IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA ANALYTICS FOR LARGE-SCALE COMPLEX DATA ENVIRONMENTS

Received June 1, 2017, accepted June 18, 2017, date of publication June 30, 2017, date of current version July 24, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2721934

Exploiting Fine-Grained Co-Authorship for Personalized Citation Recommendation

LANTIAN GUO[1], XIAOYAN CAI[1], FEI HAO[2], DEJUN MU[1], CHANGJIAN FANG[1], AND LIBIN YANG[1]

[1]School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[2]School of Computer Science, Shaanxi Normal University, Xi'an 710119, China

Corresponding author: Dejun Mu (mudejun@nwpu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61672433 and Grant 61402373, in part by the Fundamental Research Funds for the Central Universities, China, under Grant GK201703059, Grant 3102016QD009, and Grant 3102016QD010, and in part by the China Postdoctoral Science Foundation Funded Project under Grant 2017M613205.

ABSTRACT In the era of big scholarly data, citation recommendation is playing an increasingly significant role as it solves information overload issues by automatically suggesting relevant references that align with researchers' interests. Many state-of-the-art models have been utilized for citation recommendation, among which graph-based models have garnered significant attention, due to their flexibility in integrating rich information that influences users' preferences. Co-authorship is one of the key relations in citation recommendation, but it is usually regarded as a binary relation in current graph-based models. This binary modeling of co-authorship is likely to result in information loss, such as the loss of strong or weak relationships between specific research topics. To address this issue, we present a fine-grained method for co-authorship modeling that incorporates the co-author network structure and the topics of their published articles. Then, we design a three-layered graph-based recommendation model that integrates fine-grained co-authorship as well as author–paper, paper–citation, and paper–keyword relations. Our model effectively generates query-oriented recommendations using a simple random walk algorithm. Extensive experiments conducted on a subset of the anthology network data set for performance evaluation demonstrate that our method outperforms other models in terms of both Recall and NDCG.

INDEX TERMS Co-authorship, graph model, topic clustering, random walk, citation recommendation.

I. INTRODUCTION

Recommender systems, which attempt to automatically suggest items of potential or undiscovered interest to solve information overload issues, have recently attracted increased attention [12]. These systems have been successfully applied in many fields such as music [5], movies [2], e-commerce [11], mobile services [38], and others [6], [7], [35], [36]. As the production of scientific papers expands, increasingly more research papers are being published and shared in various digital databases and on academic websites. Thus, this phenomenon behind big scholarly data has led to information overload in research [1], [34]. Consequently, applying recommendation techniques to obtain a rapid, accurate and sufficient publication list can help researchers to advance their research, especially when they move into a new research field.

There is a variety of forms of citation recommendation. The instance of personalized citation recommendation in our work can be described as follows: the searcher provides a query manuscript that contains the searcher's identity and a summary that briefly describes their query, i.e., paper title/abstract or research idea descriptions. The recommender system then provides a citation list according to the submitted query manuscript.

Early work on personalized citation recommendation employed collaborative filtering (CF) and content-based filtering (CBF) methods. CF explores the rating matrices created from the researchers' readership or the publication citation network [17]; this approach is generally limited by data sparsity, cold-start and scalability problems [38]. By contrast, CBF approaches focus on the content relevance among papers [4], [27], [37]; however, CBF suffers from traditional information retrieval issues such as semantic ambiguity [26].

With the emergence and rapid development of heterogeneous information networks, graph-based recommendation approaches have received increasing attention [34]. A heterogeneous graph model is a comprehensive representation of objects with different types of relations [20]; it influences recommendation performance by utilizing the interaction

between different types of relations. Thus, a graph-based model that utilizes various relations among heterogeneous objects, e.g., paper citation and content relations, can achieve high performance with respect to recommendation results [23].

Many graph-based approaches treat citation recommendation as a citation link prediction task [28], [39]. However, it is assumed that a partial list of citation papers is already known, and these models can only predict papers that might be cited in the future. To address this problem, most recent graph-based models recommend citation papers based on a query manuscript that briefly summarizes their research work. Furthermore, these models incorporate the content information of the citations to produce citation recommendations for the submitted query [16], [20], [21]. Nevertheless, although both citation and content information are considered, co-authorship information is neglected.

The use of co-authorship provides effective information to personalized citation recommendation. Co-authorships provide useful and interesting information for analyzing the citation behavior of researchers [3], and these personal characteristics can affect the citations, e.g., researchers may cite similar papers if they collaborated previously [29]. A co-author network is an extraordinary social network due to the academic nature of co-authorship [32], but it is usually treated as a binary relation in current graph-based personalized citation recommendation [9], [14], [18].

Since many factors influence the measurement of co-authorship (e.g., the number of times that two authors have been co-authors and the topic distribution of their publications), the importance weight of co-authorship links should be different. Binary modeling of authorship is likely to lead to information loss such as missing strong or weak relationships between specific research topics. Specifically, binary modeling leads to two critical problems: (1) how to discover the importance weight of co-authorship links for co-authorship modeling and (2) how to exploit the newly modeled co-authorship relations to achieve effective citation recommendation.

The major contributions of this paper are summarized as follows.

- **Fine-grained Co-authorship Modeling**. We propose a fine-grained co-authorship modeling method that combines the co-author network structure and the topics of their publications. We define the collaboration influence distribution in different topics as the authors' collaboration features, which represent the interaction strength between different topics.
  To reveal the authors' academic features within different topics, a topic clustering model and a random walk model are applied to the co-author network to evaluate the authors' collaboration influence within each topic. Subsequently, fine-grained co-authorship relations are derived from the similarities of the collaboration influence distribution and are then employed to construct a co-authorship relation graph.
- **Recommendation Model**. We design a three-layered graph-based recommendation model that incorporates the fine-grained co-authorship graph, paper-citation graph, paper-author graph, and paper-keyword graph to produce recommendations. A personalized query-oriented recommendation task is implemented on the proposed multi-relation graph. Our method can effectively generate query-oriented recommendation results by using a simple random walk algorithm to compute the relevance between a query and the unified graph based on multi-relation information.
- **Evaluation**. We conduct extensive experiments on a subset of the Association of Computational Linguistics (ACL) Anthology Network (AAN) dataset to evaluate the impacts of the defined features and the performance of the proposed method. The experimental results demonstrate that our method outperforms the models that incorporate binary co-authorship relations in the multi-layered graph (by an average of 4.86% with respect to *Recall* and 7.68% with respect to normalized discounted cumulative gain (*NDCG*)). A parametric study and case study are also presented and analyzed.

The remainder of this paper is organized as follows. Section II reviews the related work on citation recommendation. The problem definition is given in Section III. Section IV presents our three-layered graph model, the details of our proposed fine-grained co-authorship modeling method, and the paper ranking method. Based on the proposed recommendation model, Section V describes our experimental setup and discusses our results in detail. Finally, Section VI concludes this paper.

II. RELATED WORK

State of the art of citation recommendation research can be divided into three main categories: collaborative filtering (CF), content-based filtering (CBF) and graph-based approaches.

A. COLLABORATIVE FILTERING (CF)

Citation paper recommendation employing CF focuses on the rating matrices created from the researchers' readership or the adjacency matrix associated with a citation network. For example, McNee *et al.* [17] assumed that citations by an author indicate a positive view of a paper. Their method, analogously to conventional CF, considers cited papers as corresponding to users and citations as corresponding to items.

