IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INNOVATION AND APPLICATION OF INTERNET OF THINGS AND EMERGING TECHNOLOGIES IN SMART SENSING

Received August 25, 2019, accepted September 2, 2019, date of publication September 12, 2019, date of current version October 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941022

Career Age-Aware Scientific Collaborator Recommendation in Scholarly Big Data

NA SUN, YONG LU, AND YONGCUN CAO
School of Information Engineering, Minzu University of China, Beijing 100081, China
Corresponding author: Na Sun (sunna_07@muc.edu.cn)

ABSTRACT Seeking a collaborator is one of the important academic activities of scholars because the right collaborators will help improve the quality of scholars' research and accelerate their research process. Therefore, it is becoming more and more important to recommend scientific collaborators based on big scholarly data. However, previous works mainly consider the research topic as the key academic factor, whereas many scholars' demographic characteristics such as career age, gender, etc are overlooked. It has been studied that scientific collaboration patterns may vary with scholars' career ages. It is not surprising that scholars at different career ages may have different collaboration strategies. To this end, we aim to design a scientific collaboration recommendation model that is sensitive to scholars' career age. For this purpose, we design a career age-aware scientific collaboration model. The model is mainly consisted of three parts, including authorship extraction from the digital libraries, topic extraction based on publication titles/abstract, and career age-aware random walk for measuring scholar similarity. Experimental results on two real-world datasets demonstrate that our proposed model can achieve the best performance by comparison with six baseline methods in terms of precision and recall.

INDEX TERMS Scientific collaboration, career age, collaborator recommendation.

I. INTRODUCTION

In recent years, with the continuous development of information technology, the scale of scientific collaboration network has continued to grow and develop in the academic society [1], [2]. Moreover, scientific collaboration as an important means of communication in the academic field has also attracted a large number of scholars to participate in academic cooperation. Through scientific collaboration, scholars can obtain a lot of benefits. It has been studied by previous research that collaborative scholars are more productive and may have a higher citations [3].

With the advance of online social network, scholars are free to publish information and exchange ideas with others. Due to the lack of professional academic atmosphere in general social networks, there have been specializations for designing academic specific social network [4]. Research academic social networking sites, such as ResearchGate, Academia.edu enable scholars to access to each other more easily. This enables scholars in all fields of the world to easily conduct real-time academic discussions and seek potential cooperation opportunities. However, due to the information overload problem, finding similar scholars and potential collaborators has become ever difficult. Thus, it is an effective solution to build a personalized collaborator recommendation system [5].

Seeking a collaborator is one of the important academic activities of scholars because the right collaborators will help improve the quality of scholars' research and accelerate their research process. Therefore, it is becoming more and more important to recommend scientific collaborators based on big scholarly data. The focus of scientific collaboration is the scientific collaboration networks, where there will be a link between two scholars if they have published papers together. Scientific collaboration network contains paper cooperation information and is a special social network [6].

Traditional collaboration recommendation methods are designed based on the idea of link prediction [7]. Link prediction refers to the probability that two nodes in the network that are not yet related may be connected with each other by some prediction methods according to the existing network information such as network structures [8]. Link prediction includes both predictions of existing but unknown links for two nodes that have never had a relationship, and predictions for future links that have already produced relationships.

The associate editor coordinating the review of this manuscript and approving it for publication was Honghao Gao.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

Link prediction can be classified into the data mining field and it has a lot of theoretical bases. These basic theoretical studies include Markov chain related methods and machine learning methods. There are three main methods for predicting links [9], including network topology-based approaches, probability model-based link prediction methods, and clustering-based link prediction approaches. However, scholars are involving in the academic society. Merely considering network information is not necessary for designing a collaborator recommendation system. Many works have been done to explore the academic factors to improve the performance of scientific collaboration recommendation systems [10].

However, previous works mainly consider the research topic as the key academic factor, whereas many scholars' demographic characteristics such as career age, gender, etc are overlooked. It has been studied that scientific collaboration patterns may vary with scholars' career ages [11]–[13]. It is not surprising that scholars at different career ages may have different collaboration strategies. For example, when starting a new collaboration, junior scholars are more possible to be pursuers while the senior scholars are more possible to be attractors.

To this end, we aim to design a scientific collaboration recommendation model that is sensitive to scholars' career age. For this purpose, we design a career age-aware scientific collaboration model. The model is mainly consisted of three parts, including authorship extraction from the digital libraries, topic extraction based on publication titles/abstract, and career age-aware random walk for measuring scholar similarity. Experimental results on two real-world datasets demonstrate that our proposed model can achieve the best performance by comparison with six baseline methods in terms of precision and recall.

The organization of this paper is as follows. Section 2 presents the related works. The preliminary and problem definition are given in Section 3. The details of the proposed method are presented in Section 4. Section 5 introduces the experimental setups and the experimental results are given in Section 6. Finally, Section 7 concludes this paper.

II. RELATED WORKS

Scientific collaboration has been extensively studied and has become an important research topic in the field of scholarly big data. Its study can help scholars better understand scientific collaboration. It has the potential to help scholars conduct scientific research more efficiently and expand their academic impact [14], [15].

Newman [16], [17] introduces the idea of scientific collaboration network and investigated the structure of it. The basic idea is that two scholars are considered connected if they have authored at least one paper together and the whole scientific collaboration networks are constructed by using data drawn from the digital libraries. He presents some basic characteristics of scientific collaboration networks, including mean and distribution of numbers of collaborators of authors, demonstrates the presence of clustering in the networks [18]. The small world phenomenon is also observed that randomly selected pairs of scholars are typically separated by only a short path in the whole network [16].

After his study, many works have been done to explore the nature of scientific collaboration based on the scientific collaboration network. For example, Bu *et al.* [19] perform a temporal analysis on the scientific collaboration of scholars in the field of computer science and find that collaborators of high-impact scholars tend to perform diverse research topics. Zhang *et al.* [20] present a systematic approach to analyze scientific collaboration from the perspective of homophily, transitivity, and preferential attachment. They find that scholars' willing to start new collaborations with their coauthors' neighbors is strong. If two scholars share many collaborators, they are more willing to be connected with each other. Wang *et al.* [11] studied scientific collaboration patterns from the perspective of career ages and find that scholars at different career stages may have different collaboration strategy. It has been studied that academic conferences may promote scientific collaboration [21].

