IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2019, accepted February 13, 2019, date of publication February 18, 2019, date of current version March 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2899907

Citation Recommendation as Edge Prediction in Heterogeneous Bibliographic Network: A Network Representation Approach

LIBIN YANG[1], (Member, IEEE), ZEQING ZHANG[2], XIAOYAN CAI[1], AND LANTIAN GUO[1]
[1]School of Automation, Northwestern Polytechnical University, Xi'an 710072, China
[2]School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Xiaoyan Cai (xiaoyanc@nwpu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61872296 and Grant 61772429, in part by the Ministry of Education, China, through the Project of Humanities and Social Sciences under Grant 18YJC870001, in part by the China Postdoctoral Science Foundation under Grant 2017M613205 and Grant 2017M623241, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JQ6031, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102018zy026.

ABSTRACT With an increasing number of scholarly publications, accessing and retrieving appropriate papers is becoming an essential task for researchers. Citation recommendation, which can automatically provide a reference list based on a text segment, can overcome this problem. In this paper, we first construct a heterogeneous bibliographic network and deem citation recommendation as edge prediction problem, and then we develop a network representation-based edge prediction (NREP) model, which can simultaneously learn the edge prediction knowledge and the predictive representation for efficient citation recommendation. For personalized recommendation, we incorporate author information. We conduct extensive experiments on two datasets; the experimental results show that the NREP-based approach outperforms the other four state-of-the-art baseline approaches in terms of recall, mean average precision, and normalized discounted cumulative gain.

INDEX TERMS Heterogeneous bibliographic network, citation recommendation, edge prediction, network representation.

I. INTRODUCTION

In recent years, more and more research papers have been published, researchers find it hard to track proper papers on time. One method to solve this problem is using keyword to retrieve papers, but it still requires researchers manually go over the retrieved papers and decide which paper should be selected. Another method is tracing reference papers of other scholarly publications that exert a tremendous fascination on researchers, but it limits reference papers within an specific research filed. Citation recommendation, which can efficiently find relevant research papers based on a given manuscript or a text segment, is a necessary technique to deal with this problem. Existing citation recommendation approaches are classified as global citation recommendation [1]–[4] and local citation recommendation [5]–[7]. The former returns relevant scientific papers for a given manuscript, and the latter aims to recommend reference papers for a given text segment. We study global citation recommendation in this paper.

Existing global citation recommendation approaches are categorized as: content-based filtering (CBF) [8], collaborative filtering (CF) [9] and graph-based approaches [1]. CBF recommends scientific papers similar to the ones that the researchers are interested in the past. CF relies on finding researchers with similar research interest and using their ratings to provide recommendations. Graph-based approaches deem citation recommendation as an edge prediction task and apply random walk to solve the task.

Network representation which represents each node as a low-dimensional vector, has attracted large interests in the fields of image, video and artificial intelligence. It has also been applied for citation recommendation problem [3]. Most network representation methods learn node representa-

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

tion by extracting structure feature in the network, such as DeepWalk [10] and Node2Vec [11] combine random walk and skip-gram, LINE [12] preserves both local and global network structures. Early work of network representation focus on homogeneous network. Recently, attention has been shifted to heterogeneous network. Huang and Mamoulis [13] studied heterogeneous network representation by preserving meta-path based proximities. Wang *et al.* [14] designed a semi-supervised deep model to optimize the first-order and second-order proximity. Chang *et al.* [15] applied deep learning techniques to map nodes of a heterogeneous network into unified vector representations, considering both contents and topological structures in networks. However, these approaches are learned in an unsupervised manner, so the learned node representations only has descriptive ability. When applying such techniques to edge prediction task, the users may disappointed with the results, due to these techniques have no ability on deducing hidden edges. Wang *et al.* [16] presented a predictive network representation approach, but they ignored node content information which reduces the performance of network representation.

In this work, we also deem the citation recommendation as an edge prediction problem, and develop a network representation based edge prediction (NREP) model, which can simultaneously learn the edge prediction knowledge and the predictive representation for efficient citation recommendation. We first construct a heterogeneous bibliographic network, and then propose a network representation based edge prediction (NREP) model, which integrates node content information, observed structure information and hidden edge information. We define a node-content correlation function for node content information, an observed structure function for observed structure information and a hidden edge prediction function for hidden edge information, separately. Finally, we optimize the three functions jointly to learn the constructed heterogeneous bibliographic network embedding, and the learned node representations can be used for citation recommendation. To sum up, this paper makes the following contributions:

1) A heterogeneous bibliographic network is constructed, which models different relationships.
2) A heterogeneous bibliographic network representation based edge prediction (NREP) model is proposed, which incorporates node content information, observed structure information and hidden edge information, to learn node representation which can be used for edge prediction.
3) An NREP model is developed to solve the citation recommendation task, and we conduct experiments to verify the performance of the approach.