However, CF approaches are generally limited by the cold-start and data sparsity problems [38]. To address the cold-start problem, [25] combined the intermediate recommendation results generated from the CF and CBF approaches, but that paper only employed term frequency-inverse document frequency (TF-IDF) to calculate content similarity. Liu et al. [12] employed citation context to mine co-occurrences between cited papers as association information to provide supplementary information for enhancing the CF method;

however, content and co-authorship information were not included.

B. CONTENT-BASED FILTERING (CBF)

CBF approaches attempt to retrieve papers that match a query with respect to textual content, which is not accurately modeled by CF. With demonstrated success in several text mining applications, latent Dirichlet allocation (LDA) probabilistic topic modeling and its modifications have drawn abundant attention in citation recommendation research. For example, to discover the effects of various types of context information on recommendations, Wang and Blei [27] combined probabilistic topic modeling CBF analysis and latent factor model CF using of an interpretable latent structure for researchers and papers. Tang and Zhang [24] proposed topic-based citation recommendation by training a two-layered restricted Boltzmann machine to learn topic distributions based on the citation and paper content relationship. Many works have also focused on modeling user personalization to achieve citation recommendation. These methods obtain user information and build user profiles with CBF algorithms. Yan *et al.* [37] proposed a personalized citation recommendation through a combination of users citing preference and content relevance measurements. Other CBF approaches have exploited citation context analysis; for example, Ding *et al.* proposed a next generation of citation recommendation by analyzing citation context using both syntactic and semantic techniques [4]. However, CBF approaches suffer from traditional information retrieval issues such as semantic ambiguity [26]. In addition, the consideration of various types of useful information in probabilistic topic models is increasingly complicated and represents another bottleneck.

C. GRAPH-BASED APPROACHES

Since heterogeneous objects and their mutual relations can be simply represented by a graph, graph-based methods can be easily applied to various types of heterogeneous data to produce recommendations [23]. Existing graph-based approaches often consider citation recommendation as a citation link prediction task. For example, the hierarchical clustering algorithm proposed by West *et al.* [28] based on a citation relation graph determines relevance and recommends papers based on their importance within these clusters. To address the sparsity of a single citation graph and noise in citation graph construction, Zhou *et al.* [39] proposed a new factorization strategy that combines multiple citation and author graphs to measure paper similarities for the recommendation task. However, the limitation of these works is that they mainly stress the role of citation network connections and overlook other useful information in the bibliographic network. Moreover, when the link prediction task has been used to recommend potential citation papers, consideration of the query of the searcher in the recommendation model has been overlooked.

To address these problems, most recent works recommend citation papers based on a query manuscript that briefly summarizes their research work, i.e., paper title/abstract or research topic. To provide a citation list for a manuscript, the recommender system constructs a homogeneous graph model exploiting both citation and content information. However, in many methods, the citation and content information are considered, whereas co-authorship information is ignored. For example, Pan *et al.* [20] constructed a heterogeneous graph to represent both the citation and content information within papers and then applied a graph-based similarity learning algorithm to perform the recommendation task. Liu *et al.* [15] proposed a context-rich citation recommendation method by combining the citation textual context and the importance of citation relationships to construct a heterogeneous graph for recommendation. Further, additional citation recommendation is performed by combining full-text citation analysis and graph mining algorithms in [16].

There are also many works that incorporate co-authorship information into a graph model to improve personalized citation recommendations; however, these methods treat co-authorship as a binary relation. For example, Meng *et al.* [18] developed a personalized query-based citation recommendation considering the researcher's unpublished manuscript and identity to construct a personalized query. In their work, co-authorship, citation and textual content are considered in the heterogeneous graph. Their mutual relations are represented to construct a unified recommendation model. Lao and Cohen [9] proposed a labeled directed-graph-based model that exploits not only text but also co-authorship and venues as implicit information for scientific article recommendation. Liu *et al.* [14] exploited a number of pre-designed meta-paths, including a co-authorship-based sub-meta-path, to address diversified paper ranking task.

III. PROBLEM DEFINITION

In this work, we formulate the citation recommendation problem as the problem of learning a relevance score list $s(q, p)$: $Q \times P \longmapsto R$ for a query manuscript $q \in Q$ and a candidate paper $p \in P$ based on the heterogeneous bibliographic graph of scientific papers. The learned score list $s(q, p)$ is used to rank scores between query and candidate papers to produce a recommendation.

Formally, we define the personalized citation recommendation problem as follows. Given a heterogeneous bibliographic graph G and the query keyword subset q_k from a searcher q_a (if the searcher is known by the system) to form a query manuscript $q \in Q$, we aim to establish a recommendation model specifically for q and recommend a small subset of target papers $p \in P$ as high-quality references for q by ranking the papers via the score list $s(q, p)$. Before describing our proposed approach, we first give some notations in Table 1 that will be used frequently in this paper.

IV. OUR RECOMMENDATION MODEL

This section presents an overview of the solution system and then details the corresponding approach and algorithm.

TABLE 1. Notations.

Notation	Description
A	The set of author collection
P	The set of paper collection
W	The set of keyword collection
Q	The set of query collection
D	The set of topic collection
m	The number of authors
n	The number of papers
k	The number of keywords
z	The number of topics
M_{AA}	The $m * m$ matrix indicating co-authorship graph
M_{PP}	The $n * n$ matrix indicating paper-citation graph
M_{PW}	The $n * k$ matrix indicating paper-keyword graph
M_{WP}	The $k * n$ matrix indicating keyword-paper graph
M_{PA}	The $n * m$ matrix indicating paper-author graph
M_{AP}	The $m * n$ matrix indicating author-paper graph
AT	The $m * z$ matrix indicating author-topic relation
AW	The $m * k$ matrix indicating author-word relation
WT	The $k * z$ matrix indicating word-topic relation
R	The $m * z$ matrix indicating updated author-topic relation
CS	The $m * z$ matrix indicating the final value of updated R
q_0	The $m + n + k$ element vector indicating query of searcher
q_a	The m element vector indicating author query in q_0
q_p	The n element vector indicating paper query in q_0
q_w	The k element vector indicating keyword query in q_0

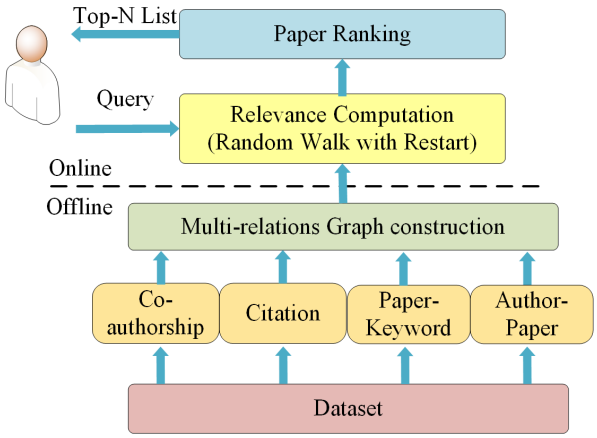


FIGURE 1. The system overview of personalized citation recommendation.

A. SYSTEM OVERVIEW

As shown in Fig. 1. Our graph-based citation recommendation framework includes the following two main stages:

1) In the offline stage, we establish the fine-grained co-authorship, author-paper, paper-citation, and paper-keyword relations in the dataset and then combine the various relationships with the constructed graph model.
2) In the online stage, the query content is used as the starting point, and a random walk algorithm is used to "walk" in the graph model with various information relations. The most relevant literature is determined, and the Top-N recommendations are implemented.