The study on scientific collaboration can help design the scientific collaboration system. Scientific collaboration recommendation is one of the widely investigated recommendation tasks in scholarly recommendation [22]. The goal is to recommend suitable potential collaborators for the target scholar so that the information overload issues can be tackled.

To design a scientific recommendation system, scholars usually take advantages of the existing approached in link prediction [9]. Various node similarity measurement indices such as common neighbors and random walk score have been explored. However, merely employing network topology is not sufficient because scholars have their own academic characteristics. Many works have introduced academic factors in designing scientific collaboration recommendation system [23], [24]. One of the most popular used metrics is the research topic [10], [25], [26]. We believe that we need also consider scholars' demographic factors such as career age.

III. PRELIMINARY AND PROBLEM DEFINITION

The focus of this paper is to design a scientific collaborator recommendation system for scholars. Before introducing our model, we will first give some preliminaries and define our questions.

A. PRELIMINARIES

We first introduce some related preliminaries for better understanding.

Definition 1 (Scientific Collaboration Network): Scientific collaboration network is a special kind of social network where nodes are scholars and links denotes the coauthorships. If two scholars have coauthored at least one publication, there will be a link between them. When constructing a scientific collaboration network, we need to first extract the author list of a certain publication from the academic digital library. Then, the links can be gained via the coauthor list.

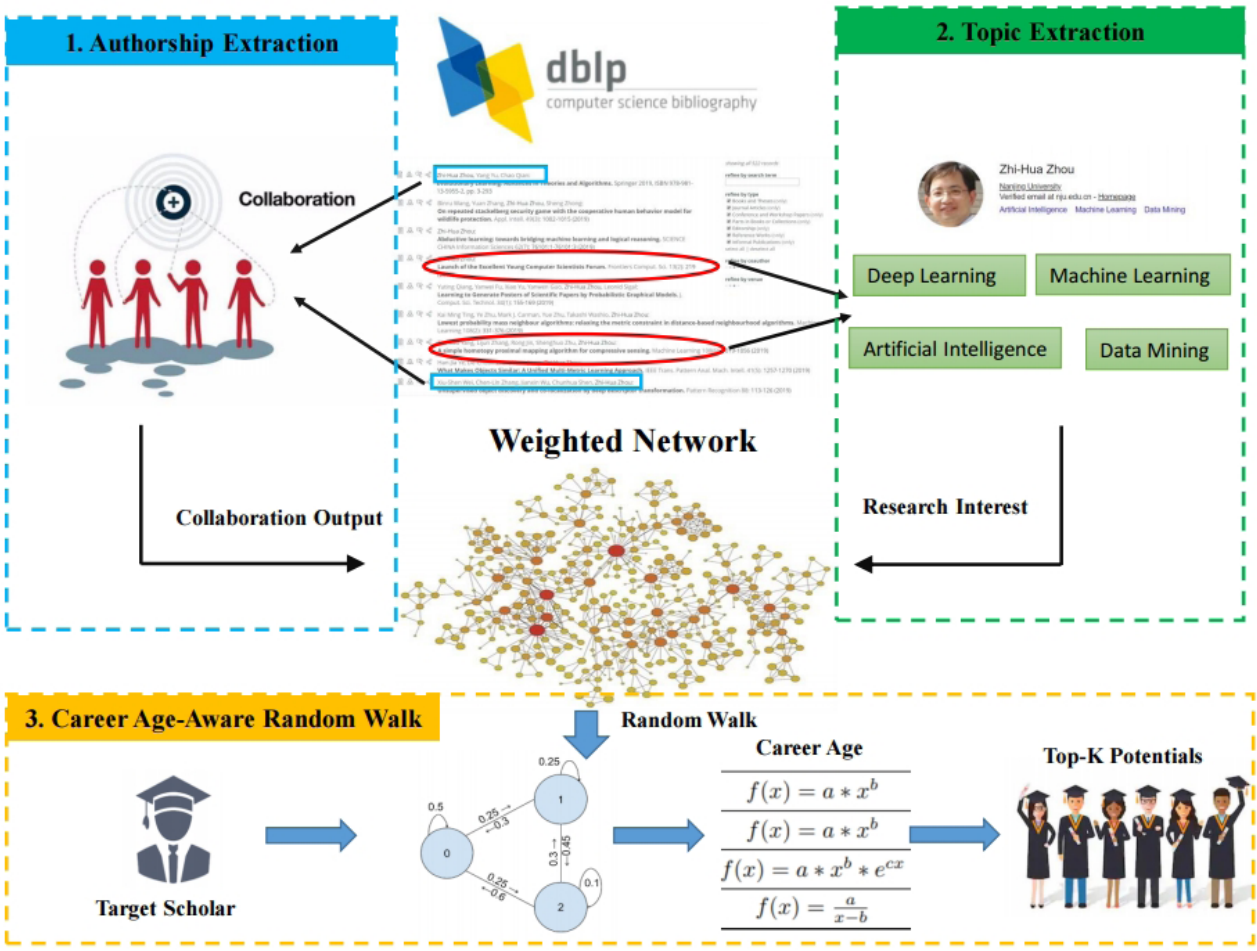


FIGURE 1. Framework of the proposed career age-aware scientific collaboration recommendation.

Definition 2 (Scholar Similarity Measurement): The key of scientific collaborator recommendation is to measure the similarity between scholars. When designing friend recommendation systems, most online social networks employ the FOF strategy (friend of friend) which assumes that people with common friends are willing to know each other. Here, the FOF denotes the basic strategy for measuring node similarity. So far, there are various network-based approaches for measuring scholar similarity.

