Collaborative Filtering with Network Representation Learning for Citation Recommendation

Wei Wang, Tao Tang, Feng Xia, *Senior Member, IEEE*, Zhiguo Gong, *Senior Member, IEEE*, Zhikui Chen, and Huan Liu, *Fellow, IEEE*

Abstract—Citation recommendation plays an important role in the context of scholarly big data, where finding relevant papers has become more difficult because of information overload. Applying traditional collaborative filtering (CF) to citation recommendation is challenging due to the cold start problem and the lack of paper ratings. To address these challenges, in this paper, we propose a collaborative filtering with network representation learning framework for citation recommendation, namely CNCRec, which is a hybrid user-based CF considering both paper content and network topology. It aims at recommending citations in heterogeneous academic information networks. CNCRec creates the paper rating matrix based on attributed citation network representation learning, where the attributes are topics extracted from the paper text information. Meanwhile, the learned representations of attributed collaboration network is utilized to improve the selection of nearest neighbors. By harnessing the power of network representation learning, CNCRec is able to make full use of the whole citation network topology compared with previous context-aware network-based models. Extensive experiments on both DBLP and APS datasets show that the proposed method outperforms state-of-the-art methods in terms of precision, recall, and MRR (Mean Reciprocal Rank). Moreover, CNCRec can better solve the data sparsity problem compared with other CF-based baselines.

Index Terms—Network Representation Learning, collaborative filtering, citation recommendation, scholarly big data.

1 Introduction

In recent years, we have witnessed the rapidly growing of scientific articles in the era of scholarly big data [1]. While certainly advantageous, the enormous and rapidly increasing volume of scientific articles brings about the problem of information overload. It is difficult for scholars to go through and digest all the articles when citing relevant and critical previous works in their manuscript. Thus, modern scholars need new ways that enable them access to relevant papers easily.

Recommender systems are one of the promising directions to address the academic information overload problem. Traditional academic search engines, such as Google Scholar and Microsoft Academic Search can recommend a list of relevant papers according to scholars' keyword-based queries. However, such results are usually biased towards the keywords. It is hard to project one's rich information needs into a few keywords. The quality and influence of papers are also overlooked. Meanwhile, query-based recommendations can hardly provide new and serendipitous articles which are essential to scientific research.

Recommending articles to scholars as citations is important, which can be illustrated from two aspects. On the one hand, it can help scholars learn about new fields and understand the development of their fields by recommending older articles; On the other hand, scholars are able to keep up with the state-of-the-art works in their disciplines by recommending new articles. More importantly, it enables scholars to access relevant articles more quickly and conveniently.

As an interesting and practical research problem, citation recommendation has been extensively explored [2]. Content-based recommendations usually find conceptually relevant papers based on topic similarity. The paper topics can be gained based on topic models, such as LDA [3]. Context-aware methods [4], [5] take advantage of local citation context, such as co-citation appearance [6] to find relevant papers. However, millions of papers may share similar topics, and the local citation context may be sparse. It is, therefore, insufficient to measure paper similarity based on topic similarity or local context.

Hybrid recommendation methods based on collaborative filtering (CF) have been proposed. In a CF-based citation recommendation method, the items are articles and the users are the scholars, as shown in Figure 1. Collaborative topic regression (CTR) [7] proposes to combine the CF

This work is partially supported by National Natural Science Foundation of China under Grants No. 61872054 and the Fundamental Research Funds for the Central Universities (DUT19LAB23). (Corresponding author: Feng Xia; E-mail: f.xia@ieee.org)

- *W. Wang is with the Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau, and State Key Laboratory of Internet of Things for Smart City, Faculty of Science and Technology, University of Macau, China.*
- *T. Tang and Z. Chen are with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China.*
- *F. Xia is with School of Engineering, IT and Physical Sciences, Federation University Australia, Ballarat, VIC 3353, Australia, and School of Software, Dalian University of Technology, Dalian 116620, China.*
- *Z. Gong is with State Key Laboratory of Internet of Things for Smart City, Faculty of Science and Technology, University of Macau, China.*
- *H. Liu is with School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281 USA.*

Manuscript received September 1, 2019.

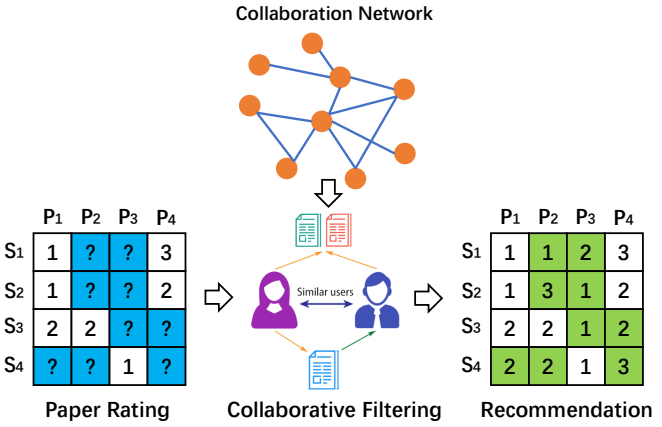


Fig. 1. Citation recommendation based on collaborative filtering. The authors are regarded as users and the scientific articles are regarded as items.

method with topic models for scientific article recommendation. However, the limitation of CTR is obvious because of the need of paper rating matrix provided by certain websites like CiteULike[1] and Mendeley[2]. Unfortunately, such information may not necessarily be available in practical citation recommendation scenario.

Other CF-based methods for citation recommendation utilize the rating matrices created from the citation network. CCF [6] and PCCF [8] propose to utilize local citation context to construct paper rating matrix to improve the accuracy of citation recommendation. However, local citation context may be sparse, causing incorrect recommendation. The whole network topology is neglected. Meanwhile, the data sparsity problem is also a challenging problem in CF-based citation recommendation.

Another limitation of previous CF-based citation recommendation is the neglect of the coauthorships among scholars. As shown in Figure 1, scholars are accomplished with a scientific collaboration network constructed based on coauthorships. Such coauthorship can be regarded as the social information of scholars. It has been proven in many social recommendation tasks that exploiting the existing social network information can enhance the performance of a recommendation system [9], [10], [11].

Recent advances in network representation learning (NRL) (or network embedding[3]) enable us to encode network topology into a low dimensional space [12], [13], [14]. In the learned vector space, similar nodes are closely related to each other. The effectiveness of NRL has been proven in many network-based tasks, such as the node classification and link prediction. It provides a new way to explore the whole citation network for creating the paper rating matrix. Until now, few works explore the potential of NRL in citation recommendation [15].

In this paper, we propose a **C**ollaborative filtering with **N**etwork representation learning framework for **C**itation **Rec**ommendation named **CNCRec**, which is designed based on the framework of user-based CF. It aims at recommending citations in heterogeneous academic information networks. To address the challenge of insufficient paper rating matrix, CNCRec creates the paper rating matrix based on attributed citation network representation learning, where the attributes are topics calculated based on paper content such as titles and abstracts. By harnessing the power of NRL, CNCRec is able to take advantages of the whole citation network topology. In order to learn citation representation, CNCRec formulates a joint optimization problem considering both the citation network and paper attribute.

To better integrate the coauthorships with the CF-based recommendation, the NRL technique is also used to select neighbor scholars in the procedure of CNCRec. Due to the sparsity of the citation network, the similarity between scholars is possibly less than 0, which will result in incorrect recommendations. CNCRec screens out neighbors whose related similarity is 0, and fulfills the neighbor list based on the most similar scholars by attributed scientific collaboration network representation learning.

Through extensive experiments on two scholarly datasets from two different disciplines, we demonstrate that our proposed CNCRec outperforms the state-of-the-art methods, with a 5.88% improvement in Precision@10, 15.13% improvement in Recall@10, and 7.18% improvement in MRR@10 over the best baseline method in DBLP dataset. Such improvements in APS dataset are 8.53% (Precision@10), 12.47% (Recall@10), and 5.1% (MRR@10), respectively. What is more, the outstanding performance of CNCRec on scholars with few publications demonstrates that CNCRec can better solve the data sparsity problem than other CF-based baselines.

Specifically, the contributions of this paper can be summarized as follows:

- We propose a novel method of creating paper rating matrix based on attributed citation network embedding to jointly consider both the network topology and text information.
- To solve the data sparsity problem in neighbor scholar selection, a novel neighbor scholar selection method is proposed based on the attributed collaboration network representation learning.
- We propose CNCRec, a novel citation recommendation algorithm, which is the first to combine the merits of collaborative filtering with network representation learning for citation recommendation.

The rest of the paper is organized as follows. We review the related work in Section 2. Section 3 provides some preliminaries on problem formulation and NRL. The details of CNCRec are presented in Section4. Section 5 introduces the experimental setup. The experimental results are discussed in Section 6. Finally, Section 7 concludes this paper.

2 Related Work

We introduce the related work from three aspects, including citation recommendation, network representation learning, and collaborative filtering.

