Hindawi
Complexity
Volume 2021, Article ID 7035467, 12 pages
https://doi.org/10.1155/2021/7035467

Research Article

Content-Enhanced Network Embedding for Academic Collaborator Recommendation

Jie Chen,[1,2] **Xin Wang**,[1,2] **Shu Zhao**,[1,2] **and Yanping Zhang**[1,2]

[1]*Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230601, Anhui, China*
[2]*School of Computer Science and Technology, Anhui University, Hefei 230601, China*

Correspondence should be addressed to Shu Zhao; zhaoshuzs@ahu.edu.cn

Received 4 June 2020; Revised 18 January 2021; Accepted 3 February 2021; Published 25 February 2021

Academic Editor: Ning Cai

It is meaningful for a researcher to find some proper collaborators in complex academic tasks. Academic collaborator recommendation models are always based on the network embedding of academic collaborator networks. Most of them focus on the network structure, text information, and the combination of them. The latent semantic relationships exist according to the text information of nodes in the academic collaborator network. However, these relationships are often ignored, which implies the similarity of the researchers. How to capture the latent semantic relationships among researchers in the academic collaborator network is a challenge. In this paper, we propose a content-enhanced network embedding model for academic collaborator recommendation, namely, CNEacR. We build a content-enhanced academic collaborator network based on the weighted text representation of each researcher. The content-enhanced academic collaborator network contains intrinsic collaboration relationships and latent semantic relationships. Firstly, the weighted text representation of each researcher is obtained according to its text information. Secondly, a content-enhanced academic collaborator network is built via the similarity of the weighted text representation of researchers and intrinsic collaboration relationships. Thirdly, each researcher is represented as a latent vector using network representation learning. Finally, top-k similar researchers are recommended for each target researcher. Experiment results on the real-world datasets show that CNEacR achieves better performance than academic collaborator recommendation baselines.

1. Introduction

During the era of big scholarly data, information overload has become a serious problem. It is challenging how to dig useful information from overloaded information [1, 2]. Prior studies show that collaboration among researchers can increase the productivity of the researcher and come up with unprecedented inspirations [3, 4]. So, academic collaborator recommendation that aims to find the proper collaborators for a target researcher has played an important role in complex academic tasks.

Academic information can be described as an academic collaborator network with attributes (as shown in Figure 1). The methods of academic collaborator recommendation are divided into three categories, including network-based recommendation, content-based recommendation, and hybrid recommendation. For network-based recommendation, the structure of the network was utilized to improve the performance of recommending the researchers [5]. The probability theory and graph theory were used to model and analyze coauthor networks [6]. Another network-based recommendation involves the classic random walk model, which can dig for useful information from the academic collaborator network. For content-based recommendation, the interest of the researcher is an important attribute that characterizes the research topics, fields, and other personalized features [7–9]. It can be analyzed and mined through papers that the researcher publishes every year, and the relationships among researchers can also be established through interest detection. Compared with the above methods which consider the academic collaborator network structure and the interests of the researcher, respectively, the

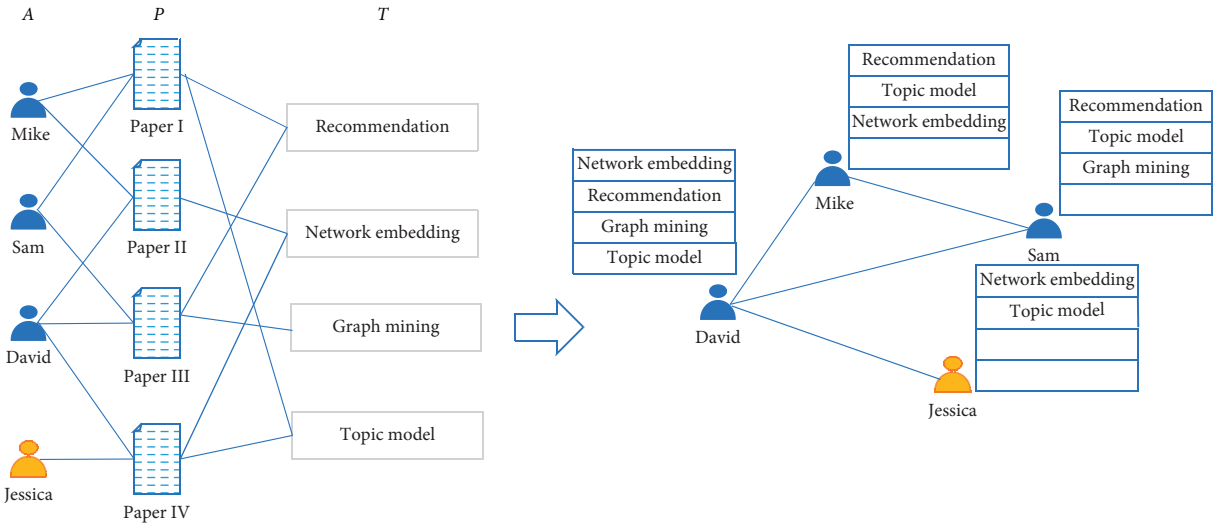

FIGURE 1: The extraction of a collaboration network with attributes from academic data. A is a list of researchers, P is a list of papers, and T is a list of topics. Each researcher has a distribution of topics. In this paper, we regard each word in the title of the paper as a topic. The figure shows that if two researchers coauthor a paper, there is a link between them, such as (David, Jessica). If a researcher has coauthored more papers, he will have more links with others, such as (David, Mike), (David, Sam), and (David, Jessica).

combination of network topology and text information is more effective. For hybrid recommendation, utilizing both text information and network structure can improve learning of the latent representation for each researcher [10–12]. Existing hybrid models always learn the feature representation of each researcher using text information and network structure independently and then combining the two feature representations into a unified latent representation. They do not utilize the complex relationships between text information and network structure [13]. The aforementioned hybrid method could improve the recommendation of academic collaboration, but the latent semantic relationships formed by the text information in an academic collaborator network were ignored.

To capture the latent semantic relationships to improve academic collaborator recommendation, we utilize text information of each researcher to build a content-enhanced network and propose the CNEacR model for academic collaborator recommendation. CNEacR builds a content-enhanced academic collaborator network that contains the intrinsic collaboration relationships and the latent semantic relationships formed by the text information. Firstly, CNEacR obtains the weighted text representation for each researcher and then builds a content-enhanced academic collaborator network based on the similarity of the feature representation of each researcher. Secondly, high-quality latent representation is obtained by network embedding. Finally, the similarity between researchers can be calculated by the cosine similarity based on high-quality latent representation. Experiment results on the real-world datasets demonstrate that CNEacR produces a better improvement on precision, recall, F1, and normalized discounted cumulative gain (NDCG) over all baseline methods.

