Identifying User Profile by Incorporating Self-Attention Mechanism Based on CSDN Data Set (Postprint)

Authors: Lu, Junru, Chen, Le, Meng, Kongming, Wang, Fengyi, Xiang, Jun, Chen, Nuo, Han, Xu, Li, Binyang, Li, Binyang

Date: 2022-11-27T00:00:00+00:00

Abstract

With the popularity of social media, there has been an increasing interest in user profiling and its applications nowadays. This paper presents our system named UIR-SIST for User Profiling Technology Evaluation Campaign in SMP CUP 2017. UIR-SIST aims to complete three tasks, including keywords extraction from blogs, user interests labeling and user growth value prediction. To this end, we first extract keywords from a user' s blog, including the blog itself, blogs on the same topic and other blogs published by the same user. Then a unified neural network model is constructed based on a convolutional neural network (CNN) for user interests tagging. Finally, we adopt a stacking model for predicting user growth value. We eventually receive the sixth place with evaluation scores of 0.563, 0.378 and 0.751 on the three tasks, respectively.

Full Text

Preamble

Identifying User Profile by Incorporating Self-Attention Mechanism Based on CSDN Data Set

Junru Lu[1], Le Chen[1], Kongming Meng[1],[2], Fengyi Wang[3], Jun Xiang[1], Nuo Chen[1], Xu Han[4] & Binyang Li[1]†

[1]School of Information Science and Technology, University of International Relations, Beijing 100091, China
[2]Deep Brain Co. Ltd., Shanghai 201200, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]College of Information Engineering, Capital Normal University, Beijing 100048, China

Keywords: User profile; Convolutional neural network (CNN); Self-attention; Keyword extraction

Citation: J. Lu, L. Chen, K. Meng, F. Wang, J. Xiang, N. Chen, X. Han & B. Li. Identifying user profile by incorporating self-attention mechanism based on CSDN data set. *Data Intelligence* 1(2019), 160-175. doi: 10.1162/dint_a_{00009}

Received: August 27, 2018; **Revised:** November 30, 2018; **Accepted:** December 6, 2018

Abstract

With the popularity of social media, user profiling and its applications have attracted increasing attention in recent years. This paper presents our system, named UIR-SIST, developed for the User Profiling Technology Evaluation Campaign in SMP CUP 2017. UIR-SIST aims to complete three tasks: keyword extraction from blogs, user interest labeling, and user growth value prediction. To achieve these objectives, we first extract keywords from a user' s blogs, considering the blog itself, blogs on the same topic, and other blogs published by the same user. We then construct a unified neural network model based on convolutional neural networks (CNN) for user interest tagging. Finally, we adopt a stacking model for predicting user growth value. Our system achieved sixth place with evaluation scores of 0.563, 0.378, and 0.751 on the three tasks, respectively.

1. Introduction

Social media platforms have recently become important venues that enable users to communicate and disseminate information. User-generated content (UGC) has been utilized for a wide range of applications, including user profiling. The Chinese Software Developer Network (CSDN) is one of China' s largest platforms where software developers share technical information and engineering experiences. Analyzing UGC on CSDN can reveal users' interests in the software development process, such as their past interests and current focus, even when user profiles are incomplete or missing. Beyond UGC, user behavior data also contains valuable information for profiling, including actions like "following," "replying," and "sending private messages," which help construct friendship networks to infer user attributes such as gender [1,2,3], age [4], political polarity [5, 6], or profession [7].

In SMP CUP 2017 [8], the competition was structured around three tasks based on CSDN blogs[1]: (1) keyword extraction from blogs, (2) user interest labeling, and (3) user growth value prediction. Our team from the School of Information Science and Technology at the University of International Relations participated in all tasks of the User Profiling Technology Evaluation Campaign. This paper describes the framework of our UIR-SIST system for the competition. We ex-

tract keywords from a user' s blogs by considering the blog itself, blogs on the same topic, and other blogs published by the same user. We then construct a unified neural network model with a self-attention mechanism for Task 2, based on multi-scale convolutional neural networks designed to capture both local and global information for user profiling. Finally, we adopt a stacking model for predicting user growth value. According to SMP CUP 2017' s metrics, our model achieved scores of 0.563, 0.378, and 0.751 on the three tasks, respectively.

This paper is organized as follows. Section 2 introduces the User Profiling Technology Evaluation Campaign in detail. Section 3 describes our system' s framework. We present the evaluation results in Section 4 and conclude with future work in Section 5.

2.1 Data Set

The dataset used in SMP CUP 2017 was provided by CSDN, one of China' s largest information technology communities. The CSDN dataset consists of all user-generated content and behavior data from 157,427 users during 2015, which can be divided into three parts: (1) 1,000,000 user blogs, including blog ID, title, and corresponding content; (2) six types of user behavior data, including posting, browsing, commenting, voting up, voting down, and adding favorites, along with corresponding date and time information; and (3) user relationships, which refer to records of following and sending private messages.

Further details about the size and type of the CSDN dataset are shown in Table 1 . Table 2 illustrates an example from the dataset.

2.2 Tasks

Task 1: Extract three keywords from each document that can effectively represent the topic or main idea of the document.

Task 2: Generate three labels to describe a user' s interests, selected from a given candidate set (42 labels total).

Task 3: Predict each user' s growth value for the next six months based on their behavior over the past year, including text content, relationships, and interactions with other users. The growth value must be scaled to [0, 1], where 0 indicates user drop-out.

2.3 Metrics

To assess system effectiveness for the above tasks, the following evaluation metrics were designed for each task.

