A Novel Personalized Citation Recommendation Approach Based on GAN

Ye Zhang(✉), Libin Yang, Xiaoyan Cai, and Hang Dai

School of Automation, Northwestern Polytechnical University,
Xi'an 710072, China
yezhang@mail.nwpu.edu.cn

Abstract. With the explosive growth of scientific publications, researchers find it hard to search appropriate research papers. Citation recommendation can overcome this obstacle. In this paper, we propose a novel approach for citation recommendation by applying the generative adversarial networks. The generative adversarial model plays an adversarial game with two linked models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability which a sample came from the training data rather than G. The model first encodes the graph structure and the content information to obtain the content-based graph representation. Then, we encode the network structure and co-authorship to gain author-based graph representation. Finally, the concatenation of the two representations will be acted as the node feature vector, which is a more accurate network representation that integrates the author and content information. Based on the obtained node vectors, we propose a novel personalized citation recommendation approach called CGAN and its variation VCGAN. When evaluated on AAN dataset, we found that our proposed approaches outperform existing state-of-the-art approaches.

Keywords: Citation recommendation · Generative adversarial network
Latent representation · Deep learning

1 Introduction

Citation recommendation is the task of providing researchers a list of references satisfying their requirements. In simple terms, a good citation recommendation system can learn publication information and help to accelerate frequent tasks by intelligently identifying references. It is thus desirable to design a system that attempts to automatically generate high-quality candidate papers given a query document.

A variety of citation recommendation systems have been proposed in recent years. These research technologies include collaborative filtering [1], topic modeling [2, 3], machine learning [4] and deep learning [5]. Collaborative filtering mainly focuses on the rating matrices created from the researchers' readership or the adjacency matrix of the citation network. The probabilistic topic models have been widely used to recommend literatures due to its sound theoretical foundation and promising performance in this decade, such as LDA [2] and PLSA [3]. Duma et al. [4] explore new

M. Ceci et al. (Eds.): ISMIS 2018, LNAI 11177, pp. 268–278, 2018.
https://doi.org/10.1007/978-3-030-01851-1_26

recommendation algorithms using classification methods. Compared with other methods, deep learning algorithms perform more effectively in feature extraction and are scalable to large-scale data [5].

Graph embedding aims to output a vector representation for each node in the graph, such that nodes close to each other in network structure have similar vector representations in a low dimensional space. It is an effective scheme to represent graph data for further analysis. Many methods for network embedding have been proposed in recent years, such as DeepWalk [6], TADW [7] and TriDNR [8]. They aim to capture various connectivity information in a network to obtain representation results. These existing methods are effective in the network with different objectives, however, they suffer from lack of additional constraints to improve the robustness of the learned representations. Recently, many researchers have proposed to learn robust and reusable representations by training generative adversarial models [9, 10]. Moreover, some models have been designed to deal with graph data [11, 12].

In this paper, we propose a novel personalized citation recommendation approach called CGAN and its variation VCGAN based Generative Adversarial Network (GAN). Our approach leverages the principle of adversarial learning [9] to learn robust network representations. It first encodes the graph structure and the content information to obtain the content-based graph representation. Then, we encode the network structure and co-authorship to gain author-based graph representation. Finally, the concatenation of the two representations will be acted as the node feature vector. Based on the obtained nodes vectors, we propose a personalized citation recommendation approach. The experimental results demonstrate that our method outperforms other models in terms of both Recall and NDCG.

2 Related Work

2.1 Citation Recommendation

According to different scenarios of usages, citation recommendation can be divided into two main categories: global citation recommendation and local citation recommendation. This difference is the different queries according to which the citation should be made. The local citation recommendation aims to recommend a short list of papers according to the given context. The given context is the query and typically consists of one to three sentences. It usually ranks papers by measuring the relevance of paper content to the citation context [13]. On the contrary, the global citation recommendation suggests a list of papers to be the references for entire paper manuscript. The global citation recommendation approaches can produce effective query-oriented recommendation using rich information, including author, title, abstract, etc., thus can provide a comprehensive view of relevant references [14, 15]. We mainly focus on global citation recommendation in this paper.

2.2 Network Embedding Methods

Network embedding is an effective manner to represent the nodes in a network as low-dimensional vectors. Many network embedding methods have been proposed in recent years. These researches can be divided into three main categories: matrix factorization methods, probabilistic models and deep learning methods. The matrix factorization methods capture different kinds of high-order proximities by processing the adjacency matrix to obtain graph embedding [16, 17]. Probabilistic models attempt to learn graph embedding by employing manifold features of the graph [6, 18]. And researchers have proved that many probabilistic algorithms are equivalent to matrix factorization approaches [7].