The remainder of this section is organized as follows. We first provide an overview of multi-layer graph construction in the offline stage of our recommendation method. Then, we illustrate the construction of the fine-grained co-authorship relations in a step-by-step manner. Finally, the graph-based ranking procedure used to generate the Top-N recommended papers in the online stage is presented.

B. MULTI-LAYERED GRAPH CONSTRUCTION

Fig. 2 shows the network structure of our graph model. Our method makes use of four types of information (co-authorship, author-paper, paper-citation, and paper-keyword) corresponding to the three types of objects and nodes in the graph model, i.e., authors, papers and keywords. Thus, the relations among them can be formulated as a three-layered graph model, in which the intra-layer links denote the relations between the same types of objects and the inter-layer links denote the links between any two different types.

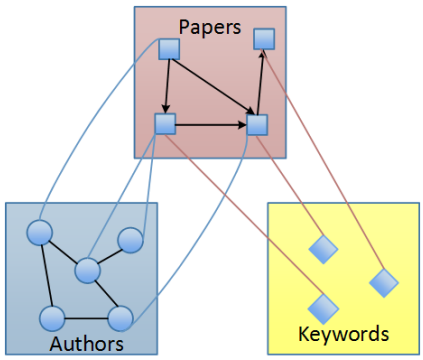


FIGURE 2. Three-layered graph *G*.

For a set of given information, our multi-layered graph model can be described as follows: Let $G = <V, E, M>$, where V is the set of vertices that consists of the author set $A = \{a_1, a_2, \ldots, a_m\}$, paper set $P = \{p_1, p_2, \ldots, p_n\}$ and keyword set $W = \{w_1, w_2, \ldots, w_k\}$, i.e., $V = A \cup P \cup W$. E is the set of edges that connect the vertices, i.e., $E = \{< v_i, v_j > | v_i, v_j \in V\}$. M is the adjacency matrix, in which the element w_{ij} represents the weight of the edge connecting v_i and v_j.

To obtain a clear understanding of our graph structure, M can be decomposed into nine blocks, i.e., M_{AA}, M_{PP}, M_{PA}, M_{AP}, M_{PW}, M_{WP}, M_{AW}, M_{WA}, and M_{WW}, each representing a sub-graph of the relation objects indicated by the subscripts. Note that M_{AW}, M_{WA}, and M_{WW} are not considered in this work and are set to zero. Finally, M can be written as Fig. 3. In the matrix M, there are four types of edges:

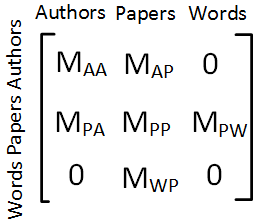


FIGURE 3. Adjacency matrix *M*.

1) $M_{PP}(i, j)$ is the paper-paper citation relation between paper p_i and paper p_j; if p_j is a reference of p_i, the edge value between them is 1. Thus, M_{PP} is also a binary graph.
2) $M_{AP}(i, j)$ is the author-paper relation between author a_i and paper p_j; if a_i is an author of p_j, the edge value between them is 1. M_{AP} is equal to M_{PA}, as the relationships between authors and papers are symmetric.
3) $M_{PW}(i, j)$ is the paper-word relation between paper p_i and keywords w_j; if w_i is a keyword in paper p_j, the edge weight is the TF-IDF value of w_j in p_i. M_{PW} is equal to M_{WP}, as the relationships between papers and keywords are symmetric.
4) $M_{AA}(i, j)$ represents the co-authorship relation between author a_i and author a_j. As mentioned in Section I, most previous recommendation research treats it as a binary relation, i.e., if the two authors have collaborated, the edge value of $M_{AA}(i, j)$ is 1; otherwise, it is 0. In this paper, we exploit a fine-grained method to model the co-authorship relation. Formally, the co-authorship relationships between author nodes are represented as weighted edges (a_i, a_j, σ_{ij}), where the weight values σ_{ij} indicate the similarities between a_i and a_j. Finally, $M_{AA}(i, j)$ is converted from a simple graph to a weighted graph.

C. MODELING FINE-GRAINED CO-AUTHORSHIP RELATIONS

Co-authorships provide very useful and interesting information for analyzing the citation behavior of researchers. Many factors influence the importance measurement of relationships between researchers. In this paper, we consider the frequency of collaboration (the number of times that two authors have been co-authors) and the topic distribution of their publications to obtain fine-grained collaboration relations. The reasons that we choose these two factors are the following:

Frequency of collaboration is an important indicator of collaboration influence between two authors; however, it is difficult to capture the global structure of the co-author network based on only the count of co-authored manuscripts. Meanwhile, considering the real academic collaboration scene, researchers often behave differently across multiple topics of interests, e.g., some authors tend to collaborate under the same research topics or different research topics. Thus, author topical similarity can serve as a reasonable indicator of the semantic relationship between researchers.

To consider the above factors in modeling co-authorship, we combine the co-author network structure and the research topics of their published articles to model the fine-grained co-authorship relations. The collaboration influence distribution under different topics is defined as the authors' collaboration features, which represent the interaction strength within different topics. The random walk with restart (RWR) recommendation model has proved to be qualified for capturing the global structure of the graph derived from co-authorship [8],

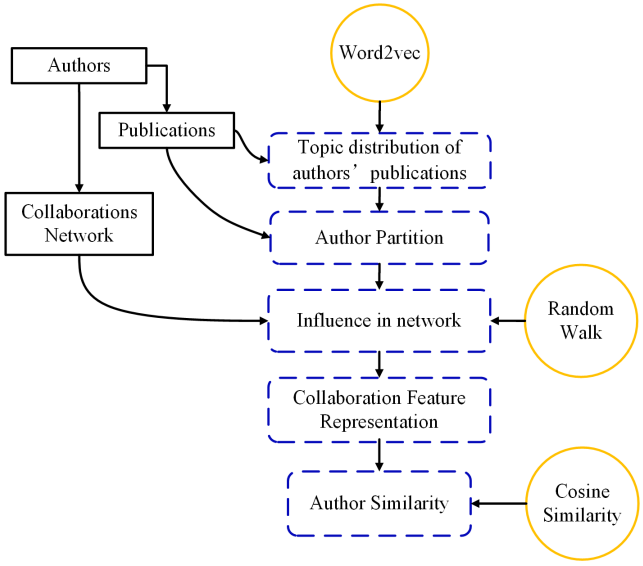


FIGURE 4. Fine-grained co-authorship modeling.

and the ranking score of each node in the co-authorship graph is used to measure the influence and interaction of authors.

Thus, to reveal authors' academic features within different topics, we employ a topic clustering method to model the authors' publication topic distribution and apply a random walk model to the co-author network to compute the authors' collaboration influence within each topic. Then, the similarities of the collaboration influence distributions can be expressed as the link importance weight between authors. Fig. 4 depicts the specific procedure of this method. Its three main steps are as follows:

Step 1 (Author Partition): In this step, a content-based method is used for author partitioning to generate various topics for the publications using a maturing natural language processing technique; then, authors are partitioned according to these topics.

First, Word2vec is executed to generate a vector representation of the words in the title and abstract corpus. Word2vec is used because it has been proven to be an efficient implementation of the skip-gram and continuous of bag-of-words architectures for generating vector representations of words [19]. It can represent a word as a low-dimensional vector to capture not only grammatical and syntactic information but also semantic features. This contextual feature representation method is widely used in topic mining applications [10].

