GCN-IA: User Profile Based on Graph Convolutional Network with Implicit Association Labels

Jie Wen[1,2], Lingwei Wei[1,2], Wei Zhou[1,2(✉)], Jizhong Han[1,2], and Tao Guo[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
zhouwei@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Abstract. Inferring multi-label user profile plays a significant role in providing individual recommendations and exact-marketing, etc. Current researches on multi-label user profile either ignore the implicit associations among labels or do not consider the user and label semantic information in the social networks. Therefore, the user profile inferred always does not take full advantage of the global information sufficiently. To solve above problem, a new insight is presented to introduce implicit association labels as the prior knowledge enhancement and jointly embed the user and label semantic information. In this paper, a graph convolutional network with implicit associations (GCN-IA) method is proposed to obtain user profile. Specifically, a probability matrix is first designed to capture the implicit associations among labels for user representation. Then, we learn user embedding and label embedding jointly based on user-generated texts, relationships and label information. On four real-world datasets in Weibo, experimental results demonstrate that GCN-IA produces a significant improvement compared with some state-of-the-art methods.

Keywords: Implicit association labels · User profile · Graph convolutional networks

1 Introduction

With the growing popularity of online social networks including Weibo and Twitter, the "information overload" [1] come up and the social media platforms took more effort to satisfy users' more individualized demands by providing personalized services such as recommendation systems. User profile, the actual representation to capture certain characteristics about an individual user [25], is the basis of recommendation system [7] and exact-marketing [2, 3]. As a result, user profiling methods, which help obtaining accurate and effective user profiles, have drawn more and more attention from industrial and academic community.

A straightforward way of inferring user profiles is leveraging information from the user's activities, which requires the users to be *active*. However, in many real-world applications a significant portion of users are *passive* ones who keep following and reading

V. V. Krzhizhanovskaya et al. (Eds.): ICCS 2020, LNCS 12139, pp. 355–364, 2020.
https://doi.org/10.1007/978-3-030-50420-5_26

but do not generate any content. As a result, label propagation user profile methods [4–6] are widely studied, which mainly use the social network information rather than user's activities. In order to obtain user profile more accurately and abundantly, multi-label is applied in many researches to describe users' attributes or interests. Different labels were assumed independently [5] in some research, while the associations among labels were ignored and some implicit label features remained hidden. Meanwhile, several researches [1, 8, 9] considered the explicit associations among labels to get user profile and achieved better performance. Besides the explicit associations, there exists implicit association among labels that is beneficial to make user profile more accurate and comprehensive. The previous work [10] leveraged internal connection of labels, which is called implicit association. However, this work only considered the relation of labels, but ignored the user and label semantic information jointly based on user-generated texts, relationships and label information, which is also important for user profile.

To take advantage of this insight, a graph convolutional networks with implicit label associations (GCN-IA) is proposed to get user profile. A probability matrix is first designed to capture the implicit associations among labels for user representation. Then, we learn user embedding and label embedding jointly based on user-generated texts, relationships and label information. Finally, we make multi-label classification based on given user representations to predict unlabeled user profiles. The main contributions of this paper are summarized as follows:

- **Insight.** We present a novel insight about combination among implicit association labels, user semantic information and label semantic information. In online social networks, due to users' personalized social and living habits, there are still certain implicit associations among labels. At the same time, user and label information from user-generated texts, relationships and label information is significant for the construction of user profile.
- **Method.** A graph convolutional networks with implicit label associations (GCN-IA) method is proposed to get user profile. We first construct the social network graph with the relationship between users and design a probability matrix to record the implicit label associations, and then combine this probability matrix with the classical GCN method to embed user and label semantic information.
- **Evaluation.** Experiments evaluating GCN-IA method on 4 real Weibo data sets of different sizes are conducted. The comparative experiments evaluate the accuracy and effectiveness of GCN-IA. The results demonstrate that the performance is significantly improved compared with some previous methods.

The following chapters are organized as follows: In Sect. 2, related works are briefly elaborated. The Sect. 3 describes the details of GCN-IA, and experiments and results are described in Sect. 4. Finally, we summarize the conclusion and future work in Sect. 5.

2 Related Works

Label propagation method shows advantages of linear complexity and less required given user's labels, and disadvantages such as low accuracy and propagation instability. The

existing label propagation methods in user profile can be divided into three parts. One is to optimize the label propagating process to obtain more stable and accurate profiles, the second part is to propagate multi-label through social network structure to get more comprehensive user profile, and the last part is to apply deep-learning methods such as GCN to infer multi-label user profile.