1. http://www.citeulike.org
2. http://www.mendeley.com
3. We do not differentiate network representation learning and network embedding in this paper

2.1 Citation Recommendation

In the age of scholarly big data, scholars are producing unprecedented amounts of publications so that it is not easy for scholars to find related papers. Various citation recommendation or scientific article recommendation approaches have been proposed to tackle such a challenge. It is worth mentioning that we do not literally differentiate citation recommendation and scientific article recommendation. The reason is that the goal of scientific article recommendation is to recommend suitable and related papers to target scholars. Only if the commended articles are cited by the target scholars, such recommendation is helpful. Some citation recommendation systems are designed based on the query of specific content, e.g. a word, a phrase, a sentence, a manuscript, etc [16], [17]. Others are designed based on the academic information network extracted from reference management tools or websites like CiteUlike and Mendeley. Our goal is different from these methods since we aim to recommend citation for scholars in academic information networks extracted from online digital libraries such as DBLP and Aminer [18].

The strategies of citation recommendation can be mainly divided into three categories, including content-based filtering (CBF), network-based approaches, and collaborative filtering [2]. The central component of CBF methods is the scholar or paper profiling process, where the interests of scholars are inferred based on their publication content or profiles [4], [5], [19]. These approaches extract words from scholar's publication information such as the title, abstract, author-provided keywords or papers' body text. In order to find related citations for recommendation based on the extracted words, the similarity between scholars and candidate citations is calculated by vector space model or cosine similarity of paper topics. However, such content-based approaches overlook the social relationships among scholars as well as the citation relationships among papers.

Network-based citation recommendation approaches are designed based on graph model. They mainly adopt the academic information network to find potential citations with social network analysis methods [20]. The academic information network is constructed based on the paper citation relationships and scholar coauthorships. Gori and Pucci [21] construct a citation network based on bibliographic references, and apply random walk model to calculate the similarity between papers for recommendation. Meng et al. [22] propose a unified graph model which contains multi-layer graph modeling, topic modeling, and random walk for paper citation recommendation. Such graph model is able to make full use of various relationships (e.g., topic, collaboration, citation, etc.) in academic information networks for recommendation. However, network-based approaches suffer from the drawback that the time complexity on large graphs is very high.

CF-based methods mainly regard citation as items, and scholars as users [6], [7], [8]. However, scholars usually do not provide ratings to certain papers, which results in the main problem of CF-based approaches, the need of scholar participation. Thus, the key challenge of the CF-based method is to find a suitable paper rating matrix. Some alternative approaches have been proposed to create the paper rating matrix. For example, McNee et al. [23] propose to use the direct citations between papers to create the rating matrix. They design offline experiments against 186,000 papers in ResearchIndex and online experiments with over 120 users to evaluate the potentials of CF on citation recommendations. The experimental results show that CF can recommend quality citations. Following this work, Liu [6] et al. adopt the co-occurring with some citing papers for paper rating matrix calculation.

2.2 Network Representation Learning

NRL is an emerging topic in social network analysis as well as graph learning [12]. Specifically, NRL tries to encode network structural information with node representation. The primary goal of NRL is to learn a mapping function that represents nodes, or the entire network into a low dimensional space. In the low dimensional space, the structure of the original graph is presented based on the geometric relationships. Finally, the learning node vectors can provide inputs for other machine learning methods. NRL is a data-driven approach to automatically learn network embedding based on machine learning methods. The effectiveness of NRL has been proven in many classical tasks such as node classification and link prediction [24], [25]. Goyal and Ferrara [24] classify NRL methods into three kinds based on embedding techniques, including factorization methods, random walks, and deep learning. They further release GEM (Graph Embedding Methods) which is a python package for network representation learning. The GEM Python library is available at https://github.com/palash1992/GEM. Another famous open source toolkit for NRL is OpenNE, which is available at https://github.com/thunlp/openne. Based on the graph embedding input, Cai et al. [25] classify NRL methods into three kinds, including homogeneous network embedding, heterogeneous network embedding, and network embedding with auxiliary information.

Recently, NRL on different network types has been proposed. Tu et al. [26] propose CANE, which can learn network embedding in context-aware networks. Here, the context refers to the interactions with different neighbors. Li et al. propose DANE [27] to learn network representations in dynamic attributed networks. Liao et al. [25] propose ASNE, which can learn network representations considering social factors by preserving both structural proximity and attribute proximity. Dong et al. [28] propose Metapath2vec to learn network representations on heterogeneous networks. Kim et al. [29] propose SIDE to learn network representations on sighed directed networks. Actually, NRL is an emerging research hot topic in the fields of complex networks and artificial intelligence [30]. More recently, some attempts have been done on citation recommendation with paper vectors by learning representations of citation network [15], [31]. For example, Granguly and Pudi [15] propose Paper2vec which aims at learning scientific paper representations by combining graph and text information.

The utilization of network embedding has many advantages: 1) it can encode network topology into a low dimensional space so that the network scale issue can be solved; 2) it can better utilized external multi-source information, i.e., node attribute and link semantics; 3) the represented node

vector can be easily used as the input of other data mining tasks, i.e., in this paper, they are used to create paper rating matrix and fining neighbor scholars.

2.3 Collaborative Filtering

Another basis of our proposed citation recommendation algorithm is collaborative filtering. CF-based recommendation method is one of the most influential recommendation algorithms [30], [32], [33]. Unlike content-based approaches, CF merely uses the item rating assigned by users for recommendation. Its basic assumption is that users who rate the same items with similar ratings are likely to have similar preferences. CF outperforms content-based recommendations when the detailed information about users and items are unclear. Meanwhile, since CF recommends items according to users' neighbors where the content of items is excluded, it can recommend users with serendipitous items.

There are mainly two categories of CF, including memory-based CF and model-based CF [32]. Memory-based CF uses historical rating data to predict a specific user's rating on target items. Such a system needs to load all the data into memory to perform recommendation. It contains two categories, i.e., user-based CF and item-based CF. User-based CF calculates the similarities between target users and other users, and takes advantages of the neighbors' ratings on a specific item to obtain the predicted ratings for recommendations. Different form user-based CF, item-based CF tries to calculate the similarity between items. Model-based CF aims at utilizing machine learning or data mining techniques to construct an offline prediction model. Some typical model-based CF include matrix factorization-based methods and clustering-based methods. Such model-based CF can recommend suitable items with less time consumption compared with memory-based CF. Our proposed model belongs to user-based CF.

However, CF-based methods can hardly do efficient recommendations for new users or items which is the so-called data sparsity problem [34]. For example, in our case, if a junior scholar never cites any paper before, we can not extract his/her citing behavior so that the paper rating matrix cannot be constructed. Another challenge in designing a CF-based recommendation system is data sparsity.

3 Preliminaries

We first formulate the task of citation recommendation in academic information networks, and then briefly introduce the NRL technique, highlighting its potentials in benefiting CF-based recommendation methods.

3.1 Problem Formulation

Figure 2 illustrates the task of citation recommendation in academic information networks. Academic information networks are heterogeneous which contain, in our case, authors and papers. Let s and $\{S\}_{t=1}^{M_1}$ denote a scholar and the whole scholar set in a specific domain, respectively; Similarly, we use p and $\{P\}_{t=1}^{M_2}$ to denote a paper and the whole paper set, respectively. Meanwhile, as shown in this figure, each paper has its topics, which can be extracted from its content. Formally, we denote g and $\{G\}_{t=1}^{M_3}$ as a topic and the whole topic set, respectively.

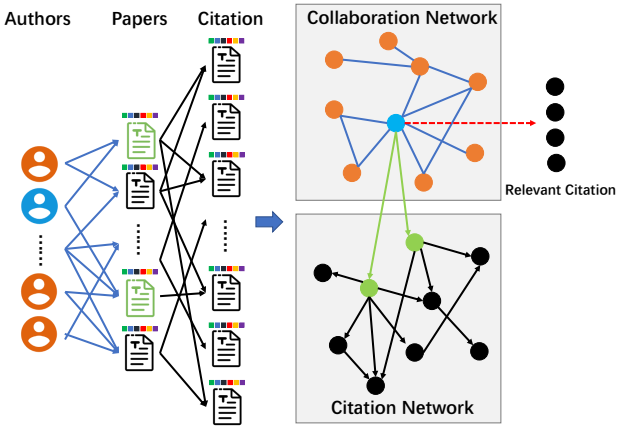


Fig. 2. Illustration of the citation recommendation task in academic information network.

TABLE 1
Description of Key Notations.

Symbol	Definiton
s	scholar s
p	paper p
g	topic g
N	number of recommended citations
k	number of nearest neighbors
L_1	collaboration network g
L_2	citation network g
$\|d\|$	dimension of embedding reuslt
M_1	number of scholars
M_2	number of papers
M_3	number of topics
$\mathbf{H} \in \mathbb{R}^{M_2 \times M_2}$	adjacency matrix of citation network
$\mathbf{A} \in \mathbb{R}^{M_2 \times M_2}$	citation attribute matrix
$\mathbf{R} \in \mathbb{R}^{M_2 \times \|d\|}$	final embedding representation
$\mathbf{Q} \in \mathbb{R}^{M_1 \times M_2}$	paper rating matrix
$sim(p_1, p_2)$	similarity between papers p_1 and p_2
$\omega(s, p)$	rating of scholar s on paper p in $\mathbf{Q}$
$Z(s, p)$	predicted rating of scholar s on paper p

Based on coauthorship, we can extract the scientific collaboration network $L_1 = (S, V_s, W_s)$, where V_s is the set of edges between scholars and W_s denotes the weight of edges. Similarly, based on citing behaviours, we can extract a citation network $L_2 = (P, V_p, W_p)$ where V_p is the set of edges between papers and W_p denotes the weight of edges. Initially, W_p is 1 since citation network is unweighted.