Second, K-means clustering [31] is performed on these word vectors to generate topic clusters. The output of K-means clustering is the corresponding topic IDs of all the words in the titles and abstracts. The research topics are clustered, and the words that they contain are mapped to each topic simultaneously. These mapping relations are denoted by the matrix WT. Hence, if a word w is related to a topic $d \in D$, where D is the set of topics, the value of $WT_{w,d}$ is 1; otherwise, it is 0.

Third, we map authors to particular research topics that contain these subject words to obtain the author-topic

representation. Specifically, authors are mapped to specific topics using the following method:

$$AT_{a,d} = AW_{a,w} * WT_{w,d} \tag{1}$$

where $AT_{a,d}$ is the author-topic relation, $AW_{a,w}$ is the author-word relation, and $WT_{w,d}$ is the word-topic relation. Then, we check all the elements in the author-topic matrix to update $AT_{a,d}$ through the following process.

$$R_{a,d} = \begin{cases} 1 & \text{that } \mathrm{AT_{a,d}} \neq 0 \\ 0 & \text{that } \mathrm{AT_{a,d}} = 0 \end{cases} \tag{2}$$

where $R_{a,d}$ represents the updated author-topic relation. This content-based feature representation for each author is employed to partition authors.

Step 2 (Collaboration Feature Representation): RWR is performed to capture the global structure of the co-authorship graph. In RWR, the ranking score of a node is determined by the voting contributions from neighbors in the graph, which can be used to represent the importance of the nodes in this network. Based on this principle, which is similar to the social-network-based academic value calculation method proposed by Kong *et al.* [8], we apply an RWR model to the co-authorship graph of a specific topic to generate the ranking score of each author in this topic. Then, the generated ranking score of each author in a specific topic is employed to denote the collaboration influence strength of the author within this topic. The equation of the RWR model is as follows:

$$R_d^{(t+1)} = (1-\alpha)N_d R_d^{(t)} + \alpha q_d \tag{3}$$

where R_d represents the ranking score vector of all authors in topic d. $R_d^{(0)}$ is the the initial state vector of R_d. The restart vector q_d is equal to $R_d^{(0)}$. $1-\alpha$ denotes the damping coefficient, which means that α is the restart probability in the RWR model. In our RWR model, α is 0.2, which was validated and confirmed by Kong *et al.* [8]. N_d is the transition matrix, which is generated from the co-authorship graph of a specific topic by normalization. N_d drives the random walker to skip to the next node with a certain probability. RWR is an iterative process. After a limited number of iterations, the vector R_d converges. We use $CS_{a,d}$ as the final value of the vector item $R_{d,u}$, which represents the collaboration influence of author a in topic d. *CS* is quantified for each author in various topics.

Step 3 (Author Collaboration Modeling): The collaboration influence distribution for different topics can represent the authors' academic interaction and interest characteristics. The similarity of collaboration influence distributions can be expressed as the link importance weight between authors. The cosine similarity is employed to define the similarity between two authors, a_1 and a_2, based on their feature vectors, F_{a_1} and F_{a_2}, to measure the similarity of their collaboration influence distributions.

$$sim(a_1, a_2) = \frac{\sum_{i=1}^{n} (F_{a_{1,i}} \cdot F_{a_{2,i}})}{\sqrt{\sum_{i=1}^{n} F_{a_{1,i}}^2} \cdot \sqrt{\sum_{i=1}^{n} F_{a_{2,i}}^2}} \tag{4}$$

This similarity value is viewed as the link importance weight between authors.

D. GRAPH-BASED PAPER RANKING

Given a multi-layered graph G, we aim to build a recommendation model specifically for the query manuscript q and to recommend a subset of target papers p as references for q by ranking the papers via the score list $s(q, p)$. Thus, the first key is to establish an adjacency matrix M for graph G after obtaining the four types of relations: fine-grained co-authorship (M_{AA}), author-papers (M_{AP}/M_{PA}), citations (M_{PP}), and paper-keywords (M_{PW}/M_{WP}).

Then, to generate the Top-N recommended candidate papers in our proposed recommendation model, an RWR algorithm is executed to obtain the relevance score list $s(q, p)$. The random walk model can be defined for a multi-layered graph by transforming its adjacency matrix M into a transition probability matrix M' using column normalization. Each vertex of the graph is assumed to be a given state of a stochastic process [33]. The transition matrix M' governs the probability of transition between states.

The details of the proposed recommendation algorithm are described below, and the corresponding pseudo-code is given in **Algorithm 1**.

Algorithm 1 Graph-Based Paper Ranking

Input: Transition probability matrix, M';
Query keyword set W_u;
Searcher ID u;
Random walk probability, β;
Minimum value of convergence, *MinDelta*;

Output: Ranking scores of all paper vertices, $S(q, p)$

Define ranking scores of all query vertices, $q_0 = [q_a, q_p, q_w]$ // q_a: all author vertices, q_p: all paper vertices, q_w: all keyword vertices
if $i = u$ **then** //author_vec_init()
    $q_{a,u}(i) \leftarrow 1$
end if
for *each* w in W_u **do** //keyword_vec_init()
    if $j = w$ **then**
    $q_{w,u}(j) \leftarrow 1$;
    end if
end for
    $diff \leftarrow 0$
    $q \leftarrow q_0$
while 1 **do**
    $q^{t+1} \leftarrow (1-\beta) * M' * q^t + \beta * q_0$
    $diff \leftarrow q^{t+1} - q^t$
    if $diff < MinDelta$ **then**
        break
    end if
end while
$PaperScore(1:m) \leftarrow q_p$; //select ranking scores of papers
$S(q, p) \leftarrow Sort(PaperScore(1:m))$
Return $S(q, p)$

First, the initial query vector q_0, which represents the starting vertices in the random walk model, is constructed using both the author and keyword information in the searcher's query **(Lines 1-7)**. Specifically, the query vector is composed of the vertices of all the authors, papers, and keywords. Given a query vector $q_0 = [q_a, q_p, q_w]$, where q_a is the author query vector corresponding to the vertices that consist of the author set $A = \{a_1, a_2, \ldots, a_m\}$ in G, q_p is the paper query vector corresponding to the paper set $P = \{p_1, p_2, \ldots, p_n\}$ in G, and q_w is the keyword query vector corresponding to the keyword set $W = \{w_1, w_2, \ldots, w_k\}$ in G.

The construction of the three subsets of the query vector is performed as follows:

- The author query vector q_a represents the searcher's identity, i.e., who entered the query text. Given a searcher u, we define the author query vector of the searcher u as $q_{a,u}$, where $q_{a,u}(i) = 1$ if $i = u$ and 0 otherwise.
- The keyword query vector q_w represents the terms entered by the searcher. Given a set of keywords $W_u = \{w_1, w_2, \ldots, w_e\}$ entered by a searcher u, we define the keyword query vector of the searcher u as $q_{w,u}$, where $q_{w,u}(j) = 1$ if $j = w_e$ and 0 otherwise.
- The paper query vector q_p represents all the candidate papers in our proposed graph. It records the probability of candidate papers when the process transitions between states. Thus, the initial value of the vector is 0.