The main contributions of this paper are summarized as follows:

(1) A context-enhanced academic collaborator network that contains not only intrinsic collaboration relationships but also the latent semantic relationships formed by the text information is built using the similarity among the weighted text representation of researchers.

(2) To get a context-enhanced academic collaborator network, the weighted text representation of each researcher is obtained from the text information. The edges are added which are between a node and its semantic similar nodes, and then a context-enhanced academic collaborator network is built.

(3) Experiment results on the datasets demonstrate the performance of CNEacR is better than other methods of academic collaborator recommendation.

2. Problem Definitions

2.1. Academic Collaborator Recommendation. Given a researcher set $A = \{a_1, a_2, \ldots, a_{|N|}\}$, where $|N|$ is the number of researchers in the set. Each researcher $a_i = (d_i, n_i)$ has text information $d_i = (d_i^1, d_i^2, \ldots, d_i^{|M|})$, where d_i^j represents the j-th term for a_i and $|M|$ is the number of terms for a_i. The structure of academic collaborator network $n_i = (n_i^1, n_i^2, \ldots, n_i^{|N|})$, where n_i^j represents the relationship between researcher a_i and researcher a_j, $n_i^j \in \{0, 1\}$. Academic collaborator recommendation aims to get a ranked researchers' list $K = \{a_1, a_2, \ldots, a_k\}$ for a given target researcher a_i, which are the most relevant k researchers from researcher collection A.

2.2. Content-Enhanced Academic Collaborator Network. The academic collaborator network can be denoted as $G = (A, E)$, where $A = \{a_1, a_2, \ldots, a_{|N|}\}$, a_i is the i-th researcher, and $E = \{e_{i,j} | a_i \in A, a_j \in A\}$. Each $e_{i,j} \in \{0, 1\}$ represents whether there exists the collaborative relationship if $e_{i,j} = 1$ denotes that there exists a collaborative relationship between researcher a_i and researcher a_j; otherwise, it does not exist. In the academic collaborator network, we firstly use the TF – IDF model to evaluate the importance of each term and then embed each term into a vector by Word2vec. The vector is the weighted text representation of each researcher. Secondly, $\mathrm{TK} = \{a_1, a_2, \ldots, a_k\}$ relevant researchers are listed based on cosine similarity of each a_i. Finally, we can get a relationship set $E' = \{E_1, E_2, \ldots, E_{|N|}\}$, $E_i = (e'_{i1}, e'_{i2}, \ldots, e'_{ik})$. If the relationship between $a_i \in \mathrm{Tk}$ and $a_j \in \mathrm{Tk}$ exists, $e'_{ij} = 1$. The content-enhanced academic collaborator network $G' = (A, E^c)$ is built, $E^c = E' \cup E$.

3. Methodology

In this section, we explain the CNEacR in detail. CNEacR builds the weighted text representation of each academic researcher and constructs a context-enhanced academic collaborator network. Then, we maximize the co-occurrence probability to obtain the high-quality latent representation of each researcher. Finally, top-k researchers are recommended for a target researcher via the similarity of high-quality latent representation. Some important notations are shown in Table 1. We summarize the framework of our proposed CNEacR in Figure 2 and show the whole algorithm framework in Algorithm 1.

As shown in Figure 2, all nodes in the dataset belong to the test set. To validate the effectiveness of our algorithm, the real collaborative relationships among nodes are divided into two classes: collaborative relationships and unknown relationships according to [13]. The collaborative relationships are edges in the academic collaborator network which means the structure of the network. The unknown relationships do not participate in the algorithm process. They are used to compare with the recommended top-k collaborative relationships. The ratio R of collaborative relationships to unknown relationships is discussed in Section 4.4.3.

3.1. Building Weighted Text Representation of the Academic Researcher. It is fundamental to represent the text in many natural language processing (NLP) tasks. There are many methods to extract the feature representation of the researcher from text information, including probabilistic latent semantic analysis (pLSA) [14], latent Dirichlet allocation (LDA) [15], Word2vec [16], and BERT [17]. Word2vec is widely used to generate more accurate feature representations based on text information in a specific scenario. So, we choose Word2vec to get weighted text representation.

Given an academic researcher set $D = \{d_1, d_2, \ldots, d_{|N|}\}$, where d_i represents the text information of the i-th researcher composed by his published paper's titles. $d_i = \{d_i^1, d_i^2, \ldots\}$, where d_i^t is the t-th term in d_i. Some similar operations apply to the set D, such as segmenting, filtering, and extracting. d_i^t is represented into vector $\mathrm{vec}_{d_i^t}$ using Word2vec. $w_{d_i^t, d_i}$ is used to describe the importance of each term to the text information of different researchers. $w_{d_i^t, d_i}$ could be defined as follows:

$$w_{d_i^t, d_i} = \mathrm{tf}_{d_i^t, d_i} \times \log \frac{|D|}{\left|\left\{d_i \in D, d_i^t \in d_i\right\}\right|}, \tag{1}$$

where D is the text information set of all researchers, $|D|$ is the total number of researchers in the dataset, d_i represents the text information of each researcher, and d_i^t represents the t-th term in d_i. $|\{d_i \in D, d_i^t \in d_i\}|$ is the total number of the text information of the i-th researcher which contains term d_i^t. $\mathrm{tf}_{d_i^t, d_i}$ stands for the term frequency of the term d_i^t in the text information of the i-th researcher, and $\log(|D|/|\{d_i \in D, d_i^t \in d_i\}|)$ is the inverse document frequency. As far as we know, the frequent occurrence of a term in the researcher's text information means that this term is important to the researcher. However, if a term appears in many researchers' text information at the same time, it indicates that this term is common to each text and is less important to each researcher. $w_{d_i^t, d_i}$ is used to weight the importance of a term in the text information of each researcher. As described above, the weights of the terms in the text information of researcher a_i can be defined as follows:

$$W_i = \left(w_{1i}, w_{2i}, \ldots, w_{|M|i}\right). \tag{2}$$

The weighted text representation of researcher a_i can be defined as follows:

$$RW_i = \frac{1}{|M|} \sum_{t=1}^{|M|} \mathrm{vec}^*_{d_i^t} w_{d_i^t, d_i}. \tag{3}$$

Since each researcher has a different amount of text information, we normalized the weighted text representation of each researcher. $|M|$ is the number of terms in the text information of each researcher, $\mathrm{vec}_{d_i^t}$ is the vector of the t-th term of the i-th researcher learned by Word2vec, and $w_{d_i^t, d_i}$ is the weight of the t-th term of the i-th researcher.

