculating the novelty indicators of scholars, learning scholar vectors based on novelty orientation, and recommending novelty-oriented collaborators.

A. Initial Encoding Module

In the Initial Encoding Module, a scholar representation vector learning process based on heterogeneous academic networks is designed to obtain the initial scholar representation vectors. The module adopts a hybrid encoding of content and structural features to fully examine the content and structural attributes of scholars in research interests. In the process of extracting research interest content features, this module uses LSTM and multi-head attention mechanisms to capture the overall and recent research interests of scholars to show the

dynamic evolution characteristics of scholars' research interests over time. In the process of extracting structural features of research interest, an embedding learning process based on meta-path graph sampling is used to generate the structural features of scholars. And the hybrid encoding process uses the attention mechanism to integrate the scholar features obtained from the content and structural dimensions to obtain the initial representation vectors of the scholars. Fig. 2 shows the process of initial encoding of the scholar vectors.

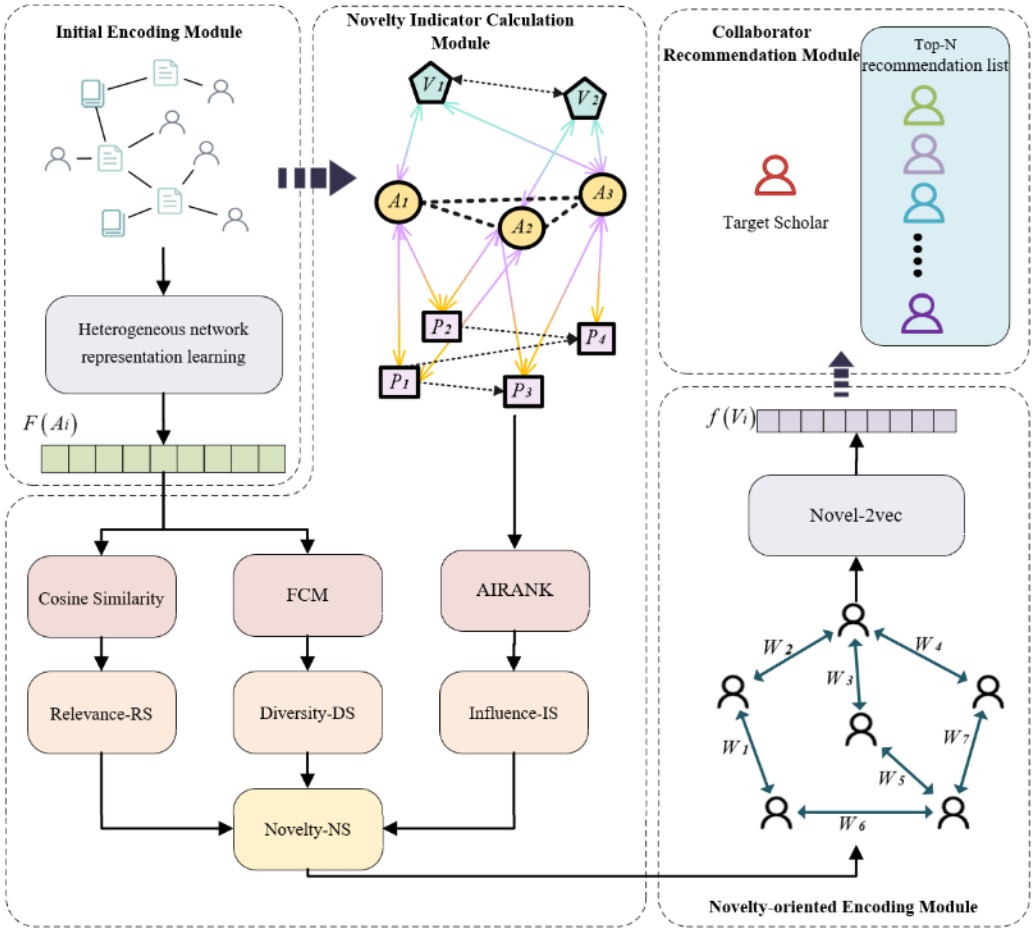

Fig. 1. The overall architecture of NovSRC model.

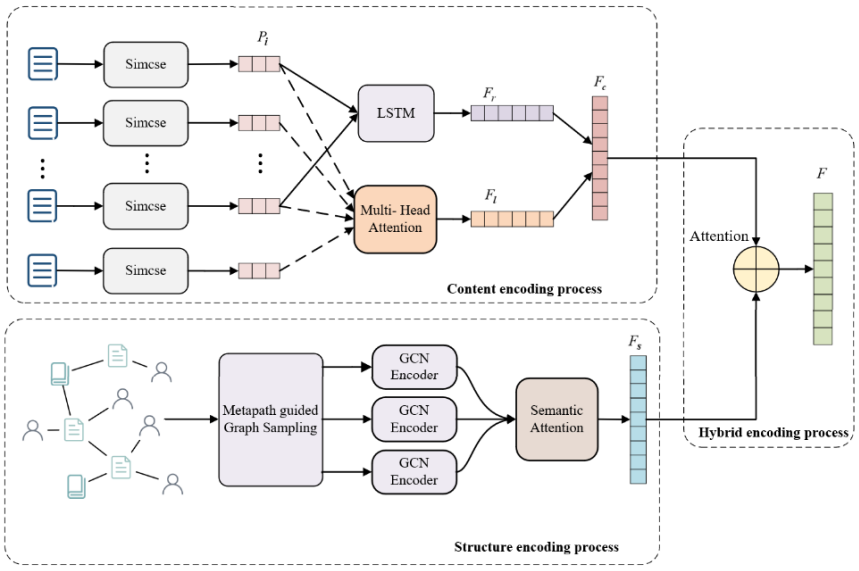

Fig. 2. The process of initial encoding of the scholar vectors.

