Embedding Identity and Interest for Social Networks

Linchuan Xu
The Hong Kong Polytechnic University
Kowloon, Hong Kong
cslcxu@comp.polyu.edu.hk

Xiaokai Wei
University of Illinois at Chicago
Chicago, IL,USA
weixiaokai@gmail.com

Jiannong Cao
The Hong Kong Polytechnic University
Kowloon, Hong Kong
csjcao@comp.polyu.edu.hk

Philip S. Yu
University of Illinois at Chicago
Chicago, IL,USA
psyu@uic.edu

ABSTRACT

Network embedding fills the gap of applying tuple-based data mining models to networked datasets through learning latent representations or embeddings. However, it may not be likely to associate latent embeddings with physical meanings just as the name, latent embedding, literally suggests. Hence, models built on embeddings may not be interpretable. In this paper, we thus propose to learn identity embeddings and interest embeddings, where user identity includes demographic and affiliation information, and interest is demonstrated by activities or topics users are interested in. With identity and interest information, we can make data mining models not only more interpretable, but also more accurate, which is demonstrated on three real-world social networks in link prediction and multi-task classification.

1. INTRODUCTION

Recently, network embedding has been utilized to fill the gap of applying tuple-based data mining models to networked datasets by learning embeddings which preserve the network structure [2, 4, 1, 8]. In this way, however, it may not be likely to associate embeddings with physical meanings as existing methods do not explicitly specify what kind of information to be embedded. Hence, models built on embeddings may not be interpretable, e.g., similarities measured on embeddings can be used to infer new interactions between nodes but one may not know why they are similar.

In this paper, we propose to embed identity and interest (EII). EII learns identity embeddings and interest embeddings separately rather than blindly fusing them into one global embedding [2, 4, 1, 8], which is important because almost all user behaviors can be explained by who the user is and what the user is interested in [3]. And models built on embeddings can be more accurate because embeddings encode user's interests besides the network structure.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054268

We can learn identity embeddings and interest embeddings because the connections between people can be similarly categorized. On the one hand, connections can easily come into being between people with similar demographic characteristics including ethnicity and education, such as family relation and schoolmate. On the other hand, connections exist between people with similar psychological characteristics including attitudes and interests and world-wide web makes it easy for people to find others in other parts of the world with similar interests.

2. THE EII MODEL

The EII embeds social networks $G(N, E, A)$ with node content, where N is the set of nodes, E is the set of weighted or unweighted, directed edges, e.g., e_{ij} is the edge from node i to j. If the relationship between users has no direction, it is replaced by two directed edges e_{ij} and e_{ji}. $A \in \mathbb{R}^{M \times L}$ is a matrix of term frequency extracted from the content where $M = |N|$ and L is the number of words or attributes.

The EII model learns identity embeddings and interest embeddings by preserving the network structure as well as user-generated content. The network structure is preserved by presenting nodes connected by edges to be close, and those not connected to be away from each other. Also, the type edges, identity-induced or interest-induced, need to be inferred, and corresponding embeddings should be adopted. Formally, the closeness between two nodes is defined as the probability that there exists an edge between them, where the probability is computed on embeddings as follows:

$$p(e_{ij}) = \pi_{ij_u} p(e_{ij}|\boldsymbol{u}_i, \boldsymbol{u}_j) + \pi_{ij_v} p(e_{ij}|\boldsymbol{v}_i, \boldsymbol{v}_j), \tag{1}$$

where $\pi_{ij_u} \in \mathbb{R}$ or $\pi_{ij_v} \in \mathbb{R}$ is the likelihood that the link is identity-induced or interest-induced, respectively, and $\pi_{ij_u} + \pi_{ij_v} = 1$. For easier subsequent optimizations, π_{ij_u} is defined as "softmax weight" as follows:

$$\pi_{ij_u} = \frac{\exp\{\xi_{ij_u}\}}{\exp\{\xi_{ij_u}\} + \exp\{\xi_{ij_v}\}}, \tag{2}$$

where $\xi_{ij_u} \in \mathbb{R}$ and $\xi_{ij_v} \in \mathbb{R}$. π_{ij_v} is similarly defined. $p(e_{ij}|\boldsymbol{u}_i, \boldsymbol{u}_j)$ is defined as follows:

$$p(e_{ij}|\boldsymbol{u}_i, \boldsymbol{u}_j) = \frac{1}{1 + \exp\{-\boldsymbol{u}_i^\top \boldsymbol{u}_j\}}, \tag{3}$$

where $\boldsymbol{u}_i \in \mathbb{R}^D$ and $\boldsymbol{u}_j \in \mathbb{R}^D$ denote identity embeddings, $D \in \mathbb{R}$ is the dimension. $p(e_{ij}|\boldsymbol{v}_i, \boldsymbol{v}_j)$ is defined similarly, and $\boldsymbol{v}_i \in \mathbb{R}^D$ and $\boldsymbol{v}_j \in \mathbb{R}^D$ are interest embeddings.

To cast the structure preserving mechanism to an optimization problem, both small probabilities of pairs of nodes connected by edges and large probabilities of pairs not connected should be penalized. EII employs the logistic loss.

To preserve the user-generated content, the interest embedding of each user should accord with his/her content. This can be achieved by regularizing interest embeddings to corresponding content, which is formulated as follows:

$$\min_{\boldsymbol{V},\boldsymbol{P}} ||\boldsymbol{V}\boldsymbol{P} - \boldsymbol{A}||_2^2, \tag{4}$$

where $\boldsymbol{V} \in \mathbb{R}^{M\times D}$ is the matrix of interest embeddings, $\boldsymbol{P} \in \mathbb{R}^{D\times L}$ is a projection matrix, and $||\cdot||_2$ is F2 norm.