Based on the terms above, we define the citation recommendation in academic information network as follows:
Input: A set of scholars S associated with a collaboration network L_1 and a set of papers P with topics G associated with a citation network L_2.
Output: A list of related citations extracted from P for a target scholar s_i.

Notions: We use boldface lowercase alphabets (e.g., $\mathbf{r}$) to denote vectors and boldface uppercase $\mathbf{R}$ to denote matrices. The i_{th} row of a matrix $\mathbf{R}$ is denoted as $\mathbf{r}_i$. The transpose of $\mathbf{R}$ is $\mathbf{R}^T$. We use $\|\cdot\|_2$ which is the Euclidean norm of a vector to denote the ℓ_2 norm of a vector. The main symbols and definitions are shown in Table 1.

3.2 Network Representation Learning

CF is one of the most influential personalized recommendation methods. However, when applying CF method in citation recommendation, there is an obvious shortcoming that we do not have the paper rating data. To address this challenge, we utilize the NRL technique to generate paper rating matrix as well as to find similar scholars for junior scholars.

Given a network $G = (V, E)$, the goal of NRL is to lean a mapping function $f : V \rightarrow \mathbb{R}^{|d|}$, where $|d|$ is a parameter denoting the number of dimensions of feature learning. A low-dimension vector of each node in the network G can be gained based on f.

Usually, NRL relies on the network structure to represent the network [14]. For each source node u, we can obtain its network neighborhood $N_S(u) \subset V$ generated through a neighborhood sampling strategy, i.e., random walks in node2vec [14]. Inspired by the Skip-Gram [35], NRL seeks to optimize the objective function that aims at maximizing the log-probability of observing $N_S(u)$ for source node u on its feature representations, which can be described as:

$$\max_{f} \sum_{\alpha \in V} \log Pr(N_S(u)|f(u)). \quad (1)$$

Traditional social network analysis heavily relies on manual effort which is time-consuming and inflexible. With the mapping function f, the original network structure can be presented in a low-dimensional vector space. After optimizing, the learned node vectors can benefit many network analysis tasks, e.g., clustering, node classification, and link prediction.

4 Design of CNCRec

The goal of citation recommendation in academic information networks is to select relevant papers from the citation network. Under the paradigm of CF-based methods, the key to address this task is on how to find the paper rating matrix. The user-item rating matrix in CF-based methods for other tasks, i.e., movies recommendation can be easily obtained based on user-generated data. Although websites like CiteUlike and Mendeley allow scholars to build their local reference library with paper ratings [7], the scale of these paper ratings is too small to recommend citations in the era of scholarly big data. An alternative approach is to generate paper ratings with local citation context [6], [8], [19], [36]. However, the global network structure and paper content are neglected.

To address the challenge of insufficient paper rating matrix, we propose an NRL-based framework to calculate a scholar's rating to a given paper based on the attributed citation network embedding. Specifically, the attributes are the papers' topics. Meanwhile, for those scholars who have few citation information, we use NRL on scientific collaboration network to find his/her similar scholars.

As shown in Figure 3, our method CNCRec follows the standard procedure of memory-based CF [37] which contains five general procedures. These procedures are 1) paper rating matrix creation based on attributed citation network with NRL; 2) similarity calculation between scholars; 3) neighbor scholar selection based on NRL with scientific collaboration networks; 4) paper rating prediction; 5) Top-N citations selection and recommendation.

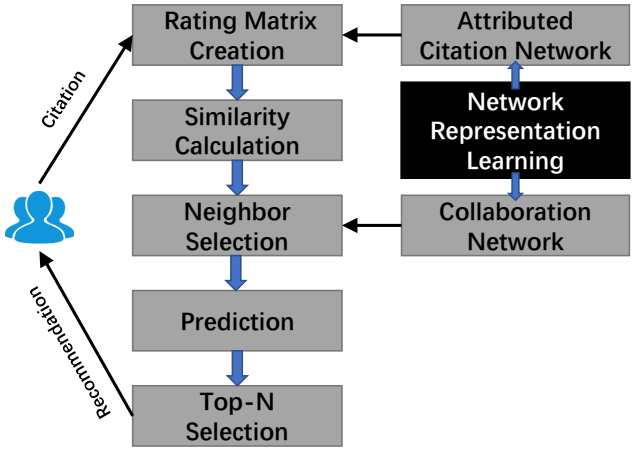


Fig. 3. Main procedure of CNCRec. CNCRec follows the standard procedure of memory-based CF. The NRL is utilized to improve the steps of rating matrix creation and nearest neighbor selection.

4.1 Rating Matrix Creation Via NRL

Inspired by previous work in CF-based citation recommendation [6], [36], [37], we use the citation information to generate paper ratings. Some researches take the number of a scholar citing the paper as the paper rating. For example, if a scholar cites a paper 2 times, the paper rating is 2. Other researches extend such method with auxiliary information, i.e., citation context [6] or potential citations [19], [37] due to the limited number of citations. On the one hand, such paper ratings are mainly calculated based on local network structure which cannot measure the correlations between the citing paper and cited paper. On the other hand, the text information of the paper is overlooked. We believe that it is necessary to generate the paper rating based on the similarity between citing and cited papers considering both the underlying network structure and paper content.

The idea of generating a scholar's rating to a given paper in CNCRec is shown in Figure 4. Given a citation network $L_2 = (P, V_p, W_p)$ with papers' raw data such as titles and abstracts, we first extract the topic information based on paper content with topic model. After performing the representation learning on the attributed citation network, we can obtain the paper vector matrix so that the similarity between papers can be calculated.

4.1.1 Paper Topic Calculation

Each paper has its topics. Although some digital libraries record the keywords information, we can hardly find out conceptually related papers merely based on word content alone, let alone the situation that many keywords information is missed in many digital libraries. The topic model is proposed to tackle such a problem. A topic model like Latent Dirichlet Allocation (LDA) [3] can help us to figure out the semantic content of papers so that a higher probability of correct match can be achieved.

We first extract the topics set $\{G\}_{t=1}^{M_3}$ of each paper with topic models based on their titles and abstracts. The goal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2020.3034976, IEEE Transactions on Big Data

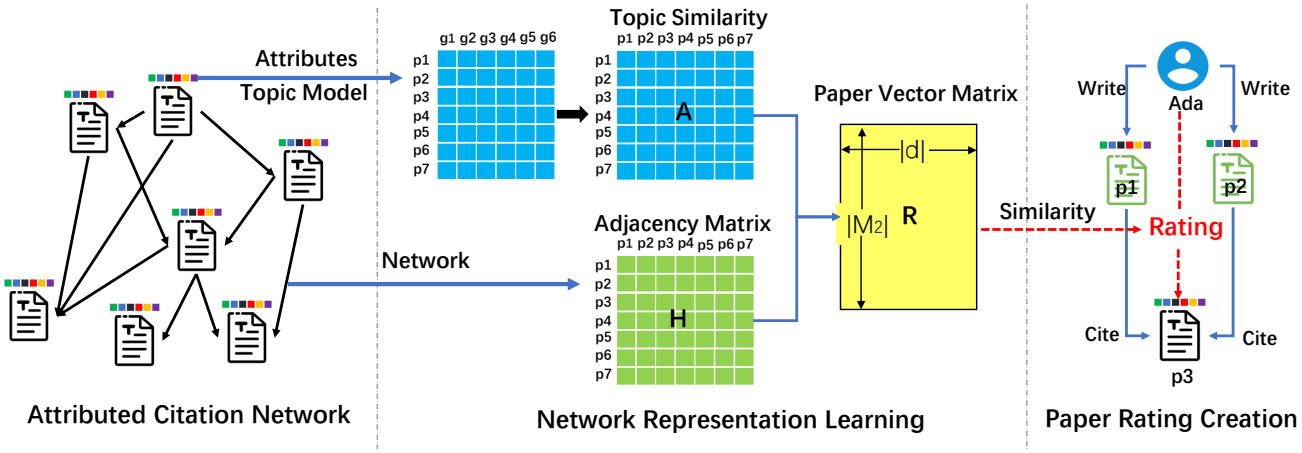


Fig. 4. Illustration of paper rating creation with attributed citation network representation learning.

of topic models is to provide a low-dimensional representation of documents over topics. They have been extensively utilized in tasks like document classification, information retrieval, and scholarly recommendation [7], [38].

Assume there are M_3 topics $g \in \{G\}_{t=1}^{M_3}$ in the paper set. Each topic is a distribution over a fixed vocabulary. Specifically, we adopt the simplest topic model, LDA [3]. For every paper p_j in the paper set $\{P\}_{t=1}^{M_2}$, the process of LDA is as follows:

- Draw topic proportions $\theta_j \sim$ Dirichlet(α)
- For each word n,
  (a) Draw topic assignment $z_{jn} \sim$ Mult(θ_j)
  (b) Draw word $w_{jn} \sim$ Mult($\beta_{z_{jn}}$)

Note that since the topic model is unsupervised, we do not know the specific topics of a given paper. In fact, we merely need the topic distribution of each paper over M_3 topics. Here, we take the topic information as the attributes of citation networks so that we can obtain an attributed citation network.