Second, RWR is implemented to compute the relevance scores for the candidate paper set q_p under the initial query vector q_0 **(Lines 8-14)**. In the RWR model, the vertices in q_0 are the starting nodes. Beginning with the starting nodes, RWR is performed by randomly jumping to another linked node according to an initial probability vector with respect to each node in the transition matrix of the whole graph G. Thus, the personalized RWR process can be defined as an iterative process represented by the following matrix equation:

$$q^{t+1} = (1 - \beta)M'q^t + \beta q_0 \tag{5}$$

where M' is the transition matrix constructed from graph G, q_0 is the initial query vector, q^t indicates the probability of visiting of each node at step t, q_0 is the initial query vector, and β is the restart probability.

At each step of the iterative process represented in Equation 5, the random walker in the RWR model may visit one of the preferred nodes in q_0 with probability β or transit to one of its linked nodes with probability $1 - \beta$ in M'. The initial state q^t in the iterative process is q_0. By iteratively applying Equation 5 until convergence, the difference value *diff* between the next state q^{t+1} and the last state q^t is smaller than the minimum of the convergence value *MinDelta* in Algorithm 1. The stationary probability value of the paper query vector q_0 represents the relevance score between candidate papers and the searcher's query. Papers that are ranked highly in $s(q, p)$ are recommended to the searcher.

Generally, our proposed recommendation method evaluates the relevance score of each candidate paper from a global perspective for the searcher's query with respect to all the papers, keywords, and authors in the whole graph.

The relevance measurements based on personalized citation recommendations are considered from two perspectives: the relevance between the query text and the paper's textual information and the relevance scores between the searcher's personal preference and the paper, the latter of which mainly relies on the co-author network. With the help of the co-author network, searchers may find personalized papers of the authors that have similar interests as them, and they may also focus on papers written by the collaborators of these authors.

V. EXPERIMENTAL EVALUATION

In this section, we describe the pre-processing of the AAN dataset, the evaluation metrics that we employed and our experimental procedure for evaluating the performance of PWFC, as well as detailed analysis of the experimental results.

We embarked on different experiments to compare PWFC with baseline models in terms of *Recall* and *NDCG* metrics. Then, to improve the accuracy and effectiveness of our PWFC model, we examined the impact of different parameters by conducting a series of experiments and optimization of some parameters. Finally, to give an intuitive understanding of the effectiveness of our citation recommendation approach, we implement case studies to show the retrieved recommended lists for a given query.

A. EXPERIMENT SETUP

1) DATA

We experimented on the AAN dataset [22], which contains the complete collection of papers included in many ACL venues. We extracted 13, 929 papers with intact titles and abstracts published from 1965 to 2012 as our experimental dataset. Each paper was then pre-processed through the following steps: (a) extract the abstract and title; (b) remove the words that consist of 3 characters or less; (c) remove stop words; and (d) stem the remaining words with a porter stemmer. To reduce the noise, we also removed the words that appeared in the dataset fewer than ten times. We obtained a total of 4, 397 distinct candidate words. To construct the paper-word network, we identified keywords from the set of candidate words using a naive method with TF-IDF as an indicator [30]: if the TF-IDF of a word was greater than a threshold, this word was selected as a keyword. A total of 3, 704 distinct keywords were identified for the set of 13, 929 papers at a TF-IDF threshold of 0.03.

We used all 12, 728 papers published before 2012 to construct the graph matrix and to train various recommendation models, and the remaining 1, 201 papers published in 2012 were used as test data. Table 2 shows some statistics for both

TABLE 2. Statistics of the experimental dataset.

	#Paper	#Author	#Cited_Papers	#Citation
Training Data	12,728	10,326	9,433	70,412
Test Data	1,201	1,264	4,508	10,842

TABLE 3. Performance comparison of different methods in terms of *Recall* and *NDCG*.

Top-N	25		50		75		100	
Metrics	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*
PW	0.2072	0.3374	0.2769	0.3502	0.3382	0.3567	0.3787	0.3582
PWBA	0.2149	0.3382	0.3029	0.3588	0.367	0.3672	0.4142	0.3715
PWFC	**0.2311**	**0.3715**	**0.3189**	**0.3869**	**0.3806**	**0.3918**	**0.4263**	**0.3951**

the training and test data, where #Cited_Papers represents the number of papers that were cited at least once and #Citation represents the total number of citations. The keywords of each query were extracted from the title and abstract of each paper in the test data. Then, following common practice [18], the reference list of each paper in the test data was adopted as the ground truth.

2) MEASURES

For the purpose of the evaluation, we employed *Recall* [13] and *NDCG* [26] as the evaluation metrics for the recommendation accuracy and quality of the predicted ranks, respectively. Both metrics are commonly used to evaluate recommendation results.

- **Recall**. *Recall* is the ratio of the number of cited papers in the Top-N recommendation list to the total number of cited articles. The ratio represents the number of hits divided by the size of each users test data. *Recall* is calculated as

$$Recall@N = \frac{1}{Q}\sum_{j=1}^{Q}\left(\frac{R_p \cap T_p}{T_p}\right) \tag{6}$$

  where Q is the number of queries and N is the length of the recommendation list. For a query in the test set, R_p is the Top-N paper list recommended based on a query of test paper p. T_p is the set of papers citing p.
- **NDCG**. The effectiveness of a recommender system is sensitive to the positions of the relevant reference papers, which cannot be fully evaluated by Recall. Intuitively, it is desirable for highly relevant references to appear higher in the Top-N list. We use (*NDCG*) to measure the ranked recommendation list. The *NDCG* value of a ranking list at a specific position is calculated as

$$NDCG@N = \frac{1}{Q}\sum_{j=1}^{Q}\left(\sum_{i=1}^{k}\frac{2^{r_i}-1}{\log_2(i+1)}/IDCG@N\right) \tag{7}$$

  where Q is the number of queries and N is the number of top items. $r(i)$ is the rating of the *i-th* document in the ranking list. $r(i) = 1$ if the paper is relevant, and $r(i) = 0$ if it is not. $IDCG@N$ is the ideal ranking of the results such that $NDCG@N = 1$ if a perfect ranking is returned (most relevant item first, followed by the second most relevant item, etc.).

B. PERFORMANCE COMPARISON

First, we construct four training matrices, M_{PP}, M_{PW}/M_{WP}, M_{AA} and M_{PA}/M_{AP}. Table 4 shows basic information and the usage of these matrices. To validate the effectiveness of the fine-grained collaboration relation among authors, we adopt two baseline methods using different types of relations for comparison. A total of three methods are used:

TABLE 4. Matrices for model training.

Matrix	#Element	Density	Usage
M_{PP}	70,412	0.043%	PW, PWBA, PWFC
M_{PW}/M_{WP}	506,652	1.07%	PW, PWBA, PWFC
M_{AA}	90,202	0.085%	PWBA, PWFC
M_{PA}/M_{AP}	33,892	0.026%	PWBA, PWFC

- **Baseline Approach 1 (PW).** Paper-Word graph (PW [20]): The two-layered graph model is constructed, i.e., the citation graph among papers and the graph between a paper and its keywords. Each query is represented solely using keywords: $q = [0, q_w]$.
- **Baseline Approach 2 (PWBA).** Paper-Word graph with Binary co-Authorships (PWBA [18]): A binary co-authorship graph is added to the PW model, and a three-layered graph model is constructed for recommendations. Thus, each query is represented as $q = [q_a, 0, q_w]$, where q_a denotes the searcher information.
- **Our Approach (PWFC).** Paper-Word graph with Fine-grained Co-authorships among authors (PWFC): This is the model proposed in this paper. Specifically, we add fine-grained collaboration among authors to the PW model as a new relation, as introduced in Section IV-C. Therefore, our model is also three layered, and the query representation is the same as in the PWBA model.

