A Fusion Model of Multi-data Sources for User Profiling in Social Media

Liming Zhang[1], Sihui Fu[1], Shengyi Jiang[1,2(✉)], Rui Bao[1], and Yunfeng Zeng[1]

[1] School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China
zhanglimingl34@foxmail.com, jiangshengyi@163.com
[2] Engineering Research Center for Cyberspace Content Security of Guangdong Province, Guangzhou, China

Abstract. User profiling in social media plays an important role in different applications. Most of the existing approaches for user profiling are based on user-generated messages, which is not sufficient for inferring user attributes. With the continuous accumulation of data in social media, integrating multi-data sources has become the inexorable trend for precise user profiling. In this paper, we take advantage of text messages, user metadata, followee information and network representations. In order to integrate seamlessly multi-data sources, we propose a novel fusion model that effectively captures the complementarity and diversity of different sources. In addition, we address the problem of friendship-based network from previous studies and introduce celebrity ties which enrich the social network and boost the connectivity of different users. Experimental results show that our method outperforms several state-of-the-art methods on a real-world dataset.

Keywords: User profiling · Social media · Multi-data sources Fusion model

1 Introduction

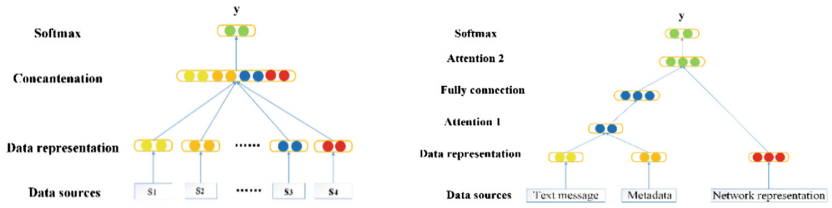
User profiling, which aims at effectively extracting user attributes from massive data information, is essentially valuable in various scientific research and business applications, such as recommendation system [1], search engine optimization [2], political position detection [3] and social network analysis [4]. Many attempts have utilized automated analysis model for user profiling tasks, such as user gender [5], age [6], geolocation [7], occupation [8], hobbies [9], personality [10] and influence [11].

With the rapid development of social media, like Twitter and Facebook, user profiling in social media obtains increasing attention. Traditional approaches [12–15] are mainly based on user-generated messages, such as tweets and micro-blogs, from which they construct a series of sophisticated features as the input of machine learning algorithms. Nevertheless, user-generated messages are usually short and full of noisy information. In addition to user's posts, recently, there have been several attempts to utilize other data sources in social media. Among them, social relationships network

© Springer Nature Switzerland AG 2018
M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 3–15, 2018.
https://doi.org/10.1007/978-3-319-99501-4_1

[16, 17] takes the most important role. Intuitively, people who become friends are likely to share the similar attributes. However, one of the existing problems in the friendship-based network is that if the training model does not include the user's friends in the network, the model may fail to predict the user's attributes.

Apart from text messages and network information, social media provides other data sources, such as users' nicknames, self-introductions, and personalized labels, which are also helpful for the identification of user attributes. Nowadays, more and more data sources are generated in social media, integrating multi-data sources has thus become the inexorable trend for precise user profiling. One naïve baseline for integration is to combine all the data representations one after another, as shown in Fig. 1(a). However, it does not consider interactions between different data sources. To address this, [18] built a hierarchical attention mechanism that unified text message, metadata and network representation for user geolocation prediction, as shown in Fig. 1(b). However, the attention mechanism may cause a certain loss of information when features derived from different data sources are combined via the summation operation. In addition, attention mechanism ignores the correlations between different data sources.



(a) Concatenation (b) Attention

Fig. 1. Previous baseline models

To better integrate different data sources, we take the advantage of bi-GRU architecture, which simultaneously captures the complementarity and diversity of each data source from bi-directions with the update gates and the reset gates. Different from the attention mechanism which integrates diverse inputs with a summation operation, we concatenate all the hidden states as the hybrid features, in order to retain the diversities of different data sources. We evaluate our model on a real-world dataset and experiment results show that our model outperforms the aforementioned methods. Besides, we also address the problem of scarcity in friendship network. Practically, we incorporate celebrity ties into the social network to enrich the information of user network representations.