Definition 3 (Carrer Age): Scholar in different academic ages may have different collaboration strategy. For example, it has been studied that junior scholars are more possible to be followers while senior scholars are more possible to be attractors. To catch the dynamic patterns of scholars' collaboration strategy, we propose the idea of career age (or academic age). Career age denotes how long a scholar has been involved in the academic society. It can be calculated by the latest publication time minus the first publication time.

Definition 3 (Scientific Collaborator Recommendation): Scientific collaborator recommendation is to recommend suitable collaborators for target scholars in the scholarly big data age where finding relevant scholars for collaboration has become increasingly difficult. Potential collaborators are selected and ranked based on the similarity to the target scholars.

B. PROBLEM DEFINITION

Our goal is to recommend collaborators via the power of scholarly big data. Since scientific collaboration network is a good way to depict coauthorships, we attempt to recommend potential scholars based on scientific collaboration network. Meanwhile, various academic factors, including career age are considered to improve the performance of our recommendation systems.

Given: a target scholar and his/her publication records in the academic digital libraries;

Recommend: a list of top-K potential scholars to the target scholar for future collaboration.

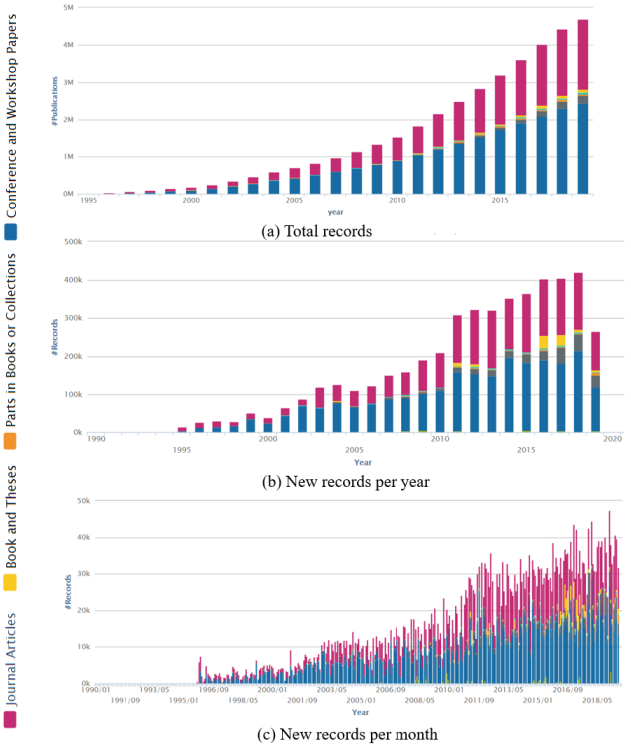
IV. METHODS

In this section, we first introduce the framework of our proposed methods which is consisted of three parts including authorship extraction from DBLP dataset, topic extraction based on publication titles/abstract, and career age-aware

random walk for measuring scholar similarity. Then, we present the details of each part.

A. MODEL FRAMEWORK

Our proposed method is mainly consisted of three parts including authorship extraction from DBLP dataset, topic extraction based on publication titles/abstract, and career age-aware random walk for measuring scholar similarity. The main ideas of each part are depicted in Figure 1. We will introduce each part in details.



(a) Total records

(b) New records per year

(c) New records per month

FIGURE 2. Statistics of DBLP dataset.

B. AUTHORSHIP EXTRACTION

There are various academic digital libraries that enable scholars access to their dataset for research purpose, i.e., APS (American Physical Society),[1] MAG (Microsoft Academic Graph),[2] DBLP (DataBase systems and Logic Programming),[3] etc. In this paper, we adopt the DBLP dataset for example. Some statistics of the DBLP dataset are illustrated in Figure 2. We can see from this figure that DBLP collects more than 4 million publications in total and the number of new publications is increasing greatly.

DBLP dataset is an XML (Extensible Markup Language) file that can be used to mine the meaningful information we want. XML is an extensible markup language and a simple data storage primitive that describes the data by using a series of simple tags that are readable, from which the meaning of the representation can be easily understood. However, the XML language cannot be identified and run by a computer and can only be parsed in other languages.

There are two commonly used XML parsing techniques including SAX based on event flow and DOM based on XML document tree structure. The dblp.xml file provided by DBLP exceeds 1G. Therefore, we use SAX's streaming parsing. SAX is based on event execution, that is, the SAX parser will generate an event stream when reading an XML document on one side, and then process each event through the corresponding method in the callback event program, such as the element's start tag and End tags. When processing the data, we use the Python language with the SAX package.

Since the object processed by our method is graph data, how to quickly acquire data and construct a computer-processable graph structure is the key issue. Taking advantages of the Redis database' key-value structure and key-set structure, we have specially designed the data structure. We use a storage model similar to the follow-up relationship employing by online social media. Among them, the key is the author and the set is all the collaborators of the author. When constructing a graph, we record it by maintaining and storing each node and all nodes directly connected to it. Such a key will cover the full picture. Of course, the disadvantage of this method is that the storage relationship is redundant, and more space is sacrificed. But the rapid construction of the graph in the calculation of the recommended algorithm is guaranteed.

During the data processing, we need the attribute information of the relationship between the two authors and the author order. In order to ensure the convenience of data processing and storage efficiency, we use a special storage method. The key is spliced by two author codes, and the value is "time: author1 order: author2 order". For each such data, the essence corresponds to the relationship between the two authors of each paper, including the paper cooperation time, the author order in the paper cooperation. Finally, we can construct a scientific collaboration network.

C. TOPIC EXTRACTION

It has been proven that research interest plays an important role in scientific collaborator recommendation. For example, Kong *et al.* [25] demonstrated that exploring publication content can benefit the collaborator recommendation system. Therefore, we need to consider the research interest similarity.

One commonly used approached for extracting scholars' research interest is the topic model [27]. Topic model is an unsupervised Bayesian model, which presents each document in a document set as a probability distribution with an unsupervised learning approach. It does not require a manually labeled training set during training. What is needed is the number of documents and the number of specified topics. Topic model is a typical word bag model which assumes that

[1] https://journals.aps.org/datasets

[2] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[3] https://dblp.dagstuhl.de/xml/

a document is a collection of words and there is no order and order relationship between words.

