

Received 27 January 2023, accepted 20 February 2023, date of publication 1 March 2023, date of current version 16 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3251189

RESEARCH ARTICLE

A Scientific Paper Recommendation Framework Based on Multi-Topic Communities and Modified PageRank

AGUNG HADHIATMA^{1,2}, AZHARI AZHARI², AND YOHANES SUYANTO²

¹Department of Informatics, Faculty of Science and Technology, Sanata Dharma University, Yogyakarta 55282, Indonesia

²Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

Corresponding author: Azhari Azhari (arison@ugm.ac.id)

ABSTRACT Personalized PageRank is a variant of PageRank, widely developed for citation recommendation. However, the personalized PageRank that works with a vast amount and rich scholarly data still results in information overload. Sometimes, junior scholars still need help to arrange queries quickly because of limited domain knowledge. Senior researchers need reference papers regarding a similar topic they intend to search for and related topics as a new insight. In this research, scientific citation recommendation aims to find the most influential papers with similar and related topics. Related topic papers in serendipitous perspectives are reference papers that are novel, diversified and unexpected to a user. The unexpectedness of recommended papers can be papers with different topics to queries but still relevant. To accomplish these challenges, we propose a framework of scientific citation recommendation with serendipitous perspectives. The framework includes feature extraction of an academic citation network, selection of multi-topic communities, and ranking papers in the selected multi-topic communities by modified PageRank. Papers in the chosen communities tend to link to similar and related papers. Modified PageRank is an extension of personalized PageRank, which works on multi-topic communities and manuscript queries. The experiments reveal that the proposed models outperform some models of personalized PageRank and some models of Content-Based Filtering. The multi-topic communities-based models work more effectively than the baselines if they run in a large dataset since the topic communities become more cohesive.


INDEX TERMS Citation recommendation, academic citation network, serendipitous perspectives, multi-topic community, personalized PageRank.

I. INTRODUCTION

Growing vast and rich scholarly data resources such as paper journals and proceedings, theses, books, patents, and presentation slides lead to exciting research in data resource management, analysis, mining, and usage [1]. Scholars use academic data resources to find the most relevant reference papers to help them to conduct research and write an article. A classical retrieval system usually initiates with a keyword query, processes the request, and then retrieves the closest similar results to the query. However, keywords-based systems still return huge papers, sometimes with diverse

unrelated contents and semantic problems. The researchers still need much time to select which papers are suitable and significant to be comprehensively read and cited [2]. Hence, using a vast amount of scholarly data causes the information overload problem.

To address the information overload problem in scholarly data, a fundamental approach, namely citation recommendation, has been established to recommend a list of reference papers relevant to the researchers' information needs [3]. Citation recommendation methods are developed not only by a query of keywords but also by a set of manuscripts [4], a user's profile [5], and context [6]. There exist three citation recommendation approaches, which are collaborative filtering (CF), content-based filtering (CBF), and graph-based

The associate editor coordinating the review of this manuscript and approving it for publication was Fu Lee Wang .

filtering (GF). CF models produce citation recommendations using correlations among papers, researchers, or venues regarded as similar research interests [7]. The limitations of CF approaches are data sparsity [8] and cold-start problems [9]. CBF approaches recommend a list of reference papers by utilizing text processing and extracting text features [10]. However, CBF's challenge is semantic ambiguity [11]. Graph-based filtering (GF) approaches often portray citation recommendation as a link prediction [12] and perform effectively to rank influential papers. However, classical GF emphasizes the structure of an academic citation network but overlooks the content or context of text information in scientific papers as an essential part of the academic citation network.

Compared with other GF methods, PageRank as a Random Walk-based model has been increasingly popular due to its flexibility in incorporating additional contextual information [13]. PageRank has been widely extended for diverse graph topologies and combined with other citation recommendation methods [14], [15]. Many papers have explored a variety of PageRank methods to deal with scientific paper recommendations, such as personalized PageRank [16], random walk with restart (RWR) [17], and a mutual reinforcement method based on PageRank+HITS [18]. Original PageRank assumes that every node has a uniform weight and all links have equal values. Unlike the original PageRank, the random walker of personalized PageRank jumps to specific nodes with certain weights driven by a bias probability vector and transition probability matrix [13]. Some researchers employed personalized PageRank to modify a bias probability vector and transition probability matrix, which corresponds to texts [19], time [20], [21], semantic [22], topic [23], [24], [25], and time + topic [26], [27]. PageRank incorporates time and topic variables to deal with biased age and field [27]. Some works applied a mutual reinforcement method considering the diversified types of relations modeled as a heterogeneous graph of various entities, including authors, venues, and papers. Then, advanced paper ranking algorithms generate a ranking list for recommendations from a heterogeneous graph [21], [28]. Mu et al. [14] introduced a citation recommendation framework of personalized PageRank by accommodating query information into the mutual reinforcement schema to improve the mutual reinforcement methods. However, a variety of PageRank-based citation recommendation methods ignore serendipitous reference papers. The serendipitous recommended articles are reference articles with closely relevant, novel, diversified, and unexpected to a user's needs [29], [30]. The unexpectedness of recommended papers can refer to documents with dissimilar topics to queries but still relevant.

Considering serendipity in a scientific paper recommendation system is necessary to cope with challenging issues. Sometimes novice scholars still need help with quickly arranging keywords of queries to search for reference papers relevant to their needs because of their limitation of domain knowledge. Meanwhile, senior researchers need reference

papers regarding a similar topic they intend to search for and related topics as a new insight even if they do not consider it yet but significant to their research. On the other hand, many research publications reveal multi-topic papers and even multidisciplinary fields. Hence, a recommender system should suggest influential recommended articles with similar and related contents (topics) to a user's queries. The related contents of recommended papers are important, novel, diversified, and unexpected to a user's needs.

To deal with serendipitous issues, we propose a new citation recommendation framework named PPR_TC, which incorporates multi-topic communities into modified PageRank. Modified PageRank is an extension of personalized PageRank, which works on multi-topic communities and manuscript queries. The framework includes some stages: feature extraction of an academic citation network, selection of multi-topic communities, and ranking of the recommended paper candidates from selected communities by modified PageRank.

We hypothesize that utilizing the modified PageRank to rank reference paper candidates in a coherent and meaningful community may answer the citation recommendation issues and improve the recommender system's performance. Communities come up with a structure with a high number of intra-cluster links and a low number of inter-cluster links showing that the intra-cluster is denser than the inter-cluster [31]. The coherent community can contain homophily relations, i.e., a tendency to link to similar nodes, and heterophily relations, i.e., a likelihood to connect to related nodes. Related nodes may be reference papers that are unforeseen but relevant to a user. Meaningful communities can represent semantic relationships among words, sentences, and phrases. It can help users to better preview and organize papers into multi-topic categories. Meaningful communities are communities with multi-topic labels generated by a topic model.

The main contributions of this paper are summarized as follows.

1. We present feature extraction of an academic citation network by converting a scientific paper dataset into an academic citation network, detecting communities over the academic citation network, and identifying multi-topic communities and papers by LDA (Latent Dirichlet Allocation) topic model.
2. We propose a method to select and merge the n -best topic communities from a set of multi-topic communities. The recommended paper candidates in selected multi-topic communities are assumed to have specific characteristics with not only closely similar but also related to manuscript queries.
3. We propose a modified PageRank method to find the k -most influential recommended paper candidates of the selected multi-topic communities. Modified PageRank is a query-personalized PageRank, which works with topic queries and multi-topic communities, namely PPR_TC,

to calculate a bias probability vector and transition probability matrix.

- For evaluation, we conduct experiments on three variants of PPR_TC (PPR_TC_A, PPR_TC_B, PPR_TC_C) compared to the baselines on the subset DBLB dataset and the whole DBLB dataset. The baselines are some models of Content-Based Filtering and personalized PageRank based on non-topic communities.

The rest of this paper is organized as follows. Section II describes related work. Section III illustrates multi-topic communities-based personalized PageRank. Section IV evaluates and analysis the experimental results, and Section V presents the conclusion and future work.

II. RELATED WORKS

The states of the art citation recommendation methods include four main categories: Collaborative Filtering, Content-Based filtering, Graph-Based filtering, and PageRank-Based methods.

A. COLLABORATIVE FILTERING METHOD

The main idea of CF approaches is the process of recommending items by using the notions of other users through establishing the knowledge of users' preferences on items captured from relevant resources such as access logs [32], user profiles [33], and questionnaires [34]. The framework operates based on the assumption that if some users choose preferences on the same items, their interests will be considered closely similar. User preferences for items are a rating matrix of users-items or items-items used to predict the possibility of a new user's option. Mcnee et al. [35] proposed CF techniques to recommend papers using the rating matrix from the citation network. However, CF may lead to a problem in accurately finding similar users or items if the corresponding users or items have less experience or no rating, namely cold start problems [9].

B. CONTENT-BASED FILTERING METHODS

CBF approaches are concerned with textual content for recommending a set of reference papers, which works based on extracting text features by applying text processing [36], topic modeling [37], and text embedding [1]. The recommendation system finds similar relationships between the features between users and items and then recommends other items similar to the specific items that the users have preferred. For academic paper recommendation, CBF approaches can use some relevant entities from text data sources, including the author profile, the information of the paper, e.g., title, abstract, full text, the label, and the user behavior of browsing, reading, downloading, etc.

There are some models applied most in CBF to extract text features, e.g., TF-IDF, topic modeling, and document embedding. For example, Kazemi and Abhari [36] compared the efficiency and usability of the two feature extraction methods: TD-IDF and word embedding in abstract extraction. Amami et al. [37] proposed an academic paper

recommendation method based on the LDA topic model on the DBLP dataset. Nevertheless, these models have very little sense of the semantics of the words and do not consider the context information of a document (i.e., order of words). To face those challenges, neural network or deep learning-based approaches have introduced text or document embedding, such as Doc2vec [38]. Doc2vec method utilizes a simple neural network model to learn distributed vectors for texts. Unlike existing language representation models, Devlin et al. [39] introduced Bidirectional Encoder Representations from Transformers (BERT), designed by considering the importance of bidirectional pre-training for representing natural language derived from the unlabeled text. Jeong et al. [40] proposed a context citation recommendation model using BERT. However, the existing CBF methods work regardless of network exploration approaches.

