SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE AND COGNITIVE COMPUTING FOR COMMUNICATION AND NETWORK

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 20, 2019, accepted June 8, 2019, date of publication June 17, 2019, date of current version July 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923293

Personalized Scientific Paper Recommendation Based on Heterogeneous Graph Representation

XIAO MA AND RANRAN WANG
School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China

Corresponding author: Xiao Ma (cindyma@zuel.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802440 and Grant 61702553, and in part by the Fundamental Research Funds for the Central Universities, ZUEL: 2722019JCT037.

ABSTRACT The accelerating rate of scientific publications makes it extremely difficult for researchers to find out the relevant papers and related works. Recommender systems that aim at solving the *information overload* problem have attracted lots of attention. However, existing paper recommendation works generally rely on the simple citation-ships between papers, which ignore the heterogeneity of the academic graphs. In this paper, we solve the personalized paper recommendation problem in the setting of heterogeneous information networks. A heterogeneous graph representation based recommendation method named HGRec is proposed. First, the author and paper profiles are constructed based on the extracted contents information. Second, we initialize the node vectors by employing the word-embedding technique. Third, we jointly update the node embeddings in the heterogeneous graph by proposing two meta-path based proximity measures. Finally, the paper recommendation is completed by calculating the similarity of the generated author and paper feature vectors. We present experiments on a real academic network, the DBLP network. The comparative results demonstrate the effectiveness of the proposed personalized recommendation approach compared to state-of-the-art methods.

INDEX TERMS Recommender systems, paper recommendation, heterogeneous information networks, graph representation, meta-paths.

I. INTRODUCTION

With the development of information science and technology, great achievements have been made in terms of electronic literatures. The rapid growth in the number of scientific papers makes it difficult for researchers to find out what they really care about. As we all know that, relevant research papers are important for researchers to keep up with the latest research progresses in their research areas.

In order to solve this problem, literature retrieval systems which are designed to help researchers find related papers have been studied in the past years. A literature retrieval system usually begins with the user's query, then a retrieval model is chosen to process the request. Finally, the search engine returns the most related results with respect to the query of users [1], [2]. Although these systems make it easier for researchers to find interesting papers, keywords based systems still return thousands or millions of relevant papers. It is time-consuming for researchers to figure out which paper to read or cite, especially for the unexperienced researchers.

Paper recommender systems which aim to recommend the most relevant papers to researchers have been studied to tackle the afore-mentioned problem [3]. As far as we know that, existing works generally rely on the citation-ships between papers to make recommendation. Some representative works include graph-based methods [4], [5], collaborative filtering based methods [6], [7]. However, as shown in Figure 1, the real academic information networks are generally heterogeneous graphs [8], which contains multiple types of entities (i.e., author, paper, venue, topic) and relationships (i.e., writing, publishing, collaborating). Simple paper-paper citation-ships are not sufficient to capture the rich semantics of the academic graphs.

Recently, some meta-pattern based graph modeling methods have been proposed to solve the paper recommendation problem by employing meta-paths or meta-graphs of various semantics in heterogeneous networks [9]–[11]. Although having been demonstrated to generate promising recommendation results, these methods still suffer the efficiency problem due to the extraction strategies for the meta-patterns [12].

With the development of deep learning techniques, graph representation learning has drawn lots of attention in the past several years. For example, DeepWalk [13] is very successful

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

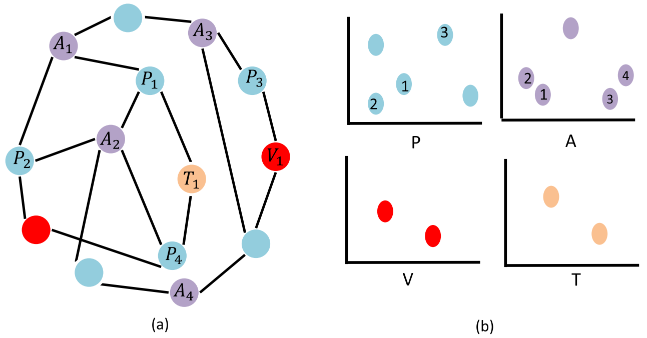

FIGURE 1. A toy example of the heterogeneous academic network. (a) Heterogeneous academic graph A, P, V, T represent author, paper, venue, term. (b) 2D projection of different types of nodes in each embedding space.

in representing large homogeneous graphs. Metapath2vec extends their work and introduces a method to learn the representations of nodes in heterogeneous graphs [14]. Inspired by the work of heterogeneous graphs representation [15], in this paper, we propose a heterogeneous graph representation based personalized scientific paper recommendation method. Unlike existing works which transfer the heterogeneous graphs into simple subgraphs [16] or homogeneous graph [15], we directly learn the embeddings of all types of nodes from the original heterogeneous graph. Besides, the contents information of papers, i.e., title, keywords, abstract, are also incorporated into the representation model to generate better recommendation results. The contributions of our work are as follows:

- We solve the problem of personalized scientific paper recommendation problem in the setting of heterogeneous information networks. A novel heterogeneous graph representation learning based recommendation method named HGRec is proposed.
- We first construct the user and paper profiles by extracting the contents information, and the Doc2vec technique is employed to initialize the node representations in the heterogeneous graphs.
- We propose two meta-path based proximities to measure the relevance of node representations in the heterogeneous graphs. HGRec obtains the representations of users and papers by jointly training and updating with respect to these two proximities.
- Compared with several baselines, we conduct experiments based on a real-world dataset and the comparative results demonstrate the effectiveness of our proposed method for the task of personalized paper recommendations.