The rest of the paper is organized as follows. In Sect. 2, we briefly review on the related work of user profiling. Then, we deliver a detailed description of our model in Sect. 3. Section 4 presents experimental results along with our analysis. Finally, we make a conclusion in Sect. 5.

2 Related Work

Most previous work in user profiling heavily relied on hand-crafted syntactic and lexical features extracted from the user-generated messages. [12] analyzed bloggers' writing styles and high frequency words at different genders and ages. [14] extracted stylistic features and lexical features from users' blogs, using SVM and logistic regression model to predict users' ages. Recently, deep learning methods have been applied in user profiling tasks and shown their effectiveness against traditional approaches. [19] devised a joint learning model with Long Short-Term Memory model to distinguish users' genders, ages, and professions.

Besides, network analysis targeting at node interactions in a connected network becomes a hot field over these years [20–22]. In network analysis methods, user profiling is treated as a node classification problem. [23] incorporated text features into network representation by matrix factorization. [4] devised a framework that preserves user profiles in the social network embeddings via a nonlinear mapping.

In social media, users are free to generate various types of data. It is found that the combination of two or more types of data is distinctly better than merely a single type in prediction tasks of user profiles. [24] established multi-scale Convolutional Neural Networks with an attention mechanism on user-generated contents, combined with user network representations. [25] unified users' tweets and other metadata (location, time zone) to predict user geolocation with a stacking approach. [18] developed a complex neural network model that joins text messages, user metadata and network representations for geolocation prediction.

In this paper, we unify user text messages, user metadata, followees' information and network representations. Different from previous methods, our model seamlessly incorporates different data sources by taking advantage of both their complementarity and diversity.

3 Model

In this section, we introduce our model, a fusion framework which joints four different types of data sources. Shown in Fig. 2, the model takes user text messages, metadata, followees' information and network representations as inputs. Then, four components are treated as a sequence while a bi-GRU layer is employed to learn their interdependency. Subsequently, all hidden units are concatenated as a new vector representation to retain their differentia, and then they are fed into the final fully connected layer. Details of the sub-models will be discussed in Sects. 3.1, 3.2, 3.3 and 3.4.

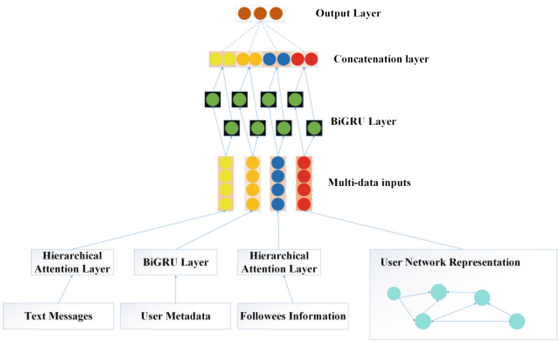


Fig. 2. Illustration of the fusion model. Hierarchical attention layer denotes hierarchical attention network. BiRNN denotes bi-recurrent neural network. Concatenation layer indicates concatenation of all hidden units learned from multi-data inputs.

3.1 Text Messages Representation

Text messages are the most important information for user profiling in social media. We take each message as one sentence formed by a sequence of words and aggregate all messages to be a document. To get the document presentation, We adopt the hierarchical attention network [26], in which there are two hierarchical layers. Figure 3 shows one typical hierarchical layer in the hierarchical attention network.