Table 3 compares the performance of the three methods. Obviously, as the length of the recommendation list N is gradually increased, the *Recall* and *NDCG* metrics increase for all three methods because a larger value of N indicates that more papers are recommended. Further, the following two observations are obtained.

Observation 1: Table 3 reveals an important conclusion: PWFC and PWBA outperform PW in terms of *Recall* and *NDCG* since co-authorship is considered in PWFC and PWBA. In other words, co-authorship information helps to generate more accurate citation recommendations in query-based personalized citation recommendation.

Observation 2: Another important observation is that as the value of N increases, PWFC always achieves larger values than PWBA in terms of *Recall* and *NDCG*. PWFC shows significant improvement over PWBA (on average 4.86% with respect to *Recall* and 7.68% with respect to *NDCG*). Furthermore, our results indicate that the proposed fine-grained co-authorsip representation approach (PWFC)

generates more accurate paper recommendations than the binary co-authorsip representation (PWBA).

C. PARAMETRIC ANALYSIS

In this section, we analyze two parameters inside our PWFC model: the walking probability β and the number of clustered topics d.

1) WALKING PROBABILITY

We conducted relevant experiments to investigate the impact of the walking probability defined in Section IV-D on the recommendation quality of the PWFC approach.

Different values of β have different impacts on the recommendation quality. Thus, we conducted relevant experiments using our proposed PWFC approach with different values of β. As stated in Section IV-D, for a node in the query vector, $(1-\beta)$ represents the probability of transitioning from the node to its neighboring nodes, and β represents the probability of transitioning from the node to the start node of the initial query vector (query paper). Larger values of β in the random walk model equation 5 indicate a greater possibility of returning to the node in the initial query. Fig. 5 shows the comparison of *Recall* and *NDCG* when β varies from 0.2 to 0.9 in increments of 0.1.

First, as shown in Fig. 5(a) and (b), *Recall@25* and *Recall@50* show an upward trend as β increases from 0.2 to 0.6 and then show a downward trend when β is 0.7. The best *Recall* values are obtained when β is 0.6, and the worst results occur for β equal to 0.7.

Second, as shown in Fig. 5(c) and (d), *NDCG@25* shows an upward trend as β increases from 0.2 to 0.7 and a downward trend when β is equal to 0.8. The best values of *NDCG@25* are obtained when β is 0.7, and the worst results occur when β is equal to 0.2. *NDCG@50* shows an upward trend as β increases from 0.2 to 0.6 and a downward trend when β is equal to 0.7. The best values of *NDCG@25* are observed when β is equal to 0.8, and the worst results occur for β equal to 0.2.

According to these figures, we can conclude that when β is equal to 0.6, both *Recall* and *NDCG* initially show upward trends as β increases and then downward trends. Through comprehensive consideration, we choose a value of 0.6 for β in our experiments.

These relevant analyses indicate that different walking probabilities have different impacts on the PWFC method; however, since β is a common parameter in PWFC and the baseline approaches, these methods can only be compared if they use the same value of β. Therefore, we assign an empirical value of 0.6 to β in our performance comparison experiments.

2) NUMBER OF TOPICS

We also conducted relevant experiments to discuss the impact of the number of clustered topics, as previously defined in Section IV-C, on the recommendation quality of the PWFC approach.

However, since there is no clear standard for setting the number of topics, we repeatedly executed experiments

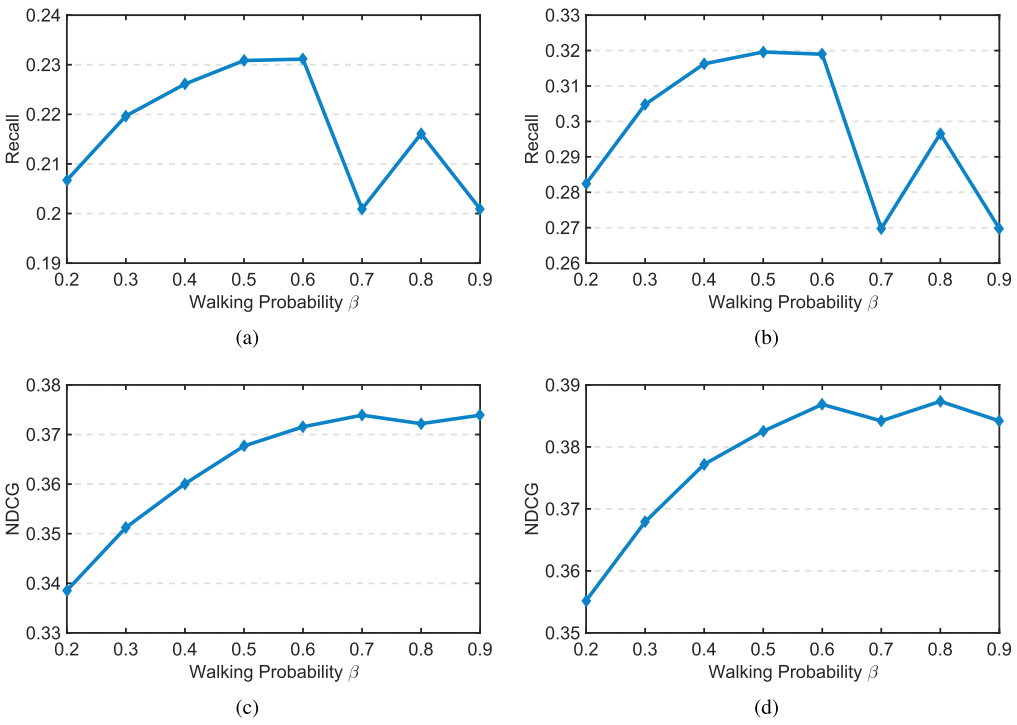


FIGURE 5. *Recall* and *NDCG* for different walking probabilities. (a) Recall@25. (b) Recall@50. (c) NDCG@25. (d) NDCG@50.

