A Hybrid Paper Recommendation Method by Using Heterogeneous Graph and Metadata

1st Shi Hui[1,2]
1.Institute of Information Engineering
Chinese Academy of Science
2.School of Cyber Security University
of Chinese Academy of Sciences
Beijing, China
shihui@iie.ac.cn

2nd Ma Wei[1]
1.Institute of Information Engineering
Chinese Academy of Science
Beijing, China
mawei@iie.ac.cn

3rd Zhang XiaoLiang[1]
1.Institute of Information Engineering
Chinese Academy of Science
Beijing, China
zhangxiaoliang@iie.ac.cn

4th Jiang JunYan[1,2]
1.Institute of Information Engineering
Chinese Academy of Science
2.School of Cyber Security
University of Chinese Academy of Sciences
Beijing, China
jiangjunyan@iie.ac.cn

5th Liu YanBing[1]
1.Institute of Information Engineering
Chinese Academy of Science
Beijing, China
liuyanbing@iie.ac.cn

6th Chen ShuJuan[1]
1.China cybersecurity review
technology and certification center
Beijing, China
chensj@isccc.gov.cn

***Abstract*—The amount of academic articles in digital libraries is increasing exponentially. This growth of scientific papers' growth made it difficult for researchers to obtain related papers from their queries. Recommendation systems can help them resolve the problem of information overload. However, existing paper recommender methods generally rely on the simple citation network, which ignores the semantic of papers and has the problem of cold start. In this paper, a hybrid paper recommendation approach AMHG is proposed which is based on a multi-level citation heterogeneous graph. Unlike existing works which only use the reference relationship, we consider the same or similar authors' papers to alleviate the cold start problem of zero-citation and newly published papers. Besides, the metadata information of papers is also incorporated into a representation model to generate better recommender results to alleviate the cold start problem. We use the authors' influence factors to reorder the candidate list outputting by MLP to obtain high-quality articles. Through experiments, we compare our model with several methods on the DBLP-REC dataset to demonstrate that AMHG outperforms state-of-the-art performance and the effectiveness of recommender.**

***Index Terms*—scientific paper recommendation, citation networks, heterogeneous information graph, hybrid recommendation method, recommender system**

I. Introduction

Recommendation systems have become popular and attracting increasing attention from both academia and industry [1]. However, compared with other recommender applications, such as those for movies, music, and news, fewer studies have examined recommendation systems for academic papers. The number of freely available scientific articles on the web have risen up-to 25 million [2]. Finding suitable papers is a time-consuming task for researchers, especially in the large and rapidly growing database of published data sciences.

A document retrieval system can solve the above problem. It can help researchers find relevant papers in recent years. The retrieval system starts with user query, then processes the request through the model, and finally returns results that are most similar to the user query [3], such as Google Scholar [4], Web of Science [6], and ScienceDirect [5]. Although these engines make it easier for researchers to find articles of interest, keyword-based systems still return hundreds or thousands of related articles. The problem with keyword-based search is that the results it returns are ambiguous and wide-ranging. The results depend on the user's ability to fine-tune query messages and user filtering capacity. The classic method used by some researchers is to follow the list of references from the papers they already possessed [7]. They use papers' references to find similar articles. Although this method might be quite effective in some cases, researchers need deep mining capabilities.

Compared with the traditional keyword-based search technique, recommendation systems are more personalized and effective for massive data [8]. The recommendation techniques can be divided into four main categories: Collaborative Filtering(CF), Content-Based Filtering(CBF), Graph-Based methods(GB) and Hybrid recommend methods. Previous research has focused on finding better recommendation systems for academic papers which related to specific research domains. One of the most common methods is CF. This method mainly focuses on the actions or ratings on the items of other users whose profiles are similar to the user's called "neighbor users" [8]. However, a lot of studies have shown that this method has inherent problems, such as data sparseness and cold start. There are many recommender methods to solve these

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

problems, such as CBF and citation analysis. CBF creates a relationship between items by analyzing their inherent characteristics, and papers use the keywords. Due to the ambiguity of natural language, the system may not be able to capture user interest, and the method assumes that the entire content of papers is freely accessible, however, this is not always true because of copyright restrictions. Beacuse the citation networks are often constructed, citation analysis is the most commonly used method in GB. Citation analysis is the comprises of co-citation analysis [9] and biblio graphic coupling [10]. They measure relevance by focusing on neighbors, making their relationships more meaningful and purposeful, but the system cannot take into account the complex relationships between papers.

In this study, we propose a hybrid approach for research paper recommendation, which is based on a multi-level heterogeneous citation network and meta-data information. Unlike existing works which only use the citation relationship, we consider the same authors' papers and similar authors' paper. Besides, the content information of papers, i.e., title, abstract, published year, are also incorporated into the representation model to generate better recommender results. The traditional recommendation approaches focus on the paper content similarity, which ignores the impact of the other information of the paper, such as author and date of publications. The research paper meta-data is available for free even if they are published in paid journals. This approach applies another filter to the recommended articles by the importance of authors, the time and venue of publication and semantic correlation.