These approaches above principally study the structure relationship or minimizing the reconstruction error. They have ignored the data distribution of the latent codes, and can easily result in poor representation in dealing with real-word sparse and noisy graph data. One common way to handle this problem is to introduce some regularization to the latent codes and enforce them to follow some prior data distribution. So generative adversarial based frameworks have been researched to learn robust latent representation. Dai et al. [11] exploit the strengths of generative adversarial networks in capturing latent features, and investigate its contribution in graph representation. And Pan et al. [12] propose an adversarial regularized graph auto-encoder for graph data.

3 Problem Definition

In this section, we introduce the definitions of problems and notations to be used in our citation recommendation system. Then, we will present details of our proposed citation recommendation approach based on generative adversarial network.

We formally define personalized citation recommendation task in this paper as follows: Given a bibliography dataset P and Q of the set of queries that contain titles and authors information. For every query, which consists of keywords, title or other information, we aim to recommend a high-quality list of references that have better correlation with the query. We first construct a bibliographic citation network, which is represented as $\mathcal{G} = (V, E, M, C)$ and $\mathcal{G}$ is a directed graph, where V is the paper nodes set, E is a set of edges representing the citation relation between the nodes. M is the adjacency matrix of graph $\mathcal{G}$ representing the network structure, C represents the node features information matrix, which $C_i \in C$ indicates the content features of the node v_i such as titles or authors information.

Given a network $\mathcal{G}$, our network embedding model aims to learn a low-dimensional vector $X_i \in R^d$ with the format as follows: f:$(M, C) \rightarrow X$, where X_i^T is the i-th row of the matrix $X \in R^{NXd}$. N is the number of nodes and d is the dimension of embedding. X is the representation matrix that encodes the structure and features information of the citation network. We construct the generative adversarial network embedding model to obtain the graph embedding matrix that it is our hard core. Then, we recommend a small subset of target $p \in P$ for a query manuscript $q \in Q$ by ranking the papers via the similarity score list s(q, p). Figure 1 shows the overview of personalized citation

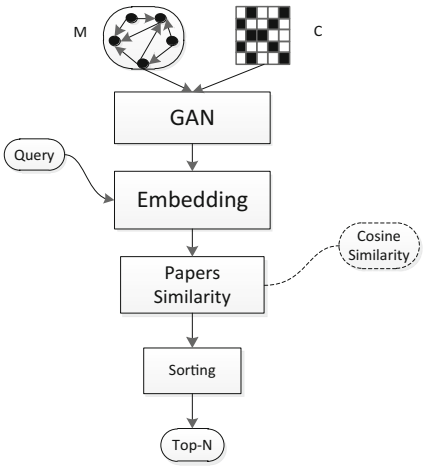

Fig. 1. Personalized citation recommendation model based on GAN

recommendation based on GAN. The details of every part will be presented in the following sections.

4 Proposed Method

4.1 Graph Convolutional Network

Graph convolutional network (GCN) [19] is an effective variant of convolutional network which operates directly on graph. We build a layer-wise convolution operation via a localized first-order approximation of spectral graph convolution. The form of this convolution propagation rule as follows:

$$X^{l+1} = \rho\left(\tilde{D}^{-\frac{1}{2}}\tilde{M}\tilde{D}^{-\frac{1}{2}}X^{l}W^{l}\right) \tag{1}$$

Here, X^l is the input of convolution layer and X^{l+1} is the output matrix in the l-th layer. The first layer input is $X^0 = C \in R^{N\times m}$($N$ nodes and m features) in our models. $\tilde{M} = M + I$ represents the adjacency matrix of graph $\mathcal{G}$ with added self-connections and I is the identity matrix, $\tilde{D}$ is the diagonal matrix, $\tilde{D}_{ii} = \sum_j M_{ij}$, and W^l is a layer-specific weight matrix that we will train in our convolution neural network. ρ denotes the activation function. In this paper, we construct a two-layer GCN and we use ReLU and linear activations in the first and second layers respectively. Consequently, $X = X^{(2)}$ is the representation matrix of the nodes, which encodes both graph structure and nodes features.

