The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Generative Adversarial Network Based Heterogeneous Bibliographic Network Representation for Personalized Citation Recommendation

Xiaoyan Cai, Junwei Han,* Libin Yang
School of Automation, Northwestern Polytechnical University, China
{xiaoyanc, jhan, libiny}@nwpu.edu.cn

Abstract

Network representation has been recently exploited for many applications, such as citation recommendation, multi-label classification and link prediction. It learns low-dimensional vector representation for each vertex in networks. Existing network representation methods only focus on incomplete aspects of vertex information (i.e., vertex content, network structure or partial integration), moreover they are commonly designed for homogeneous information networks where all the vertices of a network are of the same type. In this paper, we propose a deep network representation model that integrates network structure and the vertex content information into a unified framework by exploiting generative adversarial network, and represents different types of vertices in the heterogeneous network in a continuous and common vector space. Based on the proposed model, we can obtain heterogeneous bibliographic network representation for efficient citation recommendation. The proposed model also makes personalized citation recommendation possible, which is a new issue that a few papers addressed in the past. When evaluated on the AAN and DBLP datasets, the performance of the proposed heterogeneous bibliographic network based citation recommendation approach is comparable with that of the other network representation based citation recommendation approaches. The results also demonstrate that the personalized citation recommendation approach is more effective than the non-personalized citation recommendation approach.

Introduction

With the rapid growth in the number of scientific papers, researchers might find it hard to find appropriate and necessary work to cite. Researchers usually retrieve papers from web search engines based on certain keywords, manually review them and decide which paper should be cited. However, it is labor-intensive as well as time-consuming, and especially difficult for the beginning researchers. Citation recommendation which can recommend a list of reference papers that are relevant to the researchers' information need, is an essential technology to overcome this problem. A variety of citation recommendation approaches have been proposed in the literature (He et al. 2010; He et al. 2011; Ren et al. 2014; Huang et al. 2014). These approaches are either global or local. Global recommendation (He et al. 2011; Meng et al. 2013; Ren et al. 2014;) recommends a list of references for a given manuscript. Local recommendation (He et al. 2010), on the other hand, aims to recommend citations for specific context of each place where a citation should be made. We focus on global citation recommendation in this work.

Existing global citation recommendation approaches fall into three categories: collaborative filtering (CF) (Hernando et al. 2016), content-based filtering (CBF) (Nascimento et al. 2011) and graph-based approaches (Meng et al. 2013). CF makes citation recommendation by finding correlations among other researchers with similar research interests. CBF recommends a reference paper based on words and/or topic features of a manuscript and the identity of a researcher. Graph-based approaches often consider citation recommendation as a link prediction problem and solve the problem using properties of random walks.

Recently, network representation has been exploited for graph-based citation recommendation (Gupta and Varma 2017). Most network representation methods are based on network structure. Inspired by deep learning techniques in natural language processing (Mikolov et al. 2013), Perozzi et al. (2014) proposed DeepWalk approach, which learns feature vectors for vertices from a corpus of random walks generated from networks by employing neural network models. Grover and Leskovec (2016) proposed Node2Vec approach, which maximizes the likelihood of preserving network neighborhoods of vertices to learn feature representation. However, they only consider network structure, ignoring content information associated with each vertex. Pan et al. (2016) proposed TriDNR approach, which uses information from three parties, i.e. vertex structure, vertex

*Corresponding author

content and vertex labels, to jointly learn vertex representation. But this approach ignores inter-relationship among heterogeneous vertices.

In this paper, we propose a generative adversarial network based model to learn heterogeneous bibliographic network representation by modeling both vertex content and network structure. The distributed representation obtained using the model in turn can be used to calculate similarity scores. Finally the top ranked scientific papers are generated as the citation recommendation list. The contributions of this paper are lists as follows:

1. A bibliographic network is constructed to model different relationships among heterogeneous objects (i.e., scientific papers, authors, manuscript and author of the manuscript);
2. A generative adversarial network based heterogeneous bibliographic network representation (GAN-HBNR) model is developed, which incorporates bibliographic network structure and content of different kinds of objects to learn optimal representations of these objects;
3. A novel personalized citation recommendation approach based on the GAN-HBNR model is proposed, and the thorough experimental studies are conducted to verify the effectiveness of the proposed approach.