This method aims to deal with the following scenarios in which: (1) A student received a dissertation from his superior to start research in the topic area covered by it. (2) A researcher hopes to get more related articles from published articles and find new research points. (3) A researcher who has found an interesting paper after some initial searches hopes to get more related papers similar to it. (4) A reviewer wants to explore more based on the received paper that addresses a subject matter which he may not a specialist in. In all these cases, we assume that all citation network information, that is, references and citations are public, and most databases are generally public, which is more realistic. The main contributions of our proposed method are:

- A new research paper recommendation method is proposed, which combines the citation relationship and meta-data content of the paper to improve the accuracy of returning the candidate list.
- We reconstructed a heterogeneous graph that considers the same authors' papers and similar authors' articles to solve the cold start problem of newly published or zero-cited papers.
- We consider the authors' influence factor in order to recommend higher quality articles.
- Our method for the task of personalized paper recommendation is effective compared with several baselines.

The article is structured as follows. A literature review of existing recommended approaches is presented in Section II. Section III illustrates the proposed method. The details of the experiments and evaluations are given in section IV and the experimental results are discussed in section V.

II. Related work

The task of academic papers recommender is to offer researchers a list of papers that they are interested in. Since the recommendation systems are introduced, many recommendation algorithms have emerged, which can generally be divided into four categories: collaborative filtering(CF), content-based filtering (CBF), graph-based methods (GB), and hybrid-based recommendation methods [8]. Each method has its own rationale underlying to recommend interesting papers for researchers.

CBF first extracts and constructs user' interest models from the users' historical preferences and personal library (including published articles and cited papers), and then generates user-paper feature vectors based on the TF-IDF, keyword extraction model or language model. Finally, the similarity of the user-paper feature vector is sorted to generate a recommendation list. This method needs to get the entire content of the paper, and it too expensive and time-consuming to match the entire text. Meta-data based methods find similarities between the research papers by using the free availability of paper meta-data [12], [13]. However, if the same author publishes a paper across disciplines, the results are inaccurate. After building a heterogeneous network, we use meta-data information, i.e., title, abstract, authors and publication time, to compute the similarity in order to solve the problem.

CF is one of the most successful techniques in recommendation systems [14]. When the paper content information is not desirable, the CF method is very effective. The main idea is that if users A and B both rate some common items, they are considered to be similar. Therefore, if some papers are cited by B but not by A, these articles can be recommended to B [15]. Considering the rating history of users, CF aims to find similar users for users where users and items correspond to papers and cited papers. There are many limitations to using this method, such as cold starts and data sparseness. In addition, cited papers as metrics limit the possibility of new articles as candidates. Haruna et al. [16] proposed an improved collaborative filtering paper recommendation method, which uses public metadata to mine hidden relationships between papers, thereby providing personalized recommendations.

GF methods focus on the construction of graphs. Graphs can be citation networks or social networks. The citation network contains citation relationships between papers, the papers are nodes and the citation relationships constitute edges. We can use co-citation analysis and bibliographic coupling to analyze. It is considered that if two articles have common references or are cited by the same article, the two articles are similar [18]. The recommendation task can be transformed into a graph search [19] or a link prediction problem [20]. The GF method requires the use of graphs to represent the collected data of researchers and papers and then uses a ranking algorithm to

rank candidate articles. In general, bipartite graph networks and cross-domain recommendation systems often use random walk algorithms. For instance, Xu et al. [21] use the random walk to find similar users for the target users in a cross-domain. The cross-domain recommendation aims to build a relationship between the source domain and the target domain, which can alleviate the problems of cold start and data sparsity, improving the quality of recommendation result [22]. The GB can use information from different sources to make recommendations results diverse. However, this method not only does not consider the content information of the papers but also the complex relationships between papers.

To improve the accuracy of the recommendation results and obtain better performance, some recommendation systems combine the two or more recommendation technologies to recommend personalized papers to the researchers [23]. The hybrid methods based on content-based(CB) + CF and CB + GB have been a research focus in recent years. The CB techniques build researchers' profiles by capturing previous research interests embodied in their past publications, the CF techniques discover the potential papers through citation matrices. Finally, these systems use the content to calculate similarity for these papers to generate recommendation lists [24]. The CB + GB methods are similar to the above, using network graphs to find as many candidate articles as possible, and then using CB methods to calculate similarity [25]. This helps improve personalized recommendation results but cold start problem is still.

In this paper, we first use a multi-level heterogeneous citation graph to screen out candidate papers, then combine the articles' metadata and authors' influence factors to get the final recommendation list, which belongs to the category of hybrid recommendation methods. The data used by our proposed method are basically publicly available. Metadata and author factors are used to evaluate papers to ensure the quality of papers and alleviate the problem of cold start of newly published or zero-cited papers.