The content features encoding process aims to learn the scholars' research interests in the content dimension. Since the articles published by the scholars can directly reflect their research interests, we use the scholars' articles as a basis to capture the scholars' research interests in the content dimension.

The titles of the articles published by the scholar are inputinto the SimCSE model [18] to learn the initial vector of the article. Then the vector is input into the multi-head attention layer to learn the scholar's overall research interest feature f_l. Meanwhile, we extract the scholar's latest published article representation sequence $\{P_1, P_2, \cdots, P_r\}$ (in this paper, r=3), and the representation sequence are input into the LSTM model to obtain the scholar's recent interest features f_r; Finally, we integrate the scholar's overall and recent interest features to obtain the scholar's feature representation in the content dimension F_c. The scholar's content features represent the learning process, which are formulated as shown in Eq. (1) to (4).

$$F_c = Concat(f_l, f_r) \tag{1}$$

$$f_l = \sum_1^n Concat(SA_1, \cdots, SA_m)W^o \tag{2}$$

$$SA_i = Softmax\left(\frac{(W_Q P_i)(W_K P_i)^T}{\sqrt{d}}\right)(W_V P_i) \tag{3}$$

$$f_r = LSTM(P_1, P_2, \cdots, P_L) \tag{4}$$

where SA_irepresents the single-head attention output result of each article, m is the number of heads in attention mechanism, d represents the dimension of P_i, W represents the weight coefficient.

The structural feature encoding process aims to learn the scholar's interest vector of structural dimensions derived from the association relationships between academic entities. In our previous work [19], the authors proposed a heterogeneous network representation learning process based on meta-path subgraph sampling. We introduce the process to encode the structural features of scholars' research interests. According to the heterogeneous academic network composed of the three academic entities of scholar-paper-journal and the relationship between them, three meta-paths are selected with the scholar node as the head node and tail node: scholars-papers-scholars (APA), scholars-papers-papers-scholars (APPA), and scholars-papers-journals-papers-scholars (APVPA). Homogeneous graphs are extracted from the heterogeneous academic network based on these three meta-paths. These homogeneous graphs can reflect the meta-path level neighbor relationships between scholars, which makes the aggregated representation learning process utilize richer network semantic information. On the homogeneous subgraph mapped by a certain meta-path, the neighborhood node set of the target node is obtained using the uniform sampling method. And the Graph Convolutional Network (GCN) is used to aggregate information from the neighbors of the neighborhood node set to generate the representation vector for the target scholar node.

The process for generating the target scholar embedding vector using GCN based on the meta-path $\mathcal{P}_i$ can be formulated as shown in Eq. (5).

$$A^{\mathcal{P}_i} = \left(D^{\mathcal{P}_i - \frac{1}{2}} N^{\mathcal{P}_i} D^{\mathcal{P}_i - \frac{1}{2}}\right) X W^{\mathcal{P}_i} \tag{5}$$

where $A^{\mathcal{P}_i}$ is the embedding vector of the target scholar node in the graph sampled by the meta-path $\mathcal{P}_i$, X represents the initial feature matrix of the scholar node, $D^{\mathcal{P}_i}$ is the degree matrix under meta-path $\mathcal{P}_i$, $N^{\mathcal{P}_i}$ is the adjacency matrix under $\mathcal{P}_i$, and $W^{\mathcal{P}_i}$ is the parameter matrix.

As a result, embedded vectors are obtained for scholars under different meta-paths. The final scholar's vector in the structural dimension is obtained by aggregating the embedded vectors of scholars under different meta-paths. A semantic-level attention mechanism is introduced to quantify the weight of semantic information provided by different meta-paths, and then the scholar vectors learned from different meta-paths are aggregated to obtain the scholar's interest vector $F_s(A)$ in the structural dimension. The aggregation process is shown in Eq. (6) to (9).

$$F_s(A) = \sum_{i=1}^{P} Att_{\mathcal{P}_i} \cdot A^{\mathcal{P}_i} \tag{6}$$

$$Att_{\mathcal{P}_i} = Softmax\left(W_{\mathcal{P}_i}\right) = \frac{exp\left(W_{\mathcal{P}_i}\right)}{\sum_{j=1}^{P} exp\left(W_{\mathcal{P}_i}\right)} \tag{7}$$

$$U_{\mathcal{P}_i} = Tanh(H^{\mathcal{P}_i} W + B) \tag{8}$$

$$W_{\mathcal{P}_i} = U_{\mathcal{P}_i} \cdot Q^T \tag{9}$$

where $Att_{\mathcal{P}_i}$is normalized by using the $Softmax$ function on $W_{\mathcal{P}_i}$, $W_{\mathcal{P}_i}$ represents he weight matrix of meta-paths under the self-attention mechanism obtained by multiplying the key vector matrix $U_{\mathcal{P}_i}$ and query vector matrix Q^T. $A^{\mathcal{P}_i}$ is obtained by mapping the vector matrix $U_{\mathcal{P}_i}$ through a layer of MLP using $Tanh$ as the activation function. W, B, and Q^T are training parameters.