Related Work

Graph-based Citation Recommendation

Recent studies employed graph-based approaches to investigate the citation recommendation problem (Strohman et al. 2007; Zhou et al. 2008; Meng et al. 2013; Pan et al. 2016; Gupta and Varma 2017). Strohman et al. (2007) deemed citation recommendation as link prediction problem. They represented each paper as a vertex, the citation relationship as the link between vertices and a new paper as a vertex without any in-link and out-link. Zhou et al. (2008) measured paper similarities by combining the author-paper graph, the paper-venue graph and the paper citation graph. Then they recommended reference papers by treating some known citations as positive labels and applying semi-supervised learning on the combined graphs. Gori and Pucci (2006) proposed to recommend research papers by a random-walk based approach. Meng et al. (2013) presented a personalized citation recommendation approach, which incorporated different kinds of information, such as content of papers, authorship and citation etc., into a unified graph model. Pan et al. (2016) proposed an academic paper recommendation approach based on a heterogeneous graph containing various kinds of features. Gupta and Varma (2017) proposed a scientific paper recommendation combining distributed representations of paper's content and distributed representations of the graph constructed from the bibliographic network. We propose a heterogeneous bibliographic network representation based citation recommendation approach, which recommends the top ranked scientific papers to the author of the manuscript based on the similarity scores among the representation of different objects (manuscript, scientific papers, the author of manuscript and authors of the scientific papers).

Network Representation

Network representation was first proposed by Hoff et al. (2002), and it was later followed by many approaches include multi-dimensional scaling (MDS) (Kruskal and Wish 1978), Laplacian Eigenmap (Belkin and Niyogi 2001), local linear embedding (LLE) (Roweis and Saul 2000) and IsoMap (Tenenbaum et al. 2000), which treats eigenvectors as representations. However, these approaches are not applicable to large scale networks due to computational complexity. Perozzi et al. (2014) proposed DeepWalk by using local information from truncated random walks as input and learning latent representation of vertices in a network. Tang et al. (2015a) proposed a network embedding method called LINE to preserve the local and global network structures. Grover and Leskovec (2016) proposed Node2Vec approach to learn vertex representation by maximizing the likelihood of preserving network neighborhoods of vertices. But DeepWalk, LINE, Node2Vec only focus on homogeneous networks. PTE (Tang et al. 2015b) extends the LINE to handle with heterogeneous network embedding problem, but PTE only leverages network structure, ignoring vertex content information. Pan et al. (2016) proposed TriDNR, a tri-party deep network representation model, to simultaneously learn network structure and vertex content. But this model cannot capture highly non-linear structures. Moreover, the structure information they use is not comprehensive (e.g., ignoring relationships among heterogeneous objects).

Generative Adversarial Network based Heterogeneous Bibliographic Network Representation (GAN-HBNR) Model

Heterogeneous Bibliographic Network Construction

In this section, we first construct a heterogeneous bibliographic network containing papers and authors as G=<V, E, C>, where $V = V_P \bigcup V_A$ is the vertex set, $V_P = \{p_i\}$ ($1 \le i \le n$, n is the total number of papers), $V_A = \{a_j\}$ ($1 \le j \le m$, m is the total number of authors). $E = < E_{PP}, E_{AA}, E_{PA} >$ is the edge set, $E_{PP} = \{e_{ij}, p_i, p_j \in V_P\}$, $E_{AA} = \{e_{ij}, a_i, a_j \in V_A\}$ and

$E_{PA} = \{e_{ij}, p_i \in V_P, a_j \in V_A\}$ correspond to the edges between papers, the edges between authors and the edges between papers and authors, respectively. If p_i cites p_j, or p_i is cited by p_j, or a_i collaborates with a_j, or a_i is one of the authors of p_j, or a_j is one of the authors of p_i, then $e_{ij} = 1$; otherwise $e_{ij} = 0$. $C = C_P \bigcup C_A$ is the set of content information, let c_{p_i} denote the text content associated with the paper p_i and $C_P = \{c_{p_i}\}$ ($1 \le i \le n$), c_{a_j} denote the text content associated with the papers which are written by author a_j and $C_A = \{c_{a_j}\}$ ($1 \le j \le m$). Let B denote the adjacency matrix for the bibliographic network, and let $b_k = \{e_{1,k}, \cdots, e_{n+m,k}\}$ be an adjacency vector.

Model Description

The architecture of the proposed GAN-HBNR model is shown in Figure.1. The whole architecture consists of two main modules: the content2vec module and the generative adversarial bibliographic network module. Based on the model, we can learn an effective feature representation vector that preserves both vertex content information and network structure information and thus can be applied to personalized citation recommendation task.