4.1.2 Attributed Citation Network Representation Learning

The goal of attributed citation network representation learning is to learn a low-dimensional paper vector. There have been various models aiming at representation learning attributed network, such as Attributed Network Embedding [39], Accelerated Attributed Network Embedding (AANE) [40], Context-Aware Network Embedding (CANE) [26], and Dynamic Attributed Network Embedding (DANE) [27].

In this work, CNCRec aims to learn paper vectors based on Attributed Citation Network Embedding (ACNE). In order to represent attributed citation network, we need to satisfy the following three requirements. First, it needs to be able to handle directed networks since citation is directed. Second, it should well preserve the paper proximity both in citation network and attribute (topic) space. Finally, it needs to be scalable since the number of paper M_2 and the dimension of paper attributes M_3 may be large.

The idea of ACNE is shown in the middle part of Figure 4. To make full use of attributed citation network, ACNE contains two parts of embedding for a paper p, i.e., citation network structure embedding and attribute proximity embedding. As shown in this figure, given the attributed citation network, ACNE first decomposes attribute similarity **A** into the final paper vector matrix **R**. Meanwhile, the decomposition procedure is controlled by an edge-based penalty that connected papers are close to each other in **R**. The edge-based penalty is determined by the citation edge in **H**. The ACNE generates the paper vector based on a joint representation learning method. Thus, the objective of ACNE can be described as follows,

$$L = L_A + L_H, \tag{2}$$

where L_A denotes the objective of attribute embedding and L_H denotes the objective of citation network embedding.

Network Structure Embedding. To preserve the paper proximity, ACNE learns the mapping function based on two hypotheses [40], [41]. First, papers with similar topological structures are more likely to be similar in the represented space. Second, a network-based mapping should be smooth for those regions with high network density [42]. To achieve these two goals, the loss function of citation network embedding is designed as follows,

$$L_H = \sum_{(i,j \in p)} w_{ij} \parallel \mathbf{r}_i - \mathbf{r}_j \parallel_2, \tag{3}$$

where row vector $\mathbf{r}_i$ and $\mathbf{r}_j$ are the vector representations of papers p_i and p_j, and w_{ij} is the link weight between two papers. The key idea is to minimize the penalty $w_{ij} \parallel \mathbf{r}_i - \mathbf{r}_j \parallel_2$. Following AANE [40], ACNE takes advantages of the ℓ_2 norm as the difference metric in order to address the negative impact of outliers and missing data. The proposed loss function is in line with the fused lasso and network lasso [43].

Attribute Proximity Embedding. Papers in citation networks usually accompany text information such as title and abstract in some digital libraries. By the paper topic calculation mentioned above, we obtain the attribute information of each paper. In traditional social network analysis, it has been proven that attribute information is highly associated with network topological structure [44]. Therefore, we need

to obtain the citation network representation $\mathbf{R}$ with reserving attribute proximity. In this paper, we take the topics as attributes. Following previous work on symmetric matrix factorization [45], we approximate paper attribute affinity matrix $\mathbf{A}$ with the product $\mathbf{R}$ and $\mathbf{R}^T$. The primary goal is to make the dot product of vector representation $\mathbf{r}_i$ and $\mathbf{r}_j$ the same as the corresponding paper attribute similarity $\mathbf{a}_{ij}$. Thus, the loss function of paper attribute proximity embedding can be described as:

$$\begin{aligned} L_A =& \| \mathbf{A} - \mathbf{R}\mathbf{R}^T \|_F^2 \\ =& \sum_{i=1}^{M_2}\sum_{j=1}^{M_2}(\mathbf{a}_{ij} - \mathbf{r}_i\mathbf{r}_j^T)^2, \end{aligned} \tag{4}$$

where M_2 is the number of papers. As shown in Figure 4, the paper affinity matrix $\mathbf{A}$ is calculated based on the cosine similarity of the paper topics.

Joint Representation Learning. Now we have the loss function L_A to model the paper attribute and loss function L_H to model the citation network. We can fulfill the Eq. (2) so that the attributed citation network embedding can be transformed into the following optimization problem:

$$\min_{\mathbf{R}} L = \|\mathbf{A} - \mathbf{R}\mathbf{R}^T\|_F^2 + \lambda \sum_{(i,j\in p)} w_{ij} \parallel \mathbf{r}_i - \mathbf{r}_j \parallel_2, \tag{5}$$

where parameter λ denotes a trade-off between the influence of citation network and paper attributes.

Based on Eq. (5), we can see that ACNE aims at maximizing several conditional probabilities between $\mathbf{A}$ and $\mathbf{R}$. Since it is computationally expensive to optimize the conditional probability using softmax function [26], [35], we employ the negative sampling strategy to transform the objective function. Afterward, the Adam strategy [46] is used to optimize the transformed objective function. Finally, we can get the final paper matrix $\mathbf{R}$.

4.1.3 Paper Rating Creation

As shown in the right part of Figure 4, we generate the paper rating based on the similarity between two papers with the low-dimensional paper matrix $\mathbf{R}$. The paper rating is calculated based on two factors, including the number of citing times and the similarity between citing and cited paper pair.

Specifically, if scholar s cites the paper p with his/her papers $p1, p2, ...p_N$, the rating $\omega_{(s,p)}$ scholar s giving to the paper p is calculated as:

$$\omega_{(s,p)} = \sum_{i=1}^{N} sim(p_i, p), \tag{6}$$

where $sim(p_i, p)$ is cosine similarity between $\mathbf{r}_{p_i}$ and $\mathbf{r}_p$ in the paper matrix $\mathbf{R}$. For example, in this figure, scholar Ada has two papers p_1 and p_2 citing the target paper p_3. Assume the similarity between p_1 (p_2) and p_3 in $\mathbf{R}$ is 0.7 (0.8). Thus, the rating of Ada to p_3 is 1.5. Finally, we can obtain the paper rating matrix $\mathbf{Q}$.

4.2 Similarity Calculation

CNCRec is a kind of user-based CF where the users are scholars and items are papers. Thus, CNCRec calculates the similarity between scholars. Based on the paper rating calculated in previous section, scholars can be represented by the row of paper rating matrix. The simplest way of calculating the similarity sim_{s_i,s_j} between two scholars is the cosine similarity:

$$sim_{s_i,s_j} = \cos(\mathbf{q}_{s_i}, \mathbf{q}_{s_j}) = \frac{\mathbf{q}_{s_i} \cdot \mathbf{q}_{s_j}}{\|\mathbf{q}_{s_i}\|_2 * \|\mathbf{q}_{s_j}\|_2}, \tag{7}$$

where $\mathbf{q}_{s_i}$ is the rating vector of scholar s_i and $\mathbf{q}_{s_j}$ is the rating vector of scholar s_j.

Since paper rating scales among scholars can be different, the same rating of two scholars does not mean the same degree of interest. This problem is not considered in simple cosine similarity. Therefore, we adopt the adjusted cosine similarity method [32]. Adjusted cosine similarity method subtracts the average rating of the scholars and utilizes co-cited papers to establish the vector. It can be defined as follows,

$$sim_{s_i,s_j} = \frac{\sum_{x\in X}(\mathbf{q}_{i,x} - \overline{\mathbf{q}_i})(\mathbf{q}_{j,x} - \overline{\mathbf{q}_j})}{\sqrt{\sum_{x\in X}(\mathbf{q}_{i,x} - \overline{\mathbf{q}_i})^2}\sqrt{\sum_{x\in X}(\mathbf{q}_{j,x} - \overline{\mathbf{q}_j})^2}}, \tag{8}$$

where X is the set of the co-cited papers, $q_{i,x}$ and $q_{j,x}$ are two paper ratings from scholars i and j, and $\overline{\mathbf{q}_i}$ and $\overline{\mathbf{q}_j}$ are the average ratings of scholars i and j.

In the context of scholarly big data, there are millions of papers. The number of paper ratings of a scholar is relative small. Therefore, we adopt the set-based similarity [47] to alleviate such problem. Given the cosine similarity sim_{s_i,s_j}, the set-based similarity sim'_{s_i,s_j} between scholars s_i and s_j is calculated as:

$$sim'_{s_i,s_j} = \frac{2|Y_{s_i} \cap Y_{s_j}|}{|Y_{s_i}| + |Y_{s_j}|} sim_{s_i,s_j}, \tag{9}$$

where $|Y_{s_i}|$ and $|Y_{s_j}|$ are the numbers of papers rated by scholar s_i and scholar s_j, respectively.

4.3 Neighbor Scholar Selection

The citation predictive accuracy is highly influenced by the selected neighbors. If some dissimilar scholars are involved into the neighbor, the accuracy will be reduced. Traditional Top-k methods select k most similar neighbors. However, some scholars may have a limited number of neighbors less than k. Therefore, we propose to remove those neighbors with similarity less than 0 and fulfill the Top-k neighbors with the most similar scholars in scientific collaboration network.