Then, we obtain an inference model parameterized by the two-layer GCN, which is a framework for unsupervised learning on graph-structured data based on the variational auto-encoder (VAE) in [20]:

$$q(X|C,M) = \prod_{i=1}^{N} q(X_i|C,M) \tag{2}$$

$$q(X_i|C,M) = N\left(X_i|\mu_i, diag\left(\sigma_i^2\right)\right) \tag{3}$$

$\mu = X^{(2)}$ is the matrix of mean vectors μ_i, similarly $\log\sigma = \rho_{linear}(X^1, M|W^1)$.

Generative Model. The generative model is designed to reconstruct graph structure M. We define $p\left(\tilde{M}|X\right)$ to predict a link between two nodes by an inner product between latent variables:

$$p\left(\tilde{M}_{ij} = 1|X_i, X_j\right) = \sigma\left(X_i^T X_j\right) \tag{4}$$

where σ is the sigmoid function. And we can derive the embedding matrix X and reconstruction adjacency matrix $\tilde{M}$ as follows:

$$\tilde{M} = \sigma\left(XX^T\right) \tag{5}$$

Optimization. We optimize our network model as follows:

$$L_1 = E_{q(X|C,M)}[logp(M|X)] \tag{6}$$

or by the variational lower bound:

$$L_2 = E_{q(X|C,M)}[logp(M|X)] - KL[q(X|C,M)||p(X)] \tag{7}$$

Where $KL(q||p)$ is the Kullback-Leibler divergence between $q(\cdot)$ and $p(\cdot)$. And $p(X)$ is a Gaussian prior.

4.2 Adversarial Embedding Model

The generative adversarial model consists of a generator G and a discriminator D. The generator function is designed to capture the data distribution. In our framework, G is applied to transform the input network features to embedding vectors. The discriminator is trained to estimate the probability that a sample came from the training data rather than G. Specifically, the embedding vectors are acted as fake samples and the real data is selected from a prior distribution $p(X)$. In this process, we simultaneously train D to maximize the probability that assigns the correct label to both training examples and samples from G, and train G to minimize $\log(1 - D(G(C,M)))$. This is a

two-layer minimax game for the generator and discriminator playing against each other. The cost function $V(D,G)$ can be written as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(X)}[logD(X)] + E_{x\sim p_x(X)}[log(1-D(G(C,M)))] \quad (8)$$

where $p_{data}(X)$ is a Gaussian prior. Algorithm 1 is the detailed description of models.

Algorithm 1 Citation network embedding models based on GAN

Input: Adjacency matrix M and paper-title matrix $C_1$and paper-author matrix C_2;
The dimension of vector d;
The number of iterations K;
The number steps of iterating discriminator T.

Output: The embedding matrix $\mathbf{X}$.

for iterator = 1, 2, 3, ..., K **do**
  Get the variable matrix $X_i, i = 1$ or 2 through Eq.(1)
  for k=1,2,3, ..., T **do**
    Sample s entities x^j, from latent matrix X_i;
    Sample s entities a^j, from the prior distribution P_x;
    Update the discriminator by ascending stochastic gradient Eq.(8)
  end for
  Update the graph convolutional network with its stochastic gradient Eq.(6) (called for CGAN) or Eq.(7) (called for VCGAN).
end for
Get the representation matrix $\mathbf{X}$, which is the concatenation of X_1 and X_2.
Return $\mathbf{X}$

4.3 Citation Recommendation

After training the GAN model, we apply the embedding matrix to recommend citation papers for a query q. The cosine similarity is an effective and regular way to define the similarity between two feature vectors. Therefore, the recommending score $s(q,p_j)$ is computed by the cosine similarity of representation vector of q and a candidate paper p_j as follows:

$$s(q,p_j) = \frac{q \cdot p_j}{\|q\| \cdot \|p_j\|} \quad (9)$$

Here, $\|\cdot\|$ is the 2-norm. According to (9), we calculate $s(q,p_j)$ for all candidates and recommend papers with higher scores.

5 Experiments and Evaluation

In this section, we establish the experiments on the real-world publication dataset to evaluate the performance of our proposed methods.

5.1 Experiment Settings

Dataset. The AAN dataset [21] is a famous dataset that contains the complete information of papers in computational linguistic and natural language processing. We select a subset published from 1965 to 2012 that contains 13,394 papers as our experimental dataset and the references are also in the dataset for every paper. Then, there are 10715 papers published before 2012 to be used to train models, and the remaining 2679 papers published in 2012 are regarded as the testing data. Table 1 demonstrates the statistics of these datasets.

Table 1. Statistics of the dataset

	Papers	Authors	Citations
Train	10715	7540	63831
Test	2679	1716	16001

Evaluation Methods. To evaluate the recommendation accuracy and quality of our models, we use Recall and NDCG as evaluation metrics. Both metrics are commonly used to evaluate recommendation results.

