Effective Strategies on Representing Information Networks

Hang Li,
Haozheng Wang
CCCE&CS,
Nankai University, China
{hangl,hzwang}@mail.
nankai.edu.cn

Zhenglu Yang*,
Jin-Mao Wei
CCCE&CS,
Nankai University, China
{yangzl,weijm}@nankai.edu.cn

Masato Odagaki
Maebashi Institute of
Technology, Japan
odagaki@maebashi-
it.ac.jp

ABSTRACT

Network representation is the basis of many applications and of extensive interest in various fields such as information retrieval, social network analysis, and recommendation systems. Majority of previous methods on network representation only considered incomplete aspects of the problem, such as link structure, node information, or partial integration. The present paper proposes a comprehensive network representation model, which seamlessly integrates the text information, node label, and first-order and second-order proximity of a network. The effectiveness of the introduced strategies is experimentally evaluated. Results demonstrate that our method is better than state-of-the-art techniques.

Keywords

Information network; representation learning; classification

1. INTRODUCTION

Information network representation is an important research issue because it is the basis of many applications [2], such as document classification in citation networks, functional label prediction in protein-protein interaction networks, potential friend recommendation in social networks. However, the rich and complex information (i.e., link structure and node contents) found in information networks imposes a great challenge for effectively representing networks.

To address the issue, several deep-learning based approaches have been introduced [1, 2, 4, 5] in recent years, where the node content is usually represented as text information indicating the properties of a node [3]. Most previous studies only utilized one kind of information. The work in [2] focused on the node content whereas others [1, 4] explored link structure. Although a few previous models [3, 5] combined both content information and network structure, they did not preserve the complete network structure and the node content was only partially utilized.

In this paper, we propose effective techniques to learn network representation by modeling both node content information and network structure comprehensively. This study aims to seamlessly integrate text information, node label, first-order and second-order proximities together. The experimental evaluation demonstrates the superior performance of our strategies on the benchmark datasets.

2. THE PROPOSED MODEL

Let $G = (V, E, C, L)$ denotes a given network, where $V = \{v_i\}$ is the node set, $E = \{e_{ij}\}$ is the edge set, $C = \{c_i\}$ is the set of text information, and $L = \{l_i\}$ is the set of class labels. Our goal is to seek a low-dimensional vector for each node of a given network. The architecture of our model is shown in Fig. 1.We learn an effective feature vector representation preserving both the link structure and the node content, which will be applied to many tasks (e.g., paper classification).

As shown in Fig. 1, we first construct the first-order node relation module. Given a network, the set of connected nodes is obtained and the joint probability between v_i and v_j is $p_1(v_i, v_j) = \frac{1}{1+exp(\vec{u}_i \cdot \vec{u}_j)}$, where $\vec{u}_i$ and $\vec{u}_j$ are the vector representation of nodes v_i and v_j,respectively. The first-order proximity indicates the similarity between two vertices. The weight w_{uv} on a edge e_uv between two nodesu, v indicates the first-order proximity between u and v. In our study, the directly linked vertices are assumed to have similar representations. Therefore, the objective function

$$\mathcal{L}_1 = -\sum\nolimits_{(v_i,v_j)\in E} \log p_1(v_i, v_j) \qquad (1)$$

is minimized to preserve the first-order proximity.

For the second-order node relation module, we adopt several stunted random walks to generate the node sequences. We then apply DeepWalk [4] on these sequences to learn the node representation. The second-order proximity is the similarity between the neighborhood network structure. If two nodes share many common neighbors, we assume that they are similar, which indicates high second-order proximity. Therefore, we maximize the likelihood of neighbor nodes given a node v_i to preserve the second-order proximity. The objective function is as follows:

$$\mathcal{L}_2 = \sum_{i=1}^{N} \sum_{s\in S} \sum_{-b\leq j\leq b, j\neq 0} \log P(v_{i+j}|v_i) \qquad (2)$$

where N is the number of nodes, b is the context width (window size), s is a node sequence, and $P(v_j|v_i)$ is computed as: $P(v_j|v_i) = \frac{\exp(\vec{u}_j^T . \vec{u}_i)}{\sum_{k=1}^{|V|} \exp(\vec{u}_k^T . \vec{u}_i)}$.

We employ the state-of-the-art approach, i.e., Doc2vec [2], which utilizes text to learn the vector representation for documents as our content2vec module. Specifically, if one node contains a label, the label is treated as a word and merged

*Corresponding author

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054241

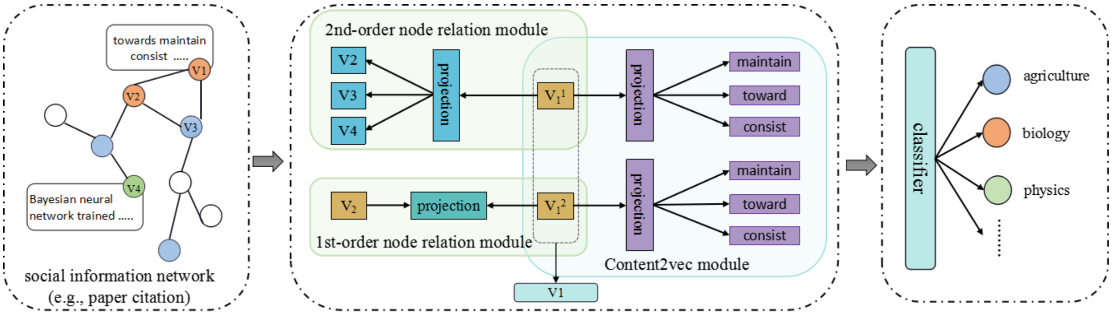
towards maintain consist .....
V1
V2
V3
V4
Bayesian neural network trained .....
social information network (e.g., paper citation)
2nd-order node relation module
V2
V3
V4
projection
V_1^1
projection
maintain
toward
consist
V_2
projection
V_1^2
1st-order node relation module
projection
maintain
toward
consist
Content2vec module
V1
classifier
agriculture
biology
physics
......