In the topic model, a topic is a probability distribution with all the words in the document as a support set, indicating how often the word appears in the topic. That is, words with high relevance to the topic have a greater probability of appearing. When a document has multiple topics, the probability distribution of each topic includes all words, but the value of one word in the probability distribution of different topics is different. A topic model attempts to embody this feature of the document with a mathematical framework. The topic model automatically analyzes each document, counts the words in the document, and based on the statistical information, determines which topics are included in the current document, and how much each topic is.

The topic model we adopt in this paper is the widely used LDA (Latent Dirichlet Allocation) model [28]. Specifically, given a paper or document d, we can sample a multinomial distribution θ_d over topics T based on a Dirichlet distribution with parameter α. For each word w_{di}'s topic t_{di} in the document d_i, its topic is picked from a topic multinomial distribution ψ_t sampled based on a Dirichlet distribution with parameter β. Therefore, we can infer the probability of a word w appearing in a document d as follows:

$$P(w|d, \theta, \psi) = \sum_{t \epsilon T} P(w|t, \psi_t) P(t|d, \theta_d) \tag{1}$$

Then, we can calculate the likelihood of corpora $\mathcal{C}$ as:

$$P(T, W, |\Theta, \Psi) = \prod_{d \epsilon D} \prod_{t \epsilon T} \theta_{dt}^{n_{dt}} \times \prod_{t \epsilon T} \prod_{w \epsilon W} \psi_{tw}^{n_{tw}} \tag{2}$$

where n_{dt} denotes the frequency that the topic t has been appeared in a document d, and n_{tw} denotes the frequency that the word w has been mentioned in a topic t.

When calculating the topic distributions of a given scholar, we need to gain the publication information of the author. Since the DBLP datasets merely include the paper title information without the abstract information. We use the extended DBLP dataset provided by Aminer system.[4] The datasets given by the Aminer system contain both the paper title and the paper abstract so that we can use the content information to calculate the topic distributions based on the LDA model. Specifically, we use the whole collections of a scholar's publication as his document. For each title and abstract, we exclude those stop words such as "the", "of" etc by the stop list provided by Google.[5]

D. CAREER AGE-AWARE RECOMMENDATION

In this section, we introduce how to calculate scholar similarity considering both the collaboration network and paper content. Meanwhile, the similarity is also related to a career age function. The whole career age-ware random walk is consisted of three steps, including:

- **Network Construction**: This step is done by the precious network construction section. The network is constructed based on the idea that there will be a link between two scholars if they had coauthored at least one publication.
- **Link Weight Calculation**: This step is to measure the link weight considering various academic factors so that the random walk can be biased to more similar scholars.
- **Random Walk and Recommendation**: Finally, the random walk with restart model is performed on the weighted collaboration network so that the similarity between scholars can be calculated for recommendation.

1) PLAIN RANDOM WALK

In our proposed model, the potential scholars are recommended to the target scholars based on the similarity between them. The more similar they are, the high possibility will the potential be recommended to the target scholar. Specifically, the similarity depends on the significance of potential scholars y to the target scholar i. Such significance is determined by the random walk score to each other. It can be calculated based on the following equation?

$$Sim_{xy}^{RWR} = \zeta_{xy} + \zeta_{yx}, \tag{3}$$

where ζ_{xy} depicts the significance of y to x and ζ_{yx} depicts the significance of x to y.

Here, the significance ζ is calculated by two factors, including the number of scholars linked to the target scholar and the importance of these scholars. It can be calculated as:

$$\zeta_{xy} = \frac{1-\theta}{N} + \theta \sum_{n_j \in M(n_y)} \frac{\zeta_j}{N(n_j)}, \tag{4}$$

where θ denotes the random walk possibility, $N(n_j)$ denotes the neighbor set of scholar n_j, and ζ_j denotes significance score of scholar n_j to the target scholar n_i. The whole procedure is iterative by the following function:

$$\zeta^{t+1} = \theta \mathbf{S} \zeta^t + (1-\theta) q, \tag{5}$$

where q is the network initial status and $\mathbf{S}$ is the transition matrix.

2) INCORPORATING RESEARCH TOPIC

As discussed before, it is not sufficient to merely consider network topology. The paper content should be considered. Therefore, we propose to use the research interest denoted by topic models to bias the random walk. Specifically, we use the topic similarity between scholars to weight the link between two scholar. Given the topic vector $\overrightarrow{\mathbf{t}_i}$ and the topic vector $\overrightarrow{\mathbf{t}_j}$ of the neighbor scholar, their topic similarity is calculated based on the cosine similarity, which can be calculated as:

$$\cos(\overrightarrow{\mathbf{t}_i}, \overrightarrow{\mathbf{t}_j}) = \frac{\overrightarrow{\mathbf{t}_i} \times \overrightarrow{\mathbf{t}_j}}{|\overrightarrow{\mathbf{t}_j}| \times |\overrightarrow{\mathbf{t}_j}|}. \tag{6}$$

Therefore, the link between two connected scholars is weighted by the $\cos(\overrightarrow{\mathbf{t}_i})$.

[4] https://aminer.org/billboard
[5] https://code.google.com/p/stop-words/

TABLE 1. Journals and conferences in the field of artificial intelligence.