As shown in Figure 2, we can extract a scientific collaboration network L_1 from the heterogeneous academic information network. In the context of scientific collaboration network, scholars also have various academic attributes so that the collaboration network is attributed. It is necessary to consider those attributes for similarity calculation. It has been proven that attributes such as academic age and reputation may have a great influence on academic relation mining [48].

While previous researches mainly consider network topology, we want to utilize various scholars' attribute for similarity calculation. Specifically, we consider four types of scholars' attributes including demographics, research,

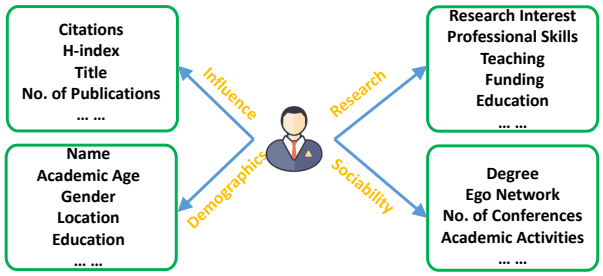


Fig. 5. Illustration of four types of scholar attributes.

influence, and sociability, as shown in Figure 5. We believe these attributes are more complete and systematic so that better accuracy can be achieved. The details of these four types of attributes are given as follows:

- **Demographics:** It has been proven that users with demographic profiles in social networks bring new insights into understanding social principles from individuals, to groups, and to societies. The demographic characteristics of scholars are of various kinds. Some popular demographics are gender, academic age, location, nationality, etc.
- **Research:** The research attributes denote the information related with scholar's studies. This kind of scholar attribute is widely considered in previous works as the research topics are provided by some popular scholarly datasets. Meanwhile, the advances of natural language processing enable us to extract scholar's research topic distribution via topic models, e.g., LDA.
- **Influence:** The influence attribute refers to the indicators denoting a scholar's academic achievement and impact. This important attribute is always overlooked in academic relationship mining tasks. In reality, junior scholars are more likely to be pursuers while senior scholars with high academic reputation are normally attractors when facing new collaborative opportunities. The influence attribute includes citations, the number of publication, h-index, academic title, etc.
- **Sociability:** The sociability attribute refers to the collaboration patterns of scholars. In modern academia, some scholars are more collaborative than others. It has been proven in many studies that collaborative scholars are more productive and influential. On the one hand, some network indicators can reflect the sociability attribute such as degree, ego network, and clustering coefficients. On the other hand, scholars' sociability is hidden in involving academic activities such as conference attending.

After the construction of attributed academic collaboration networks, we utilize the NRL technique to obtain the vector representation of scholars. Specifically, the procedure of attributed collaboration network representation learning is similar to the attributed citation network embedding. They have a different attribute matrix. For the construction of scholars' attribute matrix, we adopt the method in ACNE [49], which categorizes attributes into two kinds, including discrete attributes and continuous attributes. Finally, based on the scholars' represented low-dimensional vectors, we can calculate the similarity between scholars. It is worth mentioning that due to the limitation of the scholarly datasets, we may not obtain all the proposed attribute.

4.4 Paper Rating Prediction

With the neighbor scholars' similarity and paper ratings, CNCRec calculates the weighted average as the prediction. Specifically, the predicted paper rating of scholar s on paper p is calculated as follows,

$$Z(s,p) = \overline{\mathbf{q}_s} + \frac{\sum_{s_n \in B(s)} sim'_{s,s_n}(\mathbf{q}_{s_n,p} - \overline{\mathbf{q}_{s_n}})}{\sum_{s_n \in B(s)} sim'_{s,s_n}}, \quad (10)$$

where B_s is the set of neighbors of scholar s, $q_{s_n,k}$ denotes the scholar s_n's rating to paper p, and sim'_{s,s_n} is the adjusted similarity between scholar s and his/her neighbor s_n.

4.5 Top-N Citation Recommendation

Once the paper rating predictions are obtained, CNCRec needs to rank all the citations according to their predicted ratings. After ranking all the candidate citations, Top-N of them are recommended to the scholar where N is a parameter needed to be preset based on specific goals before the citation recommendation task.

5 Experimental Setup

5.1 Research Questions

The research questions we want to investigate are listed as follows:

- **RQ1:** How do the parameters, including $|d|$ and λ in attributed citation network representation learning affect the performance of CNCRec?
- **RQ2:** What is the performance of CNCRec in citation recommendation? Does it outperform state-of-the-art methods?
- **RQ3:** Can CNCRec address the cold start problem? Does it outperform other CF-based methods in recommending citations to scholars with few publications?

We perform extensive experiments on two scholarly datasets to investigate these three questions.

5.2 Datasets

TABLE 2
Statistics of two datasets

Datasets	DBLP	APS
Number of papers	143,358	78,391
Number of authors	140,781	80,234
Number of citations	3.1M	2.1M
paper average citations	5.78	6.72

We use two different scholarly datasets: the DBLP dataset in the field of Computer Science and APS dataset

in the field of Physics. For the DBLP dataset, we obtain the citation relationships from the AMiner project by Tang *et al.* [18]. Following the setup in [20], we screen out those papers with incomplete information or less than 5 citations. For the APS dataset, we first do the name disambiguation based on the method in [50]. Then, a subset of APS dataset is generated by the same method with DBLP dataset. According to the network schema in Figure 2, we convert these two datasets into academic information networks. The statistics of these two datasets as shown in Table 2.

The topic number M_3 is generated based on LDA model. Specifically, we first extract each paper's titles and abstracts. Then, we remove those words which consist of two characters or less. Meanwhile, we remove stopwords based on the stopwords list from https://github.com/stopwords-iso/stopwords-en. Based on these two filtering steps, we can obtain the most meaningful words for topic calculation. Finally, we stem the remaining words with porter stemmer. To reduce noise, we also remove those words that appear less than ten times in each dataset. The topic number M_3 is set as 100 for both datasets.

For the construction of attributed collaboration network, we can not obtain all the proposed attributes merely based on the information in two datasets. Specifically, the demographical attribute we use is the academic age which can be calculated by the investigated year minus the year of first publication. The research attributes are the topic distributions by performing LDA on scholars' title and abstract information. The topic number is also 100. The utilized influence attributes are the number of citations and number of publications. The sociability attribute is clustering coefficient [48].

In order to get the ground truth of recommendation results, we split the dataset into two subsets, including training set and testing set based on the paper publication year. We consider two time intervals $\tau_1 = [2001, 2010]$ and $\tau_2 = [2011, 2015]$. Papers in τ_1 are used for parameter tuning and paper rating prediction, and papers in τ_2 are used for testing and evaluation. Specifically, we assume citations from τ_2 to papers τ_1 as the ground truth. Such method evaluates the performance of citation recommendation systems based on realistic citing behaviours, and has been extensively used [50].

5.3 Baseline Methods

We compare our proposed method CNCRec with its variation CNCRec- which does not consider collaboration network embedding for neighbor scholar selection. Meanwhile, we conduct several baselines or state-of-the-art citation recommendation approaches for comparison, including content-based methods, CF-based methods, and NRL-based methods.

Content-based:

- **TopicSim:** TopcSim recommends citation to scholars based on the topic similarity between papers. The topic similarity is calculated with the simplest topic model LDA.
- **CTR:** CTR [7] considers both content and other scholars' ratings with the probabilistic topic model for citation recommendation.
- **ClusCite:** ClusCite [20] is a cluster-based citation recommendation algorithm. Recommended candidates are ranked based on paper authority to different research groups.
- **NCN:** NCN [51] uses a flexible encoder-decoder architecture for citation recommendation, which embodies a robust representation of the citation context with a max time delay neural network, further augments with an attention mechanism and author networks.

CF-based:

- **CF:** CF is the typical user-based CF method which recommends citation based on the rating matrix which is extracted directly from the citation network.
- **CCF:** CCF [6] is a context-based CF method. It uses the co-occurrence relationships between citing papers to improve the paper rating matrix.
- **PCCF:** PCCF [8] is a CF method considering potential citation papers for rating matrix construction.

NRL-based:

- **Paper2vec:** Candidates in Paper2vec [15] are ranked based on the similarity calculated based on paper vectors. Paper vectors are learned based on combining text and network embedding.
- **DeepWalk:** DeepWalk [52] employs truncated random walks on the citation network with language modeling techniques for learning paper representations.
- **TADW:** TADW [53] is a matrix factorization based approach which considers both the network structure and text information for learning paper representations.

Proposed:

- **CNCRec:** Candidates are ranked based on Eq. 10.
- **CNCRec-:** CNCRec- is a variation fo CNCRec which does not consider attributed collaboration network representation learning in selecting nearest neighbor scholars.

5.4 Evaluation Metrics

We adopt three widely used evaluation metrics in recommendation systems, including the Precison@N, Recall@N, and MRR@N. Precision@N measures the performance of citation recommendation by checking whether the ground truth citations are ranked in the recommended citations, which can be calculated as,

$$P@N = \frac{\#\ of\ true\ citations\ in\ the\ Top-N\ list}{\#\ of\ N}. \quad (11)$$

Recall@N is calculated as the ratio of real citing papers appearing in the Top-N recommendation list, which is defined as,

$$R@N = \frac{\#\ of\ true\ citations\ in\ the\ Top-N\ list}{total\ \#\ of\ new\ citations}. \quad (12)$$

MRR@N is the Mean Reciprocal Rank (MRR), which is calculated as:

$$MRR = \frac{1}{|Y|}\sum_{y\in Y}\frac{1}{rank(y)}, \quad (13)$$

TABLE 3
Parameter sensitivity of parameter |d|.