The rest of this paper is organized as follows. We first briefly review the related work of paper recommendation and graph representation learning in Section II. Afterwards, the proposed recommendation model (HGRec) is presented in Section IV. In Section V, extensive comparative experiments are conducted to validate the superiority of the proposed HGRec method. Finally, Section VI draws a conclusion with a future work.

II. RELATED WORK

A. SCIENTIFIC PAPER RECOMMENDATION

The task of scientific paper recommendation is to offer researchers a list of relevant papers that researchers would like to read or cite in the future. Generally, paper recommendation methods can be classified into three categories: content-based, collaborative filtering based and graph-based [3]. In content-based methods, the researcher profile is firstly constructed, which may include the researcher's publications and his/her cited papers. Then, a researcher and paper feature vectors are generated based on the TF-IDF model [17] or keyphrase extraction model [18]. Finally, the recommendation list can be generated by calculating the similarity between the researcher and paper feature vectors [17].

Collaborative filtering (CF) is a very popular and successful technique in recommender systems [19]–[21]. CF based paper recommendation methods are very effective in recommending relevant papers when content information is not available. The idea of CF is two users A and B give ratings on some common items, these two users are considered to be similar. Therefore, if the papers are cited by A but not by B, it is intuitive to recommend these papers to B [6], [22], [23]. For example, Yang et al. [7] propose a joint collaborative filtering based model that can exploit the latent correlation between relations and solve several tasks (i.e., author/paper/venue recommendations) in a unified way.

Graph-based paper recommendation methods firstly construct an academic graph [5]. The papers are treated as vertices, and the paper-paper citation-ships are treated as the edges. The recommendation task can be transformed into a graph search [24] or link prediction problems [25]. For example, Anand et al. [26] employ random walk on the paper citation graphs to balance both the relevance and diversity while searching for research papers.

Hybrid recommendation which combines two or more recommendation techniques have also been introduced to solve the paper recommendation tasks [3], [27]. Content-based + collaborative filtering based and content-based + graph-based hybrid methods are two representative hybrid paper recommendation strategies. In this paper, we firstly rely on the contents to build the researcher and paper profiles, and then solve the paper recommendation problem by representing the heterogeneous academic graph. Therefore, the proposed HGRec belongs to the scope of hybrid paper recommendation methods.

B. GRAPH REPRESENTATION LEARNING

Graph representation learning was firstly introduced by DeepWalk [13], which got the inspiration from the word2vec model of the natural language processing research field. Given a text corpus, Mikolov et al. designed word2vec to learn the distributed representations of words in a corpus [28]. DeepWalk was designed to map the the word-context in the text corpus into a network. Similarly, node2vec [29] shares the same idea. Both of them employ the random walks to

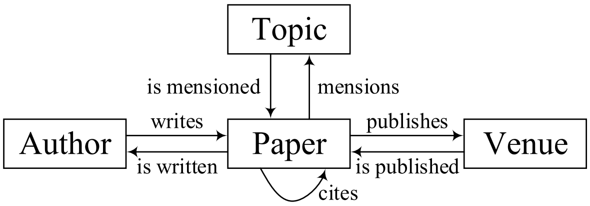

FIGURE 2. DBLP network schema.

generate the node distributions and learn the representations of node by utilizing the skip-gram model [13]. LINE [30] is also a representative graph representation method. All of these methods are designed to learn the embeddings of homogeneous graphs.

Recently, some researchers focus on the study of heterogeneous graphs representation. For example, PTE [31] relies on both the labeled and unlabeled data to learn the embeddings of text in a large-scale heterogeneous text network. Metapath2vec [14] firstly extracts the neighbors for nodes in the heterogeneous graphs with respect to each meta-path and introduces a heterogeneous-based skip-gram model to learn the representations of nodes. Shi et al. [15] propose HERec which introduces a meta-path based random walk algorithm to generate the node sequences and learn the embedding of nodes by transforming the heterogeneous node sequence into homogeneous ones. Different from HERec, our proposed HGRec learns the embeddings of various types of nodes directly from the heterogeneous graphs without transferring the sampled node sequences into homogeneous ones. Cai et al. [32] propose a deep network representation based citation recommendation method which integrates the graph structure and content information of nodes into a unified model by employing the generative adversarial network. However, they represent different types of nodes in a continuous and common vector dimension which is very space-consuming in large heterogeneous graphs.