3.2. Constructing the Context-Enhanced Academic Collaborator Network. Given an academic collaborator network $G = (A, E)$, where $A = \{a_1, a_2, \ldots, a_{|N|}\}$ is the researcher set and $E = \{e_{i,j} | a_i \in A, a_j \in A\}$ represents collaborative relationships among researchers. We calculate any two nodes' similarity using their weighted text representation by widely used cosine similarity:

$$CosSim\left(RW_i, RW_j\right) = \frac{RW_i^T * RW_j}{\left\|RW_i\right\|^* \left\|RW_j\right\|}. \tag{4}$$

So, generate relationships $E' = (E_1, E_2, \ldots, E_n)$, $E_i = (e'_{i1}, e'_{i2}, \ldots, e'_{i|N|})$; each e'_{ij} is defined as follows:

$$e'_{ij} = \begin{cases} 1, & \text{if } CosSim\left(RW_i, RW_j\right) \text{ in topKList}, \\ 0, & \text{else}, \end{cases} \tag{5}$$

Table 1: Notations.

Symbol	Description
A	Researcher set
a_i	The i-th researcher
R	The ratio of the training set
SK	The number of semantic relationships
G	The academic collaborator network
D	The text information set of all researchers
Top-k	The number of recommended collaborators
X	The feature representation set of all researchers
RW_i	The weighted text representation of researcher a_i
G'	The content-enhanced academic collaborator network
W_i	The weights of the terms in the text information of researcher a_i

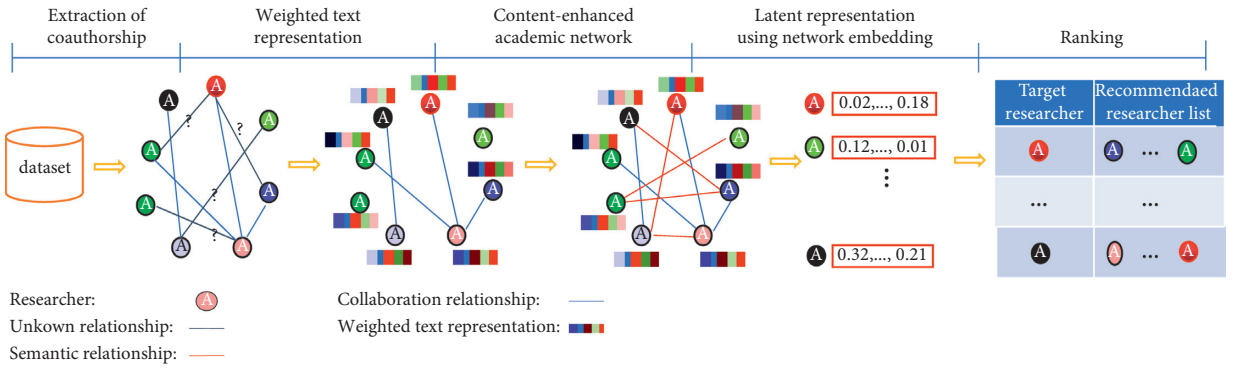

Figure 2: The framework of our proposed CNEacR.

Input: the academic collaborator network $G = (A, E)$, the text information set of all researchers $D = (d_1, d_2, \ldots, d_{|N|})$, SK, top-$k$
Output: the top-k list $K = \{a_1, a_2, \ldots, a_k\}$ for a target researcher
Traning process:
(1) **for** $a_i \in A$ **do**
(2) calculate RW_{a_i} by equation (3)
(3) **end for**
(4) **for** $a_i \in A$ **do**
(5) **for** $a_j \in A$, $a_i \neq a_j$ **do**
(6) calculate $CosSim(RW_{a_i}, RW_{a_j})$ by equation (4)
(7) **end for**
(8) choice SK similar researchers for a_i
(9) **end for**
(10) construct $G' = (A, E^c)$
(11) map G' into a low-dimensional space to get latent representation X of all researchers
(12) *Testing process*:
(13) **for** $a_i \in A$ **do**
(14) **for** $a_j \in A$, $a_i \neq a_j$ **do**
(15) calculate $CosSim(X_{a_i}, X_{a_j})$ by equation (4)
(16) **end for**
(17) $K \leftarrow$ top-k most similar collaborator for a_i
(18) **end for**

Algorithm 1: CNEacR.

where topKList is the top SK researchers in the similarity list for each researcher. SK is a hyperparameter. If $e_{ij}' = 1$, e_{ij}' is a new relationship. We add these new relationships to G, and then we will obtain a new academic collaborator network $G' = (A, E^c)$, where $E^c = E \cup E'$, which is our context-enhanced academic collaborator network.

3.3. Network Embedding. The latent representation of each researcher is the input feature of many downstream tasks, such as classification, link prediction, clustering, and visualization. To get a low-dimensional space $\mathbb{R}^d$, $d \ll |N|$, the network embedding aims to learn a function $f: N \longrightarrow \mathbb{R}^d$. Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_{|N|})$ denote the embedded vectors in the latent space. Θ maintains as much of the original network topology information as possible. There exist many network embedding methods, such as DeepWalk [18], LINE [19], Node2Vec [20], and GCN [21]. In this paper, the local information and global information are equally important for each target researcher, so DeepWalk is suitable to obtain high-quality latent representation.

Given a context-enhanced academic collaborator network, we use the DeepWalk model to represent the relationships of academic collaboration. Intuitively, for academic collaborator recommendation, it is equally important for both local information, namely neighborhood, and global information. We use latent academic collaborative relationships obtained from random walks to learn academic researcher latent representation. For each walk sequence $s = \{a_1, a_2, \ldots, a_s\}$, following skip-gram, we aim to maximize the probability of the neighbors of researcher a_i in this walk sequence as follows:

$$\max_{\phi} \Pr\left(\{a_{i-\omega}, \ldots a_{i+\omega}\} \backslash a_i | \phi(a_i)\right) = \prod_{j=i-\omega, j \neq i}^{i+\omega} \Pr\left(a_j | \phi(a_i)\right), \tag{6}$$

where ω is the window size, $\phi(a_i)$ is the current representation of researcher a_i, and $\{a_{i-\omega}, \ldots, a_{i+\omega}\} \backslash a_i$ is the local context researchers of a_i.