The rest of this paper is organized as follows. We review related work in Section II. Then we construct a heterogeneous bibliographic network and propose a network representation based edge prediction (NREP) model. After that, we illustrate the citation recommendation approach based on NREP model in Section IV. The experiments and the evaluation results are presented in Section V. Section VI illustrates the conclusions.

II. RELATED WORK

A. NETWORK REPRESENTATION BASED CITATION RECOMMENDATION

Pan *et al.* [17] developed a heterogeneous graph-based similarity learning algorithm for academic paper recommendation. Gupta and Varma [18] addressed the problem of scientific paper recommendation through a novel method, which combines the network representation from the bibliographic network with the content information. Kobayashi *et al.* [19] proposed a context-based approach, it integrates distributed representations of text and citation graphs, and learns distributed vector representations of papers, with each vector capturing a discourse facets within an article. Cai *et al.* [2] proposed a heterogeneous information network embedding (HINE) approach, it can capture inter-relationships among heterogeneous vertices, intra-relationships among homogeneous vertices and correlations between vertices and text contents simultaneously, modeling various kinds of objects as vectors in a continuous and common vector space. They also proposed a deep network representation model [3] that integrates network structure and the vertex content information into a unified framework by exploiting generative adversarial network, and represents various kinds of vertices of the heterogeneous network in a unified vector space. Based on the proposed model, they obtained heterogeneous bibliographic network representation for efficient citation recommendation. Yang *et al.* [20] presented an LSTM based approach. They first separately learned citation contexts' and scientific papers' vector representations, then they measured relevance between them based on the learned vector representation. Finally, scientific papers that have high scores are generated as recommended reference paper list. In this work, we propose an NREP model, which learns node representations with predictive ability and can enhance the performance of citation recommendation.

B. HETEROGENEOUS NETWORK REPRESENTATION

As heterogeneous networks consist of different types of nodes and edges, how to represent these nodes in a common vector space is a challenging problem. Tang *et al.* [21] proposed a PTE algorithm, which learns a distributed representation of text through embedding the heterogeneous text network into a low dimensional space. Dong *et al.* [22] developed a meta-path-guided random walk strategy in a heterogeneous network, which contains multiple types of nodes in the network. Gui *et al.* [23] used hyperedge to learn heterogeneous network embedding. Tao *et al.* [24] proposed a HIN2Vec model, which learns both node vectors and relationship vectors by maximizing the likelihood of predicting relationships among nodes jointly. Xu *et al.* [25] proposed an Embedding of Embedding model to encode the intra-network and

inter-network edges for the coupled heterogeneous network, which consists of two different but related homogeneous networks. Hu *et al.* [26] proposed a deep neural network to simultaneously obtain distributed representations of users, items and meta-path. Chen and Sun [27] proposed a task-guided and path-augmented heterogeneous network embedding framework. Jacob *et al.* [28] proposed a heterogeneous social network embedding algorithm for classifying nodes. Chang *et al.* [15] designed a deep embedding algorithm for networked data. The algorithm reflects both the local and global network structures, and makes the resulting embedding useful for a variety of data mining tasks. Huang and Mamoulis [29] studied the problem of heterogeneous network embedding for meta path based proximity, which can fully utilize the heterogeneity of the network. In our work, we construct a heterogeneous bibliographic network containing papers, venues and authors firstly, and then we propose an NREP model, finally the learned node representations based on the model can be used for citation recommendation.

III. HETEROGENEOUS BIBLIOGRAPHIC NETWORK REPRESENTATION BASED EDGE PREDICTION (NREP) MODEL

A. HETEROGENEOUS BIBLIOGRAPHIC NETWORK CONSTRUCTION

We construct a heterogeneous bibliographic network $G = \langle V, E \rangle$, where V is the set of vertices that consists of the paper set $P = \{p_1, p_2, \cdots, p_{n_p}\}$, the author set $A = \{a_1, a_2, \cdots, a_{n_a}\}$ and the venue set $V = \{v_1, v_2, \cdots, v_{n_v}\}$, i.e., $V = P \cup A \cup VN$, n_p is the number of papers, n_a is the number of authors and n_v is the number of venues. E is the edge set of G, we denote E as $E = E^+ \cup E^- \cup E^?$, representing different edge status in the network. For any node pair (v_i, v_j), if there exists an edge between them, then $(v_i, v_j) \in E^+$, if there exists no edge between them, then $(v_i, v_j) \in E^-$, if it is not sure whether there exist an edge between them, then $(v_i, v_j) \in E^?$. Based on the constructed heterogeneous bibliographic network, we propose a heterogeneous bibliographic network representation based edge prediction (NREP) model, which maps each node $v_i \in V$ into a common low-dimensional vector space, aiming to predict whether there exist an edge between node pair (v_i, v_j) in the set $E^?$.