- **Recall**, is the rate of the number of real cited papers in the Top-N recommendation list to the total number of cited articles. This metric is calculated as follows:

$$\text{Recall@N} = \frac{1}{Q}\sum_{j=1}^{Q}\left(\frac{R_q \cap T_q}{T_q}\right) \tag{10}$$

Here N is the length of the recommendation list and Q is the number of queries. R_q is the Top-N list based on a query paper q. T_q is the set of papers citing q.

- **NDCG**, normally, the highly relevant references should appear higher in the Top-N list. So, we use NDCG to measure the ranked recommendation list. It can be defined as:

$$\text{NDCG@N} = \frac{1}{Q}\sum_{j=1}^{Q}\left(\sum_{i=1}^{N}\frac{2^{r_i-1}}{log_2(i+1)}/IDCG@N\right) \tag{11}$$

where r_i is the ratio of the i-th paper in the ranking list. The paper is relevant or not when $r_i = 1$ or $r_i = 0$. IDCG@N is the ideal ranking result.

Baselines. To validate the effectiveness of our models based on GAN, we compare them with other baseline approaches as follows:

- **DeepWalk:** It is a network representation method by using random walk and Skip-gram model which encodes relations into a continuous vector space.
- **TADW:** It is a state-of-the-art approach that incorporates text information and network to learn the network representation.
- **TriDNR:** It simultaneously considers paper network structure and paper vertex content to learn paper network representation.

Parameter Settings. We compare models based on their ability to recommend references in citation recommendation. For DeepWalk, TADW and TriDNR, the parameters are provided by the authors in their original papers. For instance, an embedding dimension of 128, 10 random walks of length 80 per node and window size of 10, trained for every epoch. For CGAN and VCGAN, we create a weight variable with Glorot and Bengio [22] and optimize the models using Adam [23] with a learning rate of 0.001. The dimension of the hidden layer is 32 and the latent variables is 16 in our experiments that we use the implementation from [20].

5.2 Results Analysis

We compare the proposed models and other baselines in terms of citation recommendation performance. The comparison results of different methods are summarized in Table 2.

It is obviously to see that CGAN and VCGAN outperform other baselines in terms of Recall and NDCG. The statistics demonstrate that generative adversarial graph network can effectively improve the robustness and discrimination of citation recommendation algorithms. Furthermore, the VCGAN model achieves higher recommendation performance than CGAN (on average of 18.42% with respect to Recall and 20.20% with respect to NDCG). DeepWalk, TADW and TriDNR algorithms perform fairly poor on citation recommendation network. This is mainly because DeepWalk learns network representation using network structure only and ignores the text information. TADW is a matrix factorization algorithm that factorizes a matrix M, which is a large sparse matrix. In practice, it is very computationally expensive so that TADW only factorizes an approximate M, which will affect its recommendation efficiency. Although TriDNR considers both network structure and vertex content, it ignores inter-relationship among vertices. So it performs worse than our proposed approaches.

Table 2. Performance comparison between different methods

Top-N	25		50		75		100	
Metrics	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
DeepWalk	0.27891	0.17256	0.34751	0.19428	0.39208	0.20682	0.42224	0.21509
TADW	0.26317	0.27469	0.34156	0.26371	0.38773	0.26402	0.42472	0.26607
TriDNR	0.28102	0.27351	0.36324	0.26901	0.40522	0.26237	0.43870	0.27119
CGAN	**0.28133**	**0.26094**	**0.38931**	**0.27023**	**0.41793**	**0.27274**	**0.44609**	**0.27491**
VCGAN	**0.42774**	**0.31312**	**0.45003**	**0.28062**	**0.44982**	**0.33013**	**0.44461**	**0.33086**

5.3 Case Study

To better understand the availability of our models, we give case study in this subsection to compare the recommended results for a given query, which we use the title in our dataset as the query information. Because of the page limit, we only list top-5 retrieved papers obtained by VCGAN, CGAN and TriDNR approaches. Table 3 shows the top-5 recommended papers selected from the recommendation list.

As shown in Table 3, the results returned by the VCGAN approach have 4 records that match the ground truth citation list of the query, whereas the results returned by the CGAN and TriDNR approaches have 3 and 2 matching records, respectively. This observation demonstrates that VCGAN and CGAN obtain a better result in this case study since the manuscript author, title, as well as text information of training papers are utilized in our approaches. And the VCGAN model achieves higher performance because of the superior performance of the variational graph auto-encoders [20]. The KL-divergence term can then be interpreted as regularizing parameters, encouraging the approximate posterior to be close to the prior. It is a very effective technique to improve the accuracy of models.