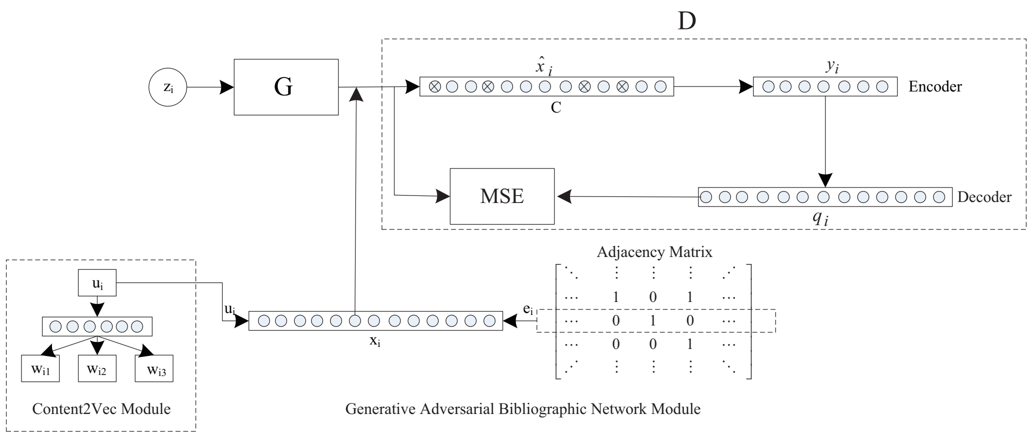

Figure.1 Architecture of the GAN-HBNR model. G is the generator, Encoder and Decoder are DAE encoder and decoder network, C is a corruption process (bypassed at test time) and D is the discriminator.

- **Content2vec Module**

We collect all the text content associated with one paper vertex, we also collect all the text content associated with papers which are written by one author. Thus each vertex in the bibliographic network contains text information. We employ doc2vec approach (Le and Mikolov 2014) as our content2vec module. The content representation of each vertex can be obtained from this module. Therefore, we can maximize the objective as follows:

$$O = \sum_{i=1}^{n+m} \log P(w_{-b} : w_b \mid v_i) \qquad (1)$$

where w is a word in the text information of v_i, b is the window size of the word sequence. After optimizing this objective, we can obtain the content representation u_i of each vertex v_i.

- **Generative Adversarial Bibliographic Network Module**

This module is the core part of our model, it integrates content information and structure information of the heterogeneous bibliographic network. As $\mathbf{e}_i$ describes link relationships among vertex v_i and other vertices in the network, the adjacency matrix E reflects the structure of the network. We extract each adjacency vector $\mathbf{e}_i$ and concatenate it with the corresponding content vector $\mathbf{u}_i$ as the input $\mathbf{x}_i$ of v_i in the GAN-HBNR model. Therefore, the content information and structure information can be learned simultaneously.

We then employ energy-based generative adversarial network (Zhao et al. 2016) to train the generative bibliographic network module. One difference to Zhao et al. (2016) is that we use a Denoising Autoencoder (DAE) as our energy function, because the DAE has been found to produce superior representations to the standard Autoencoder (Vincent et al. 2010). We define a feed-forward generator network $G(\mathbf{z})$ that takes a vector $\mathbf{z} \in \Re^{h_g}$ as input and produces a generated vector, with h_g being the number of dimensions in the input noise vector (sampled from $N(0, I)$). We also define a discriminator network $D(\mathbf{x})$, seen as an energy function, that takes vectors $\mathbf{x} \in \Re^{n+m}$ and produces an energy estimate $E \in \Re$.

During the encoding phrase of DAE, we use single layer to map the input data to a highly nonlinear latent space, so the encoding process is

$$\mathbf{y} = f(\mathbf{W}^e \mathbf{x}^c + \mathbf{b}_e) \quad (2)$$

where $\mathbf{W}^e$ is a set of learned parameters, $\mathbf{b}_e$ is a learned bias term, $\mathbf{x}^c$ is a corrupted version of $\mathbf{x}$, f is a nonlinear function, and $\mathbf{y} \in \Re^{h_d}$ is the hidden representation of $\mathbf{x}$ with size h_d.