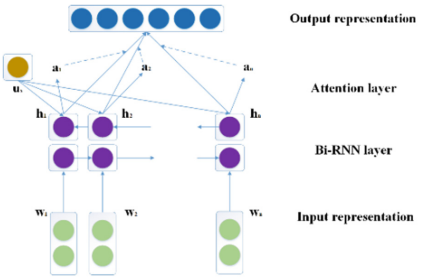


Fig. 3. The architecture of one hierarchical layer in the hierarchical attention network

The input representations can be the word embeddings, while the output comes to be a sentence embedding. Similarly, sentence embeddings can be combined into a document. To implement RNN, we use Gated Recurrent Unit (GRU) [27]. Formally, the formulations of GRU are as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$\tilde{h}_{\mathrm{t}} = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{3}$$

$$h_{\mathrm{t}} = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_{\mathrm{t}} \tag{4}$$

where z_t denotes an update gate, r_t a reset gate, $\tilde{h}_t$ a candidate state, h_t a hidden state, and x_t an input state, W_z, W_r, W_h, , U_z, U_r, U_h, b_z, b_r, b_h are model parameters, $\odot$ denotes the element-wise multiplication operator.

We get the hidden presentation h_{it} by concatenating forward hidden state $\vec{h}_{it}$ and backward hidden state $\overleftarrow{h}_{it}$. Then, a self-attention mechanism is introduced, which automatically assigns weights to different inputs. The formulation of the self-attention mechanism is defined as follows:

$$e_{\mathrm{i}} = tanh(W_w h_i + b_w) \tag{5}$$

$$\alpha_i = \frac{\exp\left(u_w^T e_{\mathrm{i}}\right)}{\sum_i \exp\left(u_w^T e_{\mathrm{i}}\right)} \tag{6}$$

$$s = \sum\nolimits_i \alpha_i h_i \tag{7}$$

where α_i is the weight of i-th of the hidden unit h_i, and u_w is the context vector, W_w and b_w are model parameters, s is the output vector.

3.2 Metadata Representation

Apart from text messages, user metadata information is also useful for inferring user attributes. In this paper, we regard *user nickname*, *self-introduction*, *education information*, *work information* and *individualized labels* as user metadata. We represent the metadata by concatenating all the elements, feeding them into a BiRNN layer and an Attention layer to the metadata representation.

3.3 Network Representation

For solving the sparsity of user friendships, most of previous works construct a 2-degree friends network. However, constructing such a network is very labor-intensive and time-consuming and most users are not well-connected. Therefore, an effective measure is to intensify the relationships among users. According to our observation, celebrity ties can be an alternative way to boost the connectivity between different users. As shown in Fig. 4, although user *B* and C do not have an explicit friendship, they have an indirect relationship as they both follow celebrity D and have interactions in blogs. Specifically, we define a social network as G = (V, E), where V represents the vertices including ordinary users as well as celebrities, and E indicates the relationships between vertices. There are three types of social relationship in our network: friendships, follower-followee relationships and blog-interacted relationships (*@mention*, *repost* and *vote*). We employ LINE [22] which involves the first-order and second-order proximities

between nodes to obtain users' vector representations in the social network, where we set all the weights of edges set to be 0 and 1.

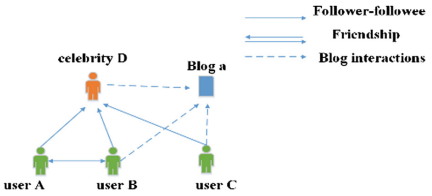


Fig. 4. An example of social network relationships

3.4 Followees' Information Representation

In addition to the construction of social network, we notice that the information of followee, especially celebrities, has strong relations with the user's traits. For example, if a user follows a certificated company that sells cosmetics, we suppose the user may be a female. In the followee list, followee usually put an explicit description of their nickname and self-introduction. Herein, we join the nickname and description to form a sentence representing one followee, and adopt another hierarchical attention network, shown as Fig. 5, to get the whole representation of followees' information.