Figure 1: Architecture of our model

into the comprehensive text information (e.g., the abstract of the paper in the citation network) as the content of the node.Therefore, we can maximize the following objective:

$$\mathcal{L}_c = \sum_{i=1}^{|L|} \log P(w_{-b} : w_b | l_i) + \sum_{i=1}^{N} \log P(w_{-b} : w_b | v_i) \quad (3)$$

where w is a word in the text information of v_i, b is the window size of word sequence, and l_i is the label of v_i.

By maximizing the objectives $O_1 = (\partial - 1)\mathcal{L}_1 + \partial\mathcal{L}_c$ and $O_2 = (1 - \partial)\mathcal{L}_2 + \partial\mathcal{L}_c$, and concatenating the two vectors learned by the two objectives as the final result of the representation of each node, we can obtain the network representation preserving the first-order proximity, second-order proximity, and the content information.

3. EXPERIMENTS

We conduct an important task, i.e., node classification, on two benchmark datasets: (1) CiteSeer-M10[1]. It contains 10 categories with 10,310 papers and 77,218 citations. Titles are treated as the text information; and (2) DBLP dataset[2]. Abstracts are treated as the text information. The setting is the same as that of [3], which has 30,422 nodes and 41,206 edges. To reduce the influence of classifiers, a common linear SVM is used in all the methods. In each network, p% nodes are randomly selected as training set, the rest are testing set. The experiments on Citeseer-M10 are independently conducted 10 times for each setting and the average values are reported. On DBLP, the experiments are conducted once for each setting[3]. As previous work did, we report Macro-F1 as the evaluation metric. Default parameters are set: dimension d=50, window size b=10, text weight ∂=0.8.

Our model is evaluated by comparing with five techniques (Table 1). DeepWalk [4] and Node2vec [1] are structure-based that exhibit inferior performance because the network is rather sparse. Doc2Vec [2] is based on text and works better on DBLP than CiteSeer-M10. TADW [5] and TriDNR are inferior to our approach, although these two methods also consider the text and structure. However, they cannot capture complete structure and utilize whole text information. Our model exhibits consistent superior performance and is much better than the state-of-the-art methods.

Each proposed technique is evaluated (i.e., Fig. 2). Considering only the first-order proximity and content information yields unsatisfactory results. Incorporating the second-order proximity can greatly improve performance.

Table 1: Performance comparison between ours and the state-of-the-art on Macro-F1

methods	DBLP			CiteSeer-M10		
%p	30%	50%	70%	30%	50%	70%
deepwalk	0.4364	0.4311	0.4337	0.2875	0.2967	0.3073
node2vec	0.4703	0.4714	0.4784	0.4507	0.4402	0.4612
TADW	0.6603	0.6717	0.6780	0.4886	0.4980	0.5030
TriDNR	0.7315	0.7425	0.7523	0.6685	0.7046	0.7272
Doc2vec	0.7082	0.7219	0.7289	0.3992	0.4052	0.4021
Ours	**0.7526**	**0.7645**	**0.7761**	**0.6982**	**0.7278**	**0.7461**

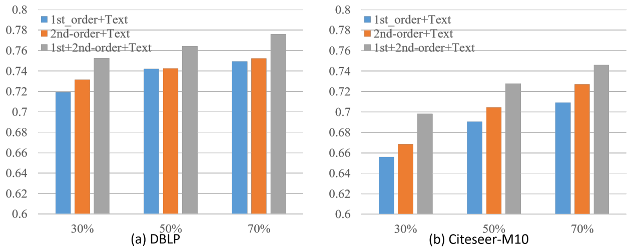

Figure 2: Performance of each strategy on different training proportion

4. CONCLUSIONS

We have introduced an effective network representation model, which comprehensively integrates text information and network structure. The experimental evaluation demonstrates the effectiveness of our strategies.

Acknowledgments: This research is supported by the National Natural Science Foundation of China: U1636116, 11431006, 61070089, Research Fund for International Young Scientists: 61650110510, Ministry of Education of Humanities and Social Science: 16YJC790123, Natural Science Foundation of Tianjin: 14JCYBJC15700, 15JCYBJC46600.

5. REFERENCES

[1] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, 2016.

[2] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[3] S. Pan, J. Wu, X. Zhu, C. Zhang, and Y. Wang. Tri-party deep network representation. In *IJCAI*, 2016.

[4] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *SIGKDD*, 2014.

[5] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Y. Chang. Network representation learning with rich text information. In *IJCAI*, 2015.

[1] http://citeseerx.ist.psu.edu/

[2] http://arnetminer.org/citation (V4 version is used)

[3] For long version please visit `http://123.56.138.34/cidewebpage/publications/lh/www_2017_extension.pdf`