III. PRELIMINARIES

In this section, we briefly introduce the notations and definitions used throughout this paper. In addition, some concepts related to the heterogeneous information networks (HIN) and the paper recommendation problem are also presented.

A heterogeneous graph is a directed graph which contains multiple types of nodes and links. More details about heterogeneous information networks can be found in [8]. Paper recommendation systems aim to provide a list of papers which are the most relevant to a given researcher.

Definition 1 (HIN Schema) [33]: The HIN schema is a meta template of heterogeneous network $G = (V, E)$ with an object type mapping function $\phi : V \rightarrow \mathcal{A}$ and the link mapping $\varphi : E \rightarrow \mathcal{R}$, which is a directed graph defined over object types $\mathcal{A}$, with edges as relations from $\mathcal{R}$, denoted as $T_G = (\mathcal{A}, \mathcal{R})$.

The HIN schema describes all the available link types between object types. Figure 2 is an example of the HIN schema with respect to the DBLP bibliographic network.

Definition 2 (Meta-Paths) [33]: A meta-path Π is a path defined on an HIN schema $T_G = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $\Pi : A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_L} A_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ between type A_1 and A_{l+1}, where $\circ$ represents the composition operator on relations.

As shown in Figure 1, A, P, T, V represent the authors, papers, terms and topics, respectively. According to Def. 1 and Def. 2, the meaningful meta-paths include APA, PAP, PTP, PVP, PAPAP et al.

Definition 3 (Heterogeneous Information Network Representation Learning): For a given network $G(V, E)$, V and E represent the set of nodes and edges, respectively. $|V|(|V| > 1)$ and $|E|(|E| > 1)$ denote the node types and edge types. The goal of heterogeneous information network representation learning is to find a mapping ψ_1 which will output a d_i dimensional vector λ_j^i to represent each node instance ${v_j}^i$ of the i-type nodes in the heterogeneous networks.

$$\psi_1({v_j}^i) \rightarrow \lambda_j^i \tag{1}$$

where ${v_j}^i$ is the j-th node in the i-type node set. d_i represents the dimension of the projection space of the i-type node.

Definition 4 (Meta-Path Based First-Order Proximity): Given a meta-path Π, if the node type of the starting point and end point is the same, we assume that the starting point and end point of any instances of the meta-path are similar.

For example, in Figure 1(a), $(P_1 - A_1 - P_2)$ is an instance of meta-path PAP. P_1 and P_2 represent two papers written by the same author in a heterogeneous information network, which indicates that papers P_1 and P_2 are similar. As we can find that nodes P_1 and P_2 are also closer than other nodes in Figure 1(b).

Definition 5 (Meta-Path Based Second-Order Proximity): For any two instances of a given meta-path Π, if the starting point and end point of the instances are the same, we assume that the center nodes of the two instances are similar.

For example, in Figure 1(a), $(P_1 - A_1 - P_2)$ and $(P_1 - A_2 - P_2)$ are two instances of the meta-path PAP, which share the same starting point P_1 and end point P_2. It can be seen from this figure that nodes A_1 and A_2 are structurally similar. Besides, nodes A_1 and A_2 are also closer than other nodes in Figure 1(b).

From these two examples we can conclude that, the meta-path based first-order proximity focuses more on the node-level similarity, while the meta-path based second-order proximity focuses more on the structural-level similarity.

IV. THE PROPOSED APPROACH

A. FRAMEWORK DESCRIPTION

The framework of this work can be found in Figure 3. Firstly, the contents information and HIN graphs are extracted from the original heterogeneous data. Then we construct the user and paper profiles with the contents information of papers. Thirdly, user and paper embeddings are generated based on the pre-trained Doc2vec techniques [34]. Afterwards,

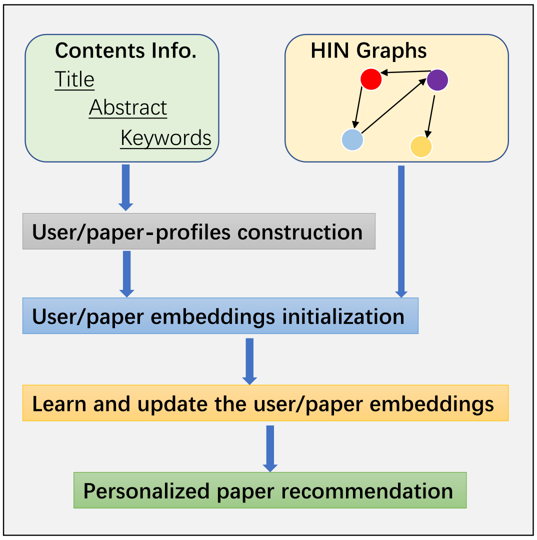
Contents Info.
Title
Abstract
Keywords
HIN Graphs
User/paper-profiles construction
User/paper embeddings initialization
Learn and update the user/paper embeddings
Personalized paper recommendation

FIGURE 3. The framework of HGRec.

the embeddings of various types of nodes are updated and learned with respect to the heterogeneous graphs by employing our HGRec method. Finally, the recommendation results are generated by calculating the cosine similarity between the final user feature vectors and paper feature vectors, and a list of top-ranked papers will be recommended to the target researchers.