In the hybrid encoding process, the attention mechanism is used to integrate the content feature vector and structural feature vector of scholars to obtain the final scholar vector representation $F(A_i)$ is shown in Eq. (10) to (13).

$$F(A_i) = W_1 \cdot F_c(A_i) + W_2 \cdot F_s(A_i) \tag{10}$$

$$W_i = Softmax(W_i) = \frac{exp(W_i)}{\sum_{j=1}^{2} exp(W_i)} \tag{11}$$

$$W_1 = Q \cdot F_c(A_i) \tag{12}$$

$$W_2 = Q \cdot F_s(A_i) \tag{13}$$

where W_1 denotes the weight matrix of scholar content features, W_2 represents the weight matrix of scholar structure features, and Q is a trainable parameter of the model.

The scholars' initial vectors obtained in the Initial encoding module are used as the basic data to calculate the similarity and diversity of scholars' research interests.

B. Novelty Indicator Calculation Module

For scientific cooperation, similarity in academic knowledge and research interests of scholars is still the cornerstone for establishing collaborative relationships, which avoids communication barriers caused by differences in

professional knowledge between scholars in cooperation. At the same time, collaborative relationships should be able to provide more perspectives to help solve scientific problems, which requires that collaborators have different and more diversified research interests than the target scholars. In addition, the scholars should have comparable academic influence, which is conducive to the development of the collaborative relationship. In summary, we evaluate the index system of novelty elements by three indicators: the similarity, the diversity of the scholars' research interests and the academic influence of the scholars.

1) Similarity score: The similarity score between scholars is obtained by calculating the cosine similarity between the scholar vectors obtained by the initial encoding module to evaluate the similarity of the scholars' research interests. The similarity score is shown in Eq. (14).

$$RS(A_i, A_j) = \frac{F(A_i) \cdot F(A_j)}{\sqrt{\|F(A_i)\| \|F(A_j)\|}} \quad (14)$$

where $F(A_i)$ and $F(A_j)$ are the representation vectors of the scholars' nodes A_i and A_j, respectively.

2) Diversity score: The Fuzzy C-means (FCM), which can divide samples into different clusters, is used to capture the diversity of scholars' research interests. In the clustering process, we first set the total number of clusters C=10, and randomly assign each scholar node probability vectors for each class of clusters. Then the cluster center of each cluster and the distance between each scholar node and the cluster center are calculated to obtain the probability vector of the scholar belonging to each cluster $\{W_i\}_{i=1}^{N}$. The FCM method is used to perform iterative calculations until the objective function converges. The process of calculating the cluster center is shown in Eq. (15).

$$c_k = \frac{\sum_{i=1}^{N} w_{i,k}^m F(A_i)}{\sum_{i=1}^{N} w_{i,k}^m} \quad (15)$$

where $m \in (1, \infty)$ is the hyperparameter, $F(A_i)$ is the scholars' vector. The probability vector w_i is calculated as shown in Eq. (16).

$$w_{i,k} = \frac{1}{\sum_{j=1}^{C} \left(\frac{\|x_i - c_k\|}{\|x_i - c_j\|} \right)^{\frac{2}{m-1}}} \quad (16)$$

where $w_{i,k}$ satisfies $\sum_{k=1}^{C} w_{i,k} = 1$. The objective function of the FCM clustering process is shown in Eq. (17).

$$J(W, C) = \sum_{i=1}^{N} \sum_{k=1}^{C} w_{i,k}^m \|x_i - c_k\|^2 \quad (17)$$

The probability matrix W of each scholar under the 10 clusters is obtained after clustering. By calculating the sum of the probability differences between the target scholar and other scholars in each cluster, the research interest diversity scores of other scholars relative to the target scholar are obtained. The diversity score can be defined as shown in Eq. (18).

$$DS(A_i, A_j) = \sum_{k=1}^{C} W_{F(A_i),k} - W_{F(A_j),k} \quad (18)$$

where, C is the number of clusters, $W_{F(A_i),k}$ represents the probability that scholar A_i is in the k-th class cluster.

3) Influence score: In our previous research [20], an algorithm for evaluating the academic influence of papers based on heterogeneous academic networks, AIRank, was proposed. By distinguishing the differences in the propagation strength of influence among node pairs and comprehensively examining the enhancement effect brought by the influence of heterogeneous neighbors, an effective evaluation of the academic influence of papers is achieved based on heterogeneous academic networks. Inspired by AIRank, we design a scholar's influence evaluation process based on heterogeneous academic networks. The step of this process can be describe as follows:

Step 1: Based on the heterogeneous academic network, a multilayer heterogeneous network consisting of three layers of homogeneous subnetworks is constructed: the collaboration subnetwork between scholars, the citation subnetwork between papers, and the citation subnetwork between journals. The connections between homogeneous subnetworks are maintained through the associative relationships between heterogeneous academic entities.