C. GRAPH - BASED FILTERING

Recently, graph-based (GB) recommender systems have become more prevalent in the citation recommendation approach. Different from the other approaches, GB can incorporate link information of single layer or multilayer graph (paper-paper, paper-author, etc.) and text information of various entities (topic, author, venue, etc.) [41] into the recommendation system to achieve a better result. GB works mainly through two steps that are graph construction and recommendation generation. A graph constructed from a scholarly dataset can represent homogenous or heterogeneous citation networks. Then algorithms like Random Walk can generate citation recommendations from the network. Ali et al. [42] discuss various methods used in graph analysis and recommendation generation, such as Factorization, Probabilistic Topic Model, Neural Network, Clustering, and Random Walk. The random Walk-based process has been increasingly popular due to its flexibility in incorporating other contextual information [13]. In extensive scholarly data usage, Random Walk (RW) approaches have been proposed to evaluate the academic impact of publication papers [21], [28] and to recommend academic papers to find out valuable and relevant published articles related to their current work [4], [14].

D. PAGERANK-BASED METHODS

The most special random walk-based algorithm in computer science areas is PageRank, introduced by [43]. PageRank, by default, has some limitations, in which it assumes that every node has a uniform value and all links have equal weight, and it also considers the global information of a graph. Many papers have explored a variety of Random Walk models to deal with its limitations, such as personalized PageRank [44], Random Walk with Restart (RWR) [17], mutually reinforcement method based on PageRank + HITS [18], and Personalized PageRank with edge weights [45].

RWR is a random walk with one added formulation: restart probability vector. A restart probability vector leads each

random walker of RWR taken in any direction to have a probability associated back to the initial starting position. Zhou et al. [46] introduced extended RWR, which specifies the time-dependent academic influence in a specific social context so that users can navigate research collaboration information to support upcoming works. This method is evaluated in the scope of the activeness of researchers and the popularity of scientific articles. Xia et al. [47] developed MVC Walker, an innovative and new method of RWR for recommending collaborators to scholars. This RWR method utilizes three academic variables, i.e., coauthor order, latest collaboration time, and times of collaboration, to define link importance in academic citation networks to raise the recommendation performance. Most previous methods only focus on one type of relationship but neglect to explore the mutual interaction among different relationships.

Mutual reinforcement has become a popular method in academic citation networks. This method's performance is based on an interaction between authority and hub entities from multiple typed links represented in a heterogeneous network [15]. Wang and Tang [18] combined the method of PageRank and HITS as a mutual style reinforcement method to rank papers by utilizing citations, authors, venues, and publication time. Zhao [28] introduced a novel ranking method, APR (Author-PageRank), which applies to heterogeneous academic networks. The Mutual reinforcement-based methods have achieved better performances than previous methods, but sometimes the results still need improvement [14]. Mu et al. [14] introduced a citation recommendation framework of personalized PageRank, which accommodates query information into the mutual reinforcement schema to achieve a more accurate result. Nevertheless, for large graphs such as an academic citation network, these mutual reinforcement-based methods face some challenges in meeting the requirements of most applications, including computational time and accuracy.

Recently, personalized PageRank (PPR) has been widely utilized in diverse computer science fields, for example, information retrieval, recommendation, and knowledge discovery [44]. Unlike the original PageRank, the random walker of PPR jumps to the specific nodes driven by a bias probability vector. A probability vector can be a node with a proximity measure, which characterizes the degree of closeness toward the user's query among nodes within a graph. To face this issue, some researchers employed query-dependent PageRank to find essential nodes in the graph which correspond to text or topic similarity. Qiao et al. [48] proposed a weighted page rank algorithm that considers the relevance of a page to the given query, which can improve the ranking accuracy. Roul and Sahoo [16] designed query-optimized PageRank approach by incorporating the TF-IDF and personalized PageRank method to produce robust ranking web pages. But, query-dependent PageRank treats all citations with equal weights ignoring the wide variety of functions that citations perform.

In academic recommendation, the citation relations between papers may vary in motivation [49] and intensity [50]. Considering the citation functions, some researchers developed personalized PageRank rank papers for recommendation by modifying the bias probability vector and the transition probabilities matrix. Wei et al [19] and Xie et al [45] computed the relationship strength between each pair of nodes by utilizing a text similarity approach to generate a directed and weighted network and then incorporate it into the PageRank algorithm. Dunaiski et al. [51] observed that personalizing PageRank with citation counts of papers decreases time bias but increases topic bias. However, The personalized PageRank based on similarities does not consider a mixture of various topics.

A topic consists of a cluster of words frequently occurring together, referring to a specific context. Some researchers utilize topic modeling to improve the PageRank matrix to receive more relevant and essential out-links and to provide a simple way to analyze large volumes of unlabeled text. Yang et al. [23] first explored Topic Modeling with LDA (Latent Dirichlet Allocation) to automatically extract topics from scientific papers and then combine them with PageRank to calculate the topic-dependent scores of papers. Ding [52] incorporated the Author Conference Topic Model (introduced by Tang et al. [24]) with a weighted PageRank to rank authors in each author's topic distribution. Jardine and Teufel [26] extended the bias and transition probabilities of PageRank, namely TPM, by considering topic distributions extracted from papers to predict scientific papers and approximate the research fields. Zhang et al. [27] presented a modified PageRank algorithm called CTPM. Different from the TPM, the CTPM implements the Correlation Topic model to extract scientific topics and their correlations. Then, the CTPM uses the topic proportions, the prestige of the paper's venue, and the correlations of scientific topics to modify a bias probability vector and a transition probability matrix of PageRank to evaluate the academic impact of papers for each extracted scientific topic. Tao et al. [25] established a paper recommendation system based on LDA, word2vec, doc2vec, and PageRank. LDA is used to calculate a probability distribution and extract topic paper keywords. Word2vec is implemented to represent a topic vector. Doc2vec is applied to describe the paper vector. And finally, the PageRank algorithm uses to retrieve the recommendation results.

However, these PageRank variants are not designed to handle serendipitous recommendations. Serendipitous items refer to items that are relevant, novel, and unexpected to a user [29], [30]. In a scientific paper recommendation, the serendipity aspect is essential because users need to find papers that are not only similar and influential but also novel, unexpected, and diversified.

Unlike the previous works, we constructed a citation recommendation framework based on multi-topic community detection and modified PageRank to deal with serendipitous recommended papers. We assume that multi-topic

communities-based a modified PageRank method can result in better performance for citation recommendation because the multi-topic communities contain homophily and heterophily relations to a user's manuscript query. Also, selected multi-topic communities may address the semantic problems and handle sparsity and scalability in big data to lower complexity.

III. MULTI-TOPIC COMMUNITY-BASED PERSONALIZED PAGERANK

The scientific paper recommendation framework shown in Fig. 1 consists of the following main steps: (1) extracting graph features, (2) retrieving recommended paper candidates through selecting communities and merging the selected communities, (3) ranking the recommended paper candidates based on modified PageRank to produce top k-recommended papers. Some notations will be used frequently to elaborate the proposed model in Table 1.

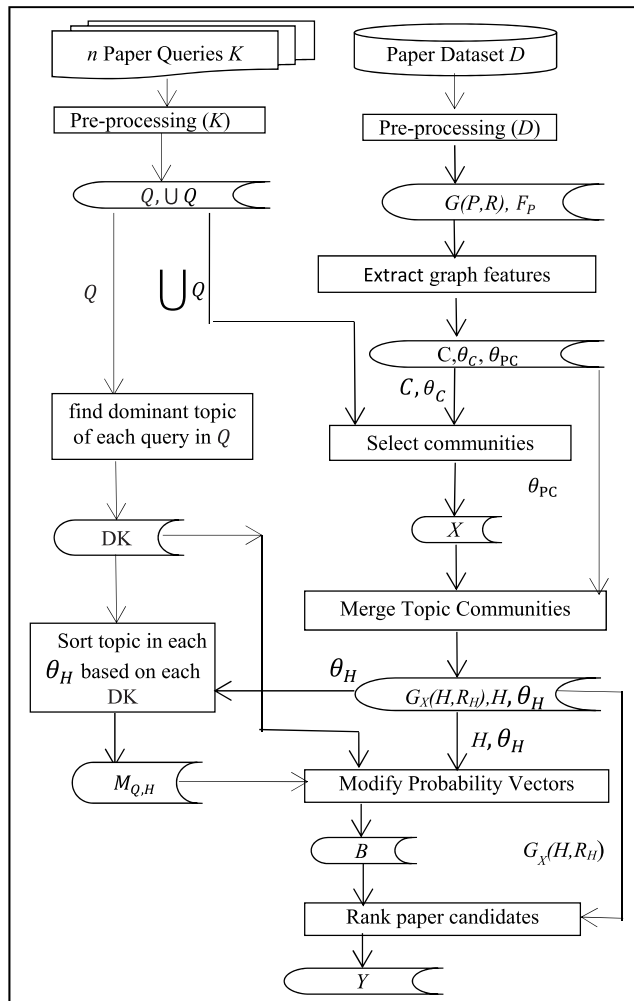


FIGURE 1. The framework of scientific citation recommendation.

A. EXTRACTION OF GRAPH FEATURES

The pre-processing dataset converts the DBLB dataset (title + abstract + citations) into an academic citation

TABLE 1. Notations.

Notations	Description
K	The set of manuscript queries
D	The set of text papers: title, abstract and citations
$G(P,R)$	The graph of Academic Citation Network
P	$P =$ The set of Bag of Word (BoW) text papers (title + abstract), $P = \{p_1, p_2, p_3, \dots, p_m\}$
R	The set of relations from p_i to p_j
F_p	The set of TF-IDF paper vectors
Q	The set of paper BoW queries $= \{q_1, q_2, q_3, \dots, q_b\}$.
TC	The topic community features: $C, \theta_{PC}, \theta_c, \varphi_T$
C	The set of communities, $C = \{c_1, c_2, c_3, \dots, c_n\}$ $C = \{G_1(P_1, R_1), G_2(P_2, R_2), \dots, G_n(P_n, R_n)\}$
θ_c	The set of community-topic distributions in C
θ_{PC}	The list of the set of paper-topic distributions in each $c_i \in C$,
φ_T	The set of topic-word proportions in C
DK	The set of dominant topics for each query $q_i \in Q$
X	The set of selected Communities
$G_x(H, R_H)$	The merged graph from the j -best topic communities in X
H	The set of filtered papers = papers in graph $G_x(H, R_H)$
θ_H	The set of paper-topic distributions in H
b	The personalized vector of PageRank
B	The list of vector b for all queries $q \in Q$
$M_{H,Q}$	The set of papers with sorted topic of θ_H for each $q_i \in Q$
Y	The top k-recommended papers

network consisting of a citation network $G(P,R)$ and a set of paper vectors F_p . A Graph $G(P, R)$ consists of P and R , where P is a paper corpus obtained from the process of tokenization, word removal, lemmatization, and bigram calculations of title + abstract in the DBLB dataset, $P = \{p_1, p_2, p_3, \dots, p_m\}$, $m =$ total text papers. A variable R is a set of relations connected from citing papers p_i to cited papers p_j , $R = \{(p_i, p_j) = \text{rel} | p_i, p_j \in P\}$, where $\text{rel} = 1$ if has reference to p_j and $\text{rel} = 0$ otherwise. Paper corpus P is calculated and converted to TF-IDF vectors defined as $F_p = \{f_1, f_2, f_3, \dots, f_m\}$.