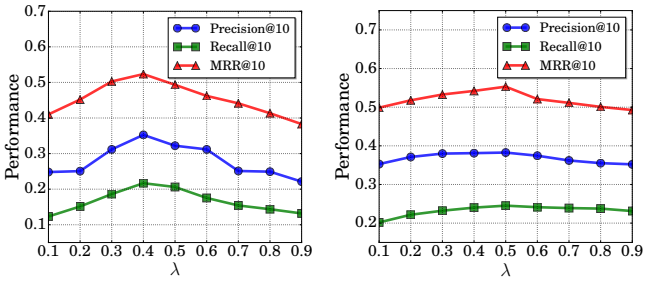
|d|	DBLP (%)			APS (%)		
	P@10	R@10	M@10	P@10	R@10	M@10
100	34.97	21.04	51.98	30.07	24.24	54.98
200	35.12	21.34	52.22	38.14	24.36	55.22
300	35.21	21.57	52.23	**38.26**	**24.52**	**55.31**
400	**35.24**	**21.67**	**52.33**	38.27	24.52	55.31
500	35.25	21.68	52.34	38.29	24.54	55.32

where Y is the testing set and $ran(y)$ is the rank of its first correct citation (positive sample). M@N is used to measure how early the ground truth citations appear in the Top-N recommended list. These metrics are widely adopted for many citation recommendation systems' evaluation, such as [6], [7], [8].

6 Results and Discussions

In this section, we present the experimental results in details by answering the previous three research questions.

6.1 Parameter Sensitivity (RQ1)



(a) DBLP (b) APS

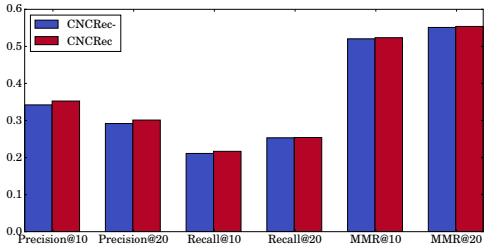
Fig. 6. Parameter sensitivity of λ.

There are two important parameters in learning paper representations: the dimension of paper vectors |d|, and λ in joint learning as shown in Eq. 5. These two parameters may have a great influence on the performance of CNCRec. Here, we investigate the influence of these two parameters and tune them for the following experiments.

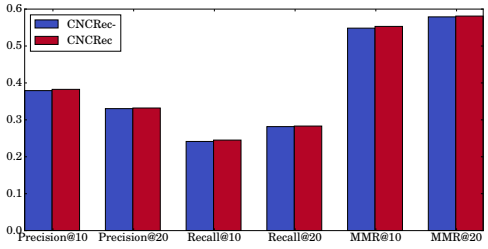
The size of the represented dimension |d| determines how large the paper matrix is. The impact of parameter |d| is shown in Table 3. In this table, we present the results with changing |d| on both DBLP and APS datasets in terms of Precision@10, Recall@10, and MRR@10, respectively. Here, we set λ as 0.4 and let |d| vary on the range [100, 200, 300, 400, 500]. We can see from this table that with the increase of |d|, the performance of CNCRec slightly increases in both datasets. Although a higher |d| may bring about better performance, such improvement is not obvious. Specifically, the accuracy on DBLP tends to saturate at 400 and the accuracy on APS tends to saturate at 300. Therefore, we set |d| as 400 and 300 for DBLP and APS, respectively in the following experiments.

Parameter λ determines the significance of papers' attributes. We plot Figure 6 to illustrate the influence of parameter λ. Specifically, we evaluate the performance of CNCRec with changing λ on DBLP and APS datasets in terms of Precision@10, Recall@10, and MRR@10, respectively. It can be seen from this figure that both DBLP and APS have similar accuracy trend that curves first increase and slightly decrease after reaching the peak. The peak for DBLP and APS is 0.4 and 0.5, respectively.

Such an observation denotes that attribute plays an important role in learning paper representations for creating paper rating matrix. Specifically, with the increase of λ, the performance of CNCRec slightly goes up. However, there is a balance between attributes and network topology. If λ is too large, the graphs show that there has been an according decrease. Therefore, we need to both consider nodes' attributes and network topology for designing academic recommendation systems. Based on Table 3 and Figure 6, we set |d| as 400 and 300, λ as 0.4 and 0.5, for DBLP and APS, respectively in the following experiments.



(a) DBLP



(b) APS

Fig. 7. Comparisons between CNCRec and its variation CNCRec- on DBLP and APS in terms of Precision@N, Recall@N, and MRR@N.

6.2 Performance and Comparison (RQ2)

In this section, we first compare CNCRec with other state-of-the-art citation recommendation methods in terms of precision, recall, and MRR.

We first compare CRCRec with nine different baselines on DBLP datasets using Precision@{5, 10, 20}, Recall@1{5, 10, 20}, and MRR@{5, 10, 20}. Here, the number of N is relatively small because recommending scholars with a large number of potential citations is meaningless. The comparison results are shown in Table 4. We can observe from this table that:

- The overall performance of our proposed method CNCRec is better than other methods over all three metrics. Notably, by comparison, CNCRec achieves a 5.88% improvement in Precision@20, 15.13% improvement in Recall@10, and 7.18% improvement in

TABLE 4
Recommendation performance comparisons on DBLP datasets in terms of Precision@N, Recall@N, and MRR@N.

DBLP	Precision@N			Recall@N			MRR@N		
	5	10	20	5	10	20	5	10	20
CTR	0.3425	0.3384	0.2832	0.1702	0.1828	0.2493	0.4987	0.5112	0.5119
TopicSim	0.2742	0.2521	0.2013	0.1463	0.1522	0.1934	0.4481	0.4421	0.4726
ClusCite	0.3407	0.3317	0.2813	0.1721	0.2022	0.2411	0.4834	0.5056	0.5157
CF	0.2578	0.2481	0.1932	0.1422	0.1623	0.2027	0.4423	0.4521	0.4832
CCF	0.3321	0.3112	0.2701	0.1678	0.2077	0.2468	0.4828	0.4921	0.5018
PCCF	0.3217	0.2983	0.2508	0.1638	0.1837	0.2235	0.4781	0.4838	0.5167
Paper2vec	0.3523	0.3402	0.2843	0.1728	0.1883	0.2428	0.5012	0.5123	0.5189
DeepWalk	0.2894	0.2522	0.2122	0.1529	0.1721	0.2115	0.4529	0.4552	0.4829
TADW	0.3087	0.2911	0.2438	0.1581	0.1812	0.2231	0.4622	0.4621	0.4937
NCN	0.3462	0.3245	0.2783	0.1783	0.2051	0.2443	0.4932	0.5034	**0.5541**
CNCRec	**0.3831**	**0.3525**	**0.3012**	**0.1932**	**0.2168**	**0.2541**	**0.5123**	**0.5234**	0.5538

TABLE 5
Recommendation performance comparisons on APS datasets in terms of Precision@N, Recall@N, and MRR@N.

APS	Precision@N			Recall@N			MRR@N		
	5	10	20	5	10	20	5	10	20
CTR	0.3837	0.3525	0.3122	0.2137	0.2408	0.2518	0.5207	0.5398	0.5529
TopicSim	0.3348	0.3174	0.2738	0.1803	0.1832	0.2273	0.4818	0.5103	0.5467
ClusCite	0.3806	0.3521	0.3067	0.2054	0.2389	0.2552	0.5173	0..5402	0.5471
CF	0.3214	0.3011	0.2548	0.1783	0.1927	0.2262		0.5056	0.5348
CCF	0.3765	0.3421	0.2981	0.2008	0.2371	0.2449	0.5124	0.5401	0.5431
PCCF	0.3674	0.3324	0.2863	0.1965	0.2247	0.2564	0.5039	0.5328	0.5446
Paper2vec	0.3892	0.3511	0.3271	0.2156	0.2421	0.2508	0.5208	0.5411	0.5511
DeepWalk	0.3418	0.3218	0.2783	0.1845	0.2085	0.2461	0.4824	0.5189	0.5408
TADW	0.3589	0.3428	0.2978	0.1922	0.2102	0.2478	0.4943	0.5212	0.5532
NCN	0.3773	0.3498	0.3044	0.2091	0.2377	0.2562	0.5272	**0.5537**	0.5497
CNCRec	**0.4122**	**0.3826**	**0.3321**	**0.2241**	**0.2452**	**0.2832**	**0.5378**	0.5531	**0.5811**

MRR@20 compared with the best baseline methods in DBLP dataset. This indicates that it is helpful to employ NRL technique for paper rating matrix creation in designing citation recommendation, which is in line with previous research [15].