Venues	Full Name	Publisher	DBLP Link
AI	Artificial Intelligence	Elsevier	http://dblp.uni-trier.de/db/journals/ai/
TPAMI	IEEE Trans on Pattern Analysis and Machine Intelligence	IEEE	http://dblp.uni-trier.de/db/journals/pami/
IJCV	International Journal of Computer Vision	Springer	http://dblp.uni-trier.de/db/journals/ijcv/
JMLR	Journal of Machine Learning Research	MIT Press	http://dblp.uni-trier.de/db/journals/jmlr/
TNNLS	IEEE Transactions on Neural Networks and learning systems	IEEE	http://dblp.uni-trier.de/db/journals/tnn/
AAAI	AAAI Conference on Artificial Intelligence	AAAI	http://dblp.uni-trier.de/db/conf/aaai/
ICML	International Conference on Machine Learning	ACM	http://dblp.uni-trier.de/db/conf/icml/
NeurIPS	Annual Conference on Neural Information Processing Systems	MIT Press	http://dblp.uni-trier.de/db/conf/nips/
IJCAI	International Joint Conference on Artificial Intelligence	Morgan Kaufmann	http://dblp.uni-trier.de/db/conf/ijcai/
ECAI	European Conference on Artificial Intelligence	IOS Press	http://dblp.uni-trier.de/db/conf/ecai/

TABLE 2. Journals and conferences in the field of data mining.

Venues	Full Name	Publisher	DBLP Link
TOIS	ACM Transactions on Information Systems	ACM	http://dblp.uni-trier.de/db/journals/tois/
TKDE	IEEE Transactions on Knowledge and Data Engineering	IEEE	http://dblp.uni-trier.de/db/journals/tkde/
TKDD	ACM Transactions on Knowledge Discovery from Data	ACM	http://dblp.uni-trier.de/db/journals/tkdd/
TWEB	ACM Transactions on the Web	ACM	http://dblp.uni-trier.de/db/journals/tweb/
DMKD	Data Mining and Knowledge Discovery	Springer	http://dblp.uni-trier.de/db/journals/datamine/
TKDD	ACM Knowledge Discovery and Data Mining	ACM	http://dblp.uni-trier.de/db/conf/kdd/
SIGIR	International Conference on Research on Development in Information Retrieval	ACM	http://dblp.uni-trier.de/db/conf/sigir/
CIKM	ACM International Conference on Information and Knowledge Management	ACM	http://dblp.uni-trier.de/db/conf/cikm/
ICDM	International Conference on Data Mining	IEEE	http://dblp.uni-trier.de/db/conf/icdm/
WSDM	ACM International Conference on Web Search and Data Mining	ACM	http://dblp.uni-trier.de/db/conf/wsdm/

3) INCORPORATING CAREER AGE

Since scientific collaboration patterns vary with scholars' career ages [11], we need to consider the career age when designing the scientific collaborator recommendation systems. We choose three different functions to measure the influence of career age because no previous study showing the influence of career age on scientific collaboration. Specifically, given the career age a_i of scholar i and the career age a_j of scholar j, we use the following four types of functions $\mathfrak{f}(\bullet)$ to measure the career age difference, including: the power law function:

$$\mathfrak{f}(x) = a \times x^b. \tag{7}$$

The exponential function:

$$\mathfrak{f}(x) = a \times x^b \times e^{cx}. \tag{8}$$

The hyperbolic function:

$$\mathfrak{f}(x) = \frac{a}{x - b}. \tag{9}$$

Therefore, the link weight is determined by both the topic similarity and the career age function. Thus, the link weight w_{ij} can be calculated as:

$$w_{ij} = \cos(\overrightarrow{\mathbf{t}_i}) \times \mathfrak{f}(\bullet) \tag{10}$$

4) RECOMMENDATION

Finally, the potential recommendation list is generated by the rank of the potential scholars to the target scholar. The top-k most similar scholars are recommended to the target scholar.

V. EXPERIMENTAL DESIGN

A. DATASET

We adopt two subset of the DBLP dataset. We investigate the scholars in two research area including the Artificial Intelligence (DBLP-AI) and the Data Mining (DBLP-DM). We extract the authors who publish papers on related journals and conferences. The journals and conference accounting for Artificial Intelligence are given in Table 1. Table 2 gives the journals and conferences for Data Mining. We first extract all the authors of these conferences/journal from 2010 to 2015 as the seed scholars. Then, we extract their first-order neighbors, second order neighbors, and third-order neighbor as the scholar set. These scholars are regarded as the nodes in the scientific collaboration networks. The links denote the collaboration relationships between them. Finally, we randomly delete 20% percents of links and try to recommend scholars accounting for the deleted links.

B. BASELINE METHODS

In order to evaluate the performance of our proposed recommendation systems, we compare our proposed method with various baseline methods, including typical link prediction approaches and existing collaboration recommendation systems:

Typical Link Prediction Approach:

- Common Neighbor (CN): The similarity between scholars are calculated based on the number of scholars that are simultaneously associated with two scholars.
- Random Walk (RW): The basic idea of RW is to traverse a graph from one or a series of vertices to get probability

distribution that depicts the probability that each vertex in the graph is accessed.

- Adamic Adar (AA): AA is a method of intimacy measurement based on co-neighbors between nodes, which can be calculated as:

$$sim_{xy} = \sum_{u \in N(x) \cap u \in N(y)} \frac{1}{\log |N(u)|} \tag{11}$$

Existing Collaborator Recommendation Approach:

- MVCWalker [29]: MVCWalker is an innovative method that stands on the shoulders of random walk with restart by considering three academic factors, i.e., coauthor order, latest collaboration time, and times of collaboration.
- SCORE [24]: SCORE is a sustainable collaborator recommendation system utilizing the weak tie relationships brought by academic conferences for recommendation.
- CACR [23]: CACR is designed by jointly representing scholars and research topics based on their mutual-dependency, and extracting scholars' underlying characters for high-quality new collaborator recommendation

Meanwhile, in order to investigate the impact of considering career age and research topic, we also compare our proposed method with two variations, including:

- **Proposed-T**: this variation does not consider the research topic when calculating the link weight.
- **Proposed-A**: this variation does not consider the career age function when calculating the link weight.

C. EVALUATION METRICS

In order the evaluate the performance of the collaborator recommendation system, we adopt two widely used evaluation metrics including Precision@k and Recall@k. Specifically, the Precision@k is given by:

$$Precision@k = \frac{|Rel_u \cap Rec_u|}{|Rec_u|}, \tag{12}$$

where, Rel_u denotes the real collaboration relationships and Rec_u denotes the recommended scholars. Specifically, we calculate the Precision@k by the average of random selected 200 scholars. The Recall@k can be calculated by:

$$Recall@k = \frac{|Rel_u \cap Rec_u|}{|Rel_u|}. \tag{13}$$

Same as the Preciison@k, we calculate the Precision@k by the average of random selected 200 scholars.