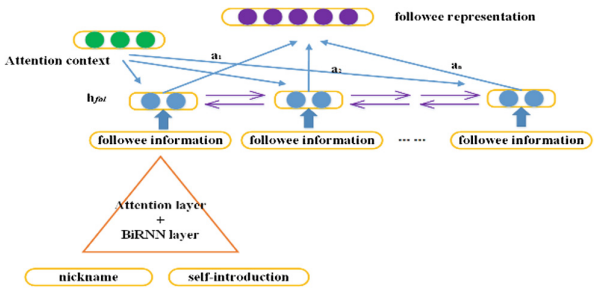


Fig. 5. Model architecture for getting followee information representation

3.5 Fusion Framework of Multi-data Sources

Given a series of data vector representations $\{s_1, s_2, s_3, s_4\}$, we adopt a Bi-GRU layer to learn their interdependencies adaptively on both forward and backward directions, resulting in corresponding hidden vectors $\{h_1, h_2, h_3, h_4\}$. Then, we concatenate all the hidden units, sending them through a fully-connected layer. In practice, the lengths of

$\{s_1, s_2, s_3, s_4\}$ may not be the same, so we adjust the hidden size of sub-models to ensure the same lengths of different inputs. Formally, a user vector representation v_u is computed by:

$$v_u = W_c[h_1 \oplus h_2 \oplus h_3 \oplus h_4] + b_c \tag{8}$$

$$h_i = f_{\text{BiGRU}}(s_i) \tag{9}$$

where W_c, b_c are the model parameters.

We adopt cross-entropy error of the predicted and true distributions as the loss function for optimization. The loss function is defined as follows:

$$L = -\sum_{i=1}^{T} \sum_{j=1}^{C} y_i^j \log\left(\widehat{y}_i^j\right) \tag{10}$$

where T denotes the number of training sets, C denotes the number of classes in user profiling tasks, and $y_i^j, \widehat{y}_i^j$ are the true labels and prediction probabilities respectively.

4 Experiments

Since there is few public benchmark data set on user profiling yet, we collect a real-word date set and evaluate our models in three user profiling tasks: gender, age and location prediction. Users for evaluation are from Sina Micro-blog[1], one of the most popular social media websites in China. We gathered user accounts from the comment lists of Micro-blog hot-searched event from March to April, 2018. Finally, we get 31,852 users with 21,884 females and 9,968 males, as females are more likely to share their comments than males. For the fair comparison, we adopt down sampling strategy that cuts off the sample number of female to 10,942. Figure 6 shows the distribution of user in age. We split ages into four intervals: [<19, 19–23, 24–27, >27], corresponding to different educational ages. User locations are in accordance with seven regions of China. Table 1 summarizes the statistics of our data set, where *NE* denotes the Northeast of China, *N* the North of China, *C* the Center of China, *E* the East of China, *NW* the Northwest of China, *SW* the Southwest of China, and *S* the South of China. We randomly select 70% as the training set, 10% the validation set and 20% the test set.

4.1 Experiment Setting

We use the pre-trained word2vec [28] vectors with a dimension of 200, and employ an open source Chinese Segmentation tool Jieba[2] to process the Chinese text. Words not included in the word2vec vectors are endowed with a uniform distribution between [−0.25, 0.25]. We adopt Adam algorithm [29] for optimization, and mini-batch size is fixed to 32. We train LINE [22] to get network representations of first-order

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

[2] https://github.com/fxsjy/jieba.

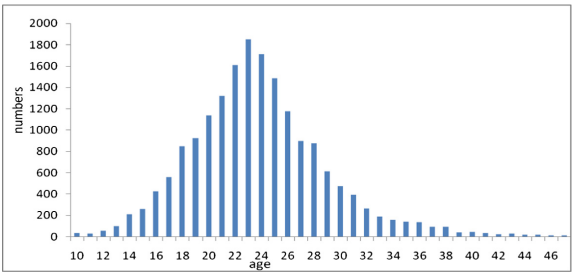


Fig. 6. Distributions of user in age

Table 1. Statistics of datasets for user profiling evaluation

Gender	Male			Female			
	9968			10942			
Age	<19		19–23	24–27		>27	
	3407		5914	5273		3716	
Location	NE	N	C	E	NW	SW	S
	2364	4359	3058	8144	1445	2951	3827

embeddings and second-order embeddings with 150 dimensions respectively. The GRU dimensions of words and sentences are set to be 50 and 150 respectively in hierarchical attention network. Since user metadata only comprises one bi-RNN layer, we set the GRU dimensions to be 150. The size of all attention units is set to 300, the dropout rate is 0.5. For evaluation, we use the metrics of F1 measures.