B. HETEROGENEOUS BIBLIOGRAPHIC NETWORK REPRESENTATION BASED EDGE PREDICTION (NREP) MODEL

The framework of the proposed heterogeneous bibliographic network representation based edge prediction (NREP) model is shown in Fig.1. The input of the model are node content set, observed structure set (E_o^+) and hidden edge set (E_h^+). We define node-content correlation objective function for node content set, define structure preservation objective function for observed structure set and define edge prediction objective function for hidden edge set, separately. Based on the vertex embedding layer of the model, we jointly optimize the above three objective functions and obtain the learned node representations which are more predictive for edge prediction.

1) VERTEX-CONTENT CORRELATION BASED NETWORK REPRESENTATION

We first collect text content of each node in the constructed heterogeneous bibliographic network, i.e. each paper node's text content, each author node's related text content, which is the text content of papers that are written by one author, and the text content related to each venue node, which is the text content of papers that are published in one venue. Then we

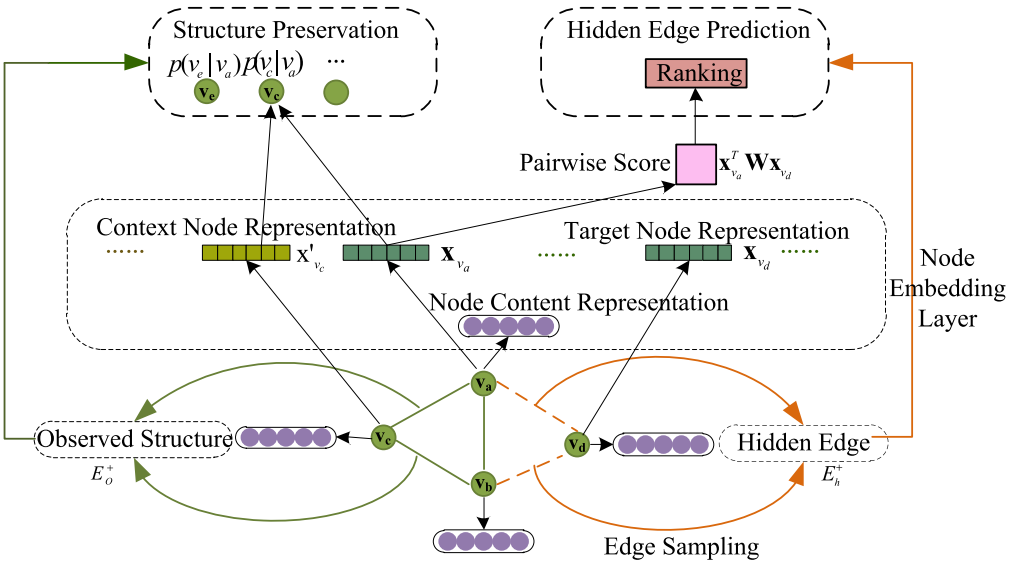


FIGURE 1. NREP model.

formulate the objective function of vertex-content correlation as:

$$L_c = \sum_{i=1}^{n_a+n_p+n_v} \sum_{d \in D} \log P(N_W(v_i)|\mathbf{x}_{v_i}) \quad (1)$$

in which D is the word corpus based on the generated biased random walk,, $N_W(v_i)$ are the current vertex's neighbor words, $\mathbf{x}_{v_i}$ is the vector representation of the vertex v_i.

2) OBSERVED STRUCTURE SET BASED NETWORK REPRESENTATION

Preserving structural proximities among nodes in a low-dimensional vector space is the learning objective of observed structure. The learning objective and Skip-gram model [30] have something in common, we deem neighbors of each node as the context of this node. The aim of the observed structure set based network representation is to maximize neighbors of each node in E_o^+. So each node v_i has two vector representations, one is node representation x_i, when it is acted as a target node, the other is context representation $\mathrm{x}_i^{'}$ when it is acted as the "context" of other nodes. Based on it, we define the learning objective function of E_o^+ as:

$$\begin{aligned} L_o &= -\sum_{(v_i,v_j)\in E_o^+} \log p(v_j|v_i) \\ &= -\sum_{(v_i,v_j)\in E_o^+} \log \frac{\exp(-\mathrm{x}_{v_j}^{'T}\mathrm{x}_{v_i})}{\sum_{h=1}^{|V|}\exp(-\mathrm{x}_{v_j}^{'T}\mathrm{x}_{v_i})} \end{aligned} \quad (2)$$

in which $p(v_j|v_i)$ illustrates the probability of the neighbor of the node v_i is the node v_j in $(v_i, v_j) \in E_o^+$.