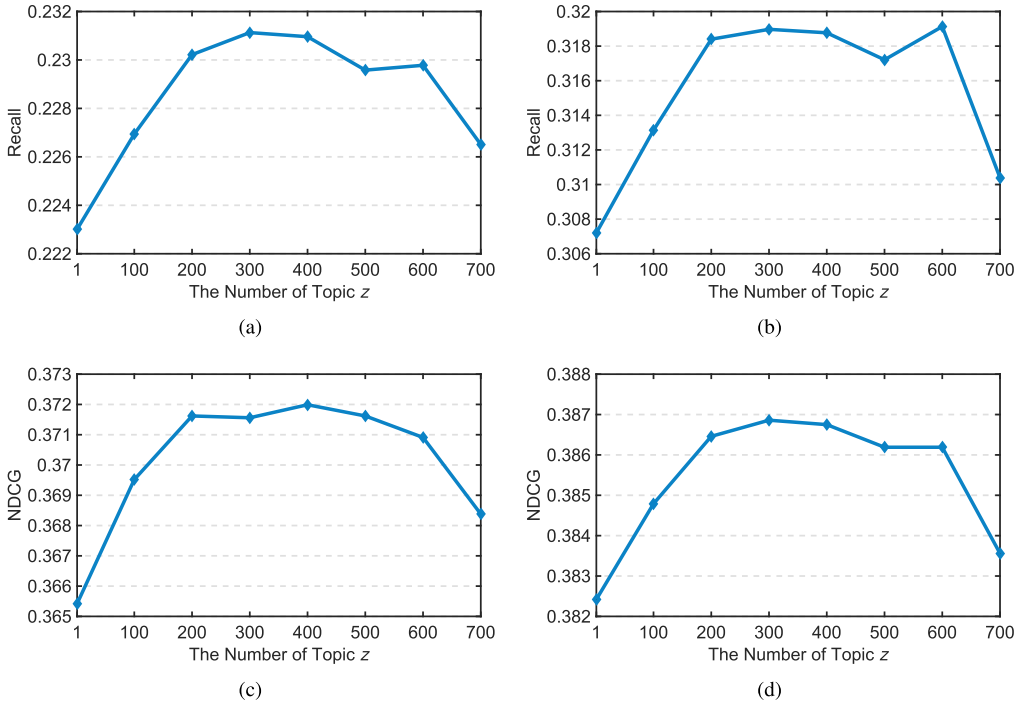


FIGURE 6. ***Recall*** **and** ***NDCG*** **for different numbers of topics. (a) Recall@25. (b) Recall@50. (c) NDCG@25. (d) NDCG@50.**

with different sizes of recommendation lists to evaluate the influence of the clustered topic number on the result.

We tested 8 candidate values for the number of clustered topics d in Section IV-C, i.e., 1, 100, 200, 300, 400, 500, 600, 700. Note that when the number of clustered topics is 1, there is no research topic clustering in the author pattern, and the random walk algorithm is performed to mine the authors' academia value. Corresponding to the 8 candidate values for the number of clustered topics d, 8 independent experiments were conducted. Fig. 6 shows the performance of PWFC in terms of *Recall* and *NDCG* as the number of topics d is gradually increased; we made two observations.

Observation 1: Both *Recall@25* and *Recall@50* increase as the topic number d increases from 1 to 300, as shown in Fig. 6(a) and (b), since the content information of published papers is considered when modeling the author similarities when d is equal to 100. *Recall@25* and *Recall@50* decrease slightly when d is equal to 400 and then significantly decrease and fluctuate when d is equal to 500. Better values of *Recall@25* and *Recall@50* are observed when d is equal to 300.

Observation 2: Both *NDCG@25* and *NDCG@50* increase as the topic number d increases from 1 to 300, as shown in Fig. 6(c) and (d). *NDCG@25* decreases when d is equal to 500, and the best values of *NDCG@25* are observed when d is equal to 400. NDCG@50 decreases when d is equal to 400, and the best values of *NDCG@25* are observed when d is equal to 300. All four figures illustrate that both *Recall* and *NDCG* achieve the worst results for d equal to 1 since no content information is considered in the co-authorship representation.

According to Fig. 6, the number of clustered topics impacts the performance of PWFC. If the number of clustered topics is appropriate, the *Recall* and *NDCG* scores show some improvement. Furthermore, compared to the analysis using only the co-authorship network (the number of clustered topics is equal to 1), combining the co-authorship network analysis and topic clustering for author partitioning enhances the co-authorship representation, which can further improve the recommendation performance.

D. CASE STUDY

In this subsection, to comprehend our citation recommendation method more intuitively, we retrieve recommended papers for a given query using our proposed approach and two baseline approaches for comparison. The query is composed of keywords extracted from a comparatively highly cited paper (cited 65 times as of April 2017) published in EMNLP 2012. We take the top 5 recommended papers from the recommendation list, and the results are shown in Table 5.

First, as shown in Table 5, the results returned by the PWFC approach have 4 records that match the groundtruth citation list of the query paper, whereas the results returned by both the PWBA (with binary co-authorship) and PW (without co-authorship) approaches only have 2 matching records. This observation demonstrates that the PWFC approach obtained a better result in this case study since fine-grained

TABLE 5. Top-5 recommended papers: matched results are (✓).

Query: An Entity-Topic Model for Entity Linking
PWFC:
1) A Generative Entity-Mention Model for Linking Entities with Knowledge Base (✓)
2) Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation
3) Large-Scale Named Entity Disambiguation Based on Wikipedia Data (✓)
4) Using Encyclopedic Knowledge for Named Entity Disambiguation (✓)
5) Forest-guided Supertagger Training (✓)
PWBA:
1) A Generative Entity-Mention Model for Linking Entities with Knowledge Base (✓)
2) Structural Semantic Relatedness: A Knowledge-Based Method to Named Entity Disambiguation
3) Construction of a Large-Scale Chinese-English Parallel Corpus
4) Word Alignment of English-Chinese Bilingual Corpus Based on Chucks
5) Large-Scale Named Entity Disambiguation Based on Wikipedia Data (✓)
PW:
1) Annotation Compatibility Working Group Report
2) Learning Field Compatibilities to Extract Database Records from Unstructured Text
3) A Generative Entity-Mention Model for Linking Entities with Knowledge Base (✓)
4) Supervised Models for Co-reference Resolution
5) Forest-guided Supertagger Training (✓)

co-authorship relations play a positive role in the PWFC approach.

Second, it can be seen that although both PWBA and PW have the same number of matching records, the paper "A Generative Entity-Mention Model for Linking Entities with Knowledge Base" is ranked higher in PWBA than in PW because the co-author network is considered in the PWBA approach.

In a word, we can conclude that by incorporating the heterogeneous bibliographic graph of personalized citation recommendation, co-authorship relations are able to help improve the recommendation performance. Further, this also validates that our proposed fine-grained co-authorship relations perform better than the binary method.

VI. CONCLUSIONS

In this paper, we focus on how to represent co-authorship in a graph to promote personalized query-based citation recommendation. To this end, we first propose a fine-grained co-authorship modeling method that combines the co-author network structure and the research topics of their publications. To reveal the authors' academic features within different topics, a topic clustering model and random walk model are applied to the co-author network to evaluate the authors' collaboration influence within each topic. Then, we design a three-layered graph-based recommendation model that incorporates the fine-grained co-authorship graph, paper-citation graph, paper-author graph, and paper-keyword graph to produce recommendations. Finally, we conduct extensive experiments on a subset of the AAN dataset to compare the performance of our PWFC method to other baseline models. Our experimental results show that PWFC outperforms the other models in terms of *Recall* and *NDCG*.

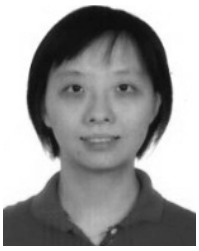
Our research reveals that the combination of network-based and content-based analysis approaches can improve the representation of academic collaboration for personalized citation recommendation. Nevertheless, there is room for future study in this direction such as the incorporation of the publication date and the location of researchers in future recommendation models.