4.2 Baseline Methods

We compare our model with several baseline methods for user profiling tasks:

Text Feature + SVM: In traditional methods, bag of words (BOW) is regularly used for extracting text feature. In our experiments, different words in BOW model are weighted by TF-IDF, and then we adopt Singular Vector Decomposition (SVD) to perform dimensionality reduction, feeding them into an SVM classifier.

Text Feature + HAN: HAN represents the hierarchical attention network. Here, we use HAN only for extracting text messages features as the baseline method.

Multi-CNN + Network Embedding (MCNE) [24]**:** Each sentence is learnt by a convolutional neural network and an attention mechanism is used to assign weights to different sentences. Both the messages embedding and the network embedding are concatenated as the user embedding.

Multi-Data + Concatenation (MDC): Four types of data inputs are simply concatenated for the fully-connected layer, with details shown in Fig. 1(a).

Multi-Data + Attention (MDA) [18]**:** Three types of text information are aggregated by an attention mechanism layer, and the output is summed up with the network vector by another attention mechanism layer, with details shown in Fig. 1(b).

4.3 Experimental Results and Analysis

We report results of our proposed model on user profiling tasks along with baseline methods in Table 2.

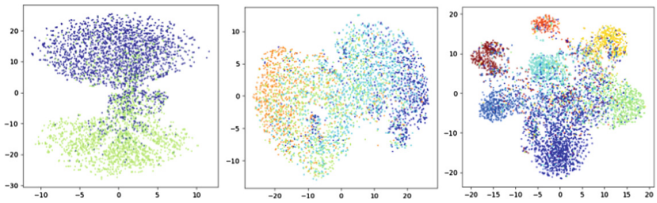
Table 2. Results of user profiling tasks on Sina micro-blogs data set

	Gender	Age	Location
SVM	76.5	43.8	48.7
HAN	81.8	48.7	61.7
MCNE	84.2	46.8	65.6
MDC	85.5	51.5	68.1
MDA	85.5	52.7	67.5
Our model	**86.6**	**56.3**	**70.2**

We can see that deep learning methods outperform remarkably better than BOW. MCNE is superior to HAN in gender and location prediction significantly, but slightly inferior in age prediction. The possible reason is that user gender and location can be inferred by local salient features, like "homeboy", "Beijing", while age requires more information of different aspects and their relations. In this respect, RNN performs better than CNN. In general, methods that unifies multi-data sources boost the performance compared to merely text-messages based methods (SVM, HAN) and the MCNE method. However, MDA does not make a significant improvement compared to the baseline MDC. The reason may be that aggregating different inputs by an attention mechanism may loss the peculiarity of each data source and information is partly lost when features are shortened to a quarter of the original inputs by the summation operation. Among all the methods, our model consistently and significantly performs the best in all user profiling tasks, especially the age identification task with 4.8% improvements of performance compared with MDC. This indicates our model effectively takes advantage of the relativity and diversity of different data sources.

Visualization: To further illustrate our proposed model, we use t-SNE [30] to make a 2-dimensional visualization of user embedding vectors in the test dataset. As shown in Fig. 7, we can observe a clear division of node distributions in gender, location tasks with 2 and 7 distinct regions respectively, where users with the same attributes are clustered tightly. Although age prediction has the lowest accuracy, we still observe two separated parts, in which the left denotes users who are above 24 years old and the right under 24. Besides, we can see a gradual change of colors from right to left (dark blue, blue, green and orange), which suggests the transition of different age groups (<19, 19–23, 24–27, 27 correspondingly). Thus, results of visualization give a strong evidence of the good performance of our model.