3) EDGE PREDICTION BASED NETWORK REPRESENTATION

Edge prediction refers to link prediction, it has also been deemed as ranking problem. If there exists a hidden edge between the node v_n and the node v_m, and there exists no hidden edge between the node v_m and the node v_l, then the ranking scores between $v_n$and v_m is higher than the ranking scores between v_m and v_l, i.e.,

$$r_{v_m,v_n} > r_{v_m,v_l}, \text{ if } (v_m, v_n) \in E_h^+ \quad \text{and } (v_m, v_l) \in E^- \quad (3)$$

Based on the node representation, we apply the bilinear product to obtain the ranking scores of node pairs as $r_{v_m,v_n} = \mathrm{x}_{v_m}^T \mathbf{W} \mathrm{x}_{v_n}$, in which $\mathbf{W}$ is a weighting matrix to be learned. Based on the above constraint, the edge prediction objective function is defined as:

$$L_h = \sum_{(v_m,v_n)\in E_h^+,(v_m,v_l)\in E^-} \max(1 - \mathbf{x}_{v_m}^T \mathbf{W}(\mathbf{x}_{v_n} - \mathbf{x}_{v_l}), 0)^2 + \frac{\lambda_W}{2}||\mathbf{W}||_2^2 \quad (4)$$

where λ_W is a regularization parameter.

4) JOINT LEARNING MODEL

We integrate the above three learning objective functions into a unified framework, and optimize the following objective function to get the node vector representation matrix X:

$$\min L = \min(\delta L_c + \theta L_o + \gamma L_h) \quad (5)$$

where δ, θ and γ are weight parameters for balancing the importance of the three learning objective functions, $\delta + \theta + \gamma = 1$.

IV. NREP BASED CITATION RECOMMENDATION

We cast the manuscript q as and manuscript text q_t and manuscript author q_a, i.e., $q = [q_a, q_t]$. q_t is acted as a testing paper, q_a is acted as a user and scientific papers are acted as training papers. Based on q, we develop a network representation based edge prediction (NREP) model to solve the citation recommendation task, which measures the similarity scores $\mathbf{r}_q = [\mathbf{r}_{qp_1}, \mathbf{r}_{qp_2}, \cdots, \mathbf{r}_{qp_l}]$ between the q and $p_i (i = 1, 2, \cdots, n_p)$. The inputs are training papers' and testing papers' word sequence, testing papers' authors and venues that have published the training papers, edge set among the training papers and edge set between the training papers and testing papers. The distributed representation of authors, papers and venues are obtained based on the NREP model. We calculate the similarity scores as $\mathbf{r}_q = \mathbf{V}_{PR}\mathbf{v}_{q_t}^T + \mathbf{V}_{AR}\mathbf{v}_{q_a}^T$, where $\mathbf{V}_{PR} = [\mathbf{v}_{p_1}; \mathbf{v}_{p_2}; \cdots; \mathbf{v}_{p_l}]$ is the vector representation of training papers, $\mathbf{v}_{q_t}$ is the vector representation of the manuscript text, $\mathbf{V}_{AR} = [\mathbf{v}_{a_1}; \mathbf{v}_{a_2}; \cdots; \mathbf{v}_{a_m}]$ is the training paper authors' vector representation, $\mathbf{v}_{q_a}$ is the manuscript author's vector representation. We rank the training papers based on the similarity scores, and generate the final reference paper recommendation list by selecting the high ranked papers. Although the distributed representation of venues and context embedding of papers, venues and authors have not

TABLE 1. NREP based citation recommendation algorithm.

Input: The heterogeneous bibliographic network $G =< V, E_t >$ consisting of the manuscript text q_t, manuscript author q_a, training papers, venues, $E_t = E^+ \bigcup E^-$, parameters α, γ, θ, δ and the recommended papers's number M.

Output: Vertex embedding matrix X, context embedding matrix X', citation recommendation list.

