MMLUP: Multi-Source & Multi-Task Learning for User Profiles in Social Network

Dongjie Zhu[1], Yuhua Wang [1], Chuiju You[2, *], Jinming Qiu[2, 3], Ning Cao[2], Chenjing Gong[4], Guohua Yang[5] and Helen Min Zhou[6]

Abstract: With the rapid development of the mobile Internet, users generate massive data in different forms in social network every day, and different characteristics of users are reflected by these social media data. How to integrate multiple heterogeneous information and establish user profiles from multiple perspectives plays an important role in providing personalized services, marketing, and recommendation systems. In this paper, we propose Multi-source & Multi-task Learning for User Profiles in Social Network which integrates multiple social data sources and contains a multi-task learning framework to simultaneously predict various attributes of a user. Firstly, we design their own feature extraction models for multiple heterogeneous data sources. Secondly, we design a shared layer to fuse multiple heterogeneous data sources as general shared representation for multi-task learning. Thirdly, we design each task's own unique presentation layer for discriminant output of specific-task. Finally, we design a weighted loss function to improve the learning efficiency and prediction accuracy of each task. Our experimental results on more than 5000 Sina Weibo users demonstrate that our approach outperforms state-of-the-art baselines for inferring gender, age and region of social media users.

Keywords: User profiles, multi-source, multi-task learning, social network.

1 Introduction

With the rapid development of the mobile Internet and the increasing network speed, accessing and producing information anytime anywhere has gradually become an indispensable lifestyle. In social network, users generate massive data in different forms every day. It contains all aspects of life and reflect the different characteristics of users. Through the massive data to establish user profiles plays an important role in providing personalized services, marketing, referral systems, and customized advertising. So how to

[1] School of Computer Science and Technology, Harbin Institute of Technology, Weihai, 264209, China.

[2] College of Information Engineering, Sanming University, 365004, Sanming, China.

[3] Fujian Province University Key Lab for Industry Big Data Analysis and Application, Fujian, China.

[4] College of Information Engineering, Qingdao Binhai University, Qingdao, China.

[5] Jiangsu Province Wireless Sensing System Application Engneering Technology Research and Development Centre, China.

[6] School of Engineering, Manukau Institute of Technology, Auckland, 2241, New Zealand.

[*] Corresponding Author: Chuiju You. Email: youchuiju@126.com.

CMC. doi:10.32604/cmc.2019.06041

mine useful information from these different forms of massive data is challenging.

Various user profiles analysis methods have been proposed [Ciot, Sonderegger and Ruths (2013); Rothe, Timofte and Van Gool (2015); Ruder, Ghaffari and Breslin (2016); Lim and Datta (2012)]. According to the type of data source they use, it can be divided into two main categories. First type of research method focuses on how to explicitly infer user profiles by analyzing user-generated text or image data. Second type of research method focuses on how to establish a reasonable user relationship by analyzing the user's social relationship, and then build a user profile model. Only one type of data source is used in these researches. But in social media platforms, users generate content in different forms, so models based on one source of information cannot generate accurate user profiles. In addition, in order to establish user profile comprehensively, the user should be described from multiple attributes. Nguyen et al. [Nguyen, Trieschnigg, Dogruoz et al. (2014)] independently processed each attribute prediction task, which ignored the influence between attributes and was susceptible to different feature sets.

In this paper, we propose a user profiles model called “Multi-source & Multi-task Learning for User Profiles in Social Network” (MMLUP). MMLUP fuses multiple social data sources for analysis and contains a multi-task learning framework that uses the dependencies between tasks. We made the following contributions. Firstly, we design their own feature extraction models for multiple heterogeneous data sources. We use two forms of data source to analysis: for the text content in social network, we propose an improved Convolutional Neural Networks for Sentence Classification (Text-CNN) [Kim (2014)] model to learn text features by integrating the Attention [Yang, Yang, Dyer et al. (2016)] mechanism to assign different weights to different text information; for friendship network, we use the network embedding method to extract relationship features by randomly traversing in the social graph. Secondly, we design a shared layer to fuse these two forms of data source as a general representation for multi-task learning. Thirdly, we design each task's own unique presentation layer for discriminant output of specific-task. Finally, we design a weighted loss function to ensure that different tasks achieve the best performance at the same time and improve overall learning efficiency and prediction accuracy.