Step 2: In each homogeneous subnetwork, the AIRank algorithm is utilized to compute the academic impact of the nodes within the subnetwork. The calculation of the scholarly node influence of the collaboration subnetwork between scholars is formulated as shown in Eq. (19) and (20).

$$AIS(A_i) = \sum_{A_j \in \tau(A_i)} \frac{W(A_i, A_j)}{\sum_{A_k \in \tau(A_i)} W(A_i, A_k)} AIS(A_j) \quad (19)$$

$$W(A_i, A_j) = Sigmod\left(DH_{A_i} - DH_{A_j}\right) \cdot e^{cos\left(F_{A_i}, F_{A_j}\right)} \quad (20)$$

where $\tau(A_i)$ represents the set of neighboring nodes of scholar node A_i, $W(A_i, A_j)$ represents the strength of influence transfer from node A_j to node A_i, DH_{A_i} and DH_{A_j} represent the academic quality values of node A_i and A_j, respectively. $cos\left(F_{A_i}, F_{A_j}\right)$ is the cosine similarity between scholar A_i and scholar A_j.

Step 3: Based on the influence of heterogeneous neighbors, the fine-tune of the scholar's influence is calculated using formula (19). Specifically, the influence of the paper nodes and journal nodes obtained in step 2 is used to adjust the transition matrix between the scholar nodes in the collaboration subnetwork. This ensures that the scholar nodes corresponding to high-impact paper nodes and journal nodes have a higher transfer probability, resulting in a positive adjustment of the influence of the scholar nodes. The revised iterative process of the scholars' academic influence is deduced as shown in Eq. (21).

$$AIS(A_i) = \sum_{A_j \in \tau(A_i)} \frac{W(A_i, A_j)}{\sum_{A_k \in \tau(A_i)} W(A_i, A_k)} \cdot \sum_{A_t \in \tau P(A_j)} \frac{PIS(A_t)}{\left|\tau P(A_j)\right|} \cdot$$

$$\sum_{V_t \in \tau V(A_j)} \frac{VIS(A_t)}{|\tau V(A_j)|} \cdot AIS(A_j) \quad (21)$$

where $\tau P(A_j)$ is the set of papers published by the scholar A_j, and $\tau V(A_j)$ is the set of journals published by the scholar A_j, $PIS(A_t)$ and $VIS(A_t)$ represent the influence values of papers and journals, respectively. The difference in academic influence of other scholars relative to the target scholar can be calculated by the tanh function, which is defined as shown in Eq. (22).

$$IS(A_i, A_j) = tanh\left(AIS(A_i) - AIS(A_j)\right) + 1 \quad (22)$$

Cooperation strength (NCS) index: We weighted and summed the three indicators of similarity, diversity, and influence to obtain the NCS, in which the weight coefficient was calculated by the entropy weight method. Assume that the authors number is n, the original data matrix is set as $X = (x_{ij})_{n\times 3}$, where x_{ij} represents the value of the i-th author on the j-th indicator. The steps for calculating the NCS using the entropy weight method are as follows:

1) Data standardization. Standardize the data for the three indicator values of similarity, diversity, and influence to avoid bias caused by different value ranges, i.e., the normalized value is calculated as shown in Eq. (13).

$$y_{ij} = \frac{x_{ij} - min_j(x_{ij})}{max_j(x_{ij}) - min_j(x_{ij})}(max_{new} - min_{new}) + min_{new} \quad (23)$$

where y_{ij} represents the normalized value, $i = 1,2,\cdots n$, $j = 1,2,3$, the mapping interval $[max_{new},\ min_{new}]$ is set to $[0,1]$.

2) The information entropy of the indicator. The information entropy of the j-th indicator is calculated as shown in Eq. (24) and (25).

$$E_j = -\ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \quad (24)$$

$$p_{ij} = y_{ij} / \sum_{i=1}^{n} y_{ij} \quad (25)$$

3) The weights of the indicators. The weight coefficient of each indicator is calculated as shown in Eq. (26).

$$W_j = \frac{1-E_j}{\sum_{j=1}^{3}(1-E_j)} \quad (26)$$

where $0 \le W_j \le 1$, $\sum_{j=1}^{3} W_j = 1$.

4) NCS between scholars can be defined as shown in Eq. (27).

$$NCS = W_1 \times RS + W_2 \times DS + W_3 \times IS \quad (27)$$

where W_i is the weight of the corresponding indicator.

C. Novelty-oriented Encoding Module

1) Constructing the novelty-oriented weighted scholar collaborative network: The traditional scholar collaboration network is undirected and unweighted, which can only show whether the collaborative relationships exist between scholars. From the analysis in the Novelty Indicator Calculation Module, the collaborative relationships between scholars will have different collaboration strength due to the differences in similarity, diversity of research interests between scholars and academic influence of scholars. Therefore, the NCS between scholars is introduced into the scholar collaboration network as the weight of the collaboration edges between scholars to distinguish the differences in the novelty-oriented collaboration strength of different scholars.