The next stage is feature extraction from an academic citation network, including community detection on the academic citation network, generating a topic model, and identifying multi-topic on the communities and papers. Figure 2 describes the feature extraction model on an academic citation network.

Community detection on a scientific citation network is the process of grouping a citation network into a set of communities C , in which intra-cluster is denser than inter-cluster, $C = \{G_1(P_1, R_1), G_2(P_2, R_2), \dots, G_n(P_n, R_n)\}$, $i = 1, \dots, n$ whole communities. Community detection on a graph $G(P, R)$ is constructed by optimizing the Newman Modularity Equation shown in Eq. 1

$$Q_{Newman} = \sum_{k=1}^n \sum_{i \in C_k, j \in C_k} \frac{1}{2m} (r_{ji} - \frac{d_j d_i}{2m}) \quad (1)$$

$\frac{1}{2m} (r_{ji} - \frac{d_j d_i}{2m})$ = a link strength, $m = |R|$, d_i = the node degree of cited paper p_i , d_j = the node degree of citing paper p_j , n = the number of communities, r_{ji} = link from p_j to p_i , $r_{ji} = 1$ if it has a reference and $r_{ji} = 0$ otherwise. To identify community using the optimization of Newman Modularity,

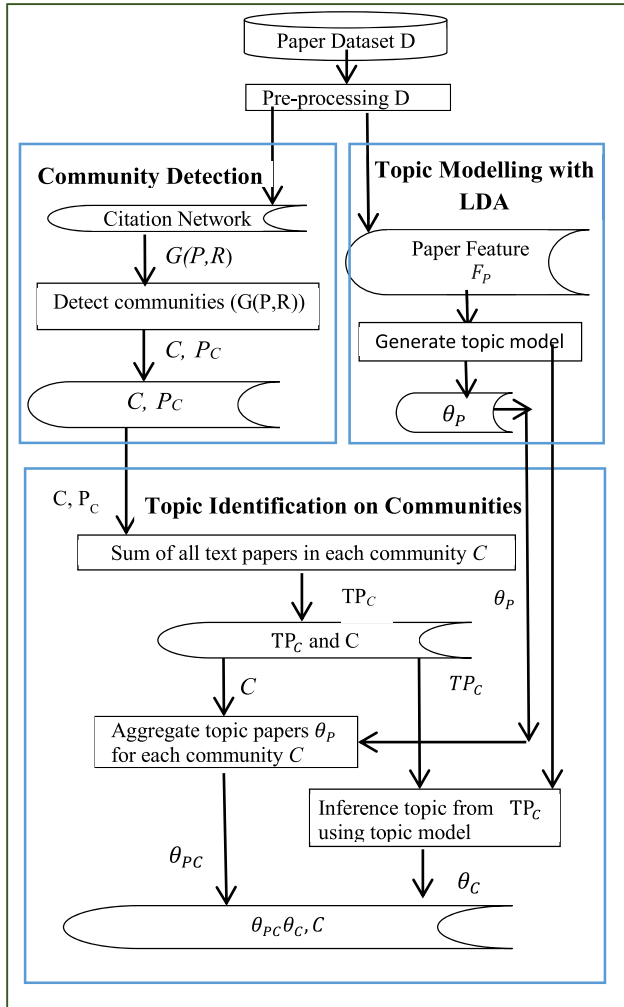


FIGURE 2. The model of feature extraction from an academic citation network.

we use the Louvain'' algorithm based on a modularity measure with a hierarchical approach [53]. A community C_i consists of $P_i =$ a group of papers in a community i , and all groups of papers for all communities are $P_c = \{P_1, P_2, P_3, \dots, P_n\}$, $n =$ the total number of communities. All text papers in a community i are summed, defined as tp_i , and for all communities defined as $TP_c = \{tp_1, tp_2, tp_3, \dots, tp_n\}$.

Topic identification on communities is by inferring each $tp_i \in TP_c$ using LDA (Latent Dirichlet Allocation) topic model to generate a set of community-topic distributions, $\theta_C = \{\theta_{C1}, \theta_{C2}, \theta_{C3}, \dots, \theta_{Cn}\}$,

The topic model by LDA (Latent Dirichlet Allocation) is a generative statistical model for discovering a set of paper-topic distributions θ_p and a set of topic-word distributions φ_T . This θ_p and φ_T are assumed as a Dirichlet distribution with parameters α and β , respectively, considering m papers and u topics, where w (word) and z (topic) are subject to multinomial distributions. Gibbs sampling is used for estimating LDA parameters θ_p and φ_T as latent feature vectors. The generation process of LDA can be described as follows:

1. Choose $\theta_{p \in P} \sim \text{Dirichlet}(\alpha)$, $p \in \{1, 2, \dots, m \text{ papers}\}$.
2. Choose $\varphi_{t \in T} \sim \text{Dirichlet}(\beta)$, $t \in \{1, 2, \dots, u \text{ topics}\}$.
3. For each of the word positions $i, j \in \{1, 2, \dots, m \text{ papers}\}$ and $j \in \{1, 2, \dots, n \text{ word length}\}$.

- 1) Choose a topic $z_{ij} \sim \text{Multinomial}(\theta_i)$
- 2) Choose a word $w_{ij} \sim \text{Multinomial}(\varphi_{z_{ij}})$.

A set of paper-topic distributions θ_p then is grouped according to each community to become $\theta_{pC} =$ a list of a set of paper-topic distributions for each community.

So, given $G(P,R)$, $u =$ the number of topics, and a set of TF-IDF paper vectors F_p , the feature extraction model for an academic citation network produces four variables C, θ_C, φ_T , and θ_{pC} . Figure 3 shows an illustration of academic citation network features. There are communities $C = \{G_1(P_1, R), G_2(P_2, R)\}$ with community-topic distributions $\theta_C = \{\theta_{C1}, \theta_{C2}\}$ and $\theta_{pC} = \{\theta_{pC1}, \theta_{pC2}\}$. Community $G_1(P_1, R)$ has $P_1 = \{p_1, p_2, p_4\}$ with paper-topic distributions $\theta_{pC1} = \{\theta_{p1}, \theta_{p2}, \theta_{p4}\}$ and Community $G_2(P_2, R)$ has $P_2 = \{p_3, p_5, p_6\}$ with paper-topic distributions $\theta_{pC2} = \{\theta_{p3}, \theta_{p5}, \theta_{p6}\}$.

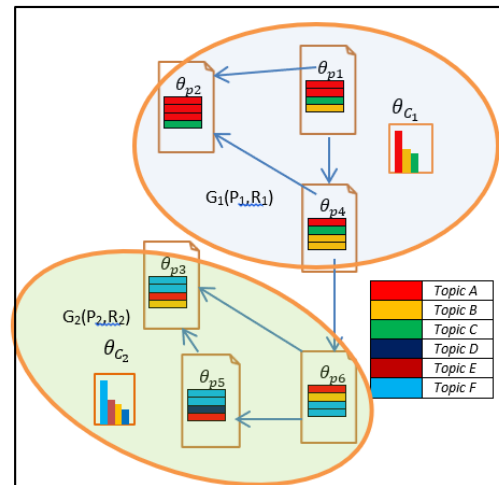


FIGURE 3. Illustration of academic citation network features.

B. SELECTION OF COMMUNITIES

Recommended paper candidates are retrieved by selecting communities from a set of communities C and merging the selected communities, as shown in Algorithm 1.

The process of finding recommended paper candidates requires some variables, which are a set of community-topic distributions θ_C , a set of manuscript queries Q , and a set of paper-topic distributions for each community $\in \theta_{pC_i}$. The selection of topic communities aims to find the j -best communities which are the most relevant to a set of manuscript queries Q in a multi-topic aspect.

The community selection needs the n -dominant topic proportion of query defined as a vector e and a sorted topic distribution in each community defined as a vector $y_{c_i \in C}$. A set of all sorted topic distributions of communities defined

Algorithm 1 Community Selection**Input:** $C, U, Q, \theta_c, \theta_{PC_i}$ **Output:** $G_X(H, R_H), H, \theta_H$

1. $ts \leftarrow$ Inference a query- topic proportion using Topic Model on U, Q
2. $e \leftarrow$ rank the n -topic proportions of ts
3. $Y_C \leftarrow$ arrange topic distributions for each $\theta_{c_i \in C}$ according to $e, i =$ from 1 to the total number of communities
4. // Filter communities
for each community c_i in C :
 $dt_i =$ get the first dominant topic of $y_{c_i \in C}$
 $S \leftarrow$ Choose $c_i \in C$ if $(dt_i == e[0]) || (dt_i == e[1]) || (dt_i == e[3])$
end for
5. // Select communities
 $X \leftarrow$ Select the j - best communities in S using Janson Shannon similarity $Sim(e, y_{c_i \in S})$, by Eq 2
6. // Merge communities
 $G_X(H, R_H) \leftarrow$ Merge X
7. H is a set of papers P in $G_X(H, R)$
8. $\theta_H \leftarrow$ select paper-topic distributions from all θ_{PC_i} , where each $p_i \in H$
9. Return $G_X(H, R_H), H, \theta_H$

as Y_C . A vector e is formed by ranking the topic proportion of vector ts in descending order. A vector ts is a query–topic proportion, inferred from accumulated manuscript queries $\sum_{i=1}^{i=b} q_i$ by the LDA topic model. Each vector $y_{c_i \in C}$ in Y_C is obtained by sorting the topic proportion of each vector $\theta_{c_i \in C}$ in θ_c by following the rank order of n - dominant topic proportion of query e .

The selection of the j -best topic communities consists of three stages: (1) communities are filtered in C if the first dominant topic community in Y_C equals $e[0]$ or $e[1]$ or $e[2]$. The set of filtered communities is defined as S . (2) the j -best communities are selected from filtered communities S by calculating a similarity between a dominant topic of query-topic proportion e and each sorted topic of community-topic distribution $y_{C_i \in S}$ using Jensen-Shannon Divergence (JSD) shown in Eq.2. JSD is a method for measuring the similarity between two probability distributions.

$$\begin{aligned}
 sim(y_{C_i \in S}, e) &= JSD(y, e) \\
 &= \frac{1}{2} KLD(y||u) + \frac{1}{2} KLD(e||u) \\
 \text{With } u &= \frac{1}{2}(y + e); KLD(y||u) = \sum_{i=1}^n y_i \ln \frac{y_i}{u_i}; \\
 KLD(e||u) &= \sum_{i=1}^n e_i \ln \frac{e_i}{u_i} \quad (2)
 \end{aligned}$$

(3) the j -best topic communities then are merged in one graph, namely graph $G_X(H, R_H)$.