2.1 Propagation Optimization

Label propagation method was optimized by leveraging more user attributes information, specifying propagation direction and improving propagation algorithm. Subelj et al. proposed balanced propagation algorithm in which an increasing propagation preferences could decide the update order certain nodes, so that the randomness was counteracted by utilizing node balancers [14]. Ren et al. introduced node importance measurement based on the degree and clustering coefficient information to guide the propagation direction [15]. Li et al. leveraged user attributes information and user attributes' similarity to increase recall ratio of user profile [5]. Huang et al. redefined the label propagating process with a multi-source integration framework that considered content and network information jointly [16]. Explicit associations among labels also have been taken into consideration in some research, Glenn et al. [1] introduced the explicit association labels and the results proved the efficiency of the method.

We innovatively introduced the implicit association labels into multi-label propagation [10], the method was proved to be convergent and faster than traditional label propagation algorithm and its performance was significantly better than the state-of-the-art method on Weibo datasets. However the research [10] ignored user embedding and label embedding jointly based on user-generated texts, relationships and label information, which seemed very important for user profile.

2.2 Multi-label Propagation

The multi-label algorithms were widely applied to get abundant profile. Gregory et al. proposed COPRA algorithm and extended the label and propagation step to more than one community, which means each node could get up to v labels [17]. Zhang et al. used the social relationship to mine user interests, and discovered potential interests from his approach [6]. Xie et al. recorded all the historical labels from the multi-label propagation process, which make the profile result more stable [18]. Wu et al. proposed balanced multi-label propagation by introducing a balanced belonging coefficients p, this method improved the quality and stability of user profile results on the top of COPRA [19].

Label propagation algorithm has been improved in different aspects in the above work, however it's still difficult to get a high accuracy and comprehensive profile due to the lack of input information and the complex community structures.

2.3 GCN Methods

GCN [20] is one of the most popular deep learning methods, which can be simply understood as a feature extractor for graphs. By learning graph structure features through

convolutional neural network, GCN is widely used in node classification, graph classification, edge prediction and other research fields. GCN is a semi-supervised learning method, which can infer the classification of unknown nodes by extracting the characteristics of a small number of known nodes and the graph structure. Due to the high similarity with the idea of label propagation, we naturally consider constructing multi-label user profile with GCN. Wu et al. proposed a social recommendation model based on GCN [21], in which both user embedding and item embedding were learned to study how users' interests are affected by the diffusion process of social networks. William et al. [22] and Yao et al. [23] applied GCN for text classification and recommendation systems respectively, with node label and graph structure considered to GCN modeling. However, the existing methods rarely consider the implicit relationships between labels in the GCN based methods.

3 Methodology

3.1 Overview

This section mainly focuses on the improvement of graph convolutional networks (GCN) based on implicit association labels. The goal of this paper is to learn user representation for multi-label user profile task by modeling user-generated text and user relationships.

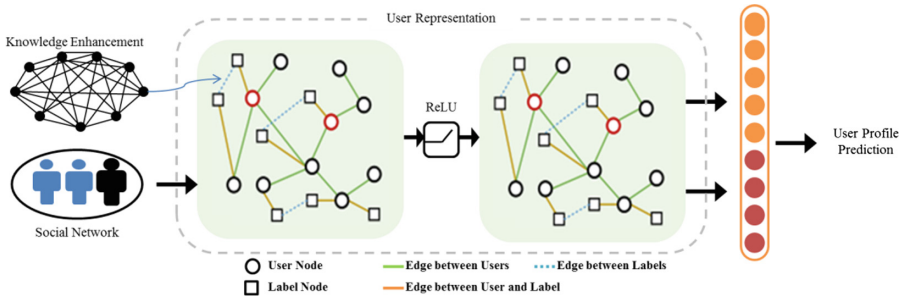

Fig. 1. Overall architecture of GCN-IA.

The overall architecture of GCN-IA is shown in Fig. 1. The model consists of three components: **Prior Knowledge Enhancement (PKE) module**, **User Representation module**, and **Classification module**. Similar with other graph-based method, we formulated the social network into a heterogeneous graph. In this graph, nodes represent the users in social network and edges represent user's multiple relationships such as following, supporting and forwarding. First, PKE captures the implicit associations among labels for user representation. Then, user representation module learns user embedding and label embedding jointly based on user-generated texts, relationships and label information. Classification module makes multi-label classification based on user representations to predict unlabeled user profiles.

3.2 Prior Knowledge Enhancement Module

Social networks are full of rich knowledge. According to [10], associations among implicit labels are very significant in user profile. In this part, we introduce the knowledge of implicit association among labels to capture the connections among users and their profile labels.