1. Initialize X, X', $E_r^+ \leftarrow E^+$;
2. Set parameter $\alpha = 0.3$;
3. While $E_r^+ \neq \phi$ do
   $E_h^+ \leftarrow$ randomly select $\alpha |E^+|$ edges from E_r^+;
   $E_h^- \leftarrow$ randomly select $\alpha |E^-|$ edges from E^-;
   $E_r^+ \leftarrow E_r^+ - E_h^+$, $E_o^+ \leftarrow E^+ - E_h^+$
4. **For** each vertex v_i **do**
   Optimization on L_c using Eq. (1);
   Optimization on L_o using Eq. (2);
   Optimization on L_h using Eq. (4);
   End for
5. Compute the score values $\mathbf{r}_q$ and rank the training papers based on $\mathbf{r}_q$;
6. Generate the citation recommendation list by selecting top ranked M training papers.

been used in the process of citation recommendation, they can enhance the vector representation of author and paper in NREP model's training process. Table 1 summarizes the whole process of the approach.

V. EXPERIMENTS

A. DATASETS

Extensive experiments are conducted on AAN[1] and DBLP[2] datasets to evaluate the performance of the proposed approach. As DBLP dataset is a large scale dataset, we select a subset of it, that is information retrieval (EACL, ECIR, ACL, EMNLP, SIGIR, CIKM, NAACL, COLING), computer security (NDSS, SP, ACSAC, FC, ARES, ISI), machine learning (PAKDD, WSDM, SIGKDD, NIPS, ICML, ICDM, ICDE), computer vision (ICIP, CVPR, ICCV, ACCV, MM, ECCV, ICPR) and networks and communications (ICC, ICNP, MOBICOM, INFOCOM, ICDCS, GLOBECOM, SIGCOMM, SECON).

We remove those papers with missing abstracts or titles in datasets, and the remaining 12,555 papers in AAN dataset and 64,332 papers of DBLP dataset are used for experiments. Each of the two datasets are divided into two parts: with regard to AAN dataset, 11, 197 papers are treated as training papers, the remaining 1,358 papers are treated as testing papers, with regard to DBLP dataset, 56,304 papers published before 2013 (included) are acted as training papers, 8,028 papers which are published in the year of 2014 and 2015 are acted as testing papers. For each paper, we only take its abstract and title as paper content.

B. EVALUATION METRICS

Recommending relevant papers to the given manuscript is the ultimate goal of the global citation recommendation. We use the following three metrics:

Recall@N: It measures the average of the proportion of papers from the test set that appear among the top N ranked list, for some given N. We set $N = \{30, 70, 100\}$ in this paper.

Mean Average Precision (MAP): It is the average precision scores for each manuscript in a set of manuscripts. We represent T_p as the testing paper set. For each paper p_i in T_p, we represent the paper p_i's ground-truth reference set as R, represent the reference paper listed generated by our proposed approach as C. The definition of MAP is illustrated as:

$$MAP = \frac{1}{|T_p|}\sum_{p_i \in T_p}\frac{1}{|R|}\sum_{r_j \in R, rank(r_j)\neq 0}\frac{q(r_j)+1}{q(r_j)}$$
$$rank(r_j) = \begin{cases} \text{the position of } r_j \text{ in C}, & \text{if } r_j \text{ is in } C \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where $r_j \in R$ indicates a correct reference paper, $q(r_j)$ is the number of the ground-truth papers that rank higher than r_j. We take top 40 papers as the recommended citations in this paper.

Normalized Discounted Cumulative Gain (NDCG): It measures accuracy of rankings when there are multiple levels of relevance judgment. Given a manuscript, NDCG at position n is defined as:

$$NDCG@n = Z_n \sum_{j=1}^{n} \frac{2^{R(j)}-1}{\log(1+j)} \quad (7)$$

in which n is the position, $R(j)$ is the score of rank j, Z_n is a normalization factor that guarantees a perfect ranking's NDCG at position n equals 1. For a manuscript that its retrieved documents is less than n, NDCG is only calculated for the retrieved scientific papers.

C. EXPERIMENTS AND DISCUSSION

In our experiments, we set the parameter α as 0.3, and set the balance parameter $\delta = \theta = \gamma = \frac{1}{3}$, as we think the three component are equally important in the NREP model.

1) PERFORMANCE OF NREP MODEL BASED CITATION RECOMMENDATION

In this experiment, we ignore the manuscript's author information, and use the manuscript's content information only. We wish to test whether scientific papers' author and venue information enhance the citation recommendation's performance. We categorize the experiments into four categories: (1) NREP-C, which uses scientific papers' content information, (2) NREP-CA, which uses scientific papers' content and author information, (3) NREP-CV, which uses scientific papers' content and venue information and (4) NREP-CAV, which uses scientific papers' content, venue and author information. The experimental results of the above four approaches on AAN and DBLP datasets are presented in Table 1.