2 Related work

Various methods of user profiling based on user-generated content have been proposed, such as predicting the user’s gender [Ciot, Sonderegger and Ruths (2013); Nguyen, Trieschnigg, Dogruoz et al. (2014)], age [Rothe, Timofte and Van Gool (2015); Nguyen, Trieschnigg, Dogruoz et al. (2014)], interests [Lim and Datta (2012)], occupations [Tu, Liu and Sun (2015)], etc. Although much progress has been made in user attribute inference, most of the work used only one type of data source from text, images or relationships to infer user attributes. But in social media platforms, users generate content in different forms which reflects user characteristics from different perspectives. Therefore, the use of one single source of information is not sufficient to infer the user’s attributes accurately.

Rui et al. [Wang, Shen, Li et al. (2018)] proposed that using feature fusion approach can achieve higher classification accuracy. Recently, great interest has been generated by constructing heterogeneous information sources. Based on the neural network, Li et al. [Li, Ritter and Jurafsky (2015)] proposed a representation learning method that combines users' Twitter texts, personal data and social relationships to improve the accuracy of user gender, occupation, geography and attention trends. Wei et al. [Wei, Zhang, Yuan et al. (2017)] proposed a heterogeneous information ensemble framework to predict users' personality traits by integrating heterogeneous information including self-language usage, avatar, emoticon, and responsive patterns. Chen et al. [Chen, Wang, Ren et al. (2018)] integrated users' profile vector, social large-scale information network embedding vector and user's tag encoding vector as user features, then applied a deep learning approach to classify users. They train a single model or an ensemble of models to perform their desired task respectively. However, a model predicts each attribute separately leads to relatively low learning efficiency. In our opinion, there must be some kind of connection between the various attributes which can be used to predict simultaneously to improve learning efficiency and accuracy.

Ruder [Ruder (2017)] defined Multi-Task Learning (MTL) and pointed out that people usually get an acceptable performance only for one task, but may ignore some information, which can help to do better on the evaluation criteria. Specifically, this information is the monitoring data of related tasks. By sharing presentation information among related tasks, the model has better generalization performance on original tasks. This method called Multi-Task Learning. MTL has been used successfully across all applications of machine learning, from natural language processing [Collobert and Weston (2008)] and speech recognition [Deng, Hinton and Kingsbury (2013)] to computer vision [Girshick (2015)].

For the problems of the above single source and single task models, we de-signed a multi-source and multi-task learning model, which combines user-generated text information and friendship network as user feature representation, to predict user's gender, age, and region simultaneously. The model achieves the purpose of learning together for each task, alleviates the burden of feature learning and improves the accuracy of user profiles according to the correlation characteristics between tasks.

3 MMLUP architecture

MMLUP consists of four layers: Multi-Source layer, Feature engineering layer, Shared layer, and Task-Specific layer. The model architecture is shown in Fig. 1. We use user-generated text content and friendship network to infer user profiles. For text content in social network, we designed a deep Text-CNN with Attention model to learn text features. For friendship network in social network, we use the network embedding method of large network structure to extract relationship features. Through the learning of neural networks, the heterogeneous information is represented as representations of low-dimensional vectors. At the same time, we design a shared layer fuses two forms of information vectors to reduce the burden of learning the features of each task. And according to the correlation characteristics between the tasks, the parameters are shared to affect each other to achieve the purpose of mutual learning. Then, we adjust the connection layers and loss function of each task to form independent task-specific representation. Finally, we design a weighted loss function to adjust the learning efficiency and prediction accuracy of each task. The details of each layer are as follows.

3.1 Multi-source layer

The multi-source layer is used to organize multiple data sources of users and preprocess the multi-source information. We use the text content generated by the user as Data Source1 and the friendship network as Data Source2. For Data Scource1, we use the microblog texts of all users as a corpus and train word vector. At the same time, we extract regional keywords in the text content and construct a regional dictionary as a feature of common learning. For Data Source2, we treat the user's friendship network as a form of node pair in the undirected graph.