VI. RESULTS AND DISCUSSIONS

We present the experimental results from the three aspects, including exploring career age function, performance comparison, and exploring research topic.

A. EXPLORING CAREER AGE FUNCTION

In the previous section, we have proposed to use three career age function to measure the career age similarity between scholars, including power law function, exponential function, and the hyperbolic function. It is not investigated which function can better depict the career age influence. Therefore, we use three functions to do recommendation and the experimental results are shown in Figure 3.

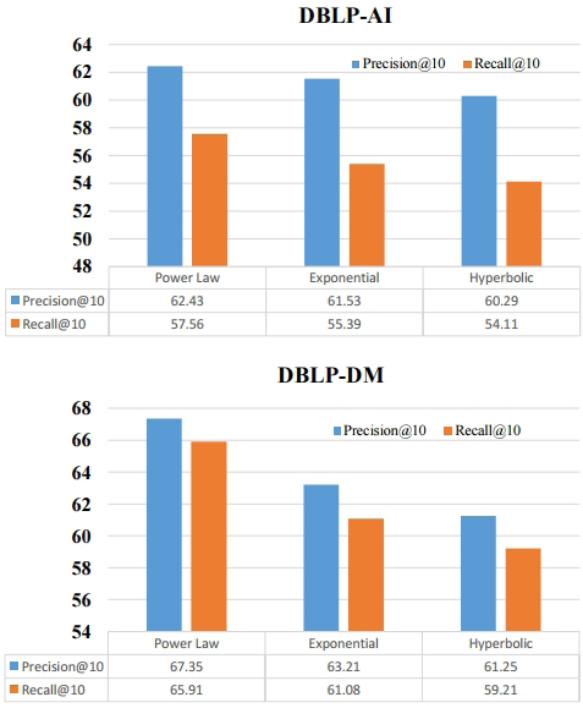


FIGURE 3. Comparison results over different career age functions.

We can see that the power law function can achieve the best performance by comparison with the other two functions. Specifically, in the DBLP-AI dataset (Figure 3(a)), the Precision@10 of the power law function is 62.43%, while the performances of exponential function and hyperbolic function are 61.35% and 60.29%, respectively.

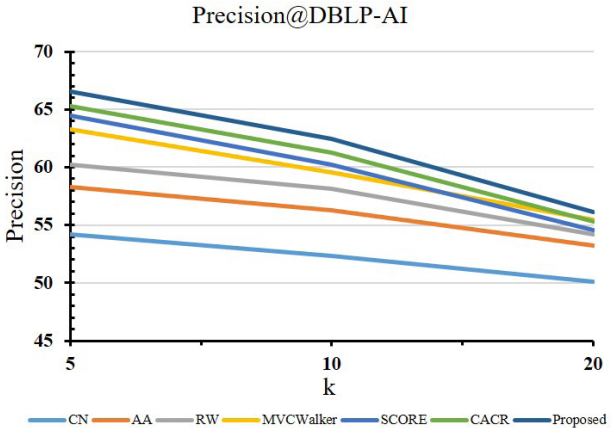
Similar results can be seen on the DBLP-DM dataset (Figure 3(b)). The Precision@10 of the power law function is 67.35%, while the performances of exponential function and hyperbolic function are 63.21% and 61.25%, respectively.

Based on the observation, we can draw the conclusion that power law function can better depict the career age influence. Thus, we adopt the power law function in the following experiments.

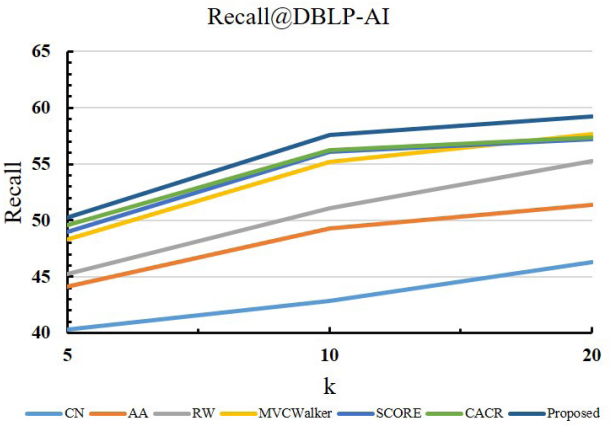
B. ACCURACY COMPARISON

In this section, we compare our propose model with six baseline methods, including CN, RW, AA, MVCWalker, SCORE, and CACR. Figure 4 shows the comparison results on the DBLP-AI dataset in terms of Precision@{5, 10, 20} and Recall@{5, 10, 20}.

We can see from Figure 4(a) that our proposed model can achieve the best performance on both precision and recall. Specifically, the Precision@5 of the proposed method is 66.58%, while the Precision@5 values of CACR, SCORE, MVCWalker, RW, AA, and CN are 65.28%, 64.47%, 63.28%,