Finally, we use hierarchical softmax [22] to obtain the embedding vector of each researcher $X = (X_1, X_2, \ldots, X_{|N|})$, $X_i = (x_{1,i}, x_{2,i}, \ldots, x_{d,i})$. The latent representation of each researcher fuses researchers' text information and network structure. d is a hyperparameter that is the dimension of the latent representation of the researcher. For each target research, we can get the top-k similar collaborators according to equation (4).

4. Experiment

In this section, we evaluate our proposed CNEacR model on two real-world datasets. We introduce datasets, baselines, evaluation criteria, and the results of experiments in detail.

4.1. Datasets. PRB (*Physical Review B*) from the APS (American Physical Society) (https://journals.aps.org/datasets) consists of some articles about the subject of physics. At first, we do name disambiguation on authors from 1893 to 2015 based on [23]. Authors who have less than 2 collaborators from 2006 to 2010 are removed. Finally, we extract 34,905 authors and 14,055 papers to evaluate our proposed CNEacR. AMiner, a larger-scale dataset, is adopted, we randomly choose 14,000 papers, and it contains 20,057 researchers, who have more than 10 papers. Table 2 shows the details of the datasets. Some necessary cleaning is done, such as removing excess code fragments, removing the stop words, tokenization, and lemmatization.

Table 2: Statistics of datasets.

Statistics	Researcher	Paper	Relationships
PRB	34,905	14,055	165,507
AMiner	20,057	14,000	168,163

To evaluate the performance of CNEacR, we assume all researchers in the dataset as target researchers. The R ratio collaborator relationships of each researcher are used as the training samples, and the $1 - R$ ratio collaborator relationships are used as the test target according to [13]. In experiments, we choose relationships with the ratio R 10 times to ensure that the selected relationships can contain as many authors as possible. All experiments are performed on a 64-bit Linux-based operation system, Ubuntu 16.04 with a 64-duo and 2.10 GHz Intel CPU, 1-T Bytes memory. All the programs are implemented with Python.

4.2. Baselines. We compare CNEacR with the following six methods, where the first is the classic method for academic collaborator recommendation. The baselines consist of the following:

(1) MVCWalker: MVCWalker [24] is a random walk model standing on the shoulder of a random walk with restart for the collaborator recommendation which combines three academic factors including coauthor order, latest collaboration time, and times of collaboration.

(2) TNERec-G: TNERec-G is a portion of TNERec which only uses the structure of the academic collaborator network to get the feature representation of the researcher for collaborator recommendation.

(3) CTPF: CTPF [25] is a probabilistic model of articles to represent researchers with their preferences for topics. It integrates two ideas: collaborative topic regression and Poisson factorization.

(4) TNERec: TNERec [13] is an academic collaborator recommendation method that learns feature representation from the interests of the researcher based on the topic model and feature representation from the structure of the academic collaborator network using network embedding, respectively, and then fuses them using a spectral technique for better collaborator recommendation.

(5) CNEacR-G: CNEacR-G is a portion of CNEacR which only uses the structure of the academic collaborator network to get the feature representation of the researcher for collaborator recommendation (does not use any semantic relationship).

(6) CNEacR-T: CNEacR-T is a portion of CNEacR which only uses the text information of the researcher to recommend the collaborator, which is based on text recommendation.

4.3. Evaluation Criteria. We use the most common evaluation criteria in information retrieval as the academic collaborator recommendation evaluation metrics.

Precision@k means the ratio of the right recommended collaborators to top-k recommended candidates when recommending k candidate collaborations for the target researcher. Precision@k is defined as follows:

$$\text{Precision@}k = \frac{1}{m}\sum_{j=1}^{m}\frac{R_a \cap T_a}{R_a}. \tag{7}$$

Recall@k means the ratio of the recommended right collaborators who are in the test set to all recommended candidates when recommending k candidate collaborations for the target researcher. The recall value is computed as follows:

$$\text{Recall@}k = \frac{1}{m}\sum_{j=1}^{m}\frac{R_a \cap T_a}{T_a}, \tag{8}$$

where m is the number of target researchers, R_a is the top-k recommended researcher list for the target researcher, and T_a is the real collaborators of the target researcher in the test set. F1 is the harmonic average of precision and recall, and F1 is defined as follows:

$$F1@k = \frac{2 * P@k * R@k}{P@k + R@k}. \tag{9}$$

IDCG represents the list of the best recommendation results. NDCG is the normalized recommended list evaluation scores. We define r_i as the rating of the i-th researcher in the recommended researcher list. If $r_i = 1$, the recommended collaborator is relevant, and $r_i = 0$, otherwise. NDCG@k is defined as follows:

$$\text{NDCG@}k = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{l}(2^{r_i}-1)/\log_2(i+1)}{\text{IDCG@}k}\right). \tag{10}$$

4.4. Experiment Results and Parameter Analysis. Table 3 demonstrates the performance comparison of CNEacR, and the results outperform all baselines on precision, recall, F1, and NDCG. Besides, we present the result of CNEacR-G and CNEacR-T in PRB. CNEacR-G only uses text information of each researcher, and CNEacR-T only uses collaborator relationships in the network. To make the results more convincing, we give the results of the experiment in AMiner and compared it with the two kinds of methods, content-based recommendation and network-based recommendation. Table 4 demonstrates the results of the experiment in AMiner.

From Tables 3 and 4, we know that CNEacR-G does not use the text information, and the results are poor. CNEacR-T does not use the network structure, and the results are not good enough. We can see that utilizing both text information and network structure plays an important role in academic collaborator recommendation. We demonstrate the performance in different recommendation lists and analyze different results when choosing different training sets of ratios in PRB. As an auxiliary experiment, we only demonstrate the performance in $R = 0.3$.

4.4.1. Parameter SK. We analyze the parameter SK used to build the relationship among researchers in two datasets. Similar to [13], set the length of the recommendation list Top $-$ k as 5, and choose SK as 0, 1, 2, 3, 4, and 5, respectively, to build the content-enhanced academic collaborator network. Figure 3 shows the comparison results of CNEacR on different SK in two datasets. From Figure 3, we can easily find that different datasets have different SK. SK of the best performance of CNEacR in PRB is 2, and SK of the best performance of CNEacR in AMiner is 1. We can see from Figure 3 that different SK have a big influence on the performance of CNEacR. With the increase of SK, the number of uncertainty relationships is increasing, which influences the performance of our proposed CNEacR to capture real collaborative relationships.