Let $G' = (V, E, W)$ be the weighted collaboration network, where V is the set of scholar nodes, E is the set of edges, and W is the set of edge weights. The edge weights represent the differences in novelty-oriented collaboration strength between the connected scholars. Based on the reconstructed weighted cooperation network, the network representation learning process is introduced to obtain embedding vectors that contain the novelty of the scholars.

2) Scholar node representation learning based on weighted cooperation networks: Node2vec is a classical biased random walk-based network representation learning method. It can simultaneously learn the homogeneity and structural equivalence of the graph. Node2vec contains two parameters, p and q, which are used to control the bias in random walks.When the value of p is small, Node2vec focuses on the structural nature of the graph, and when the value of q is small, Node2vec focuses on the homogeneity of the graph. However, the random walk process of the Node2vec algorithm does not take into account the weight of the edges between nodes, and thus cannot be applied in the weighted scholar collaboration networks. Inspired by the Node2vec+ algorithm proposed by Liu et al. [21], we designed a novelty-oriented network representation learning model Novel-2vec. In the model, collaboration edges in the weighted collaboration network are differentiated into strong and weak collaboration edges based on the weights of the edges, and a random walk process is performed based on the network.

Assume that v_a is one of the scholar nodes in the weighted collaboration network, the average weight of all edges connected to node v_a can be calculated as $\mu(v_a) = \frac{\sum_{v' \in N(v_a)} w(v_a, v')}{|N(v_a)|}$, where $N(v_a)$ is the set of neighboring nodes of v_a. Let (v_a, v_b) be an edge between scholar v_a and scholar v_b, then, if $w(v_a, v_b) < \mu(v_a)$, the edge (v_a, v_b) is considered a strong collaboration edge; otherwise, if $w(v_a, v_b) \ge \mu(v_a)$, the edge (v_a, v_b) is considered a weak collaboration edge. Let v_a be the previous walking node, v_b be the current node, and v_c be the next node in the walk, the rules for the random walk are as follows:

Rule 1: The next node that the current node v_b walks to is v_a at walk probability $\alpha(v_a, v_b, v_c) = \frac{1}{p}$.

Rule 2: If there is a strong collaboration edge between nodes v_b and v_c, and a weak collaboration edge or no edge between node v_a and node v_c, the walk probability is

$\alpha(v_a, v_b, v_c) = \frac{1}{p} + \left(1 - \frac{1}{q}\right)\frac{w(v_a,v_c)}{\mu(v_c)}$ or $w(v_a, v_c) = 0$, respectively.

Rule 3: If there is a cooperative edge between node v_b and node v_c, and a strong cooperative edge between node v_a and node v_c, the walk probability is $\alpha(v_a, v_b, v_c) = 1$.

Rule 4: If there is a weak cooperative edge between node v_b and node v_c, and also between node v_a and node v_c, the walk probability is $\alpha(v_a, v_b, v_c) = min\left\{1, \frac{1}{q}\right\}$.

Perform the process of random walk under the guidance of the above walk probability to obtain the node sequence, and the node sequence is input into the Skip-gram model to optimize the vector representation $f(v)$ of each scholar node. Compared to the scholar's initial vector obtained from the learning results in Initial Encoding module, the scholar's vector obtained by Novel-2vec is a vector representation obtained based on a full evaluation of the strength of novelty collaboration between scholars. Since the scholar vector already contains the novelty of the scholars' research interests and academic level, it can be used as a basis for recommending novelty collaborators.

D. Collaborator Recommendation Module

Based on the novelty representation vector $f(v)$ of the scholar node, the cosine similarity between node vectors can represent the novelty-oriented similarity of the scholar node, and a Top-N recommendation list is generated based on the similarity. The similarity is calculated as shown in Eq. (28).

$$sim(v_i, v_j) = \frac{f(v_i), f(v_j)}{\sqrt{|f(v_i)| \cdot |f(v_j)|}} \quad (28)$$

For a target scholar, the cosine similarities with other scholars are sorted in descending order. The top-N scholars are extracted to generate the Top-N recommendation list as the recommended collaborators for the target scholar.

IV. Experimental Results and Discussion

A. Data Preprocessing

This article focuses on Chinese research scholars in the field of "Information Science and Library Science". A search formula is constructed in the WoS Core Collection database with the criteria "WC=Information Science& Library Science AND CU=PEOPLES R CHINA", and the publication date range is set from January 1, 2008, to October 1, 2022. The search yielded 7,141 papers published by Chinese research scholars. Delete the papers missing in the title, abstract, keywords, author, or publication year, and ultimately obtain 6,952 valid papers. Extract all authors from these papers to obtain a collection of scholars for the experiment. Due to the relatively narrow and highly specialized characteristics of the "Information Science and Library Science" field, scholars of the same name from the same affiliated institution are recognized as the same person. Afterward, for scholars with the same name from different affiliated institutions, the ORCID number of the scholar was retrieved for further identity verification. A total of 16,249 scholars are collected. Extract the venue information where the papers are published to form a collection of venues for the experiment. A total of 82 venues are collected. Taking December 30, 2018, as the dividing point, the collected academic entities and their relationships are divided into training and testing sets, respectively. That is to say, the data from January 1, 2008, to December 31, 2018, are collected as the training set, and the data obtained from January 1, 2019, to October 1, 2022, are taken as the testing set. Table I shows the basic information of the data set collected.