- Two basic recommendation methods, basic CF and TopicSim have the worst performance among all methods. This indicates that hybrid recommendation systems may have a better recommendation performance. Either merely adopting CF without text information or merely considering topic similarity without CF will lead to worse performance. It is worth mentioning that Paper2vec achieves the second best performance. The reason is that, Paper2vec considers both text information and network topology in calculating paper similarity. However, it does not utilize the ability of CF so that its performance is slightly worse than CNCRec.
- What can be clearly seen in this table is the steady decrease of Precision@N with the increase of N. This is because of the definition of Precision@N. With the increase of N, more papers are recommended, whereas the number of true citation is stable. Different observations on Recall@N and MRR@N can be founded. With the increase of N, both Recall@N and MRR@N go down accordingly.

Table 5 depicts the comparison results on APS dataset. We can see from this table that CNCRec also achieves the best performance among all the recommendation approaches. The improvements compared with second best approach are 8.53% in Precision@10, 12.47% in Recall@10, and 5.1% in MRR@20, respectively. Meanwhile, hybrid recommendation methods such as, Paper2vec and CTR have a better performance compared with simple approaches such as, CF and TopicSim. These also indicates the effectiveness of our proposed method.

Meanwhile, By comparing the results on two datasets, we can find that almost all recommendation methods have a better performance on APS dataset than that on DBLP dataset. The reason is that, as shown in Table 2, the citation network of APS dataset is denser than that of DBLP dataset. The average number of citations of each paper is 6.72 in APS, compared with 5.78 in DBLP. This indicates that a citation network will benefit the citation recommendation systems.

In the step of nearest neighbor selection, we utilize NRL on attributed collaboration networks for scholar similarity calculation. In order to evaluate the effectiveness of this assumption, we compare CNCRec with its variation CNCRec-. Here, CNCRec- does not consider attributed collaboration network representation learning in selecting nearest neighbor scholars. The comparison results on both datasets in terms of Precision@N, Recall@N, and MRR@N are shown in Figure 7.

We can see from Figure 7 that the performance of CNCRec- on both DBLP and APS datasets is slightly worse than that of CNCRec. This implies that considering collaboration network representation learning in selecting neighbor scholars is useful. Although such improvement is relatively small, it exists in both datasets over all evaluation metrics. Meanwhile, we can also find that both CNCRec and CNCRec- have a better performance on APS datasets than DBLP datasets.

Based on the observations on Tables 4 and 5, and Figure 7, we can obtain the conclusions that: 1) By applying NRL with CF, better citation recommendation systems can be designed. The utilization of NRL for creating paper rating matrix is helpful; 2) The hybrid recommendation approaches, such as CNCRec, Paper2vec, CCF, and CTR always perform better than single recommendation approaches, such as CF, TopicSim, and DeepWalk. In particular, the performance of CF and TopicSim, and DeepWalk is worst. Such phenomenon tells that it is necessary to adopt auxiliary information in designing citation recommendation systems.

6.3 Data Sparsity Performance (RQ3)

One of the biggest advantages of CNCRec is using NRL technique to generate paper rating matrix. With the help of ACNE, citations can be better represented by considering network topology and content information. Meanwhile, in CNCRec, the NRL technique is employed in finding neighbor scholars for those target scholars who have few papers. We conduct NRL with attributed collaboration networks so that we can select better nearest neighbors. Therefore, we believe that CNCRec can better solve the data sparsity problem in CF-based recommendation.

In order to evaluate the performance of our proposed methods, we compare the performance of CNCRec and CNCRec- with other CF-based methods on scholars with different publication counts. Specifically, we divide target scholars into groups based on their number of publications. In this case, the data sparsity problem refers to recommending citation to scholars with few publications.

Figure 8 shows the comparing results on scholars with different paper counts in terms of Precision@10, Recall@10, and MRR@10 with DBLP dataset, respectively. Figure 9 shows the results on APS dataset. We can observe from these two figures that:

- These two figures show that there has been a gradual rise with the increase of scholars' publication counts. Specifically, for scholars with 5 publications, the Precision@10 of CNCRec, CNCRec-, CF, CCF, and PCCF on DBLP datasets is 0.25, 0.16, 0.12, 0.13, and 0.14, respectively. For scholars with 30 publications, the result is 0.35, 0.34, 0.25, 0.31, and 0.29, respectively. This indicates that all citation recommendation systems can better recommend citation for scholars with more publications.
- Overall, both CNCRec and CNCRec- outperform other CF-based methods. This observation also indicates the effectiveness of our proposed methods. What stands out in these two figures is the marked improvement of CNCRec in recommending citations to scholars with less than 5 papers. This demonstrates that CNCRec can better solve the data sparsity problem.
- Although CNCRec- performs better than baselines on scholars with less than 5 papers, such improvement is slight. Meanwhile, its performance is much worse than that of CNCRec. These indicate that considering collaboration network representation learning in selecting neighbor scholars can be beneficial in solving the data sparsity problem.

7 Conclusion

To design an effective CF-based citation recommendation system, it is crucial to account for both citation network structure and paper attribute information. Meanwhile, the missing of practical paper ratings makes CF-based citation recommendation a difficult task. To this end, we propose a citation recommendation framework named CNCRec which combines the merits of collaborative filtering and network representation learning. The NRL technique is utilized to promote collaborative filtering by creating paper rating matrix with attributed citation network embedding and selecting neighbor scholars with attributed scientific collaboration network embedding. We formulate a joint optimization problem to learn paper representations considering both citation network and paper topics. For calculation similarity between scholars, we propose to construct an attributed collaboration network based on four types of scholar attributes. Extensive results on two scholarly datasets show a significant improvement compared to state-of-the-art citation recommendation methods. Meanwhile, our proposed method can better solve the data sparsity problem.

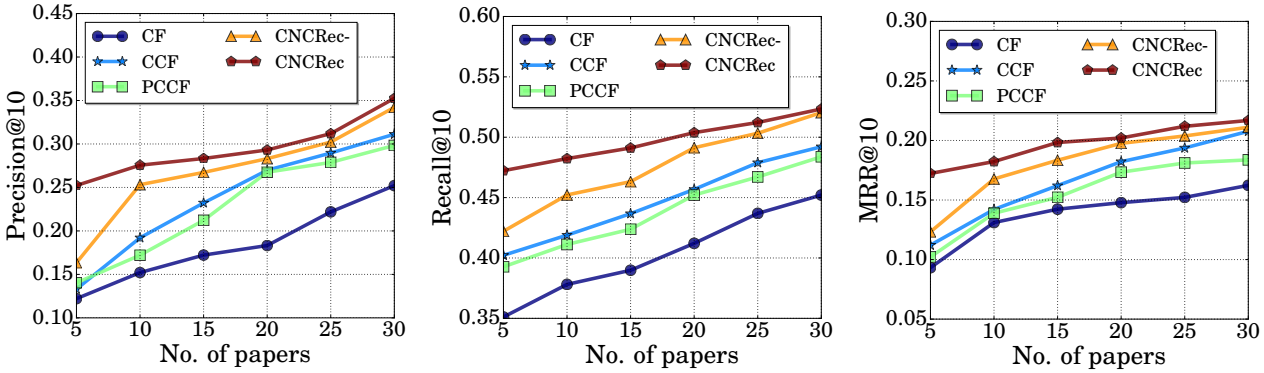
The findings reported here shed new light on other scholarly recommendation systems such as collaborator recommendation. Although this study focuses on citation recommendation, the framework of CNCRec may well have a bearing on other CF-based recommendation tasks. Our future work in this area will focus on investigating the following questions: 1) How to learn paper vectors from attributed citation network in a dynamic environment because citation network is dynamic with new papers appearing; 2) Will our proposed method has a good performance in other user-based item recommendation scenarios?

Acknowledgments

We would like to thank Tianyi Hu and Mengyi Mao for their help with experiments.

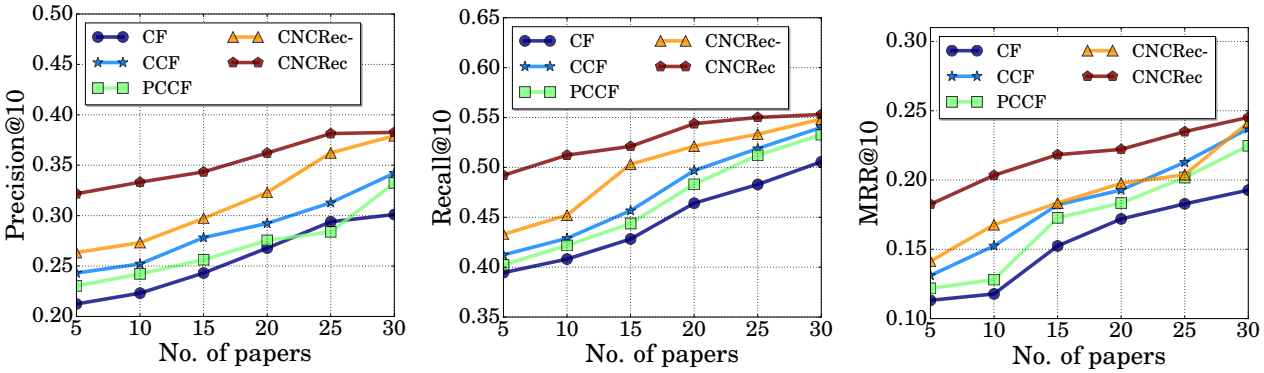
References

[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.