A priori knowledge probability matrix P is defined as Eq. (1). Probability of propagation among labels gets when higher P_{ij} gets a higher value.

$$P_{ij} = \frac{\left|\left\{t|t \in I \wedge \left(l_i, l_j\right) \subseteq t\right\}\right|}{\sum_{i=0}^{m} \sum_{j=0}^{m} \left|\left\{t|t \in I \wedge \left(l_i, l_j\right) \subseteq t\right\}\right|} \tag{1}$$

Associations in social network are complex due to uncertainty [12] or special events [13]. Therefore, we define the set of labels, where elements are sampled by co-occurrence, cultural associations, event associations or custom associations, as shown in Eq. (2).

$$I = I_1 \cup I_2 \cup I_3 \cup \ldots \tag{2}$$

Where $\mathrm{I_i}(\mathrm{i} = 1, 2, 3, \ldots)$ represents respectively a set of each user's interest label set.

3.3 User Representation Module

Generally, the key idea of GCNs is to learn the iterative convolutional operation in graphs, where each convolutional operation means generating the current node representations from the aggregation of local neighbors in the previous layer. A GCN is a multilayer neural network that operates directly on a graph and induces embedding vectors of nodes based on properties of their neighborhoods.

In the user representation module, we apply GCNs to embed users and profile labels into a vector space and learn user representation and label representation jointly from user-generated content information and social relationships. Specifically, the implicit associations as prior knowledge are introduced to improve the GCNs to model the associations among labels.

Formally, the model considers a social network $G = (V, E)$, where V and E are sets of nodes and edges, respectively. In our model, there are two types of nodes, *user node* and *label node*. The initialized embedding of user nodes and label nodes, denoted as X, is initialized with user name and their content via pre-trained word2vec model.

We build edges among nodes based on user relationships (*user-user edges*), users' profiles (*user-label edges*) and implicit associations among labels (*label-label edges*). We introduce an adjacency matrix A of G. and its degree matrix D, where $D_{ii} = \sum_{j=1,\ldots,n} A_{ij}$. The diagonal elements of A are set to 1 because of self-loops. The weight of the edges between a user node and a label node is based on user profile information, formulated as Eq. (3).

$$A_{\mathrm{ij}} = \begin{cases} 1 \text{ if the user } i \text{ is with the label } j \\ 0 \qquad\qquad \text{otherwise} \end{cases}, \textit{where } i \in \mathcal{U}_{gold}, j \in \mathcal{C} \tag{3}$$

Where $\mathcal{U}$ is the set of all users in the social network, $\mathcal{U}_{gold}$ denotes labeled users. And $\mathcal{C}$ is the set of labels of user profile.

To utilize label co-occurrence information for knowledge enhancement, we calculate weights between two label nodes as described in Sect. 3.2. The weights between two user nodes are defined as Eq. (4) according to user relationships.

$$A_{ij} = \begin{cases} 1 \times \text{Sim(i, j)} & if \left(u_i, u_j\right) \in \mathcal{R} \\ 0 & otherwise \end{cases}, where\ i, j \in \mathcal{U} \tag{4}$$

Where $\mathcal{R} = \{(u_0, u_1), (u_1, u_3), ..)\}$ is the set of relations between users and $Sim(i, j)$ indicates the similarity between user i and user j followed by [10]. The less the ratio of the value is, the closer the distance is.

GCN stacks multiple convolutional operations to simulate the message passing of graphs. Therefore, both the information propagation process with graph structure and node attributes are well leveraged in GCNs. For a one-layer GCN, the new k-dimensional node feature matrix is computed as:

$$L^{(1)} = \sigma\left(\tilde{A}XW_0\right) \tag{5}$$

Where $\tilde{A}$ ($\tilde{A} = \mathrm{D}^{-1/2}W\mathrm{D}^{-1/2}$) is a normalized symmetric adjacency matric, and W_0 is a weight matrix. $\sigma(\cdot)$ is an activation function, e.g. a ReLU function $\sigma(x) = \max(0, x)$. And the information propagation process is computed as Eq. (6) by stacking multiple GCN layers.

$$L^{(j+1)} = \sigma\left(\tilde{A}L^{(j)}W_j\right) \tag{6}$$

Where j denotes the layer number and $L^{(0)} = X$.

