SCIENCE CHINA
Information Sciences

• LETTER •

November 2019, Vol. 62 219104:1–219104:3
https://doi.org/10.1007/s11432-018-9718-9

Inferring user profiles in social media by joint modeling of text and networks

Ruifeng XU[1*], Jiachen DU[1], Zhishan ZHAO[2], Yulan HE[3], Qinghong GAO[1] & Lin GUI[3]

[1]*School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen* 518000*, China;*
[2]*Baidu Inc., Beijing* 100085*, China;*
[3]*Department of Computer Science, University of Warwick, Warwick* CV4 7AL*, UK*

Received 21 March 2018/Revised 21 October 2018/Accepted 20 December 2018/Published online 18 September 2019

Citation Xu R F, Du J C, Zhao Z S, et al. Inferring user profiles in social media by joint modeling of text and networks. Sci China Inf Sci, 2019, 62(11): 219104, https://doi.org/10.1007/s11432-018-9718-9

Dear editor,

With the rapid growth of social media platforms such as Twitter and Chinese Weibo, personalized services, such as recommendation and personalized advertising, provide more engaging experiences for the users. A prerequisite to personalized services is effective user profiling. However, many users do not want to share their profiles online due to the privacy concerns. To address this problem, automatic user profiling based on users' posted contents and their social activities online has attracted many attentions in recent years [1–3].

Existing approaches to user profiling in social media platforms, such as Twitter, Instagram and Chinese Weibo, have explored information extracted from user-generated content, social networks, and through semantic enrichment of user profiles.

In this study, we propose an alternative approach based on neural networks for user profiling by jointly leveraging the user-generated text and social networks. Evaluations on the SMP CUP 2016 dataset show that our proposed model outperforms the state-of-the-art baselines on inferring the age, gender and region of the users on the Chinese Weibo platform.

Methodology. Our model is comprised of two components: text representation learning and graph embedding learning. More concretely, for text representation learning, words in text are firstly converted to low-dimensional vectors pre-trained by word2vec [4]. In order to capture long-distance semantic information in text, we use the bidirectional long-short term memory (LSTM) model to obtain the representation of each post. As a user could post more than one post, we leverage the attention mechanism [5] to calculate the weighted sum of all posts as the final text representation for the user.

For graph embedding learning, we first construct a social network of users based on their following-followed relations, and then use the large information network embedding (LINE) [6], which preserves both the first-order and second-order proximity between users, to learn the node representation for each user.

Finally, the representations of text and the node in the social network for a given user are concatenated to jointly model the user profile. The joint representation is fed to a classifier of predefined attributes of a user like gender, age and region.

Text representation learning. Assuming that the posts of a user is denoted as $\{p^1, p^2, \ldots, p^m\}$, where m is the number of posts, each post p^j can be seen as a sequence of words $p^j = \{w_1^j, w_2^j, \ldots, w_{L_j}^j\}$, where L_j is the length of post j. The word is firstly converted into a d-dimensional vector by looking up a pre-trained word embedding

* Corresponding author (email: xuruifeng@hit.edu.cn)

matrix W_E. Thus the sequence of word embeddings $v^j = \{v_1^j, v_2^j, \ldots, v_{L_j}^j\}, v_t^j \in \mathbb{R}^d$ is used as the input to text representation learning. We use the bidirectional LSTM to capture the post representation by capturing the information from both the forward and the backward directions. The process is written by

$$\overrightarrow{h_t} = \overrightarrow{\text{LSTM}}(v_t, \overrightarrow{h_{t-1}}), \quad t \in [0, T-1],$$
$$\overleftarrow{h_t} = \overleftarrow{\text{LSTM}}(v_t, \overleftarrow{h_{t+1}}), \quad t \in [0, T-1].$$

The post representation is obtained by concatenating the forward hidden state $\overrightarrow{h_t}$ and the backward hidden state $\overleftarrow{h_t}$, i.e., $p_j = \overrightarrow{h_t} \oplus \overleftarrow{h_1}$.

In order to obtain the representation of multiple posts of a user, we firstly compute the importance score e_j for post p_j by a fully connected neural network with a tanh activation function. Then the sigmoid function is applied to get the normalized importance score of a post p_j, denoted as α_j.

$$e_j = {v_m}^{\mathrm{T}} \tanh(W_m p_j),$$

where $W_m \in \mathbb{R}^{d_p \times d_p}, v_m \in \mathbb{R}^{d_p}$ are parameters of the neural network and e_j is the importance score. To normalize the importance score for variable-size sets of posts, we use the sigmoid function to obtain the attention weight of each post by

$$\alpha_j = \frac{\exp(e_j)}{\sum_{k=1}^{T} e_j},$$

where α_j is the attention weight for post j. The whole set of posts of user i can be represented by the weighted sum of post representations by

$$u_{\text{post}}^i = \sum_{j=1}^{m} \alpha_j p_j^i,$$

where u_{post}^i is the representation of all posts generated by user i.

Graph embedding learning. We construct a social network based on the following-followed relations, in which a node denotes a user and the link connected two nodes denotes the existing of the following-followed relation. We use the large information network embedding (LINE) [6], which preserves both the first-order and second-order proximity between users, to obtain the representation of each user in the social network. The first-order proximity refers to the local pairwise proximity between the vertices in the network. To model the first-order proximity, for each edge (u_i, u_j), the first-order proximity objective function is defined as

$$O_1 = -\sum_{(i,j)\in E} \log p_1(u_i, u_j),$$

where $\boldsymbol{u}_i, \boldsymbol{u}_j \in \mathbb{R}^d$ are low-dimensional embedding vectors of node u_i and u_j, respectively, and $p_1(u_i, u_j)$ is the joint probability of nodes u_i and u_j.