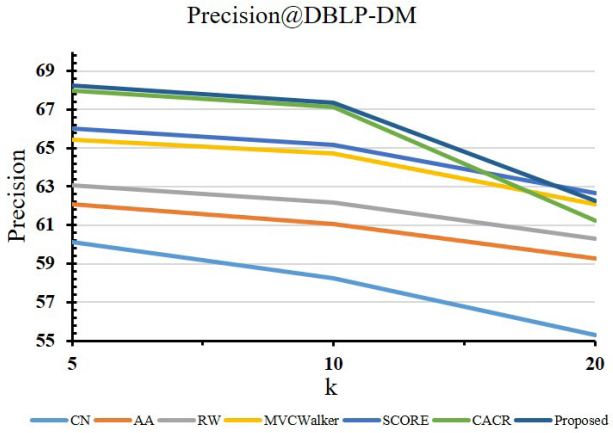


(a) Precision

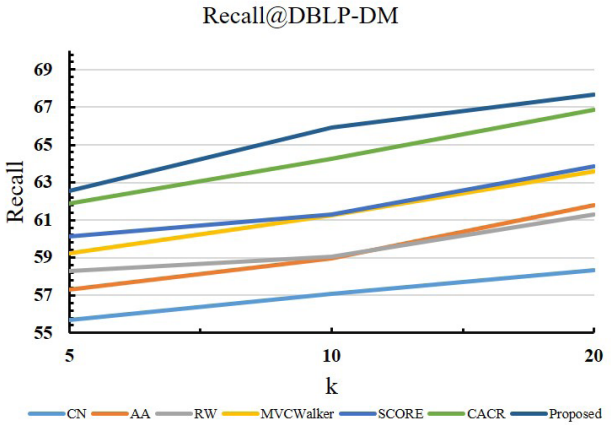


(b) Recall

FIGURE 4. Performance comparison on DBLP-AI dataset.



(a) Precision



(b) Recall

FIGURE 5. Performance comparison on DBLP-DM dataset.

60.21%, 58.28%, and 54.21, respectively. The proposed model can achieve a 2.14% improvement by comparison with the second-best approach CACR. We can also observe that the CACR, SCORE, and MVCWalker perform better the CN, AA, and RW, which demonstrates that it is beneficial to consider academic factors in designing the collaborator recommendation system. Such observation is in line with previous findings [10], [25].

Meanwhile, from Figure 4(b), we can see that the Recall@5 of the proposed model is 50.25%, while the Recal@5 values of the rest baseline methods are 40.28%, 44.12%, 45.23%, 48.28%, 49.01%, and 49.57%, respectively. Our proposed model can achieve a 1.4% improvement compared with the second-best approach CACR. Similarly, We can also observe that the CACR, SCORE, and MVCWalker have higher recall than those of the CN, AA, and RW, which is the evidence that considering academic factors can benefit designing the collaborator recommendation system.

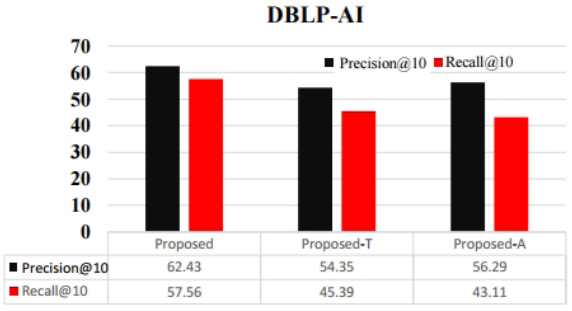
Figure 5 shows the comparison results on the DBLP-DM dataset. Specifically, we can observe from Figure 5(a) that the proposed model has the best performance on both precision and recall. The Precision@5 of the proposed method on DBLP-DM dataset is 68.25%, while the Precision@5 values of CACR, SCORE, MVCWalker, RW, AA, and CN are 67.98%, 66.01%, 65.45%, 63.05%, 62.07%, and 60.12, respectively. The improvement of the proposed method by comparison with the second-best approach CACR is 1.4%. Meanwhile, the CACR, SCORE, and MVCWalker have better performance than the CN, AA, and RW, which also demonstrates that it is beneficial to consider academic factors in designing the collaborator recommendation system.

As can be seen from Figure 5(b), the Recall@5 of the proposed model is 62.57, while the Recal@5 values of the rest baseline methods are 61.89%, 60.14%, 59.25%, 58.28%, 57.29%, and 55.68%, respectively. The proposed model can achieve a 1.1% improvement compared with the second best

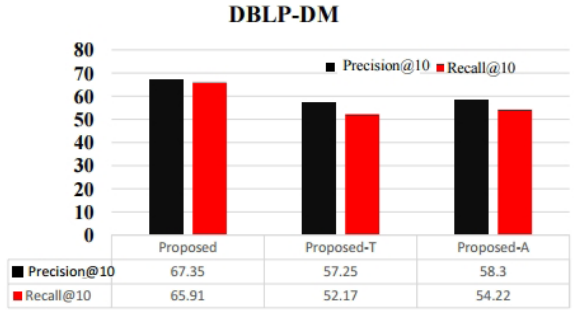
approach CACR. Similar with the results on the DBLP-AI dataset, We can also notice that the CACR, SCORE, and MVCWalker have better performance than CN, AA, and RW, which is also the evidence that it is necessary to consider academic factors when designing the collaborator recommendation system.

C. EXPLORING RESEARCH TOPIC

We have considered two academic factors when designing the recommendation system. We want to know whether these two factors are beneficial for designing the recommendation system. In this section, we compared our proposed model with its two variations, namely Proposed-T and Proposed-A. The Proposed-T variation does not consider the research topic when calculating the link weight and the Proposed-A does not consider the career age function when calculating the link weight. The results on DBLP-AI dataset and DBLP-DM dataset are shown in Figure 6.



(a) DBLP-AI



(b) DBLP-DM

FIGURE 6. Comparison results over different model variations.

We can see that our proposed method can achieve the best performance by comparison with the other two variation methods. Specifically, in the DBLP-AI dataset (Figure 6(a)), the Precision@10 of the proposed method is 62.43%, while the performances of Proposed-T and Proposed-A are 54.35% and 56.29%, respectively. This indicates that considering career age and research topic can help improve the recommendation performance.

Similar results can be seen on the DBLP-DM dataset (Figure 6(b)). The Precision@10 of the proposed method is 67.35%, while the performances of two variations are 57.25% and 58.3%, respectively. Based on the observation, we can draw the conclusion that it is necessary to consider both the research topic and career age for scientific collaborator recommendation.