The decoding phase is a reflection of the encoder, its output $\mathbf{q}_i$ should be close to the input $\mathbf{x}_i$. The decoding process is

$$\mathbf{q} = g(\mathbf{W}^d \mathbf{y} + \mathbf{b}_d) \quad (3)$$

where $\mathbf{W}^d$ and $\mathbf{b}_d$ are another learned set of weights and bias terms, g is a nonlinear function. The final energy value is the mean squared reconstruction error:

$$E = \frac{1}{n+m}\sum_{i=1}^{n+m}(\mathbf{x}_i - \mathbf{y}_i)^2 \quad (4)$$

The energy function is trained to push down on the energy of real vectors $\mathbf{x}$, and to push up on the energy of generated vectors $G(\mathbf{z})$ (Zhao et al. 2016). Given a positive margin m, a real vector $\mathbf{x}$ and a generated vector $G(\mathbf{z})$, the discriminator loss L_D and the generator loss L_G are formally defined by:

$$L_D(\mathbf{x},\mathbf{z}) = D(\mathbf{x}) + [m - D(G(\mathbf{z}))]^+ \quad (5)$$

$$L_G(\mathbf{z}) = D(G(\mathbf{z})) \quad (6)$$

where $[\cdot]^+ = \max(0,\cdot)$. Minimizing L_G with respect to the parameters of G is similar to maximizing the second term of L_D. It has the same minimum but non-zero gradients when $D(G(\mathbf{z})) \geq m$. We use the vector $\mathbf{y}$ as the new network representation of the heterogeneous bibliographic network.

One of the advantages of our GAN-HBNR model is that when new vertices enter the network, we do not need to retrain the generated adversarial bibliographic network module. When we obtain as new adjacency vector $\mathbf{e}_j$, we can feed it into the model and obtain the representation at a complexity of $O(1)$. If there exist no link between the new vertex and the network, we can exploit its content information.

GAN-HBNR Model based Personalized Citation Recommendation Approach

Given a manuscript q, we propose a heterogeneous bibliographic network representation based personalized citation recommendation approach, which aims to return top ranked scientific papers as reference papers by measuring the similarity scores between the manuscript and all the scientific papers in the dataset. In our work, we formulate the manuscript q as manuscript author q_a and manuscript text q_t, i.e., $q = [q_a, q_t]$. We deem q_t as a testing paper, q_a as an author of the manuscript and all the scientific papers in the dataset P as training papers. Algorithm 1 below summarizes the whole process of the heterogeneous bibliographic network representation based personalized citation recommendation approach.

Algorithm 1 GAN-HBNR Model based Personalized Citation Recommendation Algorithm

Input: The heterogeneous bibliographic network G=<V,E,C> consists of the manuscript text q_t, the manuscript author q_a (if available), the training papers and all the authors of the training papers, Adjacency matrix $\boldsymbol{B}$, window size w, the dimension h_d, Number Q.

Output: Citation Recommendation list

1: Train a paragraph vector model based on C, obtain the content representations $\mathbf{U}$
2: $\mathbf{X}$=Merge[$\mathbf{B}$,$\mathbf{U}$]
3: Generate negative samples using $G(\boldsymbol{z})$
4: **Repeat:**
5: **for** d-steps **do**
6: Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m+n}\sum_{i=1}^{m+n}\left[\log D(\hat{x}^{(i)}) + \log(1 - D(G(z^{(i)})))\right]$$

7: **end for**
8: **for** g-steps **do**
9: Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m+n}\sum_{i=1}^{m+n}\log(1 - D(G(\mathbf{z}^{(i)})))$$

10: **end for**
11: **Until converge**
12: Obtain the network representations $\mathbf{Y} = \{y_i\}$
13: Calculate the similarity score $\mathbf{r}_q$ for the manuscript and rank the training papers according to $\mathbf{r}_q$;
14: Select top ranked Q training papers as citation recommendation list.

The similarity scores is represented as $\mathbf{r}_q = [\mathbf{r}_{qp_1}, \mathbf{r}_{qp_2}, \cdots, \mathbf{r}_{qp_l}]$, $p_i \in P(i = 1,2,\cdots,n)$. The input to the recommendation system is the word sequence of training papers and testing papers, all the authors of the training papers and author of the testing papers, as well as the adjacency matrix based on the network. All these papers and authors are mapped into vectors based on our proposed GAN-HBNR model. Thus the similarity scores can be calculated 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 vector representation of authors related to training papers, $\mathbf{v}_{q_a}$ is the vector representation of the manuscript author. Training papers are ranked according to the similarity scores, the top ranked ones are selected as the final citation recommendation list.