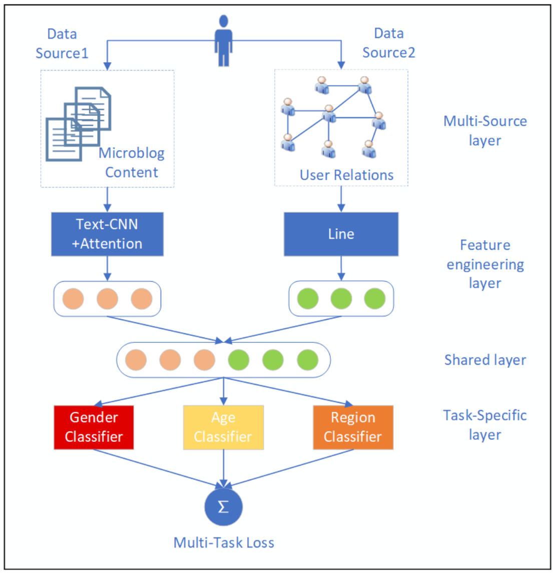

Figure 1: MMLUP architecture

3.2 Feature engineering layer

In the feature engineering layer, we extract feature of multiple sources by training the respective learning models of the heterogeneous information.

3.2.1 For data source1

We designed a deep Text-CNN with Attention model shown in Fig. 2 to extract text features using an unsupervised method.

We represent the user's $T\text{-}th$ message as Eq. (1):

$$M_T = X_1 \oplus X_2 \oplus \cdots \oplus X_n \tag{1}$$

where $X_i \in R^k$ is the k-dimensional word vector corresponding to the $i\text{-}th$ word in each message, and $\oplus$ is the concatenation operator.

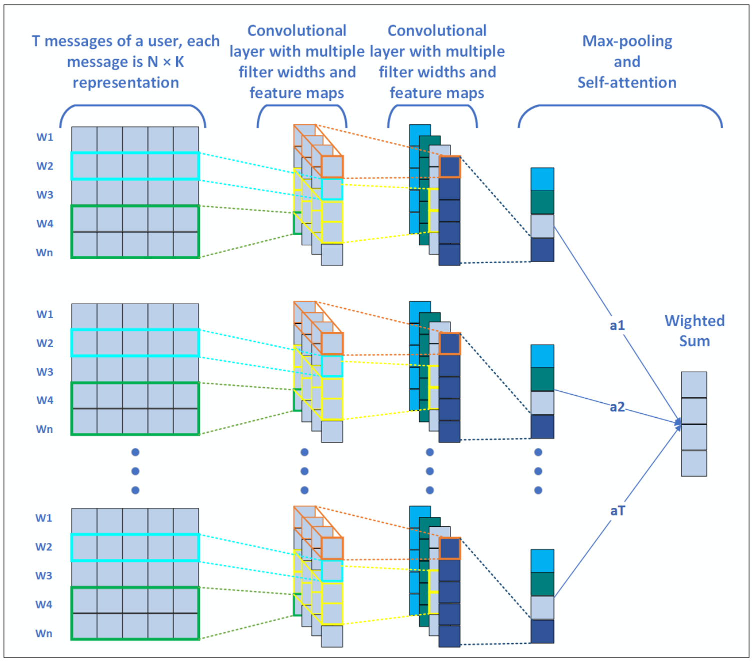
T messages of a user, each message is N × K representation
Convolutional layer with multiple filter widths and feature maps
Convolutional layer with multiple filter widths and feature maps
Max-pooling and Self-attention
W1
W2
W3
W4
Wn
a1
a2
aT
Wighted Sum

Figure 2: Deep Text-CNN with Attention model architecture

We use n′ different filter widths and m feature maps to form the first layer of convolution layer according to Eq. (2).

$$S_i = f(W \cdot X_{i:i+h-1} + b) \tag{2}$$

The features S_i learned by the different filter widths are spliced to form a vector $c \in R^{\mathrm{n'm \cdot k}}$. To learn more about the semantic relationship between n-grams in text, we designed a second-level convolution with c as input, see Eq. (3).

$$\hat{c}_j = f(W \cdot c_{j:j+h-1} + b) \tag{3}$$

Then using the max-pooling operation to capture the most important features in the text. The T-th message which output from the max-pooling layer is represented as vector h_T. In the messages published by the user, the user's different life experiences and attitudes are reflected. Each message has different degrees of importance for the determination of user attributes. We want to give each piece of information different attention, so we introduce the Attention mechanism to give each message a different weight, see Eq. (4) and Eq. (5).

$$\alpha_T = (V^T \tanh(W \cdot h_T + b)) \tag{4}$$

$$m = \sum_T \alpha_T h_T \tag{5}$$

where α_T is the weight of the user's T-th message, and we sum the weighted vector learned from the user's all messages to get the final user text feature representation m.