Table 3. The top 5 recommended citations: the ground truths are ☑

Query: sentiment classification using automatically extracted subgraph features
VCGAN
(1) Recognizing contextual polarity in phrase-level sentiment analysis ☑
(2) Sentiment analysis of citations using sentence structure-based features ☑
(3) Sentiment classification and polarity shifting ☑
(4) Sentiment classification on customer feedback data: noisy data, large feature vectors and the role of linguistic analysis ☑
(5) Opinion and suggestion analysis for expert recommendations
CGAN
(1) Recognizing contextual polarity in phrase-level sentiment analysis ☑
(2) Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis ☑
(3) Interactive annotation learning with indirect feature voting
(4) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales ☑
(5) Going beyond traditional QA systems: challenges and keys in opinion question answering
TriDNR
(1) Learning subjective nouns using extraction pattern bootstrapping
(2) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts ☑
(3) Identifying types of claims in online customer reviews
(4) Generalizing dependency features for opinion mining adding redundant features for CRFs-based sentence sentiment classification ☑
(5) User-directed sentiment analysis: visualizing the affective content of documents

6 Conclusion

Graph embedding is an effective approach to represent graph data into a low-dimensional space for further analysis. Generative adversarial network can estimate generative model via an adversarial process. In this paper, we propose a generative adversarial citation recommendation model combining network structure, content information and authorship. This framework can capture network structure and other abundant information to improve the robustness of recommendation. Experimental results demonstrate the effectiveness of our algorithms. In the future, we plan to explore the affection of the models parameters and the number of the convolution layers, and expand the construction to combine with venue information in the process of network representation.

References

1. McNee, S.M., Istvan, A., et al.: On the recommending of citations for research papers. In: Proceedings of ACM Conference on Computer Supported Cooperative Work, pp. 116–125 (2002)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. **42**, 177–196 (2001)
4. Duma, D., Liakata, M., Clare, A., Ravenscroft, J., Klein, E.: Applying core scientific concepts to context-based citation recommendation. In: Proceedings of LREC (2016)
5. Ebesu, T., Fang, Y.: Neural citation network for context-aware citation recommendation. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (2017)
6. Perozzi, B., Al-Rfou, R., Skiena, S.: DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710. ACM (2014)
7. Yang, C., Liu, Z.Y., Zhao, D.L., Sun, M.S., Chang, E.Y.: Network representation learning with rich text information. In: International Joint Conference on Artificial Intelligence (2015)
8. Pan, S.R, Wu, J., Zhu, X.Q, Zhang, C.Q, Wang, Y.: Tri-party deep network representation. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York City, NY, USA, pp. 701–710 (2016)
9. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
10. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I.: Adversarial autoencoders. In: ICLR Workshop (2016)
11. Dai, Q.Y., Li, Q., Tang, J., Wang, D.: Adversarial network embedding. arXiv preprint arXiv:1711.07838 (2017)
12. Pan, S.R., Hu, R.Q., Long, G.D., Jiang, J., Yao, L., Zhang, C.Q.: Adversarially Regularized Graph Autoencoder. arXiv preprint arXiv:1802.04407v1 (2018)
13. Bethard, S., Jurafsky, D.: Who should I cite: learning literature search models from citation behavior. In: Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM 2010), pp. 609–618 (2010)

14. Dai, T., Zhu, L., Cai, X.Y., Pan, S.R., Yuan, S.: Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network. J. Ambient Intell. Hum. Comput. **9**, 957–975 (2017)
15. Shaparenko, B., Joachims, T.: Identifying the original contribution of a document via language modeling. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 696–697 (2009)
16. Cao, S., Lu, W., Xu, Q.: Deep neural networks for learning graph representations. In: AAAI, pp. 1145–1152 (2016)
17. Ou, M., Cui, P., Pei, J., et al.: Asymmetric transitivity preserving graph embedding. In: KDD, pp. 1105–1114 (2016)
18. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: KDD, pp. 855–864 (2016)
19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, (2016)
20. Kipf, T.N., Welling, M.: Variational graph auto-encoders. In: NIPS (2016)
21. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The ACL anthology network corpus. Lang. Resour. Eval. **47**(4), 919–944 (2013)
22. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feed forward neural networks. In: AISTATS, vol. 9, pp. 249–256 (2010)
23. King, D.P., Ba, J.L. Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015)