TABLE I. Basic Information of the Dataset

Training data (2008~2018)				Testing data (2019~2022)			
Node type	Num	Edge type	Num	Node type	Num	Edge type	Num
Author	7505	A-P	11251	Author	9885	A-P	13783
Paper	3321	A-V	9183	Paper	3631	A-V	11915
Venue	76	A-A	15692	Venue	72	A-A	23643
-	-	P-V	3321	-	-	P-V	3631
-	-	P-P	4543	-	-	P-P	10310

B. Evaluation Indicators and Baseline Models

We use Precision and F1 score to evaluate the performance of the scientific research collaboration recommendation model NovSRC. *Precision@k* denotes the accuracy of the recommendation when the length of the recommendation list is *k*. The calculation formula is shown in (29), where *R* is the set of scholars in the recommendation list, and T is the set of scholars who have collaborative relationships with the target scholar in real world.

$$Precision@k = \frac{1}{N}\sum_{I=1}^{N}\frac{|R \cap T|}{|R|} \quad (29)$$

F1@k denotes the F1 score of the recommendation result when the length of the recommendation list is *k*. t can be calculated as shown in (30), where $Recall@k = \frac{1}{N}\sum_{I=1}^{N}\frac{|R \cap T|}{|T|}$.

$$F1@k = \frac{2 \times Precision@k \times Recall@k}{Precision@k + Recall@k} \quad (30)$$

Meanwhile, we also calculate the NCS of the collaborators in the recommendation list to evaluate the novelty of the collaborators in the recommendation list generated using different recommendation algorithms. *NCS@k* denotes the novelty value of the recommendation result when the length of the recommendation list is k, which is calculated as shown in Eq. (31).

$$NCS@k = \frac{\sum_{i=1}^{N} NCS(R)}{N} \quad (31)$$

To validate the performance of the NovSRC, two network representation learning models commonly used in research collaboration recommendation tasks, Deepwalk and Node2vec, are selected as baseline comparison models. Two baseline models are used to learn the representation vectors of scholars based on the initial scholar collaboration network, and generating a recommendation list of collaborators that is only guided by similarity indicators. By comparing the novelty-oriented and similarity-oriented list of recommendation, we verify the significance of introducing the novelty into the collaborator recommendation system.

1) Deepwalk [22]: Deepwalk is used to perform a random walk on the initial academic cooperation network to generate a node sequence. And the sequence is input into the Skip Gram model to learn the vector representation of scholar nodes. Finally, the similarity between scholar node vectors is calculated to obtain Top-N recommendations.

2) Node2vec [23]: Node2vec is an improved version of the Deepwalk model, where the random walk strategy is changed by hyperparameters p and q to consider both graph homogeneity and structural equivalence. Node2vec performs a random walk process on the initial academic cooperation network to generate a node sequence. Then the sequence is processed in the same way as Deepwalk to obtain the Top-N recommendations.

C. Results and Discussion

The collaborator recommendation results generated by each model are shown in Tables II and III. ΔMax represents the maximum improvement of the NovSRC model relative to the baseline models. It can be seen that the NovSRC model has achieved the best recommendation performance in both Precision and F1 metrics, and the optimal performance of NovSRC when the length of the recommendation list is k = 5. Compared with the baseline models, the Precision@5 of NovSRC has been improved by 6.9%, and the F1@5 of NovSRC has been improved by 17.3%. The experimental results show that by integrating the novelty indicators into the collaborator recommendation system, a higher precision can be achieved than the indicators that only consider similarity.

TABLE II. PRECISION@K THE RESULTS OF THE EXPERIMENT

Model	Precision@5	Precision@10	Precision@15	Precision@20	Precision@25	Precision@30
Deepwalk	0.193	0.171	0.124	0.113	0.096	0.087
Node2vec	0.259	0.217	0.175	0.131	0.103	0.093
NovSRC	**0.262**	0.243	0.179	0.145	0.117	0.098
ΔMax	0.069 ↑	0.072 ↑	0.055 ↑	0.032 ↑	0.021 ↑	0.011 ↑

TABLE III. F1@K THE RESULTS OF THE EXPERIMENT

Model	F1@5	F1@10	F1@15	F1@20	F1@25	F1@30
Deepwalk	0.246	0.192	0.163	0.151	0.136	0.129
Node2vec	0.402	0.296	0.230	0.189	0.167	0.153
NovSRC	**0.419**	0.316	0.252	0.209	0.178	0.156
ΔMax	0.173 ↑	0.124 ↑	0.089 ↑	0.058 ↑	0.042 ↑	0.027 ↑

To validate the necessity of scholars for novelty when seek collaborators, we compare the novelty indicators of collaborators recommended by the NovSRC model and the baseline models that only contains similarity. The experimental results are shown in Table IV.