B. USER AND PAPER PROFILES CONSTRUCTION

Given a paper p_i, the profile of p_i is composed of the title, keywords, abstract of this paper. All these textual contents represent the interest of paper p_i. Given a user u_j, the profile of u_j is composed of his/her published papers. That is to say, the interest of user u_j is represented by all his/her previous publications. Since the number of researchers publications follow the power-law distribution [35], which means the majority of researchers only have few publications.

Suppose a user has no publications, which means he/she is a new user in the system and has no explicit feedbacks about his/her interests. A better way to get the explicit feedback is to ask the user to specify papers that are interested by him/her. In the meantime, we can also collect the implicit feedbacks for the user. For example, the viewing logs on the abstracts and clicking links to full-text articles. For simplicity, in this paper, we assume that users have at least one publication.

C. THE PROPOSED HGREC METHOD

Following the approach in [32], the initialization of user and paper embeddings can be generated by employing the Doc2vec technique [34] based on the user and paper profiles constructed in Section IV-B.

1) THE OBJECTIVE FUNCTION

In this section, the objective function of our HGRec method will be presented. First of all, let us introduce some background knowledge about the general definition of objective function in the graph representation learning.

Suppose M_j and M_k are two nodes, λ_j and λ_k are their embeddings. Therefore, the similarity of nodes M_j and M_k can be calculated by the inner product of the embeddings, which is denoted as $S_{jk} = \lambda_j^T \lambda_k$ [36]. The larger S_{jk} is, the more similar they are. Given another node M_t, suppose M_t is less similar to M_j than M_k. Intuitively, S_{jk} will be larger than S_{jt}. That is to say, the distance between nodes M_j and M_k should be smaller than the one between nodes M_j and M_t. Specifically, we model $S_{jk} > S_{jt}$ using the logistic function $\sigma(x) = \frac{1}{1+e^x}$. Therefore, the objective function which describes the relationships of nodes in the embedding space can be defined as follows:

$$\forall (M_j, M_k, M_t) \in G, maxpro(\sigma(S_{jk} - S_{jt}) | \lambda_j, \lambda_k, \lambda_t) \qquad (2)$$

The afore-mentioned objective function is designed for homogeneous graphs. In order to measure the similarities between nodes in the heterogeneous graphs, two meta-path based proximity measures which defined in Section III will be incorporated into the definition of our objective functions.

Specifically, S'_{jk} represents the meta-path based first-order proximity, and S''_{jk} represents the meta-path based second-order proximity. Similar to Equation 2, we minimize the sum of negative log-likelihood objective functions which can be defined as follows:

$$OBJ_M = min - \ln \sigma(S'_{jk} - S'_{jt}) + \gamma_1 Reg(M) \qquad (3)$$

$$OBJ_N = min - \ln \sigma(S''_{jk} - S''_{jt}) + \gamma_2 Reg(N) \qquad (4)$$

where $\gamma_1 Reg(M)$ and $\gamma_2 Reg(N)$ are l_2-norm regularization terms to avoid overfitting. γ_1 and γ_2 are the penalty coefficients with respect to the two proximity measures. $Reg(M)$ and $Reg(N)$ are set as $\|M\|_F^2$ and $\|N\|_F^2$, respectively.

Note that Equation 3 and Equation 4 are jointly learned in the training process in order to update the embedding of all types of nodes exist in multiple meta-paths. More details about the updating can be found in Section IV-C.2. The difference between Equation 3 and Equation 4 can be briefly explained as follows: given a meta-path $\Pi(MNM)$, OBJ_M pays more attention to the starting node and ending node of Π with respect to the meta-path based first-order proximity, while OBJ_N focuses more on the centering node of Π with respect to the meta-path based second-order proximity.

2) EMBEDDING UPDATES

The stochastic gradient descent [37] optimization strategy is employed to update the node embeddings in the heterogeneous graphs. Both the positive and negative samples are used for the two similarities.

Given a meta-path $\Pi(MNM)$, $(M_j - N_j - M_k)$ is one of its instances. According to Definition 4, $(M_j - N_j - M_k)$ will be used to update the embedding of node M_j, the positive training sample M_k and negative sample M_t. More details about the positive and negative samples can be found in Section V-B. In the updating procedure, λ_k^M should be closer to λ_j^M than

TABLE 1. Selected meta-paths and explanations.