As content and edge information of papers can only provide coarse information for network representation, leading to less accuracy for calculating similarities of the scientific papers and the given manuscript, the performance of NREP-C performs poorest. The results in Table 1 shows the performance of NREP-CA is better than that of NREP-CV. We attribute it to that the venue information can provide coarse information than author information for evaluating relevance between the scientific papers and the given manuscript, and thus the performance of recommended reference papers by NREP-CA is better than that of NREP-CV. Furthermore, when fully utilizing author, content and venue information of scientific papers, we can obtain more relevance scores between scientific papers and the given manuscript, and thus can obtain high performance of citation recommendation list for the given manuscript.

We are also interested in studying whether incorporating personal information can enhance the performance of citation recommendation. We represent the personalized manuscript as $q_1 = [q_t, q_a]$, represent the non-personalized manuscript as $q_2 = [q_t]$. When a user, who does not published papers before, wish to obtain recommended papers based on a manuscript, the personalized recommendation task automatically sinks to non-personalized task, because the manuscript

[1] http://tangra.cs.yale.edu/newaan/
[2] https://dblp.uni-trier.de/

TABLE 2. Experimental results of NREP model based citation recommendation approach on the AAN and DBLP datasets.

Dataset	Approach	MAP	Recall@30	Recall@70	Recall@100	NDCG@30	NDCG@70	NDCG@100
AAN	NREP-C	0.243	0.536	0.682	0.738	0.613	0.708	0.730
	NREP -CV	0.260	0.571	0.684	0.747	0.622	0.731	0.739
	NREP -CA	0.266	0.592	0.705	0.763	0.648	0.749	0.752
	NREP -ACV	0.278	0.615	0.716	0.775	0.673	0.780	0.771
DBLP	NREP-C	0.235	0.521	0.630	0.693	0.578	0.671	0.699
	NREP -CV	0.251	0.536	0.642	0.718	0.582	0.698	0.713
	NREP -CA	0.263	0.548	0.669	0.725	0.596	0.712	0.724
	NREP -ACV	0.275	0.557	0.681	0.736	0.613	0.728	0.733

TABLE 3. Comparison of performance on personalized and Non-Personalized citation recommendation on the AAN and DBLP datasets.

Dataset	Approach	MAP	Recall@30	Recall@70	Recall@100	NDCG@30	NDCG@70	NDCG@100
AAN	NREP- ACV, q_1	0.286	0.624	0.728	0.786	0.682	0.791	0.783
	NREP- ACV, q_2	0.278	0.615	0.716	0.775	0.673	0.780	0.771
DBLP	NREP- ACV, q_1	0.284	0.565	0.690	0.744	0.625	0.739	0.746
	NREP- ACV, q_2	0.275	0.557	0.681	0.736	0.613	0.728	0.733

only contains content information q_t. Table 2 reports the experimental results.

Table 2 illustrate the non-personalized recommendation approach's performance is inferior to the personalized recommendation approach's performance. When we make deep analysis to the top-30 recommended reference papers generated by NREP-ACV with q_1 and NREP-ACV with q_2, we found the overlap of the recommended papers retrieved by NREP-ACV with q_1 and NREP-ACV with q_2 is about 77.26% on AAN dataset and 76.03% on DBLP dataset, respectively. We take a further step to compare the top-3 recommended reference papers returned by the two approaches, as for the AAN dataset, the accuracy of NREP-ACV with q_1 is about 80.59% and the accuracy of NREP-ACV with q_2 is about 79.63%. Meanwhile, as for the DBLP dataset, the accuracy of NREP-ACV with q_1 is about 80.12% and the accuracy of NREP-ACV with q_2 is about 79.45%, respectively.

2) COMPARISON WITH THE OTHER STATE-OF-THE-ART APPROACHES

We aim to get more meaningful vector representation of each node by developing network representation approach, and apply them to the citation recommendation task. So we compare the proposed approach with the four state-of-the-art network representation based citation recommendation approaches: (1) DeepWalk [10], which use information of network structure to learn representation of scientific paper network; (2) Line [12], which learns representation of paper network by preserving global and local network structure; (3) Node2Vec [11], which learns a mapping of scientific paper vertices to a low-dimensional space of features that maximizes the likelihood of preserving scientific paper network neighborhoods of scientific paper vertices; and (4) PNR [16], which learns scientific paper representations that have predictive ability. After obtaining distributed representations of scientific papers by using different baseline approaches, we can conduct citation recommendation based on it.

In general, we only focus on the content information of manuscript in the following experiments, i.e., $q_2 = [q_t]$. Table 3 below lists the experimental results of our proposed approach and the four baseline approaches. We set the parameters of Node2Vec the same as in [11] and omit for ease of representation.