So the set of recommended paper candidates H is retrieved from $G_X(H, R_H)$. θ_H is the set of topic distributions of recommended paper candidates in H .

C. CALCULATION OF A PROBABILITY VECTOR OF PAGERANK

PageRank working in an academic citation network can be described as the following recursive formulation.

$$p_i^{t+1} = (1 - \alpha) \frac{1}{m} + \alpha \sum_{j=1}^r \frac{p_j^t}{d_j} \quad (3)$$

where m is the total number of papers, p_j^t is a weight of a citing paper i at time t , p_i^{t+1} is a weight of a cited paper j at time $t + 1$, d_j is the out-degree of a citing paper and a damping factor $\alpha \in [0, 1]$. PageRank is formally defined as the stationary distribution of a random walk process over the graph shown in Eq.4. The formulation combines the paper node states and their transition probability matrix T constructed from an academic citation network. The first term of $\frac{1}{m} \vec{1}$ represents the uniform probability distributions of node states, and the second term of $T r$ represents the transition probability from the current state.

$$r = \frac{1}{m} (1 - \alpha) \vec{1} + c T r \quad (4)$$

The combination makes the formulation irreducible and aperiodic to find the stationary distribution. The combination depends on a dumping factor α .

The difference between the original PageRank and Personalized PageRank (PPR) is the first term of the formulation where $\frac{1}{m} \vec{1}$ is replaced with b , described as follows:

$$r = (1 - \alpha) b + c T r \quad (5)$$

PageRank assumes that the random walker returns to any nodes with uniform probability, whereas PPR defines that the random walker randomly returns to specific nodes with a state probability b , namely a personalized vector. A personalized vector can be defined as selected specific papers p_n with probability states w_n forming a vector $b = [w_{1,p_1}, w_{2,p_2}, \dots, w_{n,p_n}]$.

In this research, modified PageRank is a query-personalized PageRank applied with topic query and multi-topic community, namely the PPR_TC approach. Different from other personalized PageRank approaches, the PPR_TC approach does not function to the entire citation network $G(P, R)$, but to the merged graph $G_X(H, R_H)$ to modify an vector b and an transition probability matrix T . We propose PPR_TC with three variant models, which are PPR_TC_A, PPR_TC_B, and PPR_TC_C. They have three different approaches for determining specific nodes p_n of personalized vector \vec{b} .

PPR_TC_A accommodates all papers H in merged graph $G_X(H, R_H)$ as specific nodes p_n . Vector b of PPR_TC_A is specified as $[w_{1,p_1}, w_{2,p_2}, \dots, w_{n,p_n}]$, each $p_n \in H$ with w_n . A variable w_n is a similarity weight between each paper candidate feature $f_{p \in H}$ and of a query feature $k_{q \in Q}$ as shown in Eq.6.

Meanwhile, PPR_TC_B and PPR_TC_C need to filter recommended paper candidates from H to identify specific

nodes p_n .

$$\begin{aligned} & \text{TF-IDF cosine sim}(\mathbf{k}_{q \in Q}, \mathbf{f}_{p \in H}) \\ &= \frac{\mathbf{k} \cdot \mathbf{f}}{|\mathbf{k}| |\mathbf{f}|} \frac{\sum_{i=1}^n k_i f_i}{\sqrt{\sum_{i=1}^n k_i^2} \sqrt{\sum_{i=1}^n f_i^2}} \end{aligned} \quad (6)$$

PPR_TC_B utilizes θ_H , a set of paper-topic distributions in H, and TK_Q , a set of query-topic distributions to filter specific papers A_B from H to form \mathbf{b} . PPR_TC_B generates $\mathbf{b} = [w_{1,p_1}, w_{2,p_2}, \dots, w_{n,p_n}]$, each $p_n \in A_B$ with its probability weight w_n , $A_B =$ set of specific papers of PPR_TC_B, w_n is a weight calculated from TF-IDF cosine sim $(\mathbf{k}_{q \in Q}, \mathbf{f}_{p \in A_B})$. Each specific node $p_n \in A_B$ of PPR_TC_B is selected from H if the value of the similarity between each multi-topic query $\mathbf{tk}_{q \in Q}$, and each multi-topic paper $\theta_{p \in H}$ is higher than a tuned threshold trB in the range [0..1]. To calculate a multi-topic similarity between $\mathbf{tk}_{q \in Q}$ and $\theta_{p \in H}$, the PPR_TC_B model uses topic cosine sim formulation in Eq.7.

$$\begin{aligned} \text{topic cosine sim}(\mathbf{tk}_{q \in Q}, \theta_{p \in H}) &= \frac{\mathbf{tk} \cdot \theta}{|\mathbf{tk}| |\theta|} \\ &= \frac{\sum_{i=1}^n \text{tk}_i \theta_i}{\sqrt{\sum_{i=1}^n \text{tk}_i^2} \sqrt{\sum_{i=1}^n \theta_i^2}} \end{aligned} \quad (7)$$

PPR_TC_C utilizes a set of n -dominant topics of queries defined as DK_Q and a set of sorted topic distribution for all recommended paper candidates defined as $M_{H,Q}$ to filter specific papers A_C from H to form \mathbf{b} described in Algorithm 3. Variable DK_Q and $M_{H,Q}$ are generated by Algorithm 2. Variable $m_{q \in Q, p \in H}$ is sorted topics in paper-topic distributions $\theta_{p \in H}$ composed in regard with n -dominant topics of a query $dk_{q \in Q}$. PPR_TC_C generates \mathbf{b} represented in Eq.8.

$$\mathbf{b} = [w_{1,p_1}, w_{2,p_2}, \dots, w_{n,p_n}] \quad (8)$$

w_{n,p_n} is a probability weight of n^{th} paper corresponding to $p_n \in A_C$. A variable w_{n,p_n} is calculated from TF-IDF cosine sim $(\mathbf{k}_{q \in Q}, \mathbf{f}_{p \in A_C})$ in Eq. 6. Each specific node $p_n \in A_C$ of PPR_TC_C is selected from H if the level of a multi-topic similarity between each $\mathbf{dk}_{q \in Q}$, and each $\mathbf{m}_{p \in H, q \in Q}$ is lower than a tuned threshold trC in the range [0..1]. To calculate a multi-topic similarity between each $\mathbf{dk}_{q \in Q}$, and each $\mathbf{m}_{p \in H, q \in Q}$, PPR_TC_C uses Jensen Shannon Divergence (JSD) shown in Eq.9. JSD is formulated based on the Kullback–Leibler Divergence (KLD). JSD is symmetric and always has a finite value. A variable B is a list of probability vectors \mathbf{b} for all $q \in Q$.

$$\begin{aligned} \text{sim}(\mathbf{dk}_{q \in Q}, \mathbf{m}_{p \in H, q \in Q}) &= \text{JSD}(\mathbf{dk}, \mathbf{m}) \\ &= \frac{1}{2} \text{KLD}(\mathbf{dk} || \mathbf{u}) + \frac{1}{2} \text{KLD}(\mathbf{m} || \mathbf{u}) \\ \text{with } \mathbf{u} &= \frac{1}{2}(\mathbf{dk} + \mathbf{m}); \text{KLD}(\mathbf{dk} || \mathbf{u}) \\ &= \sum_{i=1}^n dk_i \ln \frac{dk_i}{u_i}; \text{ and} \\ \text{KLD}(\mathbf{m} || \mathbf{u}) &= \sum_{i=1}^n m_i \ln \frac{m_i}{u_i} \end{aligned} \quad (9)$$

Algorithm 2 Topic Sorting of Recommended Paper Candidates (PPR_TC_C)

Input: Q, H, θ_H

Output: $DK_Q, M_{Q,H}$

1. $TK_Q \leftarrow$ Inference query-topic distributions for each $q \in Q$ using LDA Topic Model
 2. //Find n -dominant topics of each query, $DK_Q = \{(tq, v), tq \in T, v = \text{proportion value}\} \leftarrow$ sort topic of each $tk_{q \in Q}$
 3. Set n
 4. //Arrange $M_{Q,H}$, a topic rank of papers according to DK_Q for all of $q \in Q$:
 - for** all of $p \in H$
 - $m_{q \in Q, p \in H} \leftarrow$ Arrange topic rank of $\theta_{p \in H}$ according to $dk_{q \in Q}$
 - end for**
 5. Return $DK_Q, M_{Q,H}$
-

Algorithm 2 generates n -dominant topic queries DK_Q and sorted topics of recommended paper candidates $M_{Q,H}$ to modify probability vector \mathbf{b} in Algorithm 3.

Algorithm 3 Calculation of Probability Vector $\mathbf{b} \in B$ (PPR_TC_C)

Input: $Q, F_P, DK_Q, M_{Q,H}$

Output: B, A_C

// Calculating personalized vector B, \vec{b} is a personalized

// vector for $q \in Q$, B is a set of \mathbf{b} for all $q \in Q$

1. Set trC
 2. $A_C = \text{list}\{\}$
 3. $B = \text{list}\{\}$ //set B as a list
 4. $K_Q \leftarrow$ Calculate tf-idf vector for each $q \in Q$
 5. **for** each $q \in Q$: // calculate \mathbf{b} for each query
 6. $\mathbf{b} = \text{list}\{\}$ //set \mathbf{b} as a list
 7. **for** each $p \in H$: // for each paper candidate
 8. $f_{p \in H} \leftarrow$ select TF-IDF vector from F_P where $p \in H$
 9. $w \leftarrow \text{Sim}(\mathbf{k}_{q \in Q}, \mathbf{f}_{p \in H})$ by Eq. 6
 10. $d2 \leftarrow \text{Sim}(\mathbf{dk}_{q \in Q}, \mathbf{m}_{q \in Q, p \in H})$ by Eq. 9
 11. **if** ($d2 < \text{trC}$): //filter a specific node p with its weight w
 12. Get p
 13. $A_C.append(p)$
 14. Get w
 15. $\mathbf{b}.append((p, w))$
 16. **End if** //end of filtering
 17. **end for**
 18. $B.append(\mathbf{b})$
 19. **end for**
 20. Return B, A_C
-

D. RANKING RECOMMENDED PAPER CANDIDATES

The modified PageRank ranks recommended paper candidates H in a topic community $G_X(H, R_H)$, resulting in the top k recommended papers y by Algorithm 4.