III. Proposed Method

A. Method overview

We propose a recommender system with a multilevel heterogeneous citation graph that combines author influence matrix and metadata information, named AMHG(as shorthand for Heterogeneous Graph combines Authors' influence factors and Metadata information). The block diagrams of the AMHG approach are shown in figure 1. After a query paper, we first build a heterogeneous graph, then we compute the score of every node and use Multilayer Perceptron(MLP) to train the model, after that, we output the candidate papers and the similarity score. Finally, an author evaluation is used to rerank the candidate recommendation list.

AMHG is based on a Multilevel Simultaneous Citation Network (MSCN) and considers the authors' relationship in papers. Besides, we use the metadata information to overcome the low precision, when either old or new papers selected as interest articles, ensuring the high quality of recommended

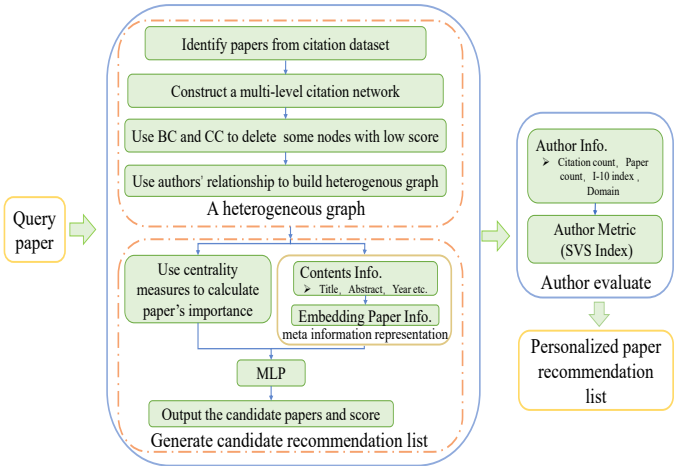

Fig. 1. Structure of AMHG

papers. We construct a multi-level citation network, then use the authors' relationship to build a heterogeneous graph. After that, we use metadata information to represent nodes. We compute the similarity of nodes by papers' embedding and evaluate the importance of nodes by using the centrality measures in order to train the model for generating a candidate list. Finally, we use the authors' impact factor to recommend papers. We will introduce in detail as following.

B. Building heterogeneous graph

Firstly, we construct a ten-level citation network by using references. The reason for using a ten-levels network and using more than ten levels may include papers not related to the paper of interest [11]. References are an essential part of a paper and generally appear at the end of the paper. The relationship between papers can be divided into "cite" and "cited". Graph describing these relationships between papers is called citation networks [26]. The first step is to construct a ten-level citation network based on the references of the paper. We consider the structural relationship of papers with the paper of interest, create a multi-levels citation network by expanding references paper in both directions. Beginning with the paper of interest I, we use the reference list of Paper I to construct the first backward layer and use the cited list to construct the first forward layer. Starting from the forward and backward layer of the network, we build a multilevel network. However, there may be many nodes, we should use methods to delete some nodes in the network. In this paper, we use bibliographic coupling (B.C) and co-citation (C.C), as shown in (1) and (2).

$$B.C(M_J, N_J) = \Sigma_{\mathcal{J}\in\mathcal{D}} B(M_J, N_J) \tag{1}$$

$$C.C(M_J, N_J) = \Sigma_{\mathcal{J}\in\mathcal{D}} C(M_J, N_J) \tag{2}$$

where $B(M_J, N_J)$ equals to (3) and $C(M_J, N_J)$ equals to (4).

$$\mathcal{B}(M_J, N_J) = \begin{cases} 0 & \text{if } \textit{M and N cite paper J} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

$$\mathcal{C}(M_J, N_J) = \begin{cases} 0 & \text{if paper } \textit{J cite M and N} \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

Secondly, we use the $\mathcal{C}_s$ to merge the two metric scores [11], where the numerator represents the similarity of the two papers based on citation information, and the denominator is the distance between the paper and other papers (that is, the number of hops in the network), shown in (5). The existing experiments indicate that the proper network size for a given problem is between 500 and 800. Therefore, in this study, we selected 800 papers with a higher C_s.

$$\mathcal{C}_s = \frac{\sum_{n=1}^{N}(B.C_{score}(M,N) + C.C_{score}(M,N))}{d(I,P)} \tag{5}$$

Finally, we add to the authors' relationship edges, figure 2 is shown. We join the same authors' papers and similar authors' papers to the network. We use the node *I* and node P_{14} are both written by A_2, so they link by the same author's edge. Node *I* and node P_4 link by similar author's edges because of A_5 is similar to A_1 and A_2.

C. Representing nodes and indentification papers

Given a heterogeneous graph, we make a correlation representation of the nodes. A paper p_i is represented as a vector, shown in equation (6).

$$p_i = [S_t, S_a, S_y, C_p] \tag{6}$$

where S_t, S_a and S_y consider the similarity of papers' metadata, C_p is computed by centrality measures.