To preserve the second-order proximity, the conditional probability $p_2(\cdot|u_i)$ should be close to the empirical distribution. The objective function is defined as

$$O_2 = -\sum_{(i,j)\in E} \log p_2(u_j|u_i).$$

The node embedding is trained by asynchronous stochastic gradient algorithm (ASGD) to optimize the joint proximity objective functions $O_1 + O_2$. In each training step, ASGD samples a mini-batch of ground-truth edges and negative samples by

$$\frac{\partial O_1}{\partial \boldsymbol{u}_i} = \frac{1}{p_1(u_i, u_j)} \frac{\partial p_1(u_i, u_j)}{\partial \boldsymbol{u}_i},$$
$$\frac{\partial O_2}{\partial \boldsymbol{u}_i} = \frac{1}{p_2(u_j|u_i)} \frac{\partial p_2(u_j|u_i)}{\partial \boldsymbol{u}_i}.$$

When the joint objective function converges to a local minimum, we get the representation of user i, u_{node}^i, in the social network.

Inferring user profile. After learning the post representation and the node representation of a given user, we feed the concatenated representations to a softmax classifier to predict the attribute of a user. The softmax classifier is computed by

$$x^i = W_{\text{cls}}(u_{\text{post}}^i \oplus u_{\text{node}}^i) + b_{\text{cls}},$$
$$y_c^i = \text{softmax}(x^i) = \frac{\exp x_c^i}{\sum_C \exp x_c^i},$$

where $\oplus$ is the vector concatenation operation. W_{cls} and b_{cls} are the parameters of the fully connected network in the softmax layer, y_c^i is the predicted probability of user belonging to class c ($c = 1, \ldots, C$) and C is the number of classes of the user attributes.

Experimental setting. The dataset used in this study is released by the SMP CUP 2016 competition, which was collected from Chinese Sina Weibo. The competition contains three tasks: predicting the gender (2 classes), age (3 classes) and region (8 classes) of the users. In our experiments, LTP[1)] is used for word segmentation. The training set of the SMP CUP 2016 dataset contains 4000 samples, and the test set consists of 1240 samples. We use the classification accuracy as the evaluation metrics, which was used in the competition.

1) http://ltp.ai/.

Experimental results. Table 1 shows the experimental results. SVM [7] and CNN [8] only use user-generated text to infer user profile. Heterogeneous graph embedding [3] learns user representation by mapping users and words in a unified heterogeneous graph. It is observed that our model performs consistently better than all of the baselines across all the three tasks. These results show that our proposed method can better predict user attributes in social media by leveraging the representations learned from both the post content and the social network. The performance of our model is better than the top teams in the SMP CUP 2016 competition [9], which relied on hand-crafted features. We also observe that the ensemble strategy further improves the performance of our proposed model for user profile inference.

Table 1 Classification accuracy of our proposed model against baselines

Baselines	Gender	Age	Region
LSTM	75.42	51.01	62.50
LINE [6]	66.30	51.82	59.99
SVM [7]	73.42	50.01	65.50
CNN [8]	76.33	51.62	62.62
Heterogeneous graph embedding [3]	81.33	74.39	60.92
TOP in SMP [9]	88.30	64.80	72.70
Our model	**90.97**	**68.98**	**73.76**
Our model (Ensemble)	**91.29**	**69.14**	**74.52**

Conclusion. In this study, we presented a novel neural network based method for microblog user profile inference. The bidirectional LSTM and the self-attention mechanism are used to capture useful features in user-generated content. Network embedding is used to extract user characteristics hidden in social networks. Incorporating both the content features and node embeddings learned from social networks, our method achieves the state-of-the-art performance on the SMP CUP 2016 dataset. For future work, we will extend our model to incorporate other information such as avatars and user names.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. U1636103, 61632011, 61876053), Shenzhen Foundation Research Funding (Grant No. 20170307150024907), Key Technologies Research and Development Program of Shenzhen (Grant No. JSGG20170817140856618), and Innovate UK (Grant No. 103652).

References

1 Cristin R, Cyril Raj V. Consistency features and fuzzy-based segmentation for shadow and reflection detection in digital image forgery. Sci China Inf Sci, 2017, 60: 082101
2 Zhao Y Y, Qin B, Liu T. Encoding syntactic representations with a neural network for sentiment collocation extraction. Sci China Inf Sci, 2017, 60: 110101
3 Gui L, Zhou Y, Xu R F, et al. Learning representations from heterogeneous network for sentiment classification of product reviews. Knowledge-Based Syst, 2017, 124: 34–45
4 Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013. 3111–3119
5 Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation. 2015. ArXiv: 1508.04025
6 Tang J, Qu M, Wang M Z, et al. Line: large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, 2015. 1067–1077
7 Wang S, Manning C D. Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 2012. 90–94
8 Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. 1746–1751
9 Zhao Z S, Du J C, Gao Q H, et al. Inferring user profile using microblog content and friendship network. In: Proceedings of Chinese National Conference on Social Media Processing, 2017. 29–39