Experiments

Datasets

In order to evaluate the quality of the proposed model, we conduct experiments on two bibliographic datasets: (1)**AAN** (ACL Anthology Network) **dataset**[1], which is established by Radev and Muthukrishnan (2009), it consists of conference papers and journal papers in the field of computational linguistics. We remove the papers which have missed titles or abstracts in the dataset, then we use the remaining 12,555 papers published from 1965 to 2013 as experimental dataset. For evaluation purpose, we divide the entire dataset into two disjoint sets, the papers published before 2013 are deemed as training set (11,197 papers) and the remaining papers fall into the testing set (1,358 papers). (2) **DBLP dataset**[2], which consists of bibliography data in computer science (Tang et al. 2008). Instead of using full dataset, we choose a subset since some samples miss complete references. We also remove the papers which have missed titles or abstracts in the dataset, then we select the papers published before 2013 as training set (33,016 papers) and papers published from 2013 to 2015 as testing set (3172 papers).

In our work, we extract the title and abstract of the papers in the two datasets as document content, and we define a manuscript as the title, abstract and an author of the manuscript.

Evaluation Methods

Our ultimate aim is to recommend more relevant reference papers to the given manuscript. We use three common metrics as follows:

Recall@*N*: It is defined as the percentage of the original reference papers that appear in the top *N* recommended papers. Here we use *N*={20,40,60,80,100} to evaluate the proposed approach.

Mean Average Precision (**MAP**): As Recall@*N* only considers the top *N* ranking results, ignoring the exact ranking position. MAP is a precision metric that emphasizes ranking relevant papers higher, which can overcome the above disadvantage. Let T_p be the set of the testing papers. For a paper p_i in T_p, the correct reference paper set of p_i is R_C, and our proposed approach returns a reference paper list R_G. We consider the top 40 recommended papers in the ranking list, so $|R_G|=40|$. The MAP is defined as:

$$MAP=\frac{1}{|T_p|}\sum_{p_i\in T_p}\frac{1}{|R_C|}\sum_{r_j\in R_C, rank(r_j)\neq 0}\frac{q(r_j)+1}{q(r_j)} \quad (7)$$

where $r_j\in R_C$ is a correct reference paper, $rank(r_j)$ is defined as the position of r_j in R_G if r_j is in R_G, otherwise $rank(r_j)$ is defined to be zero. $q(r_j)$ is set to the number of the correct reference papers which ranks higher than r_j.

Mean Reciprocal Rank (MRR): It measures how far from the top appears the first relevant reference papers. MRR is defined as:

$$MRR=\frac{1}{|T_p|}\sum_{p_i\in T_p}\left(\frac{1}{rank(p_{first})}\right) \quad (8)$$

where $rank(p_{first})$ is the position of the first relevant reference papers in the reference list R_G.

Performance on GAN-HBNR Model based Personalized Citation Recommendation

In this set of experiments, we first focus on the query text only, ignoring the query author information, i.e., $q_1=[q_t]$. q_1 also denotes the non-personalized manuscript. We set the representation size h_d, which is the size of the DAE hidden state, to 100. The generator input noise vector h_g is set to be the same size. The generator is a three-layer feedforward neural network, the first two layers use ReLU activation function and the output layer uses a sigmoid nonlinear function. The first two layers are both of size 400, with the final output layer being the same size as the vocabulary. Meanwhile, the first two layers use batch normalization (Ioffe and Szegedy 2015). The discriminator encoder consists of a single layer followed by a leaky ReLU nonlinear function (with a leak of 0.03). The decoder is a linear transformation back to the vocabulary size. We optimize both *G* and *D* using Adam (Kingma and Ba 2015) with an initial learning rate of 0.0001. Our DAE corruption process is to randomly set 40% of the input values as zero, and we use a margin size *m* of 5% of the vocabulary size. We follow the same validation procedure as Glover (2016), with a learning rate of 0.01 and using the tanh activation function.

We are interested in studying whether personalized citation recommendation can provide more appropriate and individualized recommendation results to the users than non-personalized citation recommendation. We denote a personalized manuscript by $q_2=[q_t,q_a]$. When a user who inputs the manuscript for the personalized citation recommendation has not yet published any papers, the proposed approach will be reduced to non-personalized citation rec-

[1] http://clair.eecs.umich.edu/aan/index.php
[2] http://arnetminer.org/citation

ommendation for the user, because the query information contains q_t only. Table 1 reports the experimental results.