Meta-path	Explanation
APA	Two authors collaborate on a paper.
PAP	Two papers are written by the same author.
PVP	Two papers are published on the same venue.
PTP	Two papers share the same research topic.
APAPA	Two authors share the same collaborator.
APVPA	Two authors have published papers on the same venue.
APTPA	Two authors have published papers on the same research topics.
PAPAP	Two papers are written by two authors who have collaborated before.

λ_t^M. Suppose $(M_j - N_k - M_k)$ and $(M_{j1} - N_t - M_{k1})$ are two new instances of meta-path $\Pi(MNM)$. Similarly, according to Definition 5, nodes N_k and N_t are the positive and negative training samples of node N_j. Thus it is necessary to update the λ_j^N, λ_k^N, λ_t^N in order to make them follow the second-order proximity in the original heterogeneous graphs.

According to Equation 3, the meta-path based first-order proximity is updated as follows:

$$\lambda_j^M = \lambda_j^M - \alpha \frac{\partial OBJ_M}{\partial \lambda_j^M}$$
$$\lambda_k^M = \lambda_k^M - \alpha \frac{\partial OBJ_M}{\partial \lambda_k^M}$$
$$\lambda_t^M = \lambda_t^M - \alpha \frac{\partial OBJ_M}{\partial \lambda_t^M} \tag{5}$$

where α is the learning rate. The gradients of λ_j^M, λ_k^M, λ_t^M can be computed as follows:

$$\frac{\partial OBJ_M}{\partial \lambda_j^M} = (\sigma(S'_{jk} - S'_{jt}) - 1)(\lambda_k^M - \lambda_t^M) + 2\gamma_1 \lambda_j^M$$
$$\frac{\partial OBJ_M}{\partial \lambda_k^M} = (\sigma(S'_{jk} - S'_{jt}) - 1)(\lambda_j^M) + 2\gamma_1 \lambda_k^M$$
$$\frac{\partial OBJ_M}{\partial \lambda_t^M} = (\sigma(S'_{jk} - S'_{jt}) - 1)(-\lambda_j^M) + 2\gamma_1 \lambda_t^M \tag{6}$$

Similarly, the meta-path based second-order proximity is updated as follows:

$$\lambda_j^N = \lambda_j^N - \alpha \frac{\partial OBJ_N}{\partial \lambda_j^N}$$
$$\frac{\partial OBJ_N}{\partial \lambda_j^N} = (\sigma(S''_{jk} - S''_{jt}) - 1)(\lambda_k^N - \lambda_t^N) + 2\gamma_1 \lambda_j^N \tag{7}$$
$$\lambda_k^N = \lambda_k^N - \alpha \frac{\partial OBJ_N}{\partial \lambda_k^N}$$
$$\frac{\partial OBJ_N}{\partial \lambda_k^N} = (\sigma(S''_{jk} - S''_{jt}) - 1)(\lambda_j^N) + 2\gamma_1 \lambda_k^N$$
$$\lambda_t^N = \lambda_t^N - \alpha \frac{\partial OBJ_N}{\partial \lambda_t^N} \tag{8}$$
$$\frac{\partial OBJ_N}{\partial \lambda_t^N} = (\sigma(S''_{jk} - S''_{jt}) - 1)(-\lambda_j^N) + 2\gamma_1 \lambda_t^N \tag{9}$$

3) RECOMMENDATION OF PAPERS

Up to now, the final user and paper embeddings which represent the interests of users and papers are obtained after training and updating. Following the idea in content-based paper recommendation method [17], the recommender system will compute the cosine similarities between the user feature vectors and paper feature vectors. The more similar they are, the more relevant they will be. Thus, highly relevant papers will be ranked first for recommendation.

TABLE 2. Statistics of the dataset.

NodeType	Author	Paper	Term	Venue
# of nodes	39,530	32,133	15,708	20
LinkType	# of link	Semantic meaning		
P-A/A-P	109,584	is published/publishes		
P-T/T-P	32,132	mentions/is mentioned		
P-V/V-P	32,133	is published/publishes		
P-P	67,435	cites/is cited		

V. EXPERIMENTS

In this section, we present our experiment settings and conduct series of experiments to evaluate the performance of the proposed HGRec method on DBLP-Citation-network V8[1] generated by [38].

A. DATASET

Note that the original DBLP dataset[2] does not contain any paper-paper citation relationships. Tang et al. extracted the citation information from other sources and generated a DBLP citation dataset for the purpose of research. Instead of using the entire dataset, a subset which contains 32,133 papers from 20 venues,[3] 39,530 researchers and 15,708 topics is used in the experiment [11]. More details about the dataset can be found in Table 2. The papers are published from 2000 to 2016, and the topics are extracted from paper titles.

B. POSITIVE AND NEGATIVE SAMPLES

The meta-paths that we are interested in are listed in Table 1. All the samplings and updating are guided by these given meta-paths. Each meta-path instance is used to directly update the embeddings of the neighboring or structural similar nodes with respect to Definition 4 and Definition 5.