Sequential Order Analysis: To provide more insight into our proposed fusion model, we investigate the influence of sequence order before the BiRNN-layer. Specifically, let t denote the text messages representation, i denote the user metadata representation, ne denote the network representation, and f denote the followee information. Table 3 reports results on four types of different combinations. It reveals



(a)gender (b)age (c)location

Fig. 7. 2-dimensional visualization of user embeddings in user profiling tasks

the significant differences among different orders in the sequence. The reason is mainly because of the effect of the update state and the reset state in GRU, which will reduce the information from the previous hidden state if it has little relevance to the current state. Herein, we observe that the order *ne-u-t-f* obtains the best performance, partly because the middle position of text messages can effectively capture both the information from user metadata and followee information from bi-directions since they all are the text representations and show a strong relevance intuitively. Besides, text messages preserve the most valuable information for inferring user attributes. A practical issue is how to let our model find the sequence order automatically, which we leave for our future work.

Table 3. Results of fusion framework with different orders of sequential combination.

Order	Gender	Age	Location
t_f_u_ne	85.68	55.03	68.64
t_u_f_ne	86.29	54.89	65.91
f_ne_t_u	86.24	53.46	68.59
ne_u_t_f	**86.69**	**56.38**	**70.24**

Network Analysis: To verify the effect of celebrity ties, we construct five different social networks by adding *friend nodes*, *celebrity nodes from followee lists*, *celebrity nodes from @mention behaviors*, *celebrity nodes from repost behaviors* and *celebrity nodes from vote behaviors* incrementally. We obtain the network embeddings by using LINE and feed them into the MLP classifier. Figure 8 shows the results on different network scales, from which we are apparently aware of the remarkable promotion by adding *celebrity nodes from followee lists*, with a drastic increase of 10% in F1-measure in gender and age prediction. The possible reason is that celebrity nodes boost the connectivity of different users who are not friends, and thus enrich the network structure. In addition, we also observe a significant effect of *repost* and *vote* behaviors on location predictions, probably due to the fact that users usually repost and vote the micro-blogs concerning local reports.

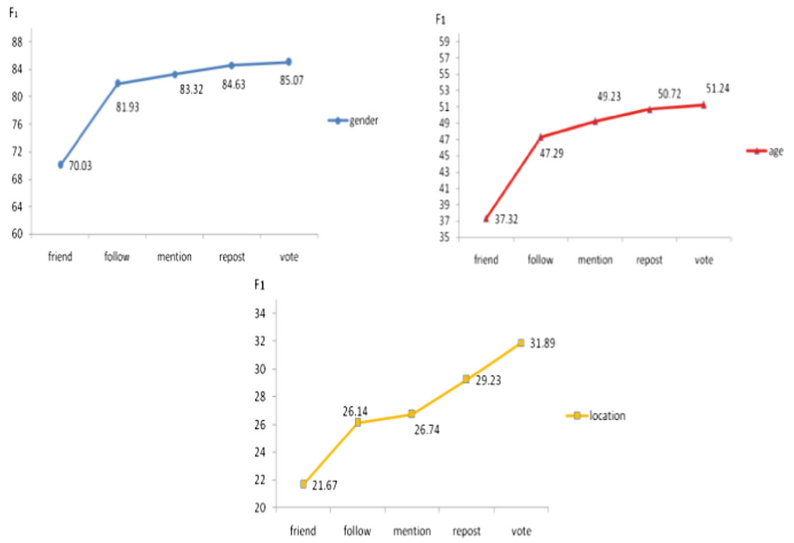


Fig. 8. Comparison on the performances of user profiling with different network scales

5 Conclusion

In this paper, we present a novel fusion model of multi-data sources for user profiling, which seamlessly integrates them by taking advantage of their relativity and diversity. Concretely, we carefully devise four different types of data sources: text messages, user metadata, followee information and network presentations, feeding them into different sub-models and integrating them via a hybrid bi-GRU framework. To alleviate the problem of weak connectivity in user-friendship based network, we innovatively incorporate celebrity nodes, noticing the indirect interactions between users via celebrity nodes. Experimental results show that our proposed model performs effectively on user profiling tasks. In the future, we will do further research on our model to implement an automatic mechanism for finding the best sequence combination of multi-data sources. In addition, we also plan to incorporate other different data sources, like images and videos that appear in user micro-blogs.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (No. 61572145) and the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (No. 2017KZDXM031). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