TABLE IV. NCS@K THE RESULTS OF THE EXPERIMENT

Model	NCS@5	NCS@10	NCS@15	NCS@20	NCS@25	NCS@30
Deep walk	0.387	0.383	0.383	0.381	0.379	0.379
Node-2vec	0.388	0.387	0.386	0.385	0.384	0.384
NovSRC	**0.420**	0.418	0.417	0.414	0.413	0.413
ΔMax	0.033 ↑	0.035 ↑	0.034 ↑	0.033 ↑	0.034 ↑	0.034 ↑

The results demonstrate that the collaborators recommended by the NovSRC model have higher novelty metric values than other two baseline models. When the length of recommendation list is 5, the recommended collaborators have the highest NCS. The results suggest that scholars are increasingly inclined to collaborate with scholars who have more diverse research interests and can provide more new research perspectives.

D. Case Analysis

Taking two scholars (ID 1024 and ID 7169) as examples, generate the recommendation lists of length 5 for these two scholars under the NovSRC model and the Node2vec model which obtains the best performance in baseline models. Based on the probability distribution results of scholars in different research fields obtained from the calculation of the diversity indicators, the topic distribution of each scholar is sorted in descending order of probability, and the probability distribution is accumulated. The topics with cumulative probability value reaches 0.8 is selected as the main research topic of interest for each scholar. By comparing the distribution of research interests between target scholars and recommended scholars, we aim to compare the differences of different models in the attention to the novelty of scholars' research interests.

Following the above calculation process, we found that the target scholar of ID 1024 is mainly interested in “Topic 5”, “Topic 1”, and “Topic 4”. Fig. 3 shows the research interest distribution of the collaborators recommended by the NovSRC and Node2vec models for the target scholar. Among them, Fig. 3(a) shows the interest distribution of collaborators using the NovSRC model. Fig. 3(b) shows the interest distribution of collaborators recommended by the Node2vec model. It can be seen that, compared to the target scholar, the collaborators recommended by the NovSRC model have a wider and more diverse distribution of research interests, with research interests

different from the target scholar accounting for 42% of the total interest distribution. Relatively, the Node2vec model focuses more on scholars with similar research interests as the target scholars. Among the 5 recommended collaborators, the only difference with the target scholar was in "Topic 3", which accounted for only 15%.

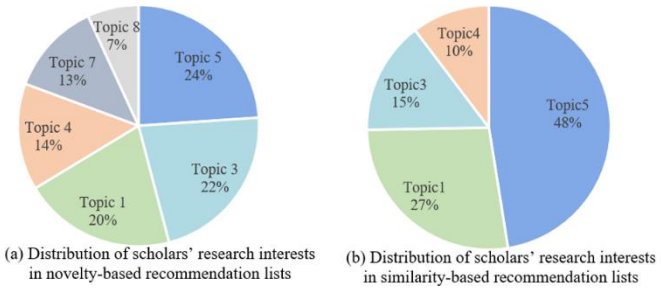

Fig. 3. Distribution of research interests of recommended collaborators (taking scholar No. 1024 as an example).

The target scholar of ID 7169 is mainly interested in "Topic1", "Topic 7", and "Topic 4". Fig. 4 shows the research interest distribution of collaborators recommended by the NovSRC and Node2vec models. Fig. 4(a) shows the interest distribution of collaborators using the NovSRC model. Fig. 4(b) shows the interest distribution of collaborators recommended by the Node2vec model. Compared with the Node2vec model, the NovSRC model recommended scholars with a wider research interest and a higher proportion of research interests that differed from those of the target scholars.

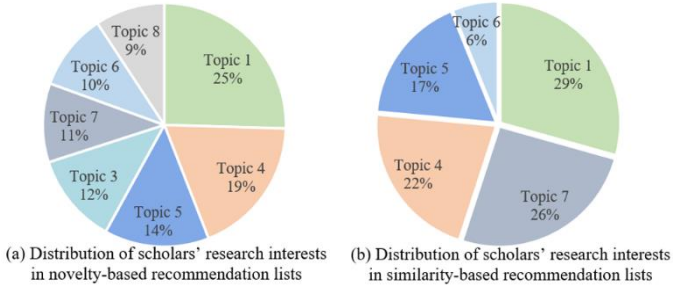

Fig. 4. Distribution of research interests of recommended collaborators (taking scholar No. 7169 as an example).

Therefore, the collaborators recommendation of oriented novelty shows a more diverse distribution of interests compared with the target scholars, which can provide more opportunities for collaboration between scholars, and may help to provide more pioneering research ideas for both sides, thus promote the joint progress of their research.

V. Discussion

In order to meet the needs of researchers for novel collaborators, this paper proposes a novel oriented scientific collaborator recommendation model NovSRC. Unlike traditional similarity-based recommendation systems, the NovSRC model fully considers the impact of novelty elements on the recommendation process, recommending collaborators with diverse research interests to target scholars, thereby improving their satisfaction and interest in the recommendation system. The experimental results indicate that compared with the baseline models that only examines the similarity of research interests among scholars, the NovSRC model recommends a wider and more diverse range of research interests among collaborators, which will inject more innovative elements into the cooperation between scholars and promote common scientific progress between both parties.