Table 1. Comparison of Performance on GAN-HBNR Model based Personalized and Non-Personalized Citation Recommendation on the AAN and DBLP datasets

Dataset	Approach	MAP	MRR	Recall@20	Recall@40	Recall@60	Recall@80	Recall@100
AAN	GAN-HBNR, q_2	0.293	0.312	0.586	0.678	0.737	0.749	0.787
	GAN-HBNR, q_1	0.280	0.299	0.563	0.665	0.712	0.726	0.769
DBLP	GAN-HBNR, q_2	0.289	0.309	0.557	0.649	0.695	0.711	0.758
	GAN-HBNR, q_1	0.273	0.296	0.541	0.632	0.673	0.695	0.734

From Table 1, we can see that the performance of non-personalized recommendation is inferior to that of personalized recommendation. The personalized recommendation achieves a gain of about 3.36% on average in the two datasets. When we compare the correct recommended papers with regard to non-personalized and personalized recommendation approaches, we observe that the personalized recommendation approach can find more papers published by co-authors. We study the distinction of the top-60 recommendation results returned by GAN-HBNR with q_1 and GAN-HBNR with q_2 on the two datasets. The overlap of the two approaches on the AAN dataset and DBLP dataset is about 77.96% and 76.25% of each, respectively. For the top-3 recommended results, the accuracy of GAN-HBNR with q_2 is about 81.25% more than that of GAN-HBNR with q_1 on the AAN dataset, meanwhile the accuracy of GAN-HBNR with q_2 is about 80.13% more than that of GAN-HBNR with q_1 on the DBLP dataset.

Comparison with Other Network Representation based Citation Recommendation Approaches

Our original intention to propose the network representation approach is to hope to obtain more meaningful vector representation of each vertex in the network, and then perform citation recommendation based on the vector representations of these vertices. So we compare our proposed bibliographic network representation approach with other five network representation approaches: (1) DeepWalk (Perozzi et al. 2014), which learns paper network representation by utilizing network structure information; (2) Line (Tang et al. 2015a), which preserves local and global network structure to learn paper network representation; (3) Doc2Vec (Le and Mikolov 2014), which maps variable length of text into a fixed length distributed vector using neural network models; (4) TriDNR (Pan et al. 2016), which simultaneously considers paper network structure and paper vertex content to learn paper network representation and (5) Node2Vec (Grover and Leskovec 2016), which learns a mapping of paper vertices to a low-dimensional space of features that maximizes the likelihood of preserving paper network neighborhoods of paper vertices. After obtaining network representation with the above different approaches, citation recommendation can then be performed.

Without loss of generality, in this set of experiments, we only focus on the manuscript text, ignoring the manuscript author information, i.e., $q_1 = [q_t]$. Table 2 below compares the performance of the other five network representation based recommendation approaches and our proposed approach on the AAN and DBLP datasets. The parameter settings for each Node2Vec entry are omitted for ease of presentation.

From Table 2, we can see that DeepWalk performs poorest, as it considers global network structure only. LINE preserves local and global network structure, so it performs better than DeepWalk. Both DeepWalk and LINE can be seen as rigid search strategies over networks. The flexible and controllable search strategy in exploring network neighborhoods of Node2Vec makes the approach can obtain better results than DeepWalk and LINE. Doc2Vec performs better than the above three approaches, we attribute it to the ability of vertex content is more important in learning network representation. Although TriDNR considers both network structure and vertex content, it ignores inter-relationship among heterogeneous vertices. So it performs worse than our proposed GAN-HBNR approach. It is glad to see that the proposed GAN-HBNR approach which simultaneously considers vertex content and network structure consistently outperforms the other five network representation approaches.

Case Study

Besides the above numerical analysis, we take an example to further illustrate the proposed recommendation approach and the limitations of existing network representation based recommendation approaches. The title of the manuscript is, Sequential Summarization: A Full View of Twitter Trending Topics, which is in DBLP dataset. This manuscript studies to provide a serial of chronological ordered short sub-summaries for a trending topic. Due to the page limit, we only list the top 5 retrieved papers obtained by the GAN-HBNR with q_2, GAN-HBNR with q_1 and TriDNR approaches. Table 4 below lists the system generated reference papers by the GAN-HBNR with q_2, GAN-HBNR with q_1 and TriDNR approaches, (✓) indicates the matched results.