A variable y is the top k recommended papers regarding a manuscript query q . A variable y represents a set of tuples = $\{(p, v), p \in H, v = \text{a weight proportion sorted}$

Algorithm 4 Ranking of Recommended Paper Candidates (Modified PageRank)**Input:** $B, G_X(H, R_H)$ **Output:** Y (k -recommended papers for all Q)

```

1. Set  $\alpha = 0.5$ , Set  $\beta$  convergence tolerance = 0.001
2.  $T \leftarrow$  convert  $G_X(H, R_H)$  to a transition matrix
3.  $r_0 \leftarrow$  initial  $1 \times |H|$  vector at iteration 0 ( $\frac{1}{|H|}$ )
   //set to  $\{\frac{1}{|H|}, \frac{1}{|H|}, \dots, \frac{1}{|H|}\}$ 
4.  $Y = []$  // set to list  $Y$ 
5. For  $i$  in 1 to  $|Q|$ 
6.    $b = B[i]$ 
7.    $r \leftarrow r_0$ 
8.   while residual  $< \beta$  do
9.      $r_{pr} \leftarrow r$ 
10.     $r = (1 - \alpha) * b + \alpha * T * r_{pr}$  // (Eq.5.)
11.    residual  $\leftarrow ||r - r_{pr} ||$ 
12.  end while
13.   $y = \{(p, v), p \in H, v = \text{proportion value}\} \leftarrow$  sort ( $r$ ) of  $H$ 
14.   $Y.append(y)$ 
15. End for
16. return  $Y$ 

```

from vector r }. A vector r is a probability distribution used to represent the likelihood of essential papers in a selected topical community. A vector r is calculated using the formulation by Eq.5 in the modified PageRank method described in Algorithm 4. Calculating r requires iterations until the result is convergent. The formulation to calculate r is constructed by modifying a bias probability vector $b \in B$ generated from Algorithm 3 and modifying transition probability matrix T converted from the graph $G_X(H, R_H)$ produced from Algorithm 1. A variable Y is a list of the top- k recommended papers y for all $q \in Q$.

IV. EXPERIMENT AND EVALUATION

In this section, we present the dataset, evaluation methods, and series of experiments to evaluate the performances of the three proposed citation recommendation models (PPR_TC_A, PPR_TC_B, and PPR_TC_C) and the baselines on The DBLB dataset (Citation-network V4).

A. DATASET

We use the DBLP-Citation-network dataset, which contains bibliography data in computer science fields. The DBLP-Citation-network dataset is a citation dataset that was collected, extracted, and introduced by Tang et al. [54]. We cleaned up the DBLB dataset version 4 (V4) with missing abstracts, titles, and citations and conducted some experiments by taking a subset dataset (46,870 papers) and a whole dataset (653,506 papers) representing the large-scale data. The subset DBLP dataset consists of five groups of computer science fields, including information retrieval (ACL, ECIR, SIGIR, COLING, and NAACL), machine learning (WSDM, ICML, ICDE, SIGKDD, and NIPS), computer vision (ACCV, CVPR, ICCV, ECCV, and ICIP), computer security (ARES, NDSS, ISI, SP, and FC) and networks and communication

(INFOCOM, ICC, SIGCOMM, GLOBECOM, and MOBICOM). Meanwhile, the large-scale dataset constitutes all journals and proceeding papers in all computer science fields. The papers published before 2010 are considered the training set, and the papers published from 2010 to 2011 are considered the testing dataset. From the testing dataset, we randomly captured a set of $50 \times 4 = 200$ papers from 4 groups of computer science fields (information retrieval, machine learning, computer vision, networks, and communication), which are considered manuscript query tests used to evaluate citation recommendation methods.

B. EVALUATION METRICS

In this research, given manuscript query tests q , the proposed models and the baselines are evaluated to return k top recommended papers from the DBLB dataset. The citation lists in manuscript query tests q are used as the ground truth. We tested three proposed models and the six baselines using the following metrics Recall@ k , MRR, and MAP, which are widely used in citation recommendation methods.

Recall in Eq.10 is a metric that measures the percentage of retrieved relevant citations in ground truth lists. Ground truth lists are the whole references in n -given manuscript query tests. Precision metric in Eq.11 measures the percentage of retrieved relevant citations in top- k recommendation lists generated by a proposed model for n -given paper query tests.

$$Recall@k = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{|Y_{q_i, k} \cap I_{q_i}|}{I_{q_i}} \quad (10)$$

$$Precision@k = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{|Y_{q_i, k} \cap I_{q_i}|}{k} \quad (11)$$

I_{q_i} is a set of references belongs a manuscript query test q_i that are considered as ground truth. $Y_{q_i, k}$ is the top k papers recommended by a proposed model given Q . A variable Q is the set of manuscript query tests. Different from Recall@ and Precision, MAP and MRR are precision metrics that consider the ranking position of retrieved relevant citations.

$$AP_{q_i}@k = \sum_{q_i \in Q} \frac{\sum_{k=1}^n Precision@k \times IF(k)}{I_{q_i}} \quad (12)$$

MAP metric is calculated in Eq.13 where AP is formulated in Eq.12 and $IF(k)$ is an indicator function, which equals 1 if a recommended reference paper is at position k in the reference list of a paper query test and equals 0 otherwise.

$$MAP@k = \frac{1}{|Q|} \sum_{q_i \in Q} AP_{q_i}@k \quad (13)$$

MRR in Eq.14 evaluates how far the first relevant reference papers are from the top, where r_{q_i} is the rank of the highest ranking of retrieved references for a manuscript query test q_i .

$$MRR = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{1}{r_{q_i}} \quad (14)$$

V. EXPERIMENT AND EVALUATION

A. EVALUATION WITH OTHER CITATION RECOMMENDATION APPROACHES

We evaluate the proposed models' performance compared to other citation recommendation approaches, i.e., content-based filtering (CBF) and content + graph-based filtering. Some CBF approaches include Doc2vect, TF-IDF Cosine Similarity, and BERT similarity. For the baselines, the random walk model as a graph-based filtering approach works by incorporating content or a dependent query [16], such as Personalized PageRank, Edge Weight Personalized PageRank, and Random Walk with Restart. In this research, we name these baselines as Personalized PageRank based on non-topic community (PPR_NonTC); meanwhile, the proposed models are named Personalized PageRank based on Topic Communities (PPR_TC). We proposed three PPR_TC models, namely PPR_TC_A, PPR_TC_B, and PPR_TC_C. We consider some of the CBF and PPR_NonTC models as baseline models for evaluating our PPR_TC models because they are developed based on the same inline approaches, i.e., content, content + graph, content + graph + community.

The baselines are defined as follows. Doc2Vec maps a variable length of paper text into a fixed length of a distributed vector using a neural network model to predict the surrounding words in contexts sampled from the paragraph [38]. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both left and right contexts in all layers for textual embedding [39]. Personalized PageRank (PPR) is a variant of PageRank methods [44] in which a random walker jumps to specific nodes determined by a bias probability vector. And the vector can be calculated using similarity scores between the document dataset and the user query [14], [16], [55]. PPR with restart is a PageRank formulation in that every random walker moves in any direction with a probability vector related to starting initial node positions [46], [47]. The edge weight PPR modifies a personalized PageRank approach, in which node weights and edge weights are recalculated to determine a bias probability vector and transition probability matrix [45]. Modifying its transition probability matrix can be conducted by calculating a similarity between pair papers in the citation network.

We tested the proposed models, and the baseline to 200 manuscript queries consisting of 4 computer science groups and 2 type datasets (a subset DBLB dataset and a whole DBLB dataset) to select and rank k top recommended papers. The parameters of the proposed models are set to optimal values tuned by experiments. Shown in Table 2 and Table 3, all of the personalized PageRank methods based on topic communities (PPR_TC_A, PPR_TC_B, PPR_TC_C) outperform the personalized PageRank methods based on non-topic community (PPR_NonTC) and also superior to the CBF methods on the subset dataset and the large dataset. The experiments are measured with Recall@ n , MAP, and MRR. In the subset dataset, the Bert similarity method is dominant in recall performances throughout the CBF methods and

among all the PPR_NonTC methods. However, in the large-scale dataset, the Edge Weight PPR method is the best method of recall performance in PPR_NonTC methods and among all the CBF methods.

The evaluation results indicate that the PPR_TC models successfully enhance the relevance of top k recommended papers. There are some explanations for the results. The rank of reference paper candidates of multi-topic community results in top k recommended papers. Reference paper candidates of a multi-topic community are constructed by merging the best n -multi-topic communities. The best n -multi-topic communities are selected from a set of multi-topic communities considered most relevant to a group of aggregated multi-topic queries. Those reference paper candidates of the multi-topic community graph are assumed to be similar and related to manuscript queries. It means that scientific papers find reference papers not only with similar topics but also with related topics. Related reference papers can contain novel, unexpected knowledge relevant to a user's need. Query-personalized PageRank is modified to work with reference paper candidates of a multi-topic community to generate the most influential recommended paper similar and related to a set of queries.

There is little difference in performance between PPR_TC_B and PPR_TC_C, and both are fairly better than PPR_TC_A. PPR_TC_A utilizes all nodes of topic communities to determine a bias probability vector of the modified PageRank. Meanwhile, PPR_TC_B and PPR_TC_C use filtered nodes of topic communities, which is more relevant to manuscript queries to determine a bias probability vector of modified PageRank. Filtered nodes are determined by calculating the similarity of multi-topics between queries and recommended paper candidates. In addition, PPR_TC_C is designed to employ n multi-topic queries.

B. EVALUATION OF CITATION RECOMMENDATION APPROACHES ON DIFFERENT DATASET VOLUMES

We compare the performance between the proposed models and the baseline models when running in 2 conditions: the whole dataset (large scale) and the subset dataset. The result shows that there are different performances of the proposed models and baselines if they perform in the large and subset dataset. Some method performances in Table 2 and Table 3 are described in Figures 4a and 4b. As shown in Figure 4, the gap between the blue line representing PPR_TC_C performance and the yellow line representing PPR performance is wider in the large-scale dataset than in the subset dataset. On the large-scale dataset, PPR_TC_C produces recall@100 at 0.435 and PPR at 0.285, while in the subset dataset, PPR_TC_C generates recall@100 at 0.535 and PPR at 0.480. The difference in performance between the two models in the large dataset is $0.435 - 0.285 = 0.15$ and in the subset dataset is only $0.535 - 0.480 = 0.05$. In conclusion, the wider performance gap between the two models in the large and subset dataset means that the topic communities-based models work more effectively than the baselines if they run in

TABLE 2. Performance comparison with other approaches on the subset DBLB dataset.