3.4 User Profile Prediction

The prediction of user profile is regarded as a multi-classification problem. After the above procedures, we obtain user representation according to user-generated content and relationships. The node embedding for user representation is fed into a *softmax* classifier to project the final representation into the target space of class probability:

$$Z = p_i(c|\mathcal{R}, \mathcal{U}; \Theta) = \textbf{softmax}\left(\tilde{A}\sigma\left(\tilde{A}XW_0\right)W_1\right) \tag{7}$$

Finally, the loss function is defined as the cross-entropy error over all labeled users as shown in Eq. (8).

$$\mathcal{L} = -\sum\nolimits_{u \in y_u} \sum\nolimits_{f=1}^{F} Y_{df} \ln Z_{df} \tag{8}$$

Where y_u is the set of user's indices with labels, and F is the dimension of the output features which is equal to the number of classes. Y is the label indicator matrix. The weight parameters W_0 and W_1 can be trained via gradient descent.

4 Experiments

4.1 Dataset

Weibo is the largest social network platform in China[1]. Followed by [10], we evaluate our method in different scale data sets in Weibo.

The datasets are sampled with different users in different time. And we select five classes as interest profiles of users, *Health, Women, Entertainment, Tourism, Society*.

The details of the datasets are illustrated in Table 1.

Table 1. The details of the datasets.

Dataset	Number of users	Health	Women	Entertainment	Tourism	Society
1#	4568	990	803	3397	1733	1031
2#	4853	1054	860	3592	1828	1088
3#	5140	1122	909	3811	1930	1163
4#	5711	1271	1014	4218	2146	1336

4.2 Comparisons and Evaluation Setting

To evaluate the performance of our method (GCN-IA), we compare it with some existing methods including textual feature-based method and relation feature-based method. In addition, to evaluate the implicit association labels for GCN, we compare GCN-IA with classical GCN. The details of these baselines are listed as follows:

SVM [26] uses the method of support vector machine to construct user profile based on user-generated context. In our experiment, we select username and blogs of users to construct user representation based on textual features. The textual features are obtained via pre-trained word2vec model.

MLP-IA [10] uses multi-label propagation method to predict user profiles. They capture relationship information by constructing probability transfer matrix. The labeled users are collected if the user is marked with a "V" which means his identity had been verified by Weibo. Analyzed by Jing et al. [24], these users were very critical in the propagation.

In the experiments, we will analyze the precision ratio (P) and recall ratio (R) of method which respectively represent the accuracy and comprehensiveness of user profile. And F1-Measure ($F1$) is a harmonic average of precision ratio and recall ratio, and it reviews the performance of the method.

[1] http://weibo.com/.

4.3 Results and Analysis

The experiment results are shown in Table 2. The results show that our method can make a significant increase in macro-F1 in all datasets.

Compared with feature-based method, our model makes a significant improvement. SVM fails since the method does not consider user relationships in the social networks. It only models the user-generated context, such as username and user's blogs.

Table 2. Experimental results of user profile task.

Method	1#	2#	3#	4#
SVM [26]	0.4334	0.2018	0.4418	0.4240
MLP-IA [10]	0.5248	0.4657	0.5030	0.5541
GCN-IA (Ours)	**0.5838**	**0.6040**	**0.5782**	**0.5708**

Compared with relation-based method, our model achieves improvements in all datasets, especially in dataset of 2#, we have improved 13.83% in macro-F1. MLP-IA [10] established user profiles based on user's relationships via label propagation. It suffers from leveraging the user-generated context, which contains semantic contextual features. Our model can represent users based on both relationships and context information via GCN module, which is more beneficial for identifying multi-label user profile task.

Fig. 2. The results of each interest class in 4# dataset.

The results of each interest class in 4# dataset are shown in Fig. 2. The results show that GCN-IA performs stably in all interest profiles, which demonstrate the good robustness of our model.

As shown in the results, the performance is little weak for the Entertainment interest class compared with baselines. In Weibo, there are much blogs with aspect to entertainment. Fake information exists in social network including fake reviews and fake accounts for specific purposes, which brings huge challenge for user profiles.

Our model constructs user profile via both textual features and relational features. The results can demonstrate that the user relationships can provide a beneficial signal for semantic feature extraction and the two features can reinforce each other.

5 Conclusion and Future Work

In this paper, we have studied the user profile by graph convolutional networks with implicit association labels, user information and label information embedding. We proposed a method to utilize implicit association among labels and then we take graph convolutional networks to embed the label and user information. On four real-world datasets in Weibo, experimental results demonstrate that GCN-IA produces a significant improvement compared with some state-of-the-art methods.

Future work will pay more attention to consider more prior knowledge to get higher performance.