REFERENCES

[1] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: A literature survey," *Int. J. Digit. Libraries*, vol. 17, no. 4, pp. 305–338, 2016.
[2] R. Catherine and W. Cohen, "Personalized recommendations using knowledge graphs: A probabilistic logic programming approach," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 325–332.
[3] J. Cui, F. Wang, and J. Zhai, "Citation networks as a multi-layer graph: Link prediction and importance ranking," Stanford Univ., Stanford, CA, USA, Tech. Rep. CS224W Project Report, 2010.
[4] Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, and C. Zhai, "Content-based citation analysis: The next generation of citation analysis," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 9, pp. 1820–1833, 2014.
[5] C. Guo, "Feature generation and selection on the heterogeneous graph for music recommendation," in *Proc. 9th ACM Int. Conf. Web Sea. Data Mining*, 2016, p. 715.
[6] F. Hao, S. Li, G. Min, H. C. Kim, S. S. Yau, and L. T. Yang, "An efficient approach to generating location-sensitive recommendations in ad-hoc social network environments," *IEEE Trans. Services Comput.*, vol. 8, no. 3, pp. 520–533, May 2015.
[7] F. Hao, G. Min, M. Lin, C. Luo, and L. T. Yang, "MobiFuzzyTrust: An efficient fuzzy trust inference mechanism in mobile social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2944–2955, Nov. 2014.
[8] 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, p. e0148492, 2016.
[9] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Mach. Learn.*, vol. 81, no. 1, pp. 53–67, 2010.
[10] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. ICML*, vol. 14. 2014, pp. 1188–1196.
[11] H. Li, R. Hong, D. Lian, Z. Wu, M. Wang, and Y. Ge, "A relaxed ranking-based factor model for recommender system from implicit feedback," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1683–1689.
[12] 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, Oct. 2015.
[13] Q. Liu, E. Chen, H. Xiong, C. H. Q. Ding, and J. Chen, "Enhancing collaborative filtering by user interest expansion via personalized ranking," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 1, pp. 218–233, Feb. 2012.
[14] X. Liu, Y. Yu, C. Guo, and Y. Sun, "Meta-path-based ranking with pseudo relevance feedback on heterogeneous graph for citation recommendation," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 121–130.
[15] X. Liu, Y. Yu, C. Guo, Y. Sun, and L. Gao, "Full-text based context-rich heterogeneous network mining approach for citation recommendation," in *Proc. 14th ACM/IEEE-CS Joint Conf. Digit. Libraries*, Sep. 2014, pp. 361–370.
[16] X. Liu, J. Zhang, and C. Guo, "Citation recommendation via proximity full-text citation analysis and supervised topical prior," in *Proc. IConf.*, 2016, pp.1–6.
[17] S. M. McNee *et al.*, "On the recommending of citations for research papers," in *Proc. ACM Conf. Comput. Supported Cooper. Work*, 2002, pp. 116–125.
[18] F. Meng, D. Gao, W. Li, X. Sun, and Y. Hou, "A unified graph model for personalized query-oriented reference paper recommendation," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 1509–1512.
[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[20] L. Pan, X. Dai, S. Huang, and J. Chen, "Academic paper recommendation based on heterogeneous graph," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Cham, Switzerland: Springer, 2015, pp. 381–392.
[21] S. Pan, J. Wu, X. Zhu, C. Zhang, and Y. Wang, "Tri-party deep network representation," *Network*, vol. 11, no. 9, p. 12, 2016.
[22] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The ACL anthology network corpus," *Lang. Resour. Eval.*, vol. 47, no. 4, pp. 919–944, 2013.
[23] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 17–37, Jan. 2017.
[24] J. Tang and J. Zhang, "A discriminative approach to topic-based citation recommendation," in *Pacific–Asia Conference on Knowledge Discovery and Data Mining*. Berlin, Germany: Springer-Verlag, 2009, pp. 572–579.
[25] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl, "Enhancing digital libraries with TechLens," in *Proc. Joint ACM/IEEE Conf. Digit. Libraries*, Jun. 2004, pp. 228–236.
[26] L. C. Totti, P. Mitra, M. Ouzzani, and M. J. Zaki, "A query-oriented approach for relevance in citation networks," in *Proc. 25th Int. Conf. Companion World Wide Web*, 2016, pp. 401–406.
[27] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 448–456.
[28] J. D. West, I. Wesley-Smith, and C. T. Bergstrom, "A recommendation system based on hierarchical clustering of an article-level citation network," *IEEE Trans. Big Data*, vol. 2, no. 2, pp. 113–123, Feb. 2016.
[29] P. Wittek and S. Darányi, and G. Nelhans, "Ruling out static latent homophily in citation networks," *Scientometrics*, vol. 110, no. 2, pp. 1–13, 2017.
[30] H. Wu, J. He, Y. Pei, and X. Long, "Finding research community in collaboration network with expertise profiling," in *Proc. Int. Conf. Intell. Comput.*, 2010, pp. 337–344.
[31] Z. Wu, G. Gao, Z. Bu, and J. Cao, "Simple: A simplifying-ensembling framework for parallel community detection from large networks," *Cluster Comput.*, vol. 19, no. 1, pp. 211–221, 2016.
[32] 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.
[33] F. Xia, H. Liu, I. Lee, and L. Cao, "Scientific article recommendation: Exploiting common author relations and historical preferences," *IEEE Trans. Big Data*, vol. 2, no. 2, pp. 101–112, Jun. 2016.
[34] 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, Jan. 2017.
[35] G. Xu, Z. Wu, Y. Zhang, and J. Cao, "Social networking meets recommender systems: Survey," *Int. J. Social Netw. Mining*, vol. 2, no. 1, pp. 64–100, 2015.
[36] G. Xu, Y. Zong, P. Jin, R. Pan, and Z. Wu, "KIPTC: A kernel information propagation tag clustering algorithm," *J. Intell. Inf. Syst.*, vol. 45, no. 1, pp. 95–112, 2015.
[37] R. Yan *et al.*, "Guess what you will cite: Personalized citation recommendation based on users preference," in *Asia Information Retrieval Symposium*. Berlin, Germany: Springer-Verlag, 2013, pp. 428–439.
[38] 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.
[39] D. Zhou *et al.*, "Learning multiple graphs for document recommendations," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 141–150.

LANTIAN GUO is currently pursuing the Ph.D. degree with the School of Automation, Northwestern Polytechnical University, China. His current research interests include big data, recommendation system, machine learning, and artificial intelligence.



XIAOYAN CAI received the Ph.D. degree from Northwestern Polytechnical University, China, in 2009. She was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, from 2009 to 2011. She is currently an Associate Professor with the School of Automation, Northwestern Polytechnical University. Her current research interests include document summarization, information retrieval, and machine learning.

FEI HAO received the Ph.D. degree in computer science and engineering from Soonchunhyang University, South Korea, in 2016. He is currently an Associate Professor with the School of Computer Science, Shaanxi Normal University, China. He has authored over 60 papers in international journals and conferences. His current research interests include social computing, ubiquitous computing, big data analysis and processing, and mobile cloud computing. He received five best paper awards from KISM 2012, GreenCom 2013, MUE 2015, UCAWSN 2015, and CUTE 2016.

DEJUN MU is the Director of Cyberspace Engineering Laboratory of Shaanxi Province, and the Director of China "985" University Key Laboratory of Control System Under Complex Network Environment. He is currently a Professor with the School of Automation, Northwestern Polytechnical University, China. His current research interests include control theory and applications, information theory, big data, and multi-source information fusion.

CHANGJIAN FANG is currently pursuing the Ph.D. degree with the School of Automation, Northwestern Polytechnical University, China. He is currently the Office Director of Jiangsu Provincial Key Laboratory of E-Business. His current research interests include social network analysis, data mining, and recommendation system.

LIBIN YANG received the Ph.D. degree from Northwestern Polytechnical University, China, in 2009. He was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, from 2009 to 2011. He is currently an Assistant Professor with the School of Automation, Northwestern Polytechnical University. His current research interests include information retrieval, computer network, and game theory.