1. Lu, Z., Pan, S.J., Li, Y., Jiang, J., Yang, Q.: Collaborative evolution for user profiling in recommender systems. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 3804–3810 (2016)
2. Zhou, M.: Gender difference in web search perceptions and behavior: does it vary by task performance? Comput. Educ. **78**(259), 174–184 (2014)
3. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: political ideology prediction of twitter users. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 729–740 (2017)
4. Zhang, D., Yin, J., Zhu, X., Zhang, C.: User profile preserving social network embedding. In: Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 3378–3384 (2017)
5. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on Twitter. In: Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309 (2011)
6. Chen, J., Li, S., Dai, B., Zhou, G.: Active learning for age regression in social media. In: China National Conference on Chinese Computational Linguistics, pp. 351–362 (2016)
7. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., Baldridge, J.: Supervised text-based geolocation using language models on an adaptive grid. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1500–1510 (2012)
8. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through Twitter content. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 1754–1764 (2015)
9. Kim, H.R., Chan, P.K.: Learning implicit user interest hierarchy for context in personalization. Appl. Intell. **28**(2), 153–166 (2008)
10. Majumder, N., Poria, S., Gelbukh, A., Cambria, E.: Deep learning-based document modeling for personality detection from text. IEEE Intell. Syst. **32**(2), 74–79 (2017)
11. Lampos, V., Aletras, N.: Predicting and characterising user impact on Twitter. In: Conference of the European Chapter of the Association for Computational Linguistics, pp. 405–413 (2014)
12. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on Blogging. In: Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp. 199–205 (2006)
13. Ciot, M., Sonderegger, M., Ruths, D.: Gender inference of Twitter users in non-english contexts. In: Conference on Empirical Methods in Natural Language Processing, pp. 1136–1145 (2013)
14. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: Conference on Empirical Methods in Natural Language Processing, pp. 158–166 (2010)
15. Marquardt, J., et al.: Age and gender identification in social media. In: Proceedings of CLEF 2014 Evaluation Labs, pp. 1129–1136 (2014)
16. Mislove, A., Viswanath, B., Gummadi, K., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Third ACM International Conference on Web Search and Data Mining, pp. 251–260 (2010)
17. Han, X., Wang, L., Crespi, N., Park, S., Cuevas, Á.: Alike people, alike interests? inferring interest similarity in online social networks. Decision Support Systems **69(C)**, 92–106 (2015)

18. Miura, Y., Taniguchi, M., Taniguchi, T., Ohkuma, T.: Unifying text, metadata, and user network representations with a neural network for geolocation prediction. In: Meeting of the Association for Computational Linguistics, pp. 1260–1272 (2017)
19. Wang, J., Li, S., Zhou, G.: Joint learning on relevant user attributes in micro-blog. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 4130–4136 (2017)
20. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 855–864 (2016)
21. Perozzi, B., Al-Rfou, R., Skiena, S.: DeepWalk: online learning of social representations bryan. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710 (2014)
22. Tang, J., Qu, M.: LINE: large-scale information network embedding categories and subject descriptors. In: International World Wide Web Conferences Steering Committee, pp. 1067–1077 (2015)
23. Yang, C., Liu, Z., Zhao, D., Sun, M., Chang, E.Y.: Network representation learning with rich text information. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 2111–2117 (2015)
24. Zhao, Z., Du, J., Gao, Q., Gui, L., Xu, R.: Inferring user profile using microblog content and friendship network. In: Communications in Computer and Information Science, pp. 29–39 (2017)
25. Han, B., Cook, P., Baldwin, T.: A stacking-based approach to twitter user geolocation prediction. In: Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 7–12 (2013)
26. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2016)
27. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Computer Science (2014)
28. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations, pp. 1–12 (2013)
29. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. Computer Science (2014)
30. Van Der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9** (2605), 2579–2605 (2008)