The results show that the performance of DeepWalk based citation recommendation is poorest, due to it only utilize global network structure. As the global and local network structure are preserved in LINE, the performance of LINE based citation recommendation is better than that of DeepWalk based citation recommendation. Node2Vec based citation recommendation approach outperforms DeepWalk and LINE based approach, as Node2Vec has its own search strategy, it investigates neighborhoods in network flexibly.PNR based approach performs better than the above three baseline approaches, as it considers learning prediction knowledge. We are happy to see that our proposed NREP based approach performs best, owing to it simultaneously learn the edge prediction knowledge and the predictive representation for efficient citation recommendation.

3) CASE STUDY

Furthermore, we select a detailed manuscript to illustrate the performance of proposed approach and other baseline

TABLE 4. Comparison of different citation recommendation approaches on the AAN and DBLP datasets.

Dataset	Approach	MAP	Recall@30	Recall@70	Recall@100	NDCG@30	NDCG@70	NDCG@100
AAN	NREP- ACV, q_2	0.278	0.615	0.716	0.775	0.673	0.780	0.771
	PNR	0.253	0.581	0.693	0.745	0.642	0.755	0.743
	Node2Vec	0.237	0.568	0.668	0.729	0.618	0.734	0.722
	LINE	0.225	0.542	0.659	0.713	0.609	0.723	0.715
	DeepWalk	0.214	0.533	0.647	0.700	0.599	0.711	0.707
DBLP	NREP- ACV, q_2	0.275	0.557	0.681	0.736	0.613	0.728	0.733
	PNR	0.249	0.538	0.667	0.715	0.592	0.695	0.701
	Node2Vec	0.234	0.532	0.641	0.690	0.577	0.672	0.683
	LINE	0.223	0.526	0.628	0.679	0.560	0.663	0.672
	DeepWalk	0.212	0.513	0.617	0.667	0.549	0.651	0.663

TABLE 5. Comparison of different approaches on DBLP dataset.

Title of the Manuscript	Approach	Top-3 System Generated Reference Papers
Sequential Summarization: A Full View of Twitter Trending Topics	NREP- ACV, q_2	(1) Earthquake shakes Twitter users: Real-time event detection by social sensors(✓) (2) TweetMotif: Exploratory search and topic summarization for Twitter(✓) (3) Event summarization using tweets
	PNR	(1) Automatic summarization of Twitter topics(✓) (2) Twitter topic summarization by ranking tweets using social influence and content quality (3) Selecting quality Twitter content for events
	DeepWalk	(1) Earthquake shakes Twitter users: Real-time event detection by social sensors(✓) (2) Detecting controversial events from Twitter (3) Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre.

approaches. The title of the manuscript is, Sequential Summarization: A Full View of Twitter Trending Topics. Due to the page limit, we list the top 3 retrieved reference papers of the NREP, DeepWalk and PNR approaches in Table 4, (✓) denotes the matching results. We find that the first two recommended papers retrieved by NREP is the same with the ground-truth reference papers. While the recommended papers retrieved by DeepWalk and PNR only have one paper, which is the same with the ground-truth reference papers. We deem DeepWalk uses global network structure only, PNR learns prediction knowledge, NREP learns edge prediction knowledge and the predictive representation simultaneously.

VI. CONCLUSION

In this paper, we focus on how to learn predictive heterogeneous bibliographic network representation, enhancing the performance of citation recommendation. To this end, we first construct a heterogeneous bibliographic network and deem citation recommendation as edge prediction problem, then we propose a heterogeneous bibliographic network representation based edge prediction (NREP) model, which can simultaneously learn the edge prediction knowledge and the predictive representation for efficient citation recommendation. Finally, we conduct extensive experiments on AAN dataset and subset of DBLP dataset to compare the performance of NREP based citation recommendation approach with the other four baseline network representation based citation recommendation approaches. The experimental results show that NREP based approach outperforms the other approaches in terms of Recall, MAP and NDCG. In our future work, we will incorporate other information to model the identity of a researcher, i.e. H-index, to study whether it can enhance the performance of citation recommendation task.

REFERENCES

[1] 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.

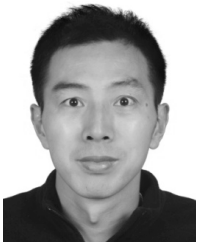
[2] X. Cai, J. Han, S. Pan, and L. Yang, "Heterogeneous information network embedding based personalized query-focused astronomy reference paper recommendation," *Int. J. Comput. Intell. Syst.*, vol. 11, pp. 591–599, Jan. 2018.