[1] https://aminer.org/billboard/citation

[2] http://dblp.uni-trier.de/

[3] 20 very significant venues in the areas of Data Mining, Database, Information Retrieval and Artificial intelligence.

Let's see a simple example. Given a meta-path *APA*, as shown in Figure 1(a), (A_1-P_2-A_2), (A_3-P_3-A_4), (A_1-P_1-A_2) are three instances of *APA*. As for (A_1-P_2-A_2), nodes A_2 and A_4 can be treated as the positive and negative training samples for node A_1 considering the meta-path based first-order proximity, and nodes P_1 and P_3 can be treated as the positive and negative training samples for node P_2 considering the meta-path based second-order proximity.

C. METHODOLOGY AND METRICS

In the experiments, 5-fold cross validation is performed. In each fold, 80% of the data is treated as the training set and the remaining 20% as the test set. Our aim is to recommend the more relevant papers for the target researchers. [4] Some common evaluation metrics are used to evaluate the performances of all the comparisons [3].

1) PRECISION

It is used to evaluate the accuracy of the recommender systems recommending relevant papers to the researchers. The equation is as follows:

$$Precision = \frac{Relevant\ papers}{Total\ recommended\ papers} \tag{10}$$

The larger the value of Precision is, the more accurate the recommendation result will be.

2) RECALL

It indicates the fraction of relevant papers in the whole set of papers appearing in the recommendation list. The equation is as follows:

$$Recall = \frac{Relevant\ papers}{Total\ relevant\ papers} \tag{11}$$

The larger value means that the recommendation system has more ability in ranking the most relevant papers at the top of the recommendation list.

3) F-MEASURE

It simultaneously considers the precision and recall, and presents a weighted harmonic average of them. The larger F value means that the paper recommendation system is more effective.

$$F = \frac{(\alpha^2 + 1)(Precision \times Recall)}{\alpha^2(Precision + Recall)} \tag{12}$$

D. COMPARATIVE METHODS

- **Content-based paper recommendation (CBR)** constructs the user and paper profiles the same as we did in Section IV-B. The difference is that the user and paper feature vectors are extracted based on the TF-IDF method.
- **Graph-based paper recommendation (GBR)** firstly extracts a homogeneous graph which is composed of the paper-paper citation relationships. Then the random walk algorithm [39] is performed to search the relevant citations for papers which are published by a query researcher.
- **MPRec** automatically extracts the interesting meta-paths of a given length from the heterogeneous graphs and trains a logistic regression model to measure the probability that a query researcher gets interested in the relevant papers [11]. For the sake of fairness, we also use the meta-paths as shown in Table 1.
- **HGRec** In this method, we firstly construct the user and paper profiles. Then the user and paper embeddings are initialized based on these profiles. Both the contents information and heterogeneous topological features are combined for the training of the representation of heterogeneous graph.
- **HGRec1** In order to verify the contributions of the contents information, a variation of HGRec named HGRec1 is designed. In this method, we randomly initialize the embeddings for users and papers, and the contents features are not considered during the embedding procedure.

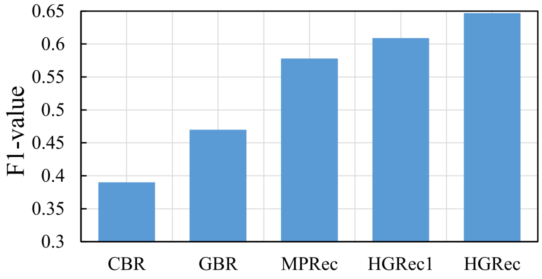

FIGURE 4. F1 values of the comparative methods.

E. EXPERIMENT RESULTS

In the experiment, α is set to 1. Thus the F-measure becomes the F1-measure. The performances of all the comparative methods on paper recommendation are as shown in Figure 4. As can be concluded from this figure that our proposed HGRec method achieves the best recommendation results in terms of F1-measure among all the competitors.

As we can see that CBR generates the worst performing results. This is because CBR only relies on the contents information to make recommendation. However, the contents information in paper titles are quite limited. Thus the feature vectors of papers extracted from the paper titles are generally very sparse. In the meantime, TF-IDF can not capture the context of words when generating the feature vectors, which also decreases the accuracy of recommendation.

Compared to CBR, GBR achives a slightly better result in terms of F1 than CBR. GBR is a pure graph-based paper recommendation method. We firstly extract the paper-paper citation-ship graph, and then perform random walk on this graph to search the relevant papers for a target user. Compared to MPRec, the performance of GBR is comparatively worse.

[4] If a paper is cited by a research, this paper is treated as his/her relevant paper.

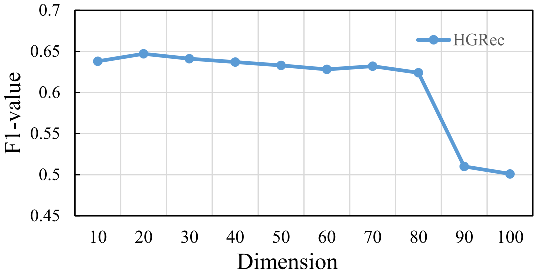

FIGURE 5. The influence of the embedding dimension on the proposed model.