The content of an article determines whether users regard it as a paper of interest, which is also research hotspots in paper recommendation systems. In this study, papers' metadata is used to calculate the similarity in the content of candidate papers. Since metadata can be obtained free of charge, and its feasibility of calculating similarity is high, it is appropriate to calculate the similarity of nodes. In detail, we extract textual information, – title and abstract, and use vector representation. S_t given by (7) is the score of Jaccard similarity on the title vector between the paper of interest and the target paper, and S_a given by (13) is the score of Jaccard similarity on the abstract vector between the paper of interest and the target paper.

$$\mathcal{S}_t = Jaccard^{v_{j_t} \to v_{j_p}} \tag{7}$$

$$\mathcal{S}_a = Jaccard^{v_{j_a} \to v_{j_a}} \tag{8}$$

Jaccard Similarity does not only measure the extent of similarity between our target paper and any of the qualified candidate papers but also their deviations [18]. S_y is the time benefit, called Freshness, which is represented by a tanh function with a value range of (-1, 1), shown in (9). Authors are more likely to be interested in newly papers, especially when finding the development of the field or looking for research hot points.

$$\begin{aligned} \mathcal{S}_y &= \tanh(\mathcal{T}_J - \mathcal{T}_P) \\ &= \frac{e^{\mathcal{T}_J - \mathcal{T}_P} - e^{-(\mathcal{T}_J - \mathcal{T}_P)}}{e^{\mathcal{T}_J - \mathcal{T}_P} + e^{-(\mathcal{T}_J - \mathcal{T}_P)}} \end{aligned} \tag{9}$$

A centrality analysis of the network is performed to examine the importance of each node[30]. Four centrality measures are applied on candidate papers to determine the most significant papers, these include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Degree centrality is the simplest way to find important nodes. It is used to measure the importance of nodes in the network, the formula is shown in (10). Considering the timeliness of paper (the publication time), this study only uses nodes' in-degree.

$$DC(P) = \frac{\mathcal{N}(p)}{n-1} \tag{10}$$

Closeness centrality reflects the closeness of a node to other nodes. If a paper is connected to many other papers, it indicates that the quality of this paper is high. The formula is shown in formula (11).

$$CC(P) = \frac{n-1}{\sum_{\mathcal{J} \neq \mathcal{P}} d(P,J)} \tag{11}$$

Betweenness centrality is used to calculate the number of shortest paths of a node to describe the importance of each node. The formula is shown in equation (12).

$$BC(P) = \sum_{J \neq V \neq P} \frac{\mathcal{G}_{JV}(P)}{\mathcal{G}_{JV}} \tag{12}$$

Eigenvector centrality is used to measure the influence of nodes in the network. The importance of nodes is measured based on the references of other nodes in the network. The formula is shown in equation (13).

$$EC(P) = \frac{1}{\lambda} \sum_{J=B_P} \mathcal{A}_{P,J} \mathcal{X}_J \tag{13}$$

where n is the total number of papers, $\mathcal{N}(P)$ represents the cited number of papers P, d(P, J) defines the distance between paper P and J, the metric $\mathcal{G}_{JV}$ provides the number of links that pass through shortest route between paper J and V, and $\mathcal{G}_{JV}(P)$ is the number of links in shortest route between J and V that pass through paper P, $\mathcal{A}_{P,J}$ is the adjacency matrix in which its element is one if J is linked to P, and zero otherwise. $\mathcal{X}_J$ is the score of the eigenvector centrality of J, and λ is the eigenvalue of P.

Synthesize the importance of the above four methods, transform their values into determinants, and use (14) to calculate the average score of each paper.

$$\mathcal{C}_p = \frac{\sum_{m=1}^{M} rank^m(P)}{\mathcal{M}} \tag{14}$$

where $rank^m(P)$ is a ranking result with mth centrality measure on paper P, $\mathcal{M}$ is the number of centrality measures.

After we get the nodes' feature representation, we feed it to a multilayer perceptron(MLP) for scoring nodes. Our model is trained by minimizing the cross-entropy loss over all training examples:

$$\mathcal{L} = -\sum_{l \in \mathcal{P}_L} \sum_{f=1}^{F} Y_{lf} ln X_{lf} \tag{15}$$