As shown in Table 4, the results returned by the GAN-HBNR with q_2 approach have four records that match the ground truth citation list of the manuscript, whereas the results returned by the GAN-HBNR with q_1 and TriDNR approaches have three and two matching records, respectively. This observation demonstrates that the GAN-HBNR with q_2 approach obtained a better result in this case study since the manuscript author, manuscript text, as well as text information and author information of training papers are utilized in this approach. There are two same citations in the top-5 results returned by the GAN-HBNR with q_2 approach and GAN-HBNR with q_1 approach, but the number of the corrected papers in the top-5 results returned by the GAN-HBNR with q_2 approach is more than that returned by the GAN-HBNR with q_1 approach. We attribute it to the GAN-HBNR with q_2 approach incorporates manuscript author information, while the GAN-HBNR with q_1 does not do it. The top 5 recommended results returned by TriDNR only contain two corrected papers, this is due to TriDNR only consider paper information, ignoring author information and author-paper information in bibliographic network representation.

Table 2. Comparison of the Network Representation based Citation Recommendation Approaches on the AAN and DBLP datasets

Dataset	Approach	MAP	MRR	Recall@20	Recall@40	Recall@60	Recall@80	Recall@100
AAN	GAN-HBNR, q_1	0.280	0.299	0.563	0.665	0.712	0.726	0.769
	TriDNR	0.259	0.273	0.551	0.648	0.689	0.702	0.753
	Doc2Vec	0.246	0.265	0.540	0.631	0.675	0.689	0.741
	Node2Vec	0.237	0.251	0.533	0.623	0.662	0.673	0.729
	LINE	0.225	0.240	0.521	0.617	0.651	0.662	0.713
	DeepWalk	0.214	0.228	0.509	0.603	0.640	0.651	0.700
DBLP	GAN-HBNR, q_1	0.273	0.296	0.541	0.632	0.673	0.695	0.734
	TriDNR	0.253	0.271	0.518	0.621	0.650	0.671	0.715
	Doc2Vec	0.245	0.263	0.509	0.613	0.641	0.659	0.703
	Node2Vec	0.234	0.250	0.497	0.601	0.632	0.647	0.690
	LINE	0.223	0.239	0.483	0.587	0.621	0.635	0.679
	DeepWalk	0.214	0.226	0.471	0.573	0.610	0.622	0.667

Table 3. Illustration of Top-5 Reference papers generated by the Three Network Representation based Citation Recommendation Approaches on DBLP dataset

Title of the Manuscript	Approaches	Top-5 System Generated Reference Papers
Sequential Summarization: A Full View of Twitter Trending Topics	GAN-HBNR with q_2	1)TweetMotif: Exploratory search and topic summarization for Twitter (✓) 2) Experiments in microblog summarization (✓) 3) Event summarization using tweets (✓) 4) Summarizing sporting events using Twitter (✓) 5) Twitter topic summarization by ranking tweets using social influence and content quality
	GAN-HBNR with q_1	1) Event summarization using tweets (✓) 2) Towards real-time summarization of scheduled events from twitter streams 3)TweetMotif: Exploratory search and topic summarization for Twitter (✓) 4) Summarizing sporting events using Twitter (✓) 5) Multi-Tweet Summarization for Flu Outbreak Detection
	TriDNR	1)TweetMotif: Exploratory search and topic summarization for Twitter (✓) 2) Event summarization using tweets (✓) 3) Towards real-time summarization of scheduled events from twitter streams 4) Twitter topic summarization by ranking tweets using social influence and content quality 5) Joint topic modeling for event summarization across news and social media streams

Conclusion

In this paper, we propose a generative adversarial network based heterogeneous bibliographic network representation model, which integrates network structure and the vertex content information into a unified framework and represents different types of vertices in the heterogeneous network in a continuous and common vector space. A personalized citation recommendation approach is developed based on the obtained network representation. We evaluate our proposed approach on the AAN and DBLP datasets, the results reveals that the combination of network-based structure and content-based analysis can improve bibliographic network representation for personalized citation recommendation. In the future, we plan to explore venue information in the process of bibliographic network representation.

Acknowledgements

The work described in this paper was supported by Nationa l Natural Science Foundation of China (61772429), China Postdoctoral Science Foundation (2017M613205) and the Fundamental Research Funds for the Central Universities (3102016QD009, 3102016QD010).

References

Belkin M. and Niyogi P. 2001. Laplacian eigenmaps andspectral techniques for embedding and clustering, In *Proceedings of Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 585-591.