4.4.2. Influence of the Recommendation List. We analyze the performance of CNEacR with different lengths of recommendation. We choose the ratio of the training set $R = 0.3$ to conduct our experiment and set the dimension of the researcher vector as 100. The parameter SK is set as 2. Figure 4 shows that our proposed model is compared with other methods of precision, recall, F1, and NDCG. With the increase of recommendation list Top $-$ k, we can see that the precision of CNEacR, CNEacR-T, CNEacR-G, TNERec, TNERec-G, and CTPF shows a downward trend. MVCWalker goes up at first and then goes down with the recommendation list increasing. The recall of all methods shows an upward trend. F1 of all methods takes on the tendency of increasing first but decreasing afterward. The NDCG of all methods keeps a steady trend. We can see that network-based and context-based collaborator recommendations can work well, respectively, and the results of experiments verify that our method which utilizes both weighted text representation and academic collaborator network can perform well compared with all the above methods.

4.4.3. Influence of Ratio R. To prevent the contingency of experimental results, we use different sizes of the training set to evaluate the performance of CNEacR over the training set. We set the ratio R varying from 20% to 80% and set recommendation list size k as 3. We also set the latent representation of the researcher as 100 and set the parameter SK as 2. Figure 5 shows the performance compared with other methods on different R in terms of precision, recall, F1, and NDCG. CNEacR outperforms other methods a lot on four metrics no matter how R is. From Figure 5, these methods have the same trends except the network-based methods including CNEacR-G and TNERec-G. We can see that CNEacR is always better than the network-based recommendation, content-based recommendation, and hybrid recommendation.

Table 3: Performance on different evaluation criteria in PRB.

Method	Precision				Recall				F1				NDCG			
	@5	@10	@15	@20	@5	@10	@15	@20	@5	@10	@15	@20	@5	@10	@15	@20
MVCWalker	0.269	0.226	0.182	0.154	0.224	0.329	0.379	0.409	0.261	0.285	0.262	0.236	0.441	0.493	0.5	0.504
TNERec-G	0.154	0.107	0.083	0.071	0.129	0.171	0.202	0.217	0.151	0.143	0.129	0.112	0.331	0.347	0.355	0.359
CTPF	0.253	0.131	0.095	0.075	0.208	0.22	0.234	0.237	0.243	0.182	0.144	0.124	0.528	0.532	0.536	0.536
TNERec	0.459	0.3	0.226	0.182	0.338	0.42	0.45	0.473	0.42	0.374	0.318	0.278	0.756	0.749	0.745	0.741
CNEacR-G	0.33	0.244	0.194	0.161	0.286	0.386	0.44	0.474	0.271	0.265	0.24	0.216	0.568	0.577	0.577	0.575
CNEacR-T	0.423	0.29	0.216	0.171	0.414	0.518	0.558	0.582	0.374	0.334	0.281	0.241	0.727	0.73	0.728	0.725
CNEacR	0.541	0.367	0.279	0.226	0.517	0.636	0.695	0.73	0.471	0.414	0.357	0.312	0.857	0.849	0.841	0.835

Table 4: Performance on different evaluation criteria in AMiner.

Method	Precision				Recall				F1				NDCG			
	@5	@10	@15	@20	@5	@10	@15	@20	@5	@10	@15	@20	@5	@10	@15	@20
CNEacR-G	0.211	0.172	0.152	0.139	0.185	0.238	0.271	0.296	0.146	0.142	0.139	0.136	0.386	0.399	0.402	0.403
CNEacR-T	0.29	0.199	0.162	0.142	0.372	0.432	0.467	0.494	0.269	0.215	0.186	0.169	0.626	0.632	0.634	0.634
CNEacR	0.365	0.259	0.213	0.187	0.458	0.537	0.579	0.611	0.329	0.27	0.236	0.215	0.725	0.728	0.726	0.725

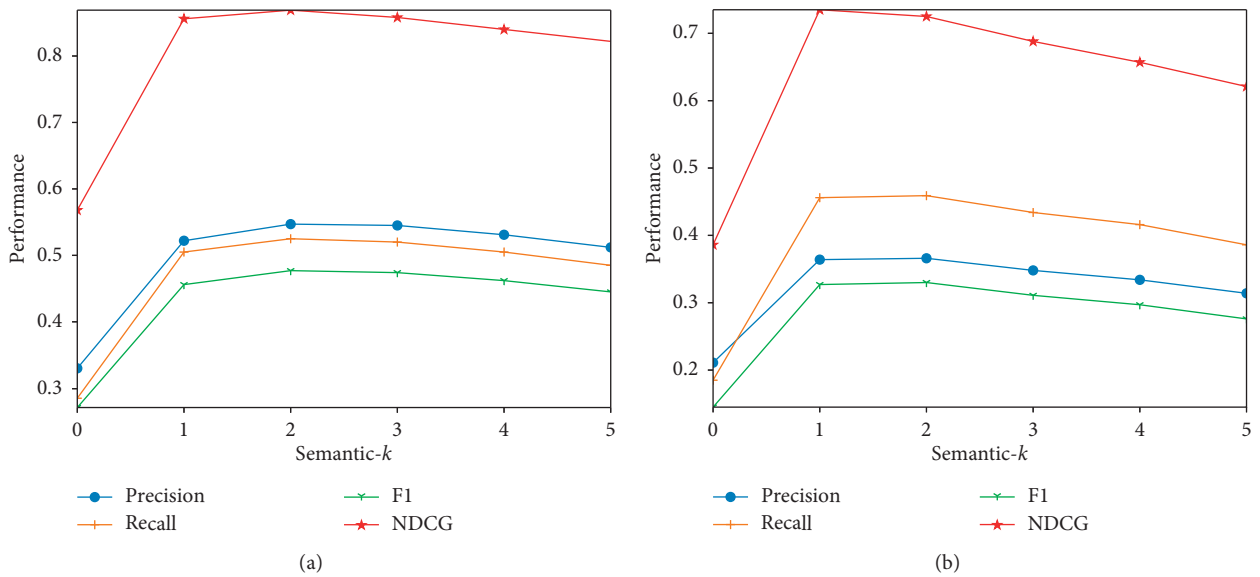

Figure 3: Performance in different SK relationships in two datasets. (a) Performance in different semantic-k relationships in PRB. (b) Performance in different semantic-k relationships in AMiner.