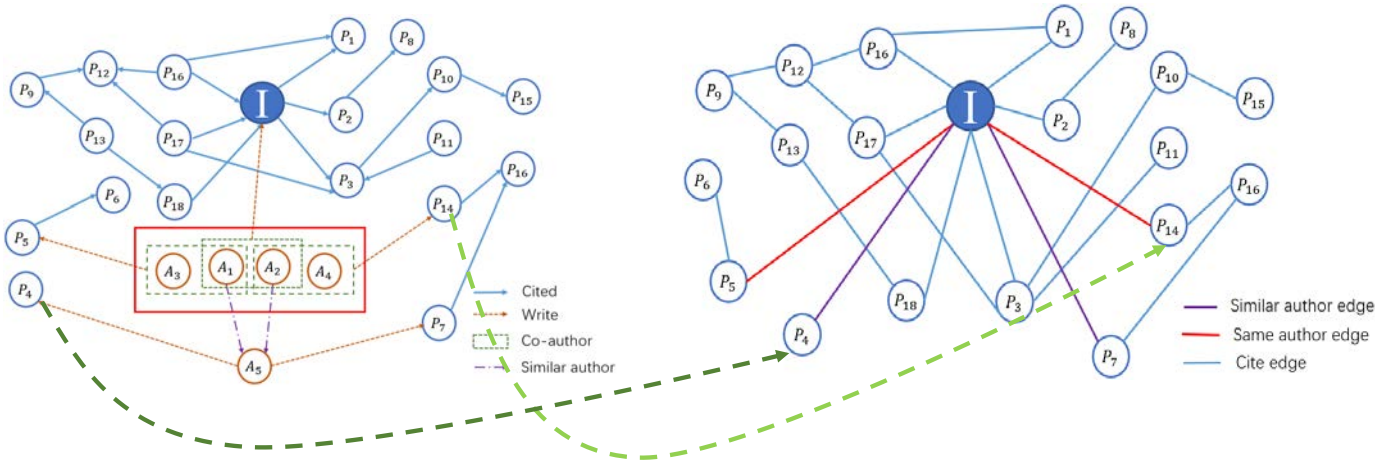

Fig. 2. An example of a heterogeneous graph. (i)Node I is the given paper;(ii)A_1 and A_2 write paper I together, A_1 and A_3 write paper P_5 together, and A_2 and A_4 write paper P_14 together;(iii)A_5 is similar to both A_1 and A_2, and write papers P_4 and P_7.

where $\mathcal{P}_L$ is the set of given papers, X_{lf} is the f-th entry of the network output for l-th given paper and Y_{lf} denotes its ground truth label.

D. Author Collaboration Score

In the researches of the paper recommendation system, there are less researches using author information for a recommendation because of the information dealing with hardly. However, in practical applications, when a user selects an article of interest, the author of the paper is used as an indicator. For example, when selecting articles of interest, users will prefer articles from experts in this research area. Because authors with higher reputations are most likely to get higher scoring standards in their research papers, such articles are also of higher quality. Therefore, this article uses authors of the paper as an evaluation index to reranking the candidate list.

We have generated an author table and calculated our own metric with the help of various factors. The following fields were included in the author table:

- Citation count: The total citation count for each author was calculated based on the citations of the paper published. Even if a paper has multiple authors, we consider that each author of the paper has the same citations.
- Number of published papers: According to the existing data, the number of papers published by each author is calculated.
- $\mathcal{I}$-10 Index: It refers to the number of papers with 10 or more citations.
- Domain Score: Domain of each author was calculated based on the domain of the papers published by them. The domain score is assigned corresponding to each author. It was calculated based on the number of papers published by all the authors in that domain. This signifies the popularity of the author in his/her domain.

According to the four fields of the author table, Author Metric (SVS Index) is generated, and the score is calculated given by (16).

$$SVS_{\mathcal{A}_i} = 0.1 * \mathcal{S}_{c/p} + 0.2 * \mathcal{S}_{I-10} + 0.7 * \mathcal{S}_{Domain} \tag{16}$$

where $\mathcal{S}_{c/p}$ indicates the citation score of each article given by (17).

$$\mathcal{S}_{c/p} = \begin{cases} log(\frac{num_c}{num_p}) + 0.5, & \text{if} \frac{num_c}{num_p} > 1 \\ e^{\frac{num_c}{num_p}} - 0.5, & \text{otherwise} \end{cases} \tag{17}$$

We had fine-tuned our parameters to change their weights by applying more complex models on our data set and checked the importance of each feature given by our training model.

$$\mathcal{S} = 0.65 * P_{score} + 0.45 * max_{A_I \in A_J}(SVS_{\mathcal{A}_i}) \tag{18}$$

where P_{score} is the nodes' score after MLP; A_J is the authors' set of a paper, parameters are obtained from multiple experiments.

For the candidate papers, combining the importance and similarity of the previous article, we add the authors' influence factor, using (18) to calculate the score, final we return the personalized recommendation results.

IV. Experiment Evaluation

In this section, we introduce our experimental setup and a series of experiments to evaluate the performance of the proposed AMHG method on the generated DBLP-REC dataset.

A. Dataset

DBLP-Citation-network V11 contains more than 4.1 million papers and more than 36 million citation relationships. It is mainly used, including papers' metadata such as title, abstract, authors and publication time. At the same time, in order to expand the useful data, we collected ScienceDirect's 50,000 available papers' metadata. Besides, we use IEEE data to fill missing data in the dataset and finally, we obtain 3.59 million papers and 35.25 million citation relationships, more details about the dataset can be found in Table I.