Belkin M., and Niyogi P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15(6): 1373-1396.

Glover, J. 2016. Modeling documents with generative adversarial networks. *Workshop on Adversarial Training, NIPS 2016*, Barcelona, Spain.

Grover A., and Leskovec J. 2016. Node2vec: scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 855-864.

Gupta S., and Varma V. 2017. Scientific article recommendation by using distributed representations of text and graph. In *Proceedings of the 26th International World Wide Web Conference*, Perth, Australia, 1267-1268.

He, Q.; Pei, J.; Kifer, D.; Mitra, P.; and Giles, L. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, North Carolina, USA, 421-430.

He, Q.; Kifer, D.; Pei, J.; Mitra, P.; and Giles, C.L. 2011. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, Hong Kong, China, 755-764.

Hernando,A.; Bobadilla J.; and Ortega F. 2016. A non-negative matrix factorization for collaborative filtering recommender systems based on a Bayesian Probabilistic model. *Knowledge based Systems*,97, 188-202.

Hoff P. D.; Raftery A.E. and Handcock M.S. 2002. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460), 1090-1098.

Huang, W.Y.; Wu, Z.H.; Mitra, P.; and Giles, C.L. 2014. RefSeer: a citation recommendation ststem. In *Proceedings of the 14th ACM/IEEE-CS Joint-Conference on Digital Libraries*. London, United Kindom, 371-374.

Ioffe, S. and Szegedy, C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint arxiv: 1502.03167*.

Kingma, D.P. and Ba, J.L. 2015. Adam: a method for stochastic optimization. *International Conference on Learning Representations*.

Kruskal J.B. and Wish M. 1978. Multidimensional scaling. *Sage publications*.

Le, Q.V. and Mikolov, T. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning*, Beijing, China, 1188-1196.

Meng, F.Q.; Gao, D.H.; Li W.J.; Sun X.; and Hou Y.X. 2013. A unified graph model for personalized query-oriented reference paper recommendation. In *Proceedings of the 22th ACM International Conference on Information & Knowledge Management*, 1509-1012. San Francisco, California, USA.

Mikolov, T.; Kai C.; Greg C.; and Jeffrey D. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301.3781*.

Nascimento,C.; Laender, A.; Silva A.; and Gonçalves M. 2011. A source independent framework for research paper recommendation. In *Proceedings of the 2011 ACM/IEEE Joint Conference on Digital Libraries*, Ottawa, Canada, 297-306.

Pan, L.; Dai, X.; Huang, S. and Chen, J. 2015. Academic paper recommendation based on heterogeneous graph. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, 381-392.

Pan, S.R.; Wu, J.; Zhu, X.Q.; Zhang, C.Q.; and Wang, Y. 2016. Tri-party deep network representation. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, New York City, New York, USA,1895-1901.

Perozzi, B.; AI-Rfou, R.; and Skiena, S. 2014. Deepwalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York City, New York, USA, 701-710.

Radev, D.R. and Muthukrishnan, V.. 2009. The ACL Anthology Network Corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, 54-61.

Ren, X.; Liu, J.L.; Khandelwal, U.; Gu, Q.Q.; Wang, L.D.; and Han, J.W. 2014. ClusCite: effective citation recommendation by information network-based clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York City, New York, USA, 821-830.

Roweis S.T., and Saul L.K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500): 2323-2326.

Strohman,T.; Croft, W.; and Jensen, D. 2007. Recommending citations for academic papers. In *Proceedings of the 30th ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, Netherland, 705-706.

Tang,J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; and Su; Z. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, Las Vegas, Nevada, USA, 990-998.

Tang, J.; Qu, W.; Wang, M.Z.; Zhang, M.; Yan, J. and Mei, Q.Z. 2015a. Line: large-scale information network embedding. In *Proceedings of the 24th International World Wide Web Conference*, Florence, Italy, 1067-1077.

Tang, J.; Qu, M. and Mei, Q.Z. 2015b. PTE: predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, 1165-1174.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol; P.A. 2010. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(3), 3371-3408.

Zhou, D.; Zhu, S.; Yu, K.; Song, X.; Tseng, B.; Zha, H. and Giles, C. 2008. Learning multiple graphs for document recommendations. In *Proceedings of the 17th International World Wide Web Conference*, Beijing China, 141-150.

Zhao, J.; Mathieu, M.; and Lecun, Y. 2016. Energy-based generative adversarial network. *arXiv: 1609.03126*.