4.5. Case Study. Table 5 shows the case study of different methods for collaborator recommendation. We randomly select a researcher (F.Ishikawa) for a demonstration from the test set. We use three methods to recommend the top 5 collaborators for the target researcher F.Ishikawa. From the table, we can see that only CNEacR-G correctly provides one collaborator, T.Naka. It indicates that CNEacR-G captures the information of the network structure. CNEacR-T correctly recommends a new collaborator, A.Matsushita, than CNEacR-G. It indicates that CNEacR-T can capture the information of semantic relationships. Our method CNEacR correctly recommends four collaborators, and it recommends two new collaborators, Y.Takaesu and T.Nakane, than CNEacR-T. It indicates that utilizing the weighted text representation and intrinsic collaborative relationships to recommend collaborators can yield better performance than context-based and network-based recommendation. CNEacR correctly recommends four collaborators including the researchers recommended in both CNEacR-G and CNEacR-T. It indicates that CNEacR can capture both the semantic relationship and the collaborative relationship to recommend latent academic collaborators for the target researcher.

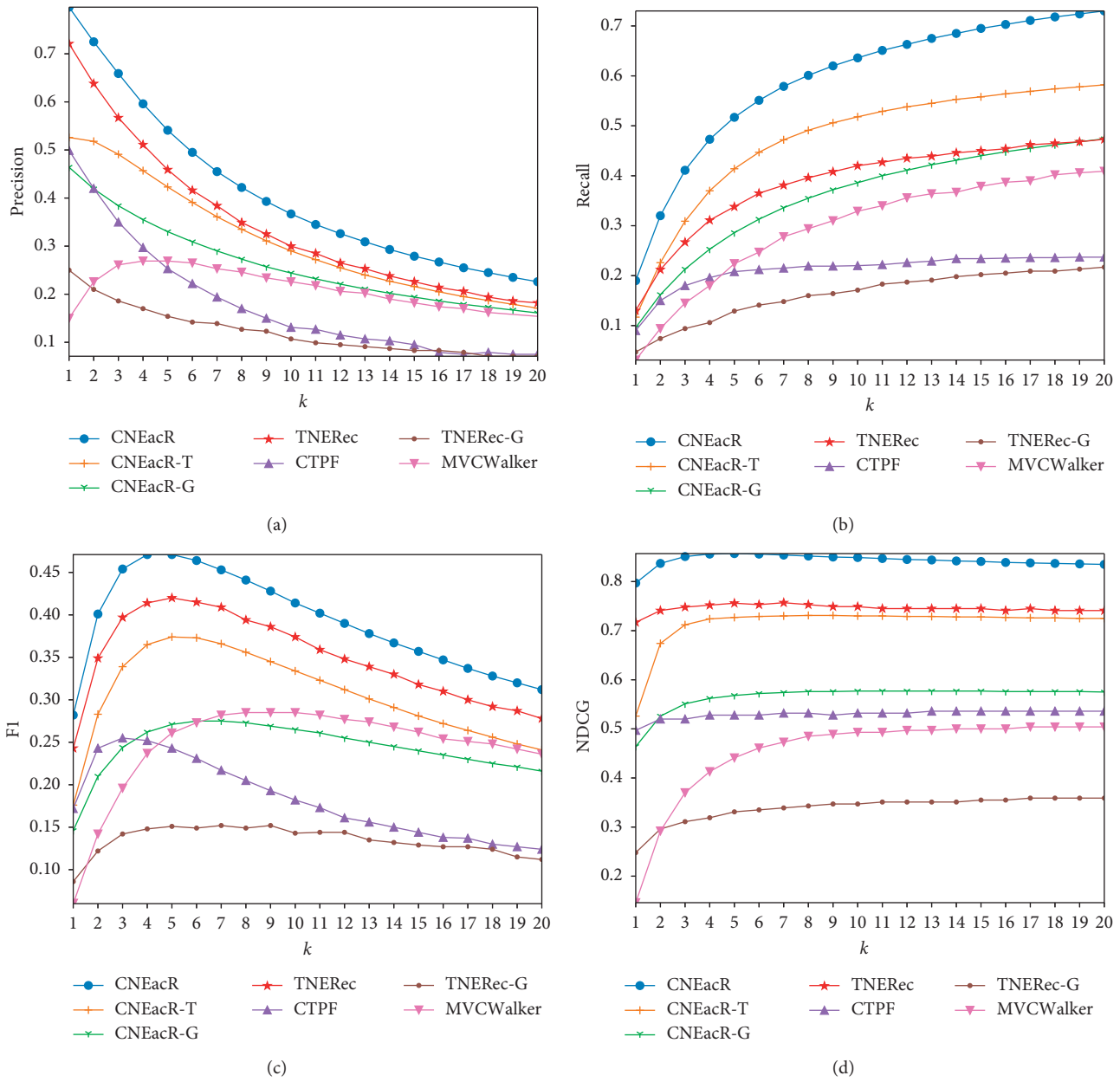

Figure 4: Performance in different k. (a) Precision in different k. (b) Recall in different k. (c) F1 in different k. (d) NDCG in different k.

5. Related Work

At present, it is common for the researcher to collaborate in research [26]. A researcher who collaborates with others has an enormous effect on scientific productivity than those who always do the research independently [27]. So, how to find proper collaborators from complex and unstructured data is essential for the researchers. Recently, lots of works have been done on how to help the researcher to find proper collaborators. These works on academic collaborator recommendations are mainly based on three categories: network-based recommendation, content-based recommendation, and hybrid recommendation.

In an academic collaborator network, academic collaborator recommendation is usually modeled as a link prediction problem. The key to predicting the relationship with structural features of the academic collaborator network is to calculate the similarity among researchers. In [28], Jeh and Widom used SimRank scores based on a simple and intuitive graph-theoretic model to measure the similarity between two researchers. However, they cannot exploit all different length paths of the network. To overcome this problem, they provided more accurate and faster friend recommendations by traversing all limited length paths [29]. Recently, new measurements such as relative entropy [30] and network motif [31] were proposed. The most popular model in the field of collaborator