TABLE I
DBLP-REC DATASET DESCRIPTION

Papers	Authors	Citation relationship
3,590,853	3,276,803	35,254,530

B. Methodlogy and Metrics

In recent years, in the research of recommendation systems for papers, there are mainly three methods for measuring the accuracy and user satisfaction of recommendation systems, which are offline evaluations, online evaluations, and user studies [17]. Offline evaluations typically measure the accuracy of a recommender system based on true value. To measure accuracy, precision at position n (P@n) is often used to express how many items of the ground-truth are recommended within the top n recommendations. Online evaluations measure the acceptance rates of recommendations in real-world recommender systems. This evaluation is expensive and time-consuming. User studies typically measure user satisfaction through explicit ratings. Aiming at the proposed AMHG method, it is compared with the two methods of MSCN [11] and CNRN [27].

For evaluation, precision(P), mean reciprocal rank(MRR) and Mean Average Precision(MAP) are used. Precision given by (19), measures the capability of proposed systems to reclaim as much relevant research papers as possible in response to the target paper request.

$$Precision = \frac{\Sigma(relevant_p) \cap \Sigma(retrieved_p)}{\Sigma(retrieved_p)} \tag{19}$$

As users often scan only documents presented at the top of the recommendation list, we use MAP and MRR to estimate the system's ability to provide useful recommendations at the top of the recommendation list. MRR given by (19), represents the ranking level at which the system returned the first relevant research paper averaged overall researchers. It measures the extent of the system to return a relevant paper at the top rank of the recommendation list.

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{rank(i)} \tag{20}$$

where rank(i) is the highest-ranking where the first relevant paper i appears, n represents the total number of target papers.

Average precision (AP) is the average of precision values at all ranks where relevant papers are found. MAP given by (20), is the average of all APs.

$$MAP = \frac{1}{\mathcal{I}}\sum_{i\in\mathcal{I}}\frac{1}{ni}\sum_{k=1}^{\mathcal{N}}\mathcal{P}(\mathcal{R}_{ik}) \tag{21}$$

where $\mathcal{P}_{ik}$ denotes the precision of returned papers from the top until paper k is reached, $\mathcal{N}$ is the length of the recommendation list, ni presents the number of relevant papers in the recommendation list, and I defines the set of papers.

C. Results and Discussions

Specifically, the results of each of the evaluation indicators in this section represent the overall average of all the researchers who conducted the trial in our dataset. Based on the most commonly used index of information retrieval, we evaluate the performance of our proposed method in the field of paper recommendation. We compare AMHG against the production baseline, as well as the other baselines, see table II. The MRR and MAP are the top ten recommender papers' result. After adding the metadata information to the network, the precision increased by more than 20 percentage points. Adding the authors' relationship to construct a heterogeneous graph, the MRR increased by more than 15 percentage points.In conclusion, our proposed AMHG method achieves the best recommendation results among all the competitors.

TABLE II
RESULTS OF PAPER RECOMENDATIONS ACROSS MODELS

Methods	***Prec@1***	***Prec@10***	***MRR***	***MAP***
MSCN	0.506	0.497	0.296	0.528
CNRN	0.532	0.508	0.408	0.583
Heterogeneous graph	0.614	0.511	0.552	0.587
MSCN+METADATA	0.767	0.653	0.553	0.571
AMHG	**0.786**	**0.678**	**0.647**	**0.649**

TABLE III
THE FIRST 10 ARTICLES ARE IN 2016-2019

Methods	***1-3 papers***	***4-6 papers***
MSCN	0.213	0.113
CNRN	0.242	0.124
Heterogeneous graph	**0.382**	**0.241**

The reranking performance of the paper candidates varies by the factors of papers. Generally, the similarity measures based on metadata information better than those that do not. The authors' relationship makes the recommendation list having more newly papers. We count the papers which are published between 2016 to 2019, from the table III, using heterogeneous graph can recommend some newly published articles in the candidate collection of about 62% of academic papers. So our proposed method can alleviate the cold start problem.

Figure 3-5 shows the comparison of the result of MSCN, CNRN and AMHG based on Precision, MRR, and MAP, respectively. The MSCN+METADATA method is the network by adding metadata on the basis of the multi-level citation network. It can be seen that the accuracy of the return result of the multi-level citation network is greatly improved after the introduction of metadata.

As can be seen in Figure 3, the accuracy of the MSCN method is not high; the CNRN method is based on MSCN and added to the author's evaluation. If the accuracy of MSCN itself is not high, then the CNRN will not be greatly improved; the MSCN+METADATA method greatly improves