Hence, jointly minimizing the logistic loss and the projection loss is the optimization objective of the EII, which is quantified with regularization as follows: $L(\boldsymbol{U},\boldsymbol{V},\boldsymbol{P},\boldsymbol{\Pi}) =$

$$\begin{aligned} &- \sum_{e_{ij}\in E} (w)_{ij} p(e_{ij}) - \sum_{e_{hk}\notin E} \log(1 - p(e_{hk})) \\ &+ ||\boldsymbol{V}\boldsymbol{P} - \boldsymbol{A}||_2^2 + \lambda||\boldsymbol{U}||_2^2 + \beta||\boldsymbol{V}||_2^2 + \gamma||\boldsymbol{P}||_2^2, \end{aligned} \tag{5}$$

where $\boldsymbol{V} \in \mathbb{R}^{M\times D}$, is the matrix of identity embeddings, λ, β and $\gamma \in \mathbb{R}$ are regularization coefficients.

$L(\boldsymbol{U},\boldsymbol{V},\boldsymbol{P},\boldsymbol{\Pi})$ is not jointly convex on the four variables, so we solve each variable iteratively by gradient descent. To obtain an appropriate initialization point for the gradient descent, $\boldsymbol{U}$ and $\boldsymbol{V}$ are pre-trained by preserving the network structure and the content, respectively. The pre-training of $\boldsymbol{U}$ is performed by solving the logistic loss without considering the type of each edge. And the pre-training of $\boldsymbol{V}$ can be performed similarly by constructing a k-nearest neighbor network of nodes, where the similarities between nodes are quantified the cosine similarity of their attributes.

3. EMPIRICAL EVALUATION

3.1 Experiment Settings

The EII model is evaluated against three embedding models, i.e., DeepWalk [2], LINE [4], and node2vec [1]. For the implementation of the algorithm to solve the EII model, the dimension of embeddings is set as 128, ξ_{iju} and ξ_{ijv} are initialized as 0.5, k of the k-NN network is set as top 1% of all the nodes, ratio of the number of e_{ij} to that of e_{hk} is set as 5 as used in LINE, λ, β, and γ are set as 1. Backtracking line search is employed to learn the descent rate of each iteration, and the relative loss that determines whether the gradient decent process converges is set as 0.001. Three studied social networks are BlogCatalog(7857 nodes, 137649 edges, 5351 attributes, 15 groups) [7], Flickr(6318, 404085, 8523, 5) [6] and DBLP(6482, 19265, 8298, 4) [5] sampled from KDD, ICDM, SDM, PAKDD, AAAI, ICML, NIPS, IJCAI, CVPR, ECML, SIGMOD, VLDB, ICDE, PODS, EDBT, WWW, SIGIR, CIKM, WSDM, ECIR from the year 2000 to 2009.

3.2 Experiment Results

In ***link prediction***, the closeness is employed as similarity measurement. For BlogCatalog and Flickr networks, we use 60% of edges as training edges and the remaining ones as test links. For the DBLP network, new co-authorships occur from 2011 to 2013 are used as test links. For all networks, the same number of negative links are randomly sampled for the evaluation purpose. For the EII model, if a particular pair of nodes does not appear during the training process, the identity weight and interest weight are both set as initialized 0.5. The performance on AUC score is presented in Table 1, and all the numbers have been multiplied by 100%.

Link Prediction	BlogCatalog	Flickr	DBLP
DeepWalk	79.32%	78.30%	73.26%
LINE(1st)	65.36%	69.10%	62.45%
LINE(2nd)	70.47%	68.66%	76.44%
node2vec	77.75%	70.71%	71.59%
EII	90.26%	82.03%	77.21%
Multi-label Classification	**BlogCatalog**	**Flickr**	**DBLP**
DeepWalk	58.42&57.86	66.52&55.71	78.48&77.46
LINE(1st)	59.55&58.72	63.95&56.92	77.62&76.51
LINE(2nd)	57.60&56.95	66.07&56.55	75.93&75.45
node2vec	55.85&55.73	57.76&48.72	76.96&75.66
EII	76.08&73.53	68.79&59.36	85.30&84.77

Table 1: Performance comparison

It shows the EII model consistently outperforms all baselines on all datasets. The superior performance is more visible on BlogCatalog and Flickr, which may suggest that interests demonstrated by user-generated content play a relatively more importance role in online social networks than in professional networks like DBLP. Besides, the EII model can provide explanations for the predictions, e.g., it is established because two users know each other in the real world if they are closer in the identity space.

In ***multi-label classification*** where groups are used as labels, 5-fold cross validation is employed as the evaluation method, and Micro-F1&Macro-F1 scores are reported. It shows similar results to the link prediction.

4. REFERENCES

[1] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[2] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[3] R. C. Stedman. Toward a social psychology of place predicting behavior from place-based cognitions, attitude, and identity. *Environment and behavior*, 34(5):561–581, 2002.

[4] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. ACM, 2015.

[5] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[6] X. Wang, L. Tang, H. Gao, and H. Liu. Discovering overlapping groups in social media. In *the 10th IEEE International Conference on Data Mining series (ICDM2010)*, Sydney, Australia, December 14 - 17 2010.

[7] X. Wang, L. Tang, H. Liu, and L. Wang. Learning with multi-resolution overlapping communities. *Knowledge and Information Systems (KAIS)*, 2012.

[8] L. Xu, X. Wei, J. Cao, and P. S. Yu. Embedding of embedding (eoe): Joint embedding for coupled heterogeneous networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 741–749. ACM, 2017.