Citation Recommendation Approach		Recall @25	Recall @50	Recall @75	Recall @100	MAP @100	MRR @50
Content Based Filtering (Content)	Doc2Vect	0.053	0.108	0.130	0.162	0.029	0.075
	Cosine TFIDF	0.138	0.191	0.231	0.267	0.058	0.093
	BERT	0.139	0.270	0.341	0.396	0.032	0.039
PPR_NonTC (Content + Graph) Based Filtering	PPR Restart	0.136	0.159	0.171	0.193	0.016	0.037
	PPR	0.305	0.383	0.439	0.480	0.133	0.107
	edge weight PPR	0.257	0.336	0.383	0.435	0.101	0.095
PPR_TC (Content + Community Graph) Based Filtering	PPR_TC_A	0.352	0.441	0.489	0.529	0.178	0.109
	PPR_TC_B	0.345	0.438	0.483	0.533	0.173	0.100
	PPR_TC_C	0.359	0.437	0.492	0.535	0.182	0.145

TABLE 3. Performance comparison with other approaches on the large scale DBLB dataset.

Citation Recommendation Approach		Recall @25	Recall @50	Recall @75	Recall @100	MAP @100	MRR @50
Content Based Filtering (Content)	Doc2Vect	0.023	0.029	0.042	0.054	0.004	0.017
	Cosine TFIDF	0.073	0.103	0.127	0.144	0.028	0.082
	BERT	0.146	0.201	0.237	0.279	0.063	0.093
PPR_NonTC (Content + Graph) Based Filtering	PPR Restart	0.088	0.114	0.131	0.143	0.033	0.069
	PPR	0.164	0.214	0.255	0.285	0.097	0.071
	Edge weight PPR	0.161	0.219	0.262	0.304	0.096	0.070
PPR_TC (Content + Community Graph) Based Filtering	PPR_TC_A	0.220	0.296	0.360	0.400	0.126	0.059
	PPR_TC_B	0.258	0.344	0.387	0.435	0.136	0.083
	PPR_TC_C	0.254	0.33	0.393	0.435	0.138	0.079

big data. The topic communities become more cohesive, and the topic map becomes more evident as the dataset becomes more massive.

We also analyze the impact of the change in the dataset volume relating to the performance of the citation recommendation models, in which the magnitude of the whole dataset is twelve times bigger than the subset dataset. The experiments show that the performances of all models will decrease if they run to the larger dataset. The decreased performance on the proposed model PPR_TC_C is less than the baseline PPR method. As shown in Table 2, PPR_TC_C tested in the subset dataset results in recall@100 at 0.535 and in the whole dataset at 0.435, so the performance reduction of PPR_TC_C in the two different volume datasets is 18%. Meanwhile, PPR results in recall@100 at 0.452 in the subset dataset and 0.2648 in the whole dataset, so the performance reduction of PPR is 40%, two times higher than the decreased performance of PPR_TC_C. Hence, this condition indicates that the PPR_TC_C's performance is less sensitive than PPR as the dataset volume increases.

C. EFFECTS OF A VARIETY OF MULTI-TOPIC COMMUNITIES ON THE PERFORMANCES OF THE PROPOSED MODEL

Tables 2 and 3 are derived from the average performance of citation recommendation experiments with 4×50 manuscript query tests from 4 topic fields in

computer science, including Information Retrieval (IR), Machine Learning (ML), Network Communication (NC), and Computer Vision (CV). Although Table 2 and Table 3 show that the average of the best performance result is the proposed models, there is a variety of the performance results of the citation recommendation methods if viewed for four different manuscript query groups represented on line diagrams in Fig. 5 and Fig. 6.

We analyze the impact of various topic communities on the performance of the proposed model PPR_TC. In this experiment, the PPR_TC_C method generates four different topic communities from four other query groups. The performance of PPR_TC_C shown in the blue line seems the highest when tested on Machine Learning (ML) query groups in the subset Dataset (Fig.5) and in the large dataset (Fig.6). PPR_TC_C works best in the ML community because of the following reasons. In PPR_TC_C, recommended paper candidates are members of multi-topic communities closely relevant to manuscript queries. A more popular and growing topic community is assumed to have many papers with denser structures than others in one community. The popular community may have a denser structure and a more coherent topic. Ranking candidate papers by the PPR_TC_C method in a denser structure and a more coherent multi-topic community will generate better performance. Hence, in this experiment, ML topic communities have a denser structure with more diversity of related topic papers than other topic communities. Evidence show that the ML field

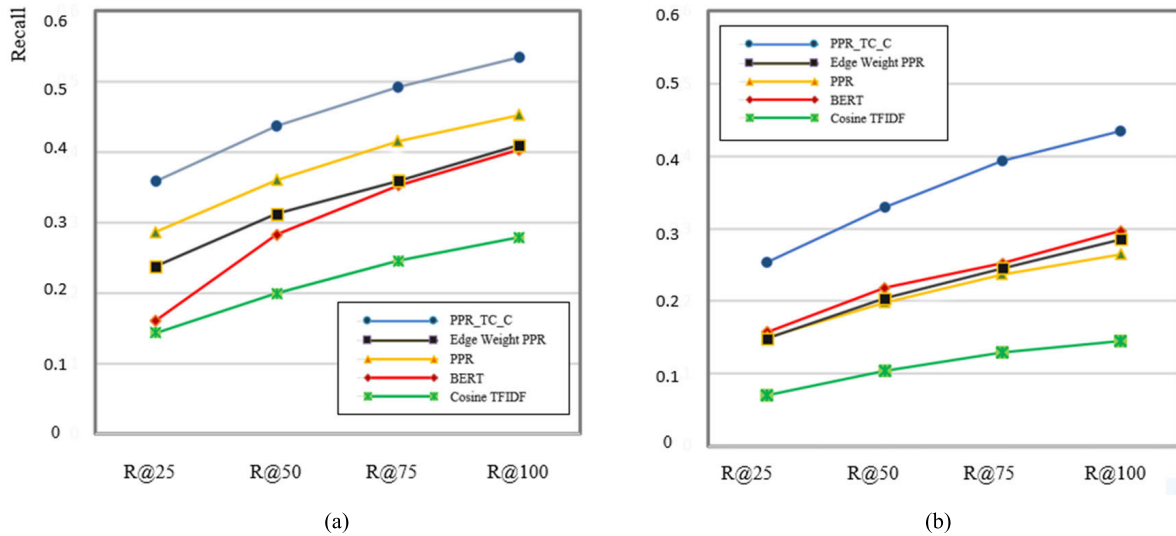


FIGURE 4. Evaluation results of PPR_TC_C and other methods on different dataset volumes: (a)The subset dataset and (b) The large scale dataset.

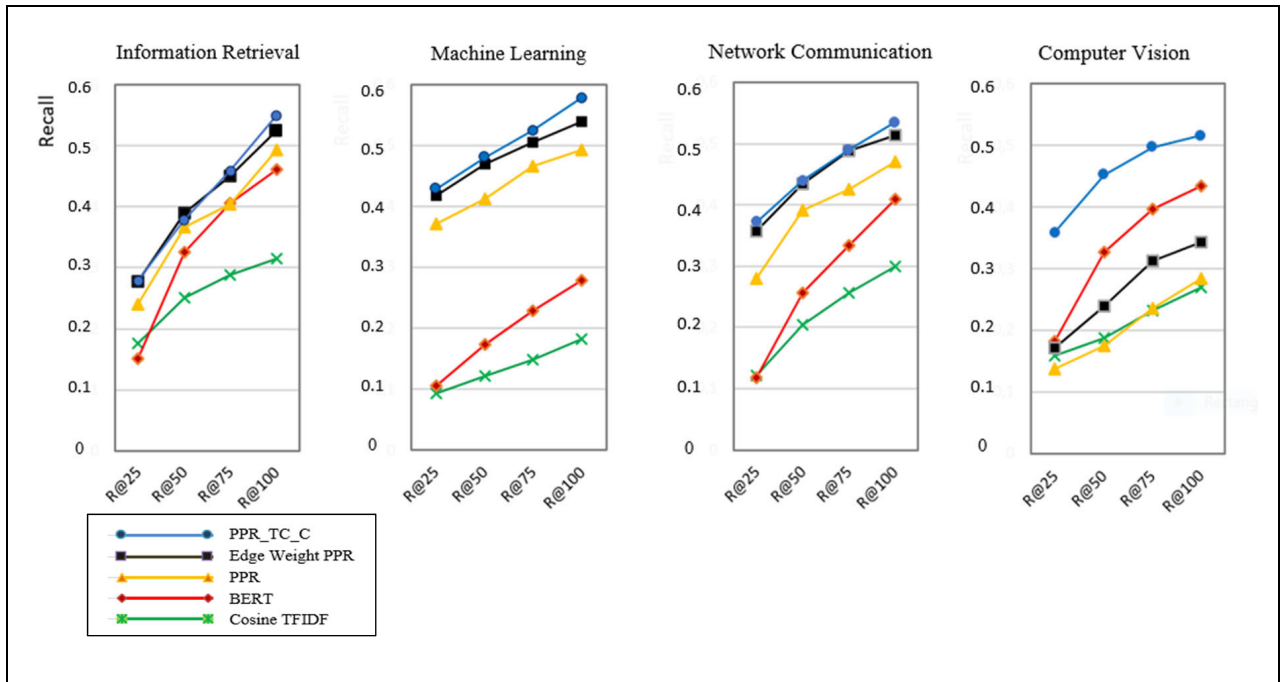


FIGURE 5. Performance comparison on four different computer science query groups in the subset dataset.

has recently been an intensive research topic and is the most related to other computer science fields and even applied to other domains such as engineering, finances, etc.

The performances of PPR_TC_C shown in the blue line in Fig 5 and Fig 6 seem the highest in most query groups except in the IR query group in the large dataset. Generally, the results reveal that the PPR_TC_C performances among most communities are consistent except for the IR. It is because the IR community structure may be less dense than the three other communities in the large dataset. The IR topic field

may be less popular for research and has fewer relations to different topics than the three fields of computer science (ML, NC, CV).