VI. Conclusion

This article fully integrates novelty elements into the recommendation process of scientific research collaborators and proposes a novel oriented collaborator recommendation model, NovSRC. This model can recommend collaborators to target scholars, and help them to effectively expand their research perspectives and promote their scientific research process. Based on the similarity and diversity of research interests among scholars, as well as the differences in academic influence among scholars, NovSRC quantifies the strength of innovation collaboration among scholars. By using the strength indicator as the edge weight of the collaborative network between scholars, the encoding process of scholar vectors is fully established under the guidance of novelty elements, which makes the collaborators recommended by the NovSRC model can bring more innovative academic ideas for the target scholars. Although the research in this paper has achieved certain results, the initial modeling process of scholars only extracted the characteristics of the scholar's research content and network structure, and lacked the impact of factors such as region and institution on the collaborator recommendation task. Therefore, future research will try to introduce other entities such as regions and institutions into heterogeneous academic networks to achieve more comprehensive scholar feature extraction, thereby further exploring the effectiveness of novelty collaborator recommendations.

Acknowledgment

This work was supported the National Natural Science Foundation of China (Grant No. 71473034), and the Heilongjiang Provincial Natural Science Foundation of China (Grant No. LH2019G001).

References

[1] X. Kong, M. Mao, J. Liu et al., "TNERec: Topic-aware network embedding for scientific collaborator recommendation," 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 1007-1014, 2018.

[2] J. Chen, X. Li, H. Ji et al., "Content Recommendation Algorithm Based on Double Lists in Heterogeneous Network," Communications and Networking: 14th EAI International Conference, ChinaCom 2019, Shanghai, China, November 29–December 1, 2019, Proceedings, Part II 14, pp. 140-153, 2020.

[3] X. Zhang, Y. Wen, and H. Xu, "A Prediction Model with Network Representation Learning and Topic Model for Author Collaboration," Data Analysis and Knowledge Discovery, vol. 5, no. 3, pp. 88-100, 2020.

[4] T. Pradhan, and S. Pal, "A multi-level fusion based decision support system for academic collaborator recommendation," Knowledge-Based Systems, vol. 197, pp. 105784, 2020.

[5] D. Hu, and H. Ma, "Collaborator recommendation integrating author's cooperation strength and research interests on attributed graph," Advances in Computational Intelligence, vol. 1, no. 4, pp. 2, 2021.

[6] B. Kumara, K. Banujan, S. Prasanth et al., "Constructing global researchers network using google scholar profiles for collaborator recommendation systems," 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), pp. 274-279, 2021.

[7] J. Du, H. Xiong, and N. Wang, "Research Collaborator Recommendation Research on fusion of Multivariate Networks and Network Representation Learning," Information and Documentation Services, vol. 43, no. 4, pp. 27-35, 2022.

[8] O. Du, and Y. Li, "Academic Collaborator Recommendation Based on Attributed Network Embedding," Journal of Data and Information Science, vol. 7, no. 1, pp. 37-56, 2022.

[9] V. Jagadishwari, R. James, and R. Abraham, "Research Collaborator Recommendation System based on citations and Influential citations," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1095-1099, 2023.

[10] X. Liu, K. Wu, B. Liu et al., "HNERec: Scientific collaborator recommendation model based on heterogeneous network embedding," Information Processing & Management, vol. 60, no. 2, pp. 103253, 2023.

[11] M. De Gemmis, P. Lops, G. Semeraro et al., "An investigation on the serendipity problem in recommender systems," Information Processing & Management, vol. 51, no. 5, pp. 695-717, 2015.

[12] R. J. Ziarani, and R. Ravanmehr, "Serendipity in recommender systems: a systematic literature review," Journal of Computer Science and Technology, vol. 36, pp. 375-396, 2021.

[13] M. Zhang, Y. Yang, R. Abbas et al., "SNPR: A serendipity-oriented next POI recommendation model," Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2568-2577, 2021.

[14] R. J. Ziarani, and R. Ravanmehr, "Deep neural network approach for a serendipity-oriented recommendation system," Expert Systems with Applications, vol. 185, pp. 115660, 2021.

[15] J. Gao, F. Liang, W. Fan et al., "On community outliers and their efficient detection in information networks," Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 813-822, 2010.

[16] Z. Xu, Y. Yuan, H. Wei et al., "A serendipity-biased Deepwalk for collaborators recommendation," PeerJ Computer Science, vol. 5, 2019.

[17] F. Ding, "Research on paper recommendation algorithm basedon novelty and influence," South China University of Technology, 2020.

[18] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," Conference on Empirical Methods in Natural Language Processing, 2021.

[19] H. Zhong, M. Wang, and X. Zhang, "Unsupervised Embedding Learning for Large-Scale Heterogeneous Networks Based on Metapath Graph Sampling," Entropy, vol. 25, no. 2, pp. 297, 2023.

[20] M. Wang, X. Zhang, H. Zhong et al., "AIRank: An algorithm on evaluating the academic influence of papers based on heterogeneous academic network," Journal of Information Science, 2023.

[21] R. Liu, M. Hirn, and A. Krishnan, "Accurately modeling biased random walks on weighted networks using node2vec+," Bioinformatics, vol. 39, no. 1, pp. btad047, 2023.

[22] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701-710, 2014.

[23] A. Grover, and J. Leskovec, "node2vec: Scalable feature learning for networks," Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 855-864, 2016.