In our consideration, MPRec solves the paper recommendation problem in the setting of heterogeneous information networks and the rich semantics underlying the heterogeneous graph have been considered.

Similar to MPRec, HGRec is also a heterogeneous graph based recommendation method. The major difference is the generation strategy of feature vectors. MPRec obtains the feature vectors by computing the proximities of nodes with respect to each meta-path [11], which is a traditional feature vectors extraction strategy in machine learning. While HGRec generates the feature vectors by employing the heterogeneous graph representation technique, which is more effective and efficient in capturing the complex interactive heterogeneous features in the heterogeneous graphs [15].

In order to verify the contributions of the contents information used in the initialization of node embeddings, the performances of HGRec and HGRec1 are also compared. The difference between HGRec and HGRec1 lies in the initialization stage. HGRec relies on the user and paper profiles to initialize the feature vectors, while HGRec1 just randomly initializes the node embeddings. From Figure 4 we can find that, the performance of HGRec in terms of F1-measure is better than HGRec1, which demonstrates the effectiveness of the contents information in initialization.

F. EMBEDDING DIMENSION ANALYSIS

In order to validate the influence of embedding dimensions to the recommendation performance of our HGRec method, a comparative experiment has been performed. Specifically, we set the number of embedding dimension from 10 to 100 with an interval of 10. Figure 5 shows the recommendation results. It can be concluded from the figure that as the embedding dimension increases, the recommendation performance in terms of F1-value first increases and then decreases. In our consideration that, when the embedding dimension is too small, the embedding representation capability is not sufficient. However, when the embedding dimension is too large, the proposed embedding model may overfit the data, leading to the unsatisfactory recommendation performances.

VI. CONCLUSION

In this paper, we develop a new heterogeneous graph representation based paper recommendation method, which not only takes into account the heterogeneous entities and relationships of the academic graphs, but also incorporates the contents information of papers into the representation of user and paper feature vectors. First, we extract the user and paper profiles based on the contents of papers (i.e., title, keywords, abstract). Then we employ a pre-trained word-embedding technique to initialize the user and paper feature vectors considering the contents information included in the user and paper profiles. Thirdly, two meta-path based proximity measures are proposed, which are used to evaluate the neighboring similarity and structural similarity between nodes in the heterogeneous graphs. Then the node embeddings are jointly updated with respect to these two proximities. Finally, the paper recommendation is generated by computing the similarities of the user and paper feature vectors. Substantial experiments have been conducted on the DBLP dataset, the results of which clearly demonstrate the effectiveness of our HGRec method in terms of F1-measure compared with several baselines.

Although it is empirically validated that our proposal has shown great potentials for paper recommendation, there still remains some aspects to be improved. For example, the meta-paths used in this paper are manually designed. In the future, some meta-pattern discovery methods can be included to automatically generate the meta-paths. In addition, some advanced techniques, e.g, Neural Factorization Machines [40], can be used to model the higher-order and non-linear feature interactions. What's more, we will also validate our proposed method on some other available datasets.

ACKNOWLEDGMENT

Thanks to our anonymous reviewers for the helpful comments and suggestions for making this paper better.

REFERENCES

[1] K. Zhou, J. Zeng, Y. Liu, and F. Zou, "Deep sentiment hashing for text retrieval in social CIoT," *Future Gener. Comput. Syst.*, vol. 86, pp. 362–371, Sep. 2018.

[2] R. Deveaud, J. Mothe, M. Z. Ullah, and J.-Y. Nie, "Learning to adaptively rank document retrieval system configurations," *ACM Trans. Inf. Syst.*, vol. 37, no. 1, p. 3, 2019.

[3] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, "Scientific paper recommendation: A survey," *IEEE Access*, vol. 7, pp. 9324–9339, 2019.

[4] G. Tian and L. Jing, "Recommending scientific articles using bi-relational graph-based iterative RWR," in *Proc. 7th ACM Conf. Recommender Syst.*, Oct. 2013, pp. 399–402.

[5] M. Amami, R. Faiz, F. Stella, and G. Pasi, "A graph based approach to scientific paper recommendation," in *Proc. Int. Conf. Web Intell.*, Aug. 2017, pp. 777–782.

[6] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 448–456.

[7] Z. Yang, D. Yin, and B. D. Davison, "Recommendation in academia: A joint multi-relational model," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2014, pp. 566–571.

[8] 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.

[9] X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2012, pp. 1119–1130.

[10] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han, "ClusCite: Effective citation recommendation by information network-based clustering," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 821–830.

[11] X. Ma, Y. Zhang, and J. Zeng, "Newly published scientific papers recommendation in heterogeneous information networks," *Mobile Netw. Appl.*, vol. 24, no. 1, pp. 69–79, 2019.