VII. CONCLUSION

Due to the information overload problem, finding similar scholars and potential collaborators has become ever difficult. Thus, it is an effective solution to build a personalized collaborator recommendation system. While various scientific collaboration recommendation systems have been designed, few of them have considered scholars' demographic characteristics such as career age. It has been studied that scholars may have different collaboration patterns at different career ages. To this end, this paper aims to design a career age-aware scientific collaboration model. The model is mainly consisted of three parts, including authorship extraction from the digital libraries, topic extraction based on publication titles/abstract, and career age-aware random walk for measuring scholar similarity. Experimental results on two real-world datasets demonstrate that our proposed model can achieve the best performance by comparison with six baseline methods in terms of precision and recall. Our work may shed light on designing scientific collaborator recommendation systems via scholars' demographic characteristics.

REFERENCES

[1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, Mar. 2017.
[2] S. Khan, X. Liu, K. A. Shakil, and M. Alam, "A survey on scholarly data: From big data perspective," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 923–944, Jul. 2017.
[3] Y. Bu, Y. Ding, X. Liang, and D. S. Murray, "Understanding persistent scientific collaboration," *J. Assoc. Inf. Sci. Technol.*, vol. 69, no. 3, pp. 438–448, 2018.
[4] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, "Academic social networks: Modeling, analysis, mining and applications," *J. Netw. Comput. Appl.*, vol. 132, pp. 86–103, Apr. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804519300438
[5] G. R. Lopes, M. M. Moro, L. K. Wives, and J. P. M. de Oliveira, "Collaboration recommendation on academic social networks," in *Proc. Int. Conf. Conceptual Modeling*. Berlin, Germany: Springer, 2010, pp. 190–199.
[6] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles, "Collabseer: A search engine for collaboration discovery," in *Proc. 11th Annu. Int. ACM/IEEE Joint Conf. Digit. Libraries*, 2011, pp. 231–240.
[7] N. Benchettara, R. Kanawati, and C. Rouveirol, "A supervised machine learning link prediction approach for academic collaboration recommendation," in *Proc. 4th ACM Conf. Recommender Syst.*, 2010, pp. 253–256.
[8] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, "Simplicial closure and higher-order link prediction," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 48, pp. E11221–E11230, 2018.
[9] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 243–252.
[10] W. Wang, B. Xu, J. Liu, Z. Cui, S. Yu, X. Kong, and F. Xia, "CSTeller: Forecasting scientific collaboration sustainability based on extreme gradient boosting," *World Wide Web*, pp. 1–22, Jul. 2019. doi: 10.1007/s11280-019-00703-y.
[11] W. Wang, S. Yu, T. M. Bekele, X. Kong, and F. Xia, "Scientific collaboration patterns vary with scholars' academic ages," *Scientometrics*, vol. 112, no. 1, pp. 329–343, 2017.
[12] A. M. Petersen, "Quantifying the impact of weak, strong, and super ties in scientific careers," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 34, pp. E4671–E4680, 2015.

[13] A. E. Bayer and J. E. Dutton, "Career age and research-professional activities of academic scientists: Tests of alternative nonlinear models and some implications for higher education faculty policies," *J. Higher Educ.*, vol. 48, no. 3, pp. 259–282, 1977.
[14] J. M. Fagan, K. Eddens, J. Dolly, N. Vanderford, H. Weiss, and J. Levens, "Assessing research collaboration through co-authorship network analysis," *J. Res. Admin.*, vol. 49, no. 1, pp. 76–99, Spring 2018.
[15] G. Abramo, C. A. D'Angelo, and F. Di Costa, "The collaboration behavior of top scientists," *Scientometrics*, vol. 118, no. 1, pp. 215–232, 2019.
[16] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 2, pp. 404–409, 2001.
[17] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Proc. Nat. Acad. Sci. USA*, vol. 101, pp. 5200–5205, Apr. 2004.
[18] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 1, 2001, Art. no. 016132.
[19] Y. Bu, D. S. Murray, J. Xu, Y. Ding, P. Ai, J. Shen, and F. Yang, "Analyzing scientific collaboration with 'giants' based on the milestones of career," *Proc. Assoc. Inf. Sci. Technol.*, vol. 55, no. 1, pp. 29–38, 2018.
[20] C. Zhang, Y. Bu, Y. Ding, and J. Xu, "Understanding scientific collaboration: Homophily, transitivity, and preferential attachment," *J. Assoc. Inf. Sci. Technol.*, vol. 69, no. 1, pp. 72–86, Jan. 2018.
[21] W. Wang, X. Bai, F. Xia, T. M. Bekele, X. Su, and A. Tolba, "From triadic closure to conference closure: The role of academic conferences in promoting scientific collaborations," *Scientometrics*, vol. 113, no. 1, pp. 177–193, Oct. 2017.
[22] K. Sugiyama and M.-Y. Kan, "Scholarly paper recommendation via user's recent research interests," in *Proc. 10th Annu. Joint Conf. Digit. Libraries*, 2010, pp. 29–38.
[23] Z. Liu, X. Xie, and L. Chen, "Context-aware academic collaborator recommendation," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1870–1879.
[24] W. Wang, J. Liu, Z. Yang, X. Kong, and F. Xia, "Sustainable collaborator recommendation based on conference closure," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 2, pp. 311–322, Apr. 2019.
[25] 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, 2016, Art. no. e0148492.
[26] Q. Zhang, R. Mao, and R. Li, "Spatial–temporal restricted supervised learning for collaboration recommendation," *Scientometrics*, vol. 119, no. 3, pp. 1497–1517, 2019.
[27] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertainty Artif. Intell.* Arlington, VA, USA: AUAI Press, 2004, pp. 487–494.
[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[29] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "MVCWalker: Random walk-based most valuable collaborators recommendation exploiting academic factors," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 364–375, Sep. 2014.

NA SUN received the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications, in 2010. She is currently a Lecturer with the School of Information Engineering, Minzu University of China. Her current research interest include parallel algorithm and intelligent systems.

YONG LU received the master's degree, in 2005. He is currently an Engineer with the School of Information Engineering, Minzu University of China. His research interests include parallel algorithm and intelligent systems.

YONGCUN CAO received the bachelor's degree, in 1986. He is currently a Professor with the School of Information Engineering, Minzu University of China. His current research interest include big data, parallel algorithm, and intelligent systems.