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

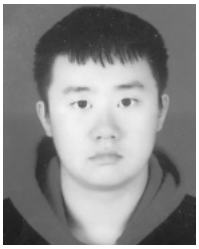
[4] L. Yang, Y. Zheng, X. Cai, S. Pan, and T. Dai, "Query-oriented citation recommendation based on network correlation," *J. Intell. Fuzzy Syst.*, vol. 35, no. 2, pp. 1–8, 2018. doi: 10.3233/JIFS-172039.

[5] X. Tang, X. Wan, and X. Zhang, "Cross-language context-aware citation recommendation in scientific articles," in *Proc. 27th ACM SIGIR Conf.*, 2014, pp. 817–826.

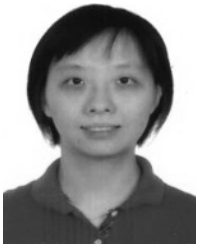
[6] W. Huang, Z. Wu, C. Liang, P. Mitra, and C. Lee Giles, "A neural probabilistic model for context based citation recommendation," in *Proc. 29th AAAI Conf.*, 2015, pp. 2404–2410.
[7] T. Ebesu and Y. Fang, "Neural citation networks for context-aware citation recommendation," in *Proc. 40th SIGIR Conf.*, 2017, pp. 1093–1096.
[8] C. Nascimento, A. Laender, A. H. da Silva, and M. A. Gonçalves, "A source independent framework for research paper recommendation," in *Proc. 11th JCDL Conf.*, 2011, pp. 297–306.
[9] 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.
[10] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 701–710.
[11] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proc. 22nd SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 855–864.
[12] J. Tang, W. Qu, M. Z. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 1067–1077.
[13] Z. Huang and N. Mamoulis. (2017). "Heterogeneous information network embedding for meta path based proximity." [Online]. Available: https://arxiv.org/abs/1701.05291
[14] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proc. 22nd SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1225–1234.
[15] S. Chang, W. Han, J. Tang, G.-J. Qi, C. C. Aggarwal, and T. S. Huang, "Heterogeneous network embedding via deep architectures," in *Proc. 21st SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 119–128.
[16] Z. Wang, C. Chen, and W. Li, "Predictive network representation learning for link prediction," in *Proc. 40th SIGIR Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 969–972.
[17] L. Pan, X. Dai, S. Huang, and J. Chen, "Academic paper recommendation based on heterogeneous graph," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Cham, Germany: Springer, 2015, pp. 381–392.
[18] S. Gupta and V. Varma, "Scientific article recommendation by using distributed representations of text and graph," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 1267–1268.
[19] Y. Kobayashi, M. Shimbo, and Y. Matsumoto, "Citation recommendation using distributed representation of discourse facets in scientific articles," in *Proc. 18th JCDL*, 2018, pp. 243–251.
[20] 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.
[21] J. Tang, M. Qu, and Q. Mei, "PTE: predictive text embedding through large-scale heterogeneous text networks," in *Proc. 21st SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1165–1174.
[22] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. 23rd SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 135–144.
[23] H. Gui *et al.*, "Embedding learning with events in heterogeneous information networks," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2428–2441, Nov. 2017.
[24] T. Fu, W. C. Lee, and Z. Lei, "Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proc. 26th CIKM*, 2017, pp. 1797–1806.
[25] L. Xu, X. Wei, J. Cao, and P. S. Yu, "Embedding of embedding (EOE): Joint embedding for coupled heterogeneous networks," in *Proc. 10th WSDM*, 2017, pp. 741–749.
[26] B. Hu, C. Shi, W. X. Zhao, and P. Yu, "Leveraging meta-path based context for top-N recommendation with a neural co-attention model," in *Proc. 24th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1531–1540.
[27] T. Chen and Y. Sun, "Task-guided and path-augmented heterogeneous network embedding for author identification," in *Proc. 10th WSDM*, 2017, pp. 295–304.
[28] Y. Jacob, L. Denoyer, and P. Gallinari, "Learning latent representations of nodes for classifying in heterogeneous social networks," in *Proc. 7th WSDM*, 2014, pp. 373–382.
[29] Z. Huang and N. Mamoulis. (2017). "Heterogeneous information network embedding for meta path based proximity." [Online]. Available: https://arxiv.org/abs/1701.05291
[30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS*, vol. 16, 2013, pp. 3111–3119.



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



ZEQING ZHANG is currently pursuing the bachelor's degree with the School of Telecommunications Engineering, Xidian University. His current research interest includes information retrieval.



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

LANTIAN GUO is currently pursuing the Ph.D. degree with the School of Automation, Northwestern Polytechnical University, China. He was a Visiting Student with the School of Computing, Queen's University, Canada, from 2014 to 2015. His current research interests include big data, recommendation systems, graph model, and artificial intelligence.