3.2.2 For data source2

We treat each user in the social network as a vertex, and the following relation-ship between users is considered an edge. In our opinion, people with the same attributes are more likely to pay attention to each other. The friendship between users has an important impact on the judgment of user attributes.

The social network is defined as $G = (V, E)$, where V is a set of points composed by users, and E is a set of edges formed by the following relationship between users. We define the first-order proximity in the social network as the local similarity between the two users. If there is a direct following relationship between the two users, the edge weight is 1, otherwise it is 0. The first-order proximity usually implies the similarity of two users in the real network, but if the similar two users do not directly follow, the first-order proximity is regarded as 0, so the first-order proximity is not enough to preserve the real network structure. Further, we define the second-order proximity in the social network as the similarity of the two-user neighbor network structure. For example, if two users do not have a direct following relationship but have some common friends, then the two users are also very similar.

In this study, based on the idea of Line [Tang, Qu, Wang et al. (2015)], we learn the implicit information in social network, and represent each user node V in low-dimensional space R^d ,while retaining the first-order and second-order proximity to accurately represent the real-world social network structure. Finally, the user vectors we learned in social network relationships are represented as $g \in R^d$.

3.3 Shared layer

The goal of building a multi-source data model is to integrate multiple heterogeneous data sources, represented as a single unified representation, and to provide a more complete description than a single data source. We consider the correlation between different data sources, learn different information sources through their respective models and encode them into vectors. Then connect the vectors as the final user features $c(u)$, see Eq. (6).

$$c(u) = m_u \oplus g_u \tag{6}$$

where m_u and g_u respectively represent the text feature representation and social network representation of the user u, and $\oplus$ is the concatenation operator.

The data source is integrated with nonlinear functions to enhance the learning process. Through joint learning, one data source is supplemented with additional information from another data source, and different sources of information are combined to obtain better decision representations.

At the same time, the shared layer is used as the common representation of tasks in the lower layer, which reduces the burden of feature learning. In the process of back propagation learning, the parameters learned by the shared layer are affected together, and the user characteristics are more abundant. The shared layer not only integrates different information sources, but also contacts the implicit relationship between tasks, achieving the goal of mutual benefit and win.

3.4 Task-specific layer

Each task has its own unique feature. At the task-specific layer, we design their own connection layer and loss function for each task. For each classification we define as Eq. (7), Eq. (8) and Eq. (9).

$$y_g = softmax(W_g U_g + b_g) \tag{7}$$

$$y_a = softmax(W_a U_a + b_a) \tag{8}$$

$$y_r = softmax(W_r U_r + b_r) \tag{9}$$

where y_g, y_a, y_r represent the gender, age, and region probability distribution of the output, W is the specific weight matrix of each task to learn, b is the bias, and U is the output vector of the shared layer vector passing through the full connection layer of each task.

In order to train MMLUP, each task is a classification task, we use the cross-entropy loss function on each task, see Eq. (10).

$$J = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{n_c}(\hat{y}_i^j log y_i^j) \tag{10}$$

where N represents the size of the training sample, n_c represents the number of classes, $\hat{y}_i^j$ represents the true probability distribution of the $i\text{-}th$ instance belonging to class j, and y_i^j represents the predicted probability distribution of the $i\text{-}th$ instance belonging to class j.J_g, J_a, and J_r represent the loss function of each task of gender, age, and region respectively.

The total loss function is a linear combination of the above three task loss functions, see Eq. (11).

$$J_{total} = C_g J_g + C_a J_a + C_r J_r \tag{11}$$

where C is a hyperparameter, used to adjust the loss function of the three tasks to ensure that the three tasks achieve the best performance at the same time.

4 Experiments

4.1 Datasets

This paper uses the dataset of the SMP CUP 2016 competition[1]. This competition provides Sina Weibo data (including user personal information, user microblog text and user fan list) to infer user profiles, including the following three tasks:

Task 1: Infer the gender of the user (2 labels in total: m/f)

Task 2: Infer the age of the user (3 labels in total: -1979/1980-1989 / 1990+)

Task 3: Infer the region of the user (8 labels in total: Northeast China/North China/Central China/East China/Northwest China/ Southwest China/ South China/Overseas).