[12] F. Zhu, Q. Qu, D. Lo, X. Yan, J. Han, and P. S. Yu, "Mining top-*k* large structural patterns in a massive network," *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 807–818, 2011.

[13] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 701–710.

[14] Y. Dong, N. V. Chawla, and A. Swami, "Metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 135–144.

[15] C. Shi, B. Hu, W. X. Zhao, and P. S. Yu, "Heterogeneous information network embedding for recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 357–370, Feb. 2018.

[16] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, "HARP: Hierarchical representation learning for networks," Jun. 2017, *arXiv:1706.07845*. [Online]. Available: https://arxiv.org/abs/1706.07845

[17] P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, "A framework for tag-based research paper recommender system: An IR approach," in *Proc. IEEE 24th Int. Conf. Adv. Inf. Netw. Appl. Workshops*, Perth, WA, Australia, Apr. 2010, pp. 103–108.

[18] C. Caragea, F. A. Bulgarov, A. Godea, and S. D. Gollapalli, "Citation-enhanced keyphrase extraction from research papers: A supervised approach," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1435–1446.

[19] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[20] Y. Zhang, M. Chen, D. Huang, D. Wu, and Y. Li, "iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization," *Future Gener. Comput. Syst.*, vol. 66, pp. 30–35, Jan. 2017.

[21] Y. Zhang, X. Ma, S. Wan, H. Abbas, and M. Guizani, "CrossRec: Cross-domain recommendations based on social big data and cognitive computing," *Mobile Netw. Appl.*, vol. 23, no. 6, pp. 1610–1623, Dec. 2018.

[22] X. Ma, H. Lu, Z. Gan, and Q. Zhao, "An exploration of improving prediction accuracy by constructing a multi-type clustering based recommendation framework," *Neurocomputing*, vol. 191, pp. 388–397, May 2016.

[23] X. Ma, H. Lu, Z. Gan, and J. Zeng, "An explicit trust and distrust clustering based collaborative filtering recommendation approach," *Electron. Commerce Res. Appl.*, vol. 25, pp. 29–39, Sep./Oct. 2017.

[24] 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.

[25] V. Martínez, F. Berzal, and J.-C. Cubero, "A survey of link prediction in complex networks," *ACM Comput. Surv.*, vol. 49, no. 4, p. 69, Feb. 2017.

[26] A. Anand, T. Chakraborty, and A. Das, "Fairscholar: Balancing relevance and diversity for scientific paper recommendation," in *Proc. Eur. Conf. Inf. Retrieval*, Springer, 2017, pp. 753–757.

[27] Y. Qian, Y. Zhang, Y. Ma, H. Yu, and L. Peng, "EARS: Emotion-aware recommender system based on hybrid information fusion," *Inf. Fusion*, vol. 46, pp. 141–146, Mar. 2019.

[28] T. Mikolov, G. S. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent. (ICLR)*, Scottsdale, AZ, USA, May 2013, pp. 1–12.

[29] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 855–864.

[30] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 1067–1077.

[31] J. Tang, M. Qu, and Q. Mei, "PTE: Predictive text embedding through large-scale heterogeneous text networks," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1165–1174.

[32] X. Cai, J. Han, and L. Yang, "Generative adversarial network based heterogeneous bibliographic network representation for personalized citation recommendation," in *Proc. 32th AAAI Conf. Artif. Intell.*, 2018, pp. 5747–5754.

[33] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-*k* similarity search in heterogeneous information networks," *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[34] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2014, pp. 1188–1196.

[35] C. Zang, P. Cui, C. Faloutsos, and W. Zhu, "On power law growth of social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1727–1740, Sep. 2018.

[36] B. Zhang and M. Al Hasan, "Name disambiguation in anonymized graphs using network embedding," in *Proc. 26th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 1239–1248.

[37] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Springer, 2010, pp. 177–186.

[38] 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*, Aug. 2008, pp. 990–998.

[39] H. Tong, C. Faloutsos, and J.-Y. Pan, "Fast random walk with restart and its applications," in *Proc. IEEE 6th Int. Conf. Data Mining*, Dec. 2006, pp. 613–622.

[40] X. He and T. S. Chua, "Neural factorization machines for sparse predictive analytics," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 355–364.

XIAO MA received the Ph.D. degree from the Huazhong University of Science and Technology, in 2017. She is an Assistant Professor with the School of Information and Safety Engineering, Zhongnan University of Economics and Law (ZUEL), China. From 2015 to 2017, she was visiting the University of Illinois at Urbana-Champaign. She has published more than ten prestigious conference and journal papers. Her research interests include recommendation systems, data mining, machine learning, etc.

RANRAN WANG is currently pursuing the M.S. degree with the School of Information and Safety Engineering, Zhongnan University of Economics and Law (ZUEL). Her research interests include recommendation systems and data mining.