References

1. Boudaer, G., Loeckx, J.: Enriching topic modelling with users' histories for improving tag prediction in Q and A systems. In: 25th International Conference Companion on World Wide Web, pp. 669–672. ACM, New York (2016)
2. Paulo, R.S., Frederico, A.D.: RecTwitter: a semantic-based recommender system for twitter users. In: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web. ACM, New York (2018)
3. Nurbakova, D.: Recommendation of activity sequences during distributed events. In: Conference on User Modeling, Adaptation and Personalization, pp. 261–264 (2018)
4. Chang, P.S., Ting, I.H., Wang, S.L.: Towards social recommendation system based on the data from microblogs. In: International Conference on Advances in Social Networks Analysis and Mining, pp. 672–677. IEEE, Washington, D.C. (2011)
5. Li, R., Wang, C., Chang, C.C.: User profiling in an ego network: co-profiling attributes and relationships. In: International Conference on World Wide Web, pp. 819–830. ACM, New York (2014)
6. Zhang, J.L.: Application of tag propagation algorithm in the interest map of Weibo users. Programmer **5**, 102–105 (2014)
7. Fernando, A., Ashok, C., Tony, J., et al.: Artwork personalization at Netflix. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp. 487–488. ACM, New York (2018)
8. Liang, S.S., Zhang, X.L., et al.: Dynamic embeddings for user profiling in twitter. In: Proceedings of the 24th SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2018), UK, pp. 1764–1773 (2018)

9. Roberto, P.S., Frederico, A.D.: RecTwitter: a semantic-based recommender system for Twitter users. In: Proceedings of the 24th Brazilian Symposium on Multimedia and the Web (WebMedia 2018), pp. 371–378. ACM, New York (2018)
10. Wei, L., Zhou, W., Wen, J., Lin, M., Han, J., Hu, S.: MLP-IA: multi-label user profile based on implicit association labels. In: Rodrigues, J.M.F., et al. (eds.) ICCS 2019. LNCS, vol. 11536, pp. 548–561. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22734-0_40
11. Huai, M., Miao, C., Li, Y., et al.: Metric learning from probabilistic labels. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2018), pp. 1541–1550. ACM, New York (2018)
12. Peng, P., Wong, R.C.-W., Yu, P.S.: Learning on probabilistic labels. In: Proceedings of the 2014 SIAM International Conference on Data Mining, SIAM, pp. 307–315 (2014)
13. Iyer, A.S., Nath, J.S., Sarawagi, S.: Privacy-preserving class ratio estimation. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 925–934. ACM (2016)
14. Šubelj, L., Bajec, M.: Robust network community detection using balanced propagation. Eur. Phys. J. B **81**(3), 353–362 (2011). https://doi.org/10.1140/epjb/e2011-10979-2
15. Ren, Z.-M., Shao, F., Liu, J.-G., et al.: Node importance measurement based on the degree and clustering coefficient information. Acta Phys. Sin **62**(12), 128901 (2013)
16. Huang, Y., Yu, L., Wang, X., Cui, B.: A multi-source integration framework for user occupation inference in social media systems. World Wide Web **18**(5), 1247–1267 (2014). https://doi.org/10.1007/s11280-014-0300-6
17. Gregory, S., et al.: Finding overlapping communities in networks by label propagation. New J. Phys. **12**(10), 103018 (2010)
18. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state-of-the-art and comparative study. ACM Comput. Surv. **45**(4), 1–35 (2011)
19. Wu, Z., Lin, Y., Gregory, S., et al.: Balanced multi-label propagation for overlapping community detection in social networks. J. Comput. Sci. Technol. **27**, 468–479 (2012)
20. Niepert, M., Ahmed, M., Kutzkov, K.: Learning convolutional neural networks for graphs. In: Proceedings of the 33rd International Conference on Machine Learning. ICML (2016)
21. Wu, L., Sun, P., Hong, R., Fu, Y., Wang, X., Wang, M.: SocialGCN: an efficient graph convolutional network based model for social recommendation. arXiv preprint arXiv:1811.02815 (2018)
22. Rex, Y., He, R., Chen, K., et al.: Graph convolutional neural networks for web-scale recommender systems. In: Proceedings of the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD (2018)
23. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. AAAI (2019)
24. Jing, M., Yang, X.X.: The characterization and composition analysis of Weibo"V". News Writ. **2**, 36–39 (2014)
25. Piao, G., Breslin, J.G.: Inferring user interests in microblogging social networks: a survey. User Model. User-Adap. Inter. **28**, 277–329 (2018). https://doi.org/10.1007/s11257-018-9207-8
26. Song, R., Chen, E., Zhao, M.: SVM based automatic user profile construction for personalized search. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 475–484. Springer, Heidelberg (2005). https://doi.org/10.1007/11538059_50
27. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. ICLR (Poster). arXiv preprint arXiv:1609.02907 (2017)