D. IMPACT OF N-MULTI-TOPIC QUERIES TO PERFORMANCES OF PROPOSED METHOD (PPR_TC_C)

We analyze the impact of adjusting the n-multi-topic parameter to the performance of PPR_TC_C, which is designed to handle n-multi-topic queries for citation recommendation. Experiments on PPR_TC_C’s performance relating to n-

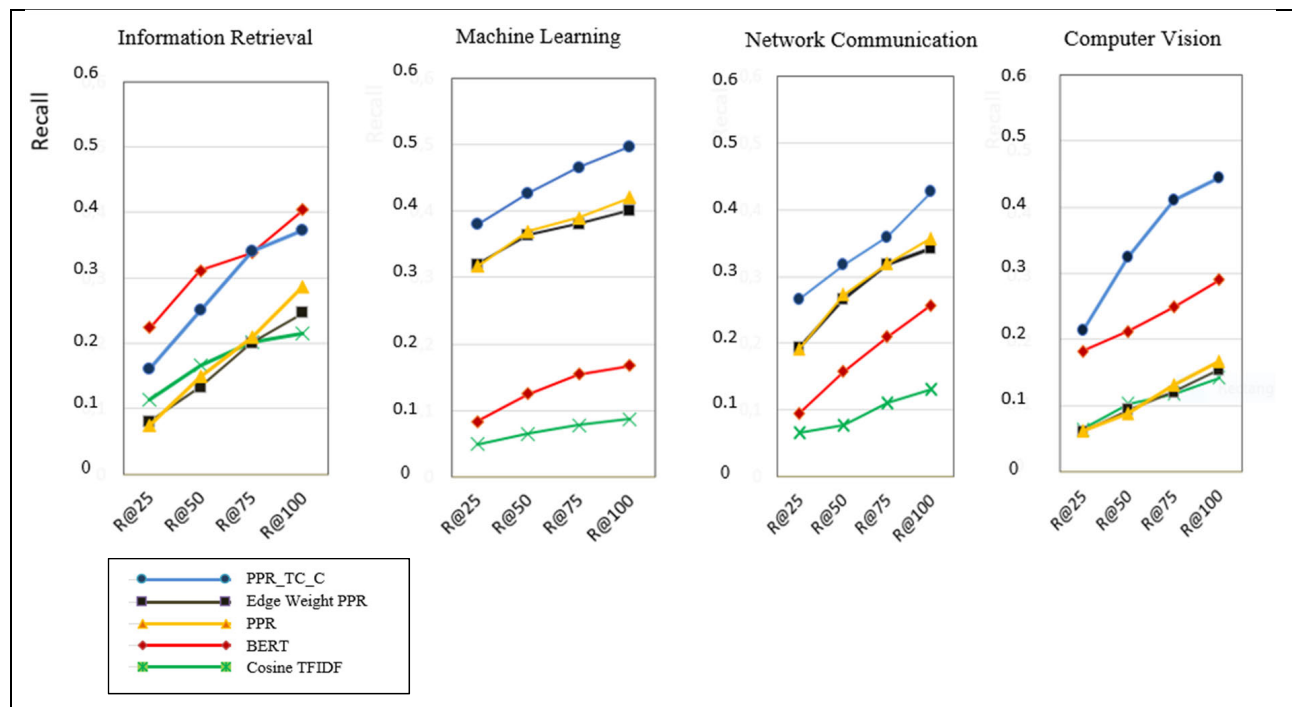


FIGURE 6. Performance comparison on four different computer science query groups in the large dataset.

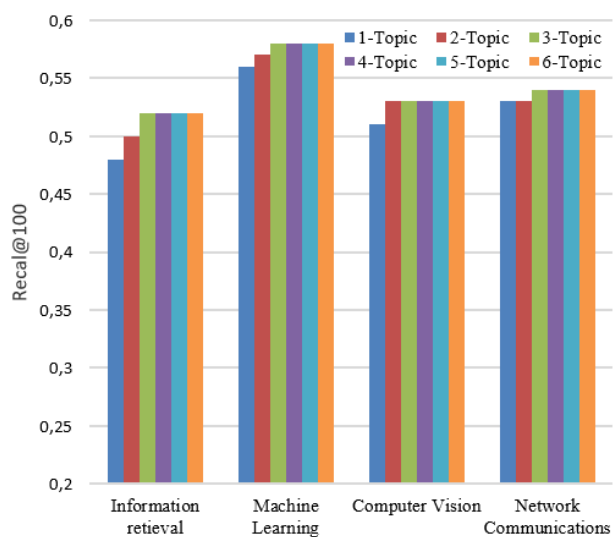


FIGURE 7. Impact of n-multi-topic queries on four different computer science query groups.

multi-topic queries are conducted on four different query groups in the subset DBLB dataset and measured by recall metric. Figure 7 indicates a variety of performances at recall@100 in each group of queries as n-multi-topic query values are set up from $n = 1$ to $n = 6$. The parameter thresholds (TrC) of PPR_TC_C to each query group are placed at each optimum value. In groups of IR, ML, and NC, the optimum performance is at $n = 3$ multi-topic, while in CV is at

$n = 2$ multi-topic. In conclusion, with an appropriate parameter n of multi-topic queries, the PPR_TC_C model results in optimum recall performance. It indicates that paper journals or proceedings in computer science fields are composed of n -multi-topic reference papers with a value of $n > 1$.

Figure 7 describes that the performance of PPR_TC_C varies within four communities generated by the four group queries. The PPR_TC_C with ML communities in the subset dataset produces the highest recall performance and then is followed by NC, CV, and IR. This support the experiment described in Figure 5 and Figure 6 that the most popular community is ML and the least is IR.

VI. CONCLUSION

In this paper, we propose the scientific paper recommendation framework, which consists of three main stages, i.e., feature extraction of an academic citation network, multi-topic community selection, and ranking recommended paper candidates from selected communities by modified PageRank. Citation candidates as members of a coherent and meaningful community can fulfill the recommended paper criteria containing homophily relations, i.e., the tendency to link to similar papers, and heterophily relations, i.e., the tendency to link to related papers. The framework results in the most influential recommended papers with not only a similar topic but also related topics relevant to a user’s queries. Generally, the three proposed models outperform the personalized PageRank models based on non-topic community and CBF models as baselines. The performance results are also

consistent when tested on 4 query groups of computer science. The proposed multi-topic communities-based models work more effectively than the baselines if they run in the big data since the topic communities become more cohesive as the dataset becomes more massive.

In the future, there are many potential ways for this work to construct multi-topic communities-based PageRank in an academic citation network by applying different methods and considering other aspects, such as time and heterogeneous networks. A new model can use time and heterogeneous network aspects to generate dynamic and complex topic communities, which can incorporate into a random walk approach. Personalized PageRank based on topic communities can deal with serendipity in the recommender system and enhance recommendation performance.