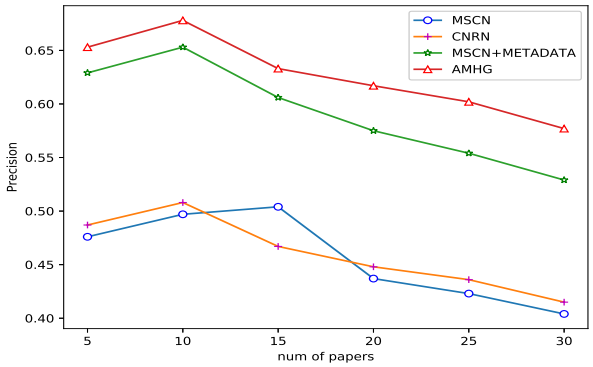

Fig. 3. Precision performance on the dataset

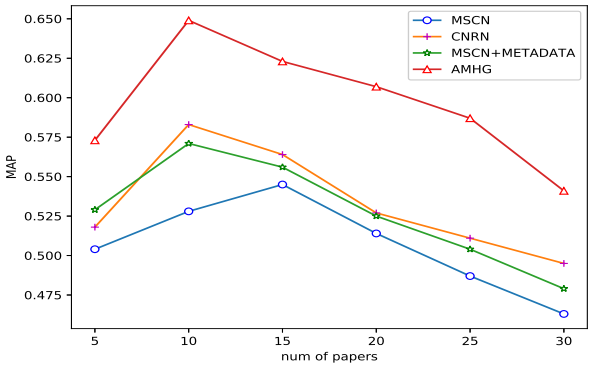

Fig. 5. MAP performance on the dataset

the accuracy of the returned results; AMHG considers the accuracy of MSCN is not high, and the content similarity of the article is considered when building the citation network so as to delete the papers with less relevance to the target paper.

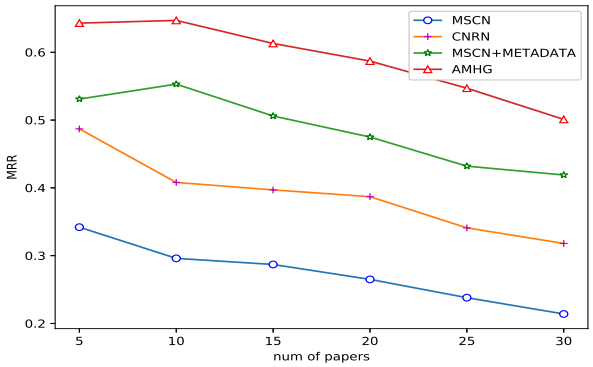

Fig. 4. MRR performance on the dataset

Because readers usually look at the first few articles of the recommendation list, they need to evaluate the situation of the first n papers returned by the recommendation system. Figures 4 and 5 show the MRR and MAP scenarios for several methods. It can be seen that among the several methods, our proposed method is obviously optimal. And in returning the first 10 articles, the effect is the best. As the number of papers increases, due to data limitations, the use of multi-level citation networks and chain-based methods may not be able to screen out those related but low citations article.

As we have pointed out earlier, all these improvements are largely based on the use of metadata for similarity calculations for candidate papers. This ensures the systems' accuracy. The system combines the authors' influence factor to increase its ability to return relevant and useful recommendations at the top of the recommendation list. Therefore, this improves user satisfaction and personalization.

V. Conclusion

Because of the challenge of dealing with big data, it is very important to extract relevant files using a recommendation system. This paper aims at the task, which is to find more useful papers for users. In this paper, we use the citation relationship to mine the implied relationship between the paper and its references and use metadata information to view the content similarity of the candidate papers in the multi-level citation network. We then introduce author evaluation factors and finally recommend a related paper to the researchers.

As demonstrated by the data sets used, our paper recommendation method outperforms the baseline method in terms of overall performance and the ability to return relevant and research worthy papers. Based on the three most commonly used indicators, our proposed method has greatly improved the accuracy, and we have also significantly improved the recommended ranking compared with the baseline. However, because the data in this paper are offline data, we can't get the user's use log. Otherwise, the recommendation list based on an article combined with the user's history can be a better-personalized recommendation. At the same time, when using metadata, we think that we can also improve our proposed method by considering the author's organization and published journal ranking and other information. If we use the method of Hin combined with GNN to classify the nodes after building the citation network graph, it may be more helpful to improve the experimental operability. We will focus on this issue in the future.

Acknowledgment

This work is supported by the National Key R&D Problem of China(NO.Y850371101). We thank all authors for the contributions and all anonymous reviewers for their constructive comments.

References

[1] J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, "Recommender system application developments: a survey," Decis. Support Syst. 74 (2015) 12–32.