(a) Precision@10 (b) Recall@10 (c) MRR@10

Fig. 8. Comparison between CNCRec and baseline CF-based methods in terms of Precision@10, Recall@10, and MRR@10 over scholars with different paper counts on DBLP dataset. We set |d| as 400, and λ as 0.4.



(a) Precision@10 (b) Recall@10 (c) MRR@10

Fig. 9. Comparison between CNCRec and baseline CF-based methods in terms of Precision@10, Recall@10, and MRR@10 over scholars with different paper counts on APS dataset. We set |d| as 300, and λ as 0.5.

[2] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. 1, pp. 993–1022, 2003.

[4] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proceedings of the 19th international conference on world wide web*, 2010, pp. 421–430.

[5] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach, "Recommending citations: translating papers into references," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 1910–1914.

[6] H. Liu, X. Kong, X. Bai, W. Wang, T. M. Bekele, and F. Xia, "Context-based collaborative filtering for citation recommendation," *IEEE Access*, vol. 3, pp. 1695–1703, 2015.

[7] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, 2011, pp. 448–456.

[8] K. Sugiyama and M.-Y. Kan, "A comprehensive evaluation of scholarly paper recommendation using potential citation papers," *International Journal on Digital Libraries*, vol. 16, no. 2, pp. 91–109, 2015.

[9] C. Hsu, M. Yeh, and S. Lin, "A general framework for implicit and explicit social recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2228–2241, Dec 2018.

[10] L. Guo, X. Cai, F. Hao, D. Mu, C. Fang, and L. Yang, "Exploiting fine-grained co-authorship for personalized citation recommendation," *IEEE Access*, vol. 5, pp. 12 714–12 725, 2017.

[11] 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, 2019.

[12] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *arXiv preprint arXiv:1801.05852*, 2017.

[13] C. Shi, B. Hu, W. X. Zhao, and S. Y. Philip, "Heterogeneous information network embedding for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357–370, 2019.

[14] 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*, 2016, pp. 855–864.

[15] S. Ganguly and V. Pudi, "Paper2vec: combining graph and text information for scientific paper representation," in *European Conference on Information Retrieval*, 2017, pp. 383–395.

[16] T. Dai, T. Gao, L. Zhu, X. Cai, and S. Pan, "Low-rank and sparse matrix factorization for scientific paper recommendation in heterogeneous network," *IEEE Access*, vol. 6, pp. 59 015–59 030, 2018.

[17] L. Yang, Y. Zheng, X. Cai, H. Dai, D. Mu, L. Guo, and T. Dai, "A lstm based model for personalized context-aware citation recommendation," *IEEE Access*, vol. 6, pp. 59 618–59 627, 2018.

[18] 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*, 2008, pp. 990–998.

[19] K. Sugiyama and M.-Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, 2013, pp. 153–162.

[20] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han, "Cluscite: Effective citation recommendation by information network-based clustering," in *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, 2014, pp. 821–830.

[21] M. Gori and A. Pucci, "Research paper recommender systems:

A random-walk based approach," in *IEEE/WIC/ACM International Conference on Web Intelligence*, 2006, pp. 778–781.

[22] F. Meng, D. Gao, W. Li, X. Sun, and Y. Hou, "A unified graph model for personalized query-oriented reference paper recommendation," in *Proceedings of the 22nd ACM international Conference on Information and Knowledge Management*, 2013, pp. 1509–1512.

[23] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl, "On the recommending of citations for research papers," in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, 2002, pp. 116–125.

[24] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," *Knowledge-Based Systems*, vol. 151, pp. 78–94, 2018.

[25] H. Cai, V. W. Zheng, and K. C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1616–1637, Sept 2018.

[26] C. Tu, H. Liu, Z. Liu, and M. Sun, "Cane: Context-aware network embedding for relation modeling," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2017, pp. 1722–1731.

[27] J. Li, H. Dani, X. Hu, J. Tang, Y. Chang, and H. Liu, "Attributed network embedding for learning in a dynamic environment," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 387–396.

[28] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 135–144.

[29] J. Kim, H. Park, J.-E. Lee, and U. Kang, "Side: representation learning in signed directed networks," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018, pp. 509–518.

[30] J. Liu, X. Kong, F. Xia, X. Bai, L. Wang, Q. Qing, and I. Lee, "Artificial intelligence in the 21st century," *IEEE Access*, vol. 6, pp. 34 403–34 421, 2018.

[31] H. Tian and H. H. Zhuo, "Paper2vec: Citation-context based document distributed representation for scholar recommendation," *arXiv preprint arXiv:1703.06587*, 2017.

[32] Z. Yang, B. Wu, K. Zheng, X. Wang, and L. Lei, "A survey of collaborative filtering-based recommender systems for mobile internet applications," *IEEE Access*, vol. 4, pp. 3273–3287, 2016.

[33] W. Wang, J. Liu, Z. Yang, X. Kong, and F. Xia, "Sustainable collaborator recommendation based on conference closure," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 311–322, April 2019.

[34] J. Lin, K. Sugiyama, M.-Y. Kan, and T.-S. Chua, "Addressing cold-start in app recommendation: latent user models constructed from twitter followers," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 2013, pp. 283–292.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[36] K. Haruna, M. A. Ismail, D. Damiasih, J. Sutopo, and T. Herawan, "A collaborative approach for research paper recommender system," *PloS one*, vol. 12, no. 10, p. e0184516, 2017.

[37] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer Communications*, vol. 41, pp. 1–10, 2014.

[38] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai, "Jointly learning word embeddings and latent topics," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 375–384.

[39] S. Wang, C. Aggarwal, J. Tang, and H. Liu, "Attributed signed network embedding," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 137–146.

[40] X. Huang, J. Li, and X. Hu, "Accelerated attributed network embedding," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, 2017, pp. 633–641.

[41] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert, "Clusterpath an algorithm for clustering using convex fusion penalties," in *28th international conference on machine learning*, 2011, p. 1.

[42] A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan, "Asymptotic behavior of \ell_p-based laplacian regularization in semi-supervised learning," in *Conference on Learning Theory*, 2016, pp. 879–906.

[43] D. Hallac, J. Leskovec, and S. Boyd, "Network lasso: Clustering and optimization in large graphs," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 387–396.

[44] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, 2009, pp. 807–816.

[45] D. Kuang, C. Ding, and H. Park, "Symmetric nonnegative matrix factorization for graph clustering," in *Proceedings of the 2012 SIAM international conference on data mining*, 2012, pp. 106–117.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[47] Z. Zheng, H. Ma, M. R. Lyu, and I. King, "Wsrec: A collaborative filtering based web service recommender system," in *IEEE International Conference on Web Services*, 2009, pp. 437–444.

[48] 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.

[49] L. Liao, X. He, H. Zhang, and T.-S. Chua, "Attributed social network embedding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2257–2270, 2018.

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

[51] T. Ebesu and Y. Fang, "Neural citation network for context-aware citation recommendation," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, pp. 1093–1096.

[52] 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*, 2014, pp. 701–710.

[53] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Y. Chang, "Network representation learning with rich text information." in *IJCAI*, 2015, pp. 2111–2117.

Wei Wang received the B.Sc. degree from Shenyang University, Shenyang, China, in 2012, and the Ph.D. degree in Software Engineering from Dalian University of Technology, Dalian, China, in 2018. He is now with Macau University of Science and Technology and University of Macau, China. His research interests include big scholarly data, social network analysis, and computational social science.

Tao Tang received the Bachelor Degree from the Chengdu College, University of Electronic Science and Technology of China, Chengdu, China in 2019. He is currently a research assistant at The Alpha Lab, School of Software, Dalian University of Technology. His research interests include data science, big data analytics and visualization.

Feng Xia (M'07-SM'12) received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently an Associate Professor and Discipline Leader in School of Engineering, IT and Physical Sciences, Federation University Australia, being on leave from School of Software, Dalian University of Technology, China, where he is a Full Professor. Dr. Xia has published 2 books and over 300 scientific papers in international journals and conferences. His research interests include data science, social computing, and systems engineering. He is a Senior Member of IEEE and ACM.

Zhiguo Gong (M'10-SM'16) received the PhD degree in the Department of Computer Science, Institute of Mathematics, Chinese Academy of Science, and the MSc degree from Peking University, Beijing, China, in 1988. He is currently a Professor and the Head in the Department of Computer and Information Science, University of Macau, Macau, China. His research interests include Machine Learning, Data Mining, Database, and Information Retrieval. He is a senior member of the IEEE.

Zhikui Chen received the B.E. degree in mathematics from Chongqing Normal University, Chongqing, China, in 1990, and the Ph.D. degree in solid mechanics from Chongqing University, Chongqing, China, in 1998. He is currently a Professor at Dalian University of Technology, Dalian, China. His research interests include Internet of Things and big data.

Huan Liu (F'12) received the B.Eng. degree in computer science and electrical engineering from Shanghai Jiaotong University and the Ph.D. degree in computer science from the University of Southern California. He is currently a Professor of computer science and engineering at Arizona State University. His research interests include data mining, machine learning, social computing, and artificial intelligence. His well-cited publications include books, book chapters, and encyclopedia entries and conference, and journal papers. He is a Fellow of IEEE, ACM, AAAS, and AAAI.