REFERENCES

- [1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, Mar. 2017, doi: [10.1109/TBDDATA.2016.2641460](https://doi.org/10.1109/TBDDATA.2016.2641460).
- [2] T. Dai, L. Zhu, Y. Wang, H. Zhang, X. Cai, and Y. Zheng, "Joint model feature regression and topic learning for global citation recommendation," *IEEE Access*, vol. 7, pp. 1706–1720, 2019, doi: [10.1109/ACCESS.2018.2884981](https://doi.org/10.1109/ACCESS.2018.2884981).
- [3] 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, doi: [10.1109/TBDDATA.2016.2555318](https://doi.org/10.1109/TBDDATA.2016.2555318).
- [4] X. Cai, Y. Zheng, L. Yang, T. Dai, and L. Guo, "Bibliographic network representation based personalized citation recommendation," *IEEE Access*, vol. 7, pp. 457–467, 2019, doi: [10.1109/ACCESS.2018.2885507](https://doi.org/10.1109/ACCESS.2018.2885507).
- [5] K. Sugiyama and M. Kan, "Towards higher relevance and serendipity in scholarly paper recommendation," in *Proc. ACM Int. Conf. Digit. Libraries*, 2015, pp. 1–15, doi: [10.1145/2719943.2719947](https://doi.org/10.1145/2719943.2719947).
- [6] L. Yang, Y. Zheng, X. Cai, H. Dai, D. Mu, and L. Guo, "A LSTM based model for personalized context-aware citation recommendation," *IEEE Access*, vol. 6, pp. 59618–59627, 2018, doi: [10.1109/ACCESS.2018.2872730](https://doi.org/10.1109/ACCESS.2018.2872730).
- [7] A. Hernando, J. Bobadilla, and F. Ortega, "A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model," *Knowl.-Based Syst.*, vol. 97, pp. 188–202, Apr. 2016, doi: [10.1016/j.knsys.2015.12.018](https://doi.org/10.1016/j.knsys.2015.12.018).
- [8] X. Luo, M. Zhou, S. Li, Y. Xia, Z. You, Q. Zhu, and H. Leung, "An efficient second-order approach to factorize sparse matrices in recommender systems," *IEEE Trans. Ind. Inform.*, vol. 11, no. 4, pp. 946–956, Aug. 2015, doi: [10.1109/TII.2015.2443723](https://doi.org/10.1109/TII.2015.2443723).
- [9] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2002, pp. 253–260.
- [10] C. Bhagavatula, S. Feldman, R. Power, and W. Ammar, "Content-based citation recommendation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 238–251.
- [11] P. Lops, M. De Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*. Boston, MA, USA: Springer, 2011, pp. 73–105.
- [12] X. Li and H. Chen, "Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach," *Decis. Support Syst.*, vol. 54, no. 2, pp. 880–890, Jan. 2013, doi: [10.1016/j.dss.2012.09.019](https://doi.org/10.1016/j.dss.2012.09.019).
- [13] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong, "Random walks: A review of algorithms and applications," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 4, no. 2, pp. 95–107, Apr. 2019, doi: [10.1109/TETCI.2019.2952908](https://doi.org/10.1109/TETCI.2019.2952908).
- [14] D. Mu, L. Guo, X. Cai, and F. Hao, "Query-focused personalized citation recommendation with mutually reinforced ranking," *IEEE Access*, vol. 6, pp. 3107–3119, 2017, doi: [10.1109/ACCESS.2017.2787179](https://doi.org/10.1109/ACCESS.2017.2787179).
- [15] X. Cai, J. Han, W. Li, R. Zhang, S. Pan, and L. Yang, "A three-layered mutually reinforced model for personalized citation recommendation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6026–6037, Dec. 2018, doi: [10.1109/TNNLS.2018.2817245](https://doi.org/10.1109/TNNLS.2018.2817245).
- [16] R. K. Roul and J. K. Sahoo, "A novel approach for ranking web documents based on query-optimized personalized pagerank," *Int. J. Data Sci. Analytics*, vol. 11, no. 1, pp. 37–55, Jan. 2021, doi: [10.1007/s41060-020-00232-2](https://doi.org/10.1007/s41060-020-00232-2).
- [17] H. Park, J. Jung, and U. Kang, "A comparative study of matrix factorization and random walk with restart in recommender systems," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2017, pp. 756–765, doi: [10.1109/BigData.2017.8257991](https://doi.org/10.1109/BigData.2017.8257991).
- [18] Y. Wang and Y. Tong, "Ranking scientific articles by exploiting citations, authors, journals, and time information," in *Proc. 27th AAAI Conf. Artif. Intell.*, 2013, pp. 933–939.
- [19] Y. Wei, F. Yi, X. Cui, and F. Chen, "An improved PageRank algorithm based on text similarity approach for critical standards identification in complex standard citation networks," *Complexity*, vol. 2021, pp. 1–15, Mar. 2021, doi: [10.1155/2021/8825947](https://doi.org/10.1155/2021/8825947).
- [20] Z.-M. Ren, "Age preference of metrics for identifying significant nodes in growing citation networks," *Phys. A, Stat. Mech. Appl.*, vol. 513, pp. 325–332, Jan. 2019, doi: [10.1016/j.physa.2018.09.001](https://doi.org/10.1016/j.physa.2018.09.001).
- [21] Y. Zhang, M. Wang, F. Gottwalt, M. Saberi, and E. Chang, "Ranking scientific articles based on bibliometric networks with a weighting scheme," *J. Informetrics*, vol. 13, no. 2, pp. 616–634, May 2019, doi: [10.1016/j.joi.2019.03.013](https://doi.org/10.1016/j.joi.2019.03.013).
- [22] A. Lamurias, P. Ruas, and F. M. Couto, "PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking," *BMC Bioinf.*, vol. 20, pp. 1–12, Oct. 2019, doi: [10.1186/s12859-019-3157-y](https://doi.org/10.1186/s12859-019-3157-y).
- [23] Z. Yang, J. Tang, J. Zhang, J. Li, and B. Gao, "Topic-level random walk through probabilistic model," in *Advances in Data and Web Management* (Lecture Notes in Computer Science), vol. 5446. Berlin, Germany: Springer, 2009.
- [24] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 1055–1060, doi: [10.1109/ICDM.2008.71](https://doi.org/10.1109/ICDM.2008.71).
- [25] M. Tao, X. Yang, G. Gu, and B. Li, "Paper recommend based on LDA and PageRank," in *Proc. Int. Conf. Artif. Intell. Secur.*, 2020, pp. 571–584.
- [26] J. Jardine and S. Teufel, "Topical PageRank: A model of scientific expertise for bibliographic search," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 501–510.
- [27] Y. Zhang, J. Ma, Z. Wang, B. Chen, and Y. Yu, "Collective topical PageRank: A model to evaluate the topic-dependent academic impact of scientific papers," *Scientometrics*, vol. 114, no. 3, pp. 1345–1372, Mar. 2018, doi: [10.1007/s11192-017-2626-1](https://doi.org/10.1007/s11192-017-2626-1).
- [28] F. Zhao, Y. Zhang, J. Lu, and O. Shai, "Measuring academic influence using heterogeneous author-citation networks," *Scientometrics*, vol. 118, no. 3, pp. 1119–1140, Mar. 2019, doi: [10.1007/s11192-019-03010-5](https://doi.org/10.1007/s11192-019-03010-5).
- [29] D. Kotkov, J. Veijalainen, and S. Wang, "How does serendipity affect diversity in recommender systems? A serendipity-oriented greedy algorithm," *Computing*, vol. 102, no. 2, pp. 393–411, Feb. 2020, doi: [10.1007/s00607-018-0687-5](https://doi.org/10.1007/s00607-018-0687-5).
- [30] D. Kotkov, S. Wang, and J. Veijalainen, "A survey of serendipity in recommender systems," *Knowl.-Based Syst.*, vol. 111, pp. 180–192, Nov. 2016, doi: [10.1016/j.knsys.2016.08.014](https://doi.org/10.1016/j.knsys.2016.08.014).
- [31] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
- [32] Z. Kang, C. Peng, and Q. Cheng, "Top-N recommender system via matrix completion," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2016, vol. 30, no. 1, pp. 179–185.
- [33] K. Chandrasekaran, S. Gauch, P. Lakkaraju, and H. P. Luong, "Concept-based document recommendations for CiteSeer authors," in *Adaptive Hypermedia and Adaptive Web-Based Systems* (Lecture Notes in Computer Science book series), vol. 5149. Cham, Switzerland: Springer, 2008, pp. 83–92.
- [34] T. Y. Tang and G. McCalla, "A multidimensional paper recommender: Experiments and evaluations," *IEEE Internet Comput.*, vol. 13, no. 4, pp. 33–41, Aug. 2009, doi: [10.1109/MIC.2009.73](https://doi.org/10.1109/MIC.2009.73).

- [35] 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 *Proc. ACM Conf. Comput. Supported Cooperat. Work*, Nov. 2002, pp. 116–125.
- [36] B. Kazemi and A. Abhari, "A comparative study on content-based paper to paper," in *Proc. 20th Commun. Netw. Symp. (ACM)*, vol. 5, 2017, pp. 1–10.
- [37] M. Amami, G. Pasi, F. Stella, and R. Faiz, "An LDA-based approach to scientific paper recommendation," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.*, in Lecture Notes in Computer Science, vol. 9612, 2016, pp. 200–210.
- [38] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. ICML*, 2014, pp. 1188–1196.
- [39] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Language Technol.*, 2019, pp. 4171–4186, doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- [40] C. Jeong, S. Jang, H. Shin, E. Park, and S. Choi, "A context-aware citation recommendation model with BERT and graph convolutional networks," 2019, *arXiv:1903.06464*.
- [41] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, "A survey of heterogeneous information network analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 17–37, Jan. 2017, doi: [10.1109/TKDE.2016.2598561](https://doi.org/10.1109/TKDE.2016.2598561).
- [42] Z. Ali, G. Qi, P. Kefalas, W. A. Abro, and B. Ali, "A graph-based taxonomy of citation recommendation models," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 5217–5260, Oct. 2020, doi: [10.1007/s10462-020-09819-4](https://doi.org/10.1007/s10462-020-09819-4).
- [43] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab, Tech. Rep., 1998, doi: [10.1109/IISWC.2012.6402911](https://doi.org/10.1109/IISWC.2012.6402911).
- [44] S. Park, W. Lee, B. Choe, and S.-G. Lee, "A survey on personalized PageRank computation algorithms," *IEEE Access*, vol. 7, pp. 163049–163062, 2019, doi: [10.1109/ACCESS.2019.2952653](https://doi.org/10.1109/ACCESS.2019.2952653).
- [45] W. Xie, D. Bindel, A. Demers, and J. Gehrke, "Edge-weighted personalized PageRank: Breaking a decade-old performance barrier," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1325–1334, doi: [10.1145/2783258.2783278](https://doi.org/10.1145/2783258.2783278).
- [46] X. Zhou, W. Liang, K. I.-K. Wang, R. Huang, and Q. Jin, "Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 1, pp. 246–257, Jan. 2021, doi: [10.1109/TETC.2018.2860051](https://doi.org/10.1109/TETC.2018.2860051).
- [47] 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, doi: [10.1109/TETC.2014.2356505](https://doi.org/10.1109/TETC.2014.2356505).
- [48] S. Qiao, T. Li, H. Li, and Y. Zhu, "SimRank: A page rank approach based on similarity measure," in *Proc. IEEE Int. Conf. Intell. Syst. Knowl. Eng.*, no. 9, Nov. 2010, pp. 390–395, doi: [10.1109/ISKE.2010.5680842](https://doi.org/10.1109/ISKE.2010.5680842).
- [49] J. Beck, B. Neupane, and J. M. Carroll, "A study of citation motivations in HCI research," SocArXiv, 2018, doi: [10.31235/osf.io/me8zd](https://doi.org/10.31235/osf.io/me8zd).
- [50] X. Zhu, P. Turney, D. Lemire, and A. Vellino, "Measuring academic influence: Not all citations are equal," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 2, pp. 408–427, Feb. 2015, doi: [10.1002/asi.23179](https://doi.org/10.1002/asi.23179).
- [51] M. Dunaiski, J. Geldenhuys, and W. Visser, "On the interplay between normalisation, bias, and performance of paper impact metrics," *J. Informetrics*, vol. 13, no. 1, pp. 270–290, Feb. 2019, doi: [10.1016/j.joi.2019.01.003](https://doi.org/10.1016/j.joi.2019.01.003).
- [52] Y. Ding, "Topic-based PageRank on author cocitation networks," *J. Assoc. Inf. Sci. Technol.*, vol. 62, no. 3, pp. 449–466, 2011, doi: [10.1002/asi.21467](https://doi.org/10.1002/asi.21467).
- [53] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Oct. 2008, Art. no. P10008, doi: [10.1088/1742-5468/2008/10/P10008](https://doi.org/10.1088/1742-5468/2008/10/P10008).
- [54] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 990–998, doi: [10.1145/1401890.1402008](https://doi.org/10.1145/1401890.1402008).
- [55] S. Hatakenaka and T. Miura, "Query and topic sensitive PageRank for general documents," in *Proc. 14th IEEE Int. Symp. Web Syst. Evol. (WSE)*, Sep. 2012, pp. 97–101, doi: [10.1109/WSE.2012.6320539](https://doi.org/10.1109/WSE.2012.6320539).



AGUNG HADHIATMA received the B.S. and M.S. degrees in electrical engineering and information technology, Universitas Gadjah Mada, Indonesia, where he is currently pursuing the Ph.D. degree in computer science with the Faculty of Mathematics and Natural Sciences. His research interests include social network analysis, natural language processing, and machine learning.



AZHARI AZHARI received the bachelor's degree in statistics from the Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia, in 1987, the master's degree in software engineering from the Bandung Institute of Technology, Bandung, West Java, Indonesia, in 1999, and the Ph.D. degree in computer science from the Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, in 2010. Since 2021, he has been the Head of the Undergraduate Program in Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada. His research interests include natural language processing, machine learning, community detection, question-answer systems, machine translation, fraud detection, music composition, intelligence agent, and big data analytics.



YOHANES SUYANTO received the bachelor's degree in physics from the Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, Indonesia, in 1987, the master's degree in computer science from Universitas Indonesia, Jakarta, Indonesia, in 1992, and the Ph.D. degree in computer science from the Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, in 2014. Since 2021, he has been the Head of the Undergraduate Program in Electronics and Instrumentation, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada. His research interests include natural language processing, speech synthesis, and sound synthesis.

...