[2] S. Mukherjee, D. M. Romero, B. Jones, and B. Uzzi, "The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot," Sci. Adv., vol. 3, no. 4, 2017,Art. no. e1601315.
[3] Deveaud R, Mothe J, Ullah M Z, et al. Learning to Adaptively Rank Document Retrieval System Configurations[J]. ACM Transactions on Information Systems (TOIS), 2018, 37(1): 3.
[4] Google Scholar. 2018. Retrieved from https://scholar.google.com/
[5] ScienceDirect. 2015. Retrieved from https://sciencedirect.com.
[6] Li Xinyi, Chen Yifan ,Pettit Benjamin, Rijke, Maarten. (2019). "Personalised Reranking of Paper Recommendations Using Paper Content and User Behavior," ACM Transactions on Information Systems. 37. 1-23. 10.1145/3312528.
[7] Liu H., Kong X., Bai X., Wang W., Bekele T. M., and Xia F., "Context-based collaborative filtering for citation recommendation," IEEE Access, vol. 3, pp. 1695–1703, 2015.
[8] Bai, Xiaomei, Wang, Mengyang, Lee, Ivan, Yang, Zhuo, Kong, Xiangjie and Xia, Feng. "Scientific Paper Recommendation: A Survey.." IEEE Access 7 (2019): 9324-9339.
[9] N. J. van Eck and L. Waltman, "Citation-based clustering of publications using CitNetExplorer and VOSviewer," Scientometrics, vol. 111, no. 2,pp. 1053–1070, 2017.[10] D. Yu, Z. Xu, W. Pedrycz, and W. Wang, "Information sciences1968–2016: A retrospective analysis with text mining and bibliometric," Inf. Sci., vols. 418–419, pp. 619–634, Dec. 2017.
[10] D. Yu, Z. Xu, W. Pedrycz, and W. Wang, "Information sciences 1968–2016: A retrospective analysis with text mining and bibliometric," Inf. Sci., vols. 418–419, pp. 619–634, Dec. 2017.
[11] J. Son and S. B. Kim, "Academic paper recommender system using mul-tilevel simultaneous citation networks," Decision Support Syst., vol. 105,pp. 24–33, Jan. 2018.
[12] Doerfel S, Jäschke R, Hotho A, et al. Leveraging publication metadata and social data into folkrank for scientific publication recommendation[C]//Proceedings of the 4th ACM RecSys workshop on Recommender systems and the social web. ACM, 2012: 9-16.
[13] M. T. Afzal, N. Kulathuramaiyer, and H. A. Maurer, " Creating links intothe future," J. Universal Comput. Sci., vol. 13, no. 9, pp. 1234–1245, 2007.
[14] K. Haruna, M. A. Ismail, D. Damiasih, J. Sutopo, and T. Herawan, "A col-laborative approach for research paper recommender system," PLoS ONE,vol. 12, no. 10, 2017, Art. no. e0184516.
[15] Ma, Xiao, Wang, Ranran. (2019). "Personalized Scientific Paper Recommendation Based on Heterogeneous Graph Representation," IEEE Access. 7. 1-1. 10.1109/ACCESS.2019.2923293.
[16] Haruna K, Akmar Ismail M, Damiasih D, Sutopo J, Herawan T (2017), "A collaborative approach for research paper recommender system," PLOS ONE 12(10): e0184516.
[17] Beel J , Gipp B , Langer S , et al. "Research-paper recommender systems: A literature survey[J]," International Journal on Digital Libraries, 2015:1-34.
[18] N. J. van Eck and L. Waltman, "Citation-based clustering of publicationsusing CitNetExplorer and VOSviewer," Scientometrics, vol. 111, no. 2,pp. 1053–1070, 2017.
[19] Z. Huang, W. Chung, T.-H. Ong, and H. Chen, "A graph-based recommender system for digital library," in Proc. 2nd ACM/IEEE-CS Joint Conf.Digit. Libraries, Jul. 2002, pp. 65–73.
[20] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction incomplex networks," ACM Comput. Surv., vol. 49, no. 4, p. 69, Feb. 2017.
[21] Z. Xu, H. Jiang, X. Kong, J. Kang, W. Wang, and F. Xia, "Cross-domainitem recommendation based on user similarity," Comput. Sci. Inf. Syst.,vol. 13, no. 2, pp. 359–373, 2016.
[22] J. Niu, L. Wang, X. Liu, and S. Yu, "FUIR: Fusing user and iteminformation to deal with data sparsity by using side information inrecommendation systems," J. Netw. Comput. Appl., vol. 70, pp. 41–50,Jul. 2016.
[23] A. Tsolakidis, E. Triperina, C. Sgouropoulou, and N. Christidis, "Research publication recommendation system based on a hybridapproach," in Proc. ACM 20th Pan-Hellenic Conf. Inform., 2016,pp. 78–83.
[24] K. Sugiyama and M.-Y. Kan, "Exploiting potential citation papers inscholarly paper recommendation," in Proc. 13th ACM/IEEE-CS Joint-Conf. Digit. Libraries, 2013, pp. 153–162.
[25] J. Beel, A. Aizawa, C. Breitinger, and B. Gipp, "Mr. DLib: Recommendations-as-a-service (RaaS) for academia," in Proc.ACM/IEEE Joint Conf. Digit. Libraries (JCDL), Jun. 2017,pp. 1–2.
[26] L. Egghe, R. Rousseau, "Co-citation, bibliographic coupling and a characterization oflattice citation networks," Scientometrics 55 (3) (2002) 349–361.
[27] Waheed W , Imran M , Raza B , et al. "A Hybrid Approach towards Research Paper Recommendation using Centrality Measures and Author Ranking[J]," IEEE Access, 2019:1-1.
[28] B. Carrera, J. Lee, J.Y. Jung, "Discovery of information diffusion process in online social networks," Int. J. Ind. Eng. Theory Appl. Pract. 23 (4) (2016).