The detailed description of the dataset is shown in Tab. 1. The training data set, valid data

[1] https://biendata.com/competition/smpcup2016/

set and test data set all contain four files: user information file ‘info.txt’, each line represents a user, each user contains three attributes and is separated by ‘||’; user label file ‘labels.txt’, each line represents a user, each user contains four attributes and is separated by ‘||’; user relationship file ‘links.txt’, each line represents a user's fan list, consisting of multiple user ids separated by spaces, and from the second user to the last user is a fan of the first user; microblog text file ‘sta-tus.txt’, each line represents a user microblog, consisting of 6 attributes and separated by commas. A detailed description of the raw data format in each file is shown in Tab. 2.

Table 1: Detailed description of the dataset

Dataset	User number	Friendship net-work edges	Users with content	Users with label
Training Data	2,565,000	550,000,000	44,000	3,200
Valid Data	1,000	150,000	1,267	1,267
Test Data	1.000	130,000	944	944

Table 2: Detailed description of the raw data format in each file

File	Format	Sample
info.txt	uid || screen_name || avatar_large	1053604635||CagePaPa||http://tp4.sinaimg.cn/1053604635/180/5615393189/1
labels.txt	uid || gender || birthday || location	1053604635||f||1972||江苏 南京
links.txt	uid fan1 fan2 fan3...	1053604635 1998321847 24969
status.txt	uid, retweet count, review count, source, time, content	1053604635,0,0,小米手机 2S,2014-09-07 21:12:16,忙活了一晚上的成果，当过节了！

4.2 Experimental setup

For the microblog text, we first remove the duplicate microblog from each user, and then use the Chinese text segmentation Jieba[1] to preprocess the dataset. We use glove[2] training word vector, the dimension of the word vector is set to 100 dimensions, the vocab minimum count is 5, and the window size is 8. For the user's friendship network, we retain the user's first-order proximity and second-order proximity. Both the first-order proximity and the second-order proximity are set to 150 dimensions and then merged into a 300-dimensional embedded vector.

In this experiment, we use Adam algorithm to calculate adaptive learning rate and use RELU activation function. Dropout is set to 0.5, learning rate is initially 1e-3, batch size is 50, CNN filter window size is set to 1, 3, 5, 7, 9, filters number is 64 and the number of hidden layers for self-attention is 64. Each specific task layer has fully connected layer

[1] https://pypi.org/project/jieba/

[2] https://nlp.stanford.edu/projects/glove/

with units of 128, 64, and 32. C_g, C_a, C_r are respectively set to 1, 3, 1.

4.3 Results

The evaluation for SMP CUP 2016 competition is calculating the accuracy of the three tasks A_1, A_2 and A_3 at first. Then according to the difficulty of the task, the weights of Task 1, Task 2 and Task 3 are set to 0.2, 0.3 and 0.5 respectively, and the final weighted average accuracy $A = 0.2 * A_1 + 0.3 * A_2 + 0.5 * A_3$. The weighted average accuracy A is the key basis for the evaluation ranking. This paper also uses such evaluation method.

We constructed three independent Text-CNN models with the same network structure as our proposed MMLUP model as baselines, but they do not share any parameters. We compare MMLUP with a single-task model, a single-source mod-el and a model with self-attention removed. The results from Tab. 3 show that our model performs very well and exceeds all models. In addition, 565 players entered the competition and a total of 197 teams were formed in the SMP CUP 2016 competition. We compared the average accuracy of the top Seven in the final score ranking list who won the competition. As can be seen from Tab. 4, the accuracy of our model is basically the same as them, and even exceeds the third place. Moreover, compared with their models, MMLUP has the advantage of not using artificial explicit extraction features on the basis of improving the ac-curacy of user profiles.