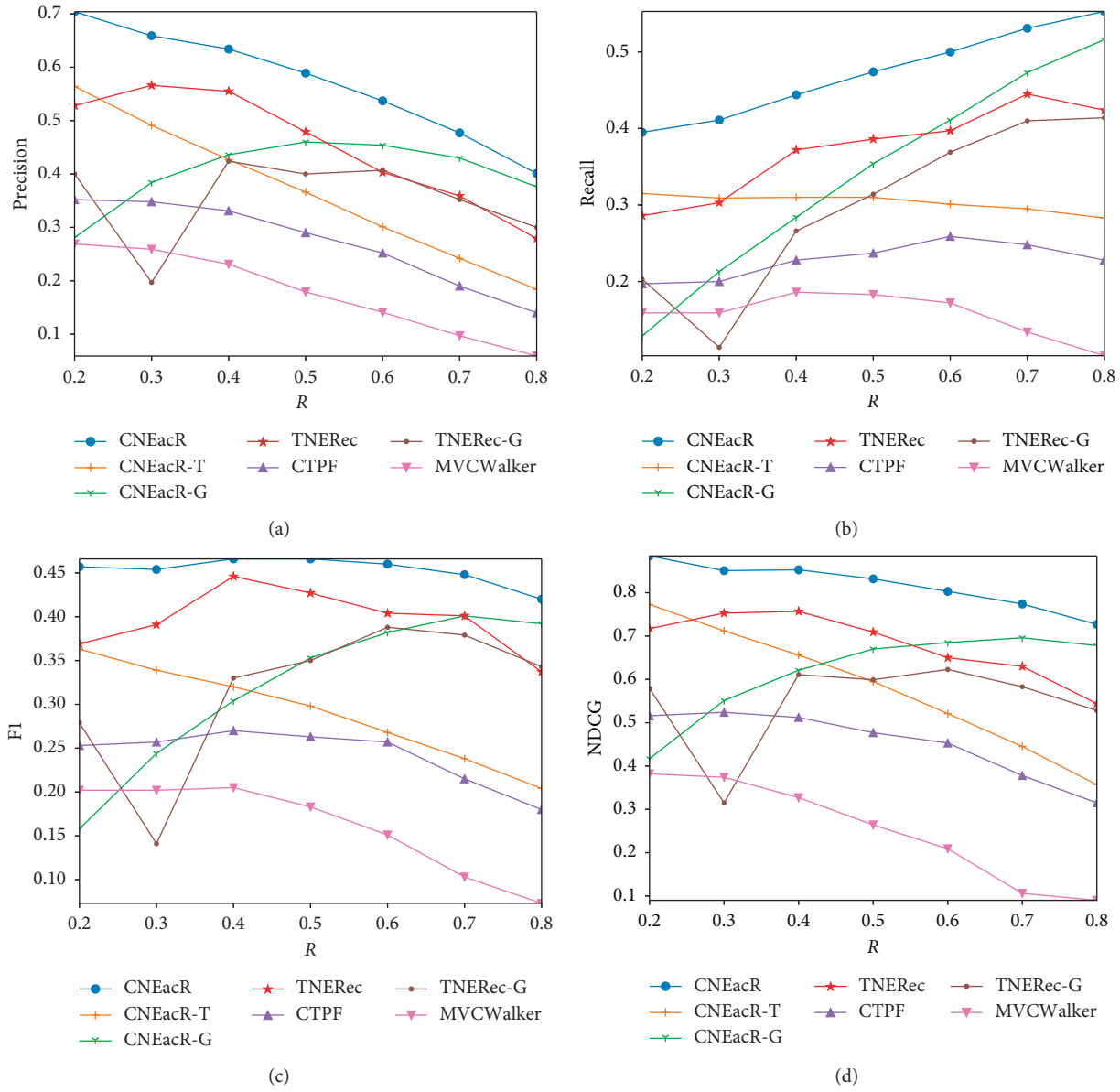

Figure 5: Performance in different R. (a) Precision in different R. (b) Recall in different R. (c) F1 in different R. (d) NDCG in different R.

recommendation was random walk [32]. There exist some works that stand on the shoulder of random walk for academic collaborator recommendation, which had been proved to be competent for calculating the rank score of researchers in the academic collaborator network [33]. These methods completely utilized the weight on edge to guide Random Walker on the academic collaborator network [24, 34, 35]. These values of weight were composed of the affiliated institution of the researcher or the academic factors, such as coauthor order, latest collaboration time, and the times of collaboration. MVCWalker used the rich information of both nodes and links to dig out the similarity structure of the academic collaborator network based on probability [33, 36]. However, Random Walker can merely extract information from the academic collaborator network.

Using structural features is not sufficient for academic collaborator recommendation. The proposed models for computing similarity between two researchers were based on expertise profiles extracted from their publications and academic home pages [7]. Kong et al. held that the interest of each researcher was very important for academic collaborator recommendation. The topic model was used to mine the text information of researchers each year to obtain the topic information and then cluster the topics as the researchers' interests [8]. The cross-domain topic learning model used topic layers to replace author layers to alleviate the sparseness issue and topic skewness for different discipline collaborations [9].

The text information and network structural information are equally important to academic collaborator recommendations. There exist some hybrid recommendation models.

Table 5: Case study of different methods for collaborator recommendation.

Target researcher: F. Ishikawa			
Method	Rank	Returned academic collaborators	
CNEacR-G	1	T. Naka	√
	2	J. C. P. Klaasse	×
	3	A. Gasparini	×
	4	J. Hartbaum	×
	5	Y. K. Huang	×
CNEacR-T	1	T. Nakama	×
	2	T. Naka	√
	3	A. Mastsushita	√
	4	ClaudiaFelser	×
	5	HitoshiOsawa	×
CNEacR	1	Y. Takaesu	√
	2	T. Naka	√
	3	A. Mastsushita	√
	4	H. Tsukagoshi	×
	5	T. Nakane	√

They combined the structural information and user-generated content. And then, a generative model was introduced to help people find friends on Twitter and Flickr [10]. CCRec clustered the topics of each researcher's text information and utilized the structure of the academic collaborator network to find the most relevant and latent collaborator [11]. A hybrid algorithm with eight measures was proposed to recommend latent academic collaborators under different disciplines [37]. It generated high-quality researchers' profiles by integrating researchers' expertise, coauthor network characteristics, and researchers' institutional connectivity into a unified framework with SVM-rank [38]. It was applied in the ScholarMate system, which is a virtual academic community for promoting researchers' collaboration. They predicted coauthor relationships based on content, social, and hybrid recommendation algorithms [12]. Kong et al. thought that the fusing topic model and academic relationships could improve the performance of academic collaborator recommendations [13]. However, the topic model showed the probability distribution of words and documents, which only demonstrated their implied topics. The title of a paper was always short, but it contained the main idea of the whole paper which can distinctly express the research field of a researcher. Word2vec [16] was based on text information (i.e., semantic and syntactic) of a researcher, which can express the researchers' feature representation in specific application scenarios. In this paper, we use the weighted text representation to represent each researcher. Then, a context-enhanced network was built according to the similarity between every two researchers to predict collaborative relationships.

6. Conclusion

In this paper, we propose a novel CNEacR method to recommend academic collaborators. CNEacR utilizes the weighted text representation to build a content-enhanced academic collaborator network that contains not only intrinsic collaborative relationships but also the latent semantic relationships formed by the text information. From this network, we use network embedding to get high-quality latent representation, which captures the latent semantic relationships among researchers. Extensive experiments on the real-world datasets demonstrate the effectiveness of CNEacR and its superiority over several existing methods.