Table 3: The results of MMLUP and baseline method comparison

Models	Gender (Accuracy)	Age (Accuracy)	Region (Accuracy)	Average accuracy
Text-CNN (single task)	0.846	0.599	0.275	0.4864
MMLUP (single task)	0.855	0.642	**0.700**	0.7136
MMLUP (without friendship network)	0.827	0.616	0.617	0.6587
MMLUP (without attention)	0.798	0.593	0.279	0.4772
MMLUP (full architecture)	**0.867**	**0.650**	0.694	**0.7154**

Table 4: The results of MMLUP and top Seven in the final score ranking list of the SMP CUP 2016 comparison

Models	Average accuracy
HLT_HITSZ	0.73516
DUTIR_TONE	0.73129
卢泓宇（Luhongyu）	0.70540
Neptune	0.69992
Cappuccino	0.69532
tobe1	0.68831
clclc	0.68831
MMLUP (our model)	**0.71540**

5 Conclusion and future work

In this paper, we have proposed a new method MMMLU to infer user profiles. MMLUP combines multiple heterogeneous data sources and predicts multiple tasks simultaneously. It infers user profiles on the dataset of SMP CUP 2016 competition, and achieves the best performance compared with the baseline method in predicting three subtasks: gender, age and region. It proves that multi-source and multi-task learning can effectively improve the accuracy of user pro-files. In future research, we will integrate more data sources, such as pictures and videos, to predict more attributes of users.

Acknowledgement: This work is supported by State Grid Science and Technology Project under Grant No. 520613180002, 62061318C002, the Fundamental Research Funds for the Central Universities (Grant No. HIT.NSRIF.201714), Weihai Science and Technology Development Program (2016DXGJMS15) and Key Research and Development Program in Shandong Provincial (2017GGX90103), Sanming Science and Technology Project. Grant No. 2015-G-6, Shandong province vocational education educational reform research project. Grant No. 2017209, Study and Development of Smart Agriculture Control System Based on Spark Big Data Decision (2017N0029), Jiangsu Province industrial Communication Technology Application Technology Innovation Team Project.

References

Chen, X.; Wang, J.; Ren, Y.; Liu, T.; Lin, H. (2018): NLPCC 2018 shared task user profiling and recommendation method summary by DUTIR_9148. *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 420-428.

Ciot, M.; Sonderegger, M.; Ruths, D. (2013): Gender inference of Twitter users in non-English contexts. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1136-1145.

Collobert, R.; Weston, J. (2008): A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th International Conference on Machine Learning*, pp. 160-167.

Deng, L.; Hinton, G.; Kingsbury, B. (2013): New types of deep neural network learning for speech recognition and related applications: an overview. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8599-8603.

Girshick, R. (2015): Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440-1448.

Kim, Y. (2014): Convolutional neural networks for sentence classification. https://www.aclweb.org/anthology/D14-1181.pdf.

Li, J.; Ritter, A.; Jurafsky, D. (2015): Learning multi-faceted representations of individuals from heterogeneous evidence using neural networks. arXiv:1510.05198.

Lim, K. H.; Datta, A. (2012): Finding twitter communities with common interests using following links of celebrities. *Proceedings of the 3rd International Workshop on Modeling Social Media*, pp. 25-32.

Nguyen, D.; Trieschnigg, D.; Dogruoz, A. S.; Gravel, R.; Theune, M. et al. (2014): Why gender and age prediction from tweets is hard: lessons from a crowdsourcing experiment. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1950-1961.

Rothe, R.; Timofte, R.; Van Gool, L. (2015): Dex: deep expectation of apparent age from a single image. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 10-15.

Ruder, S.; Ghaffari, P.; Breslin, J. G. (2016): Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. arXiv:1609.06686.

Ruder, S. (2017): An overview of multi-task learning in deep neural networks. arXiv:1706.05098.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J. et al. (2015): Line: large-scale information network embedding. *Proceedings of the 24th International Conference on World Wide Web*, pp. 1067-1077.

Tu, C.; Liu, Z.; Sun, M. (2015): Prism: profession identification in social media with personal in-formation and community structure. *Chinese National Conference on Social Media Processing*, pp. 15-27.

Wang, R.; Shen, M.; Li, Y.; Gomes, S. (2018): Multi-task joint sparse representation classification based on fisher discrimination dictionary learning. *Computers, Materials & Continua*, vol. 57, no. 1, pp. 25-48.

Wei, H.; Zhang, F.; Yuan, N. J.; Cao, C.; Fu, H. et al. (2017): Beyond the words: Predicting user personality from heterogeneous information. *Proceedings of the Tenth ACM international Conference on Web Search and Data Mining*, pp. 305-314.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A. et al. (2016): Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the As-sociation for Computational Linguistics: Human Language Technologies*, pp. 1480-1489.