We just pay attention to these strong relationships (the paper content and academic relationships), while the weak-tie relationship such as conference or journal is also supposed to be considered. Because the two papers from the same conference or journal share the same research field, researchers are likely to build a collaborative relationship in the future. Thus, we will take the weak-tie relation into account in the next job.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant nos. 61876001, 61602003, and 61673020), National High Technology Research and Development Program (Grant no. 2017YFB1401903), the Provincial Natural Science Foundation of Anhui Province (Grant no. 1708085QF156), the Major Program of the National Social Science Foundation of China (Grant no. 18ZDA032), and the Recruitment Project of Anhui University for Academic and Technology Leader.

References

[1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: a survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[2] S. Khan, X. Liu, K. A. Shakil, and M. Alam, "A survey on scholarly data: from big data perspective," *Information Processing & Management*, vol. 53, no. 4, pp. 923–944, 2017.

[3] S. Lee and B. Bozeman, "The impact of research collaboration on scientific productivity," *Social Studies of Science*, vol. 35, no. 5, pp. 673–702, 2005.

[4] J. N. Cummings and S. Kiesler, "Collaborative research across disciplinary and organizational boundaries," *Social Studies of Science*, vol. 35, no. 5, pp. 703–722, 2005.

[5] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 287–296, Hong Kong, China, February 2011.

[6] T. Huynh, K. Hoang, and D. Lam, "Trend based vertex similarity for academic collaboration recommendation," in *Proceedings of the International Conference on Computational Collective Intelligence*, pp. 11–20, Craiova, Romania, September 2013.

[7] S. D. Gollapalli, P. Mitra, and C. L. Giles, "Similar researcher search in academic environments," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 167–170, Washington, NJ, USA, June 2012.

[8] X. Kong, H. Jiang, W. Wang, T. M. Bekele, Z. Xu, and M. Wang, "Exploring dynamic research interest and academic influence for scientific collaborator recommendation," *Scientometrics*, vol. 113, no. 1, pp. 369–385, 2017.

[9] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1293, Beijing, China, August 2012.

[10] N. Barbieri, F. Bonchi, and G. Manco, "Who to follow and why: link prediction with explanations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1266–1275, New York, NY, USA, August 2014.

[11] X. Kong, H. Jiang, Z. Yang, Z. Xu, F. Xia, and A. Tolba, "Exploiting publication contents and collaboration networks for collaborator recommendation," *PLoS One*, vol. 11, no. 2, Article ID e0148492, 2016.

[12] D. H. Lee, P. Brusilovsky, and T. Schleyer, "Recommending collaborators using social features and mesh terms," *Proceedings of the American Society for Information Science and Technology*, vol. 48, no. 1, pp. 1–10, 2011.

[13] X. Kong, M. Mao, J. Liu, B. Xu, R. Huang, and Q. Jin, "Tnerec: topic-aware network embedding for scientific collaborator recommendation," in *Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing*, pp. 1007–1014, Guangzhou, China, May 2018.

[14] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval-SIGIR '99*, pp. 211–218, Tokyo, Japan, 2017.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[16] Y. Goldberg and O. Levy, "Word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," https://arxiv.org/abs/1402.3722.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," https://arxiv.org/abs/1810.04805.

[18] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: online learning of social representations," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, New York, NY, USA, August 2014.

[19] X. Huang, J. Li, and X. Hu, "Label informed attributed network embedding," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 731–739, Cambridge, UK, February, 2017.

[20] A. Grover and J. Leskovec, "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, San Francisco, CA, USA, August 2016.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," https://arxiv.org/abs/1609.02907.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," https://arxiv.org/abs/1301.3781.

[23] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, Article ID aaf5239, 2016.

[24] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "Mvcwalker: random walk-based most valuable collaborators recommendation exploiting academic factors," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 364–375, 2014.

[25] P. K. Gopalan, L. Charlin, and D. Blei, "Content-based recommendations with Poisson factorization," *Advances in Neural Information Processing Systems*, vol. 4, pp. 3176–3184, 2014.

[26] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 990–998, Las Vegas, NV, USA, August 2008.

[27] B. Bozeman and C. Boardman, *Research Collaboration and Team Science: A State-Of-The-Art Review and Agenda*, Springer, Berlin, Germany, 2014.

[28] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 538–543, Edmonton, Canada, July 2002.

[29] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Friendlink: link prediction in social networks via bounded local path traversal," in *Proceedings of the 2011 International Conference on Computational Aspects of Social Networks (CASoN)*, pp. 66–71, Salamanca, Spain, October 2011.

[30] Q. Zhang, M. Li, and Y. Deng, "Measure the structure similarity of nodes in complex networks based on relative entropy," *Physica A: Statistical Mechanics and Its Applications*, vol. 491, pp. 749–763, 2018.

[31] F. Aghabozorgi and M. R. Khayyambashi, "A new similarity measure for link prediction based on local structures in social networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 501, pp. 12–23, 2018.

[32] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[33] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in *Proceedings of the Fourth ACM International Conference on*

Web Search and Data Mining, pp. 635–644, Hong Kong, China, February 2011.

[34] J. Li, F. Xia, W. Wang, Z. Chen, N. Y. Asabere, and H. Jiang, "Acrec: a co-authorship based random walk model for academic collaboration recommendation," in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 1209–1214, Seoul, Korea, April 2014.

[35] X. Zhou, L. Ding, Z. Li, and R. Wan, "Collaborator recommendation in heterogeneous bibliographic networks using random walks," *Information Retrieval Journal*, vol. 20, no. 4, pp. 317–337, 2017.

[36] C. Yang, J. Ma, T. Silva, X. Liu, and Z. Hua, "A multilevel information mining approach for expert recommendation in online scientific communities," *The Computer Journal*, vol. 58, no. 9, pp. 1921–1936, 2013.

[37] P. Chaiwanarom and C. Lursinsap, "Collaborator recommendation in interdisciplinary computer science using degrees of collaborative forces, temporal evolution of research interest, and comparative seniority status," *Knowledge-Based Systems*, vol. 75, pp. 161–172, 2015.

[38] C. Yang, J. Sun, J. Ma, S. Zhang, G. Wang, and Z. Hua, "Scientific collaborator recommendation in heterogeneous bibliographic networks," in *Proceedings of the 48th Hawaii International Conference on System Sciences*, pp. 552–561, Kauai, HI, USA, January 2015.