Score1 calculates the overlapping ratio between extracted keywords and standard answers, computed using Equation (1):

$$Score_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{|K_i \cap K_i^*|}{|K_i|}$$

where N is the size of the validation or test set, Ki is the extracted keyword set from document i, and Ki* is the standard keyword set for document i. Note that |Ki| = 3 and |Ki*| = 5.

Score2 denotes the overlapping ratio between model-generated tags and ground truth tags, expressed by Equation (2):

$$Score_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{|T_i \cap T_i^*|}{|T_i|}$$

where Ti is the automatically generated tag set for user i, and Ti* is the standard tag set for user i, with |Ti| = 3 and |Ti*| = 3.

Score3 is calculated based on the relative error between predicted and actual user growth values, expressed by Equation (3):

$$Score_3 = \begin{cases} 1 - \frac{|v_i - v_i^*|}{v_i^*} & \text{if } v_i^* > 0 \\ 1 & \text{if } v_i^* = 0 \text{ and } v_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

where vi is the predicted growth value for user i, and vi* is the actual growth value.

The overall score is computed by Equation (4):

$$Score = \frac{Score_1 + Score_2 + Score_3}{3}$$

3. System Overview

The overall architecture of UIR-SIST is described in Figure 1 [Figure 1: see original paper]. The system comprises four modules:

1. **Preprocessing Module:** Reads all blogs from the training and test sets, performing word segmentation, part-of-speech (POS) tagging, named entity recognition, and semantic role labeling.
2. **Keyword Extraction Module:** Extracts three keywords to represent a blog' s main idea from three aspects: the blog content itself, other blogs published by the same user, and blogs on the same topic (shown in green).
3. **User Interests Tagging Module:** Constructs a neural network combining user content embedding with keyword and user tag embeddings for interest tagging (shown in red).

4. **User Growth Value Prediction Module:** Incorporates user interaction information and behavior features into a supervised learning model for growth value prediction (shown in blue).

3.1 Keywords Extraction

The objective of Task 1 is to extract three keywords from each blog that represent its main idea. We believe the main idea can be captured from three aspects: the blog itself, other blogs by the same user, and blogs on the same topic. Based on this assumption, we adopt three models to capture each aspect and generate candidate keyword sets: tf-idf, TextRank, and LDA, which have proven effective for related tasks. We then extract three keywords from the candidate set using different rules.

We first employ the classic tf-idf term weighting scheme to reflect blog content, ranking keywords by tf-idf scores and selecting the top 100 to form the candidate set. For blogs on the same topic, we adopt the TextRank approach [9] to cluster them while weighting all keywords, selecting the top 300. Additionally, we utilize topic information for keyword extraction. Since Task 2 provides 42 tag categories, we assume these topics are extracted from all blogs. Therefore, we use Latent Dirichlet Allocation (LDA) [10] to extract the top 100 keywords for each category from the 1,000,000 blogs, obtaining inter-topic distribution information for these 4,200 topic keywords.

In summary, we consider three aspects to reflect blog content and obtain three independent candidate keyword sets via tf-idf, TextRank, and LDA models. We then retain only the intersection. In our Task 1 training set, approximately 5,000 keywords were collected after extraction and deduplication.

A drawback of the classic tf-idf model is its assumption that rarer words in the corpus are more important and contribute more to the text' s main idea. However, when analyzing a group of articles that primarily use the same keywords and describe similar concepts, this approach yields many errors. This is why we use tf-idf for individual short blogs while employing TextRank for longer blog collections published by the same user.

To enhance cross-topic analysis capability, we borrow ideas from the 2016 Big Data & Computing Intelligence Contest sponsored by the China Computer Federation (CCF)[2] and improve traditional tf-idf calculations using Equation (5):

$$S\text{-}TFIDF(w) = TFIDF(w) \times \log\left(\frac{42}{C_w}\right)$$

where Cw is the frequency of word w appearing across the 42 categories.

3.2 User Interests Tagging

This task aims to tag user interests with three labels from the 42 provided. We model this with neural networks, as shown in Figure 2 [Figure 2: see original paper]. Each blog is represented by a blog embedding [11] through convolution and max-pooling layers. We obtain a user' s content embedding from the weighted sum of all their blog embeddings, where weights are determined by a self-attention mechanism. Content embedding and keyword embedding are concatenated as user embedding and fed to the output layer.

We construct a CNN model rather than RNN for blog representation to capture more global information for indicating user interests and improve time efficiency. Multi-scale convolutional neural networks [12] have demonstrated outstanding achievements in computer vision [13], and TextCNNs designed by vertically arraying word embeddings have shown high effectiveness for NLP tasks [14].

In our CNN model, we treat a blog as a word sequence x = [x_1, x_2, ···, x], where each word is represented by its embedding vector, producing a feature matrix S. The narrow convolution layer uses a kernel W ^{k×d} of width k, a nonlinear function f, and bias variable b, as described by Equation (6):

$$c_i = f(W \cdot x_{i:i+k-1} + b)$$

where x_{i:j} denotes the concatenation of word vectors from positions i to j. We use multiple kernel sizes to obtain diverse local contextual feature maps, then apply max-over-time pooling [15] to extract the most important features.

This yields a low-dimensional, dense, quantified representation for each blog. We then compute each user' s relevant blogs. Initially, we average blog vectors to obtain content embedding c(u) for user u:

$$c(u) = \frac{1}{T}\sum_{i=1}^{T} h_i$$

where T is the total number of user u' s related blogs.

However, different blog sources imply varying degrees of user interest in different topics. For example, a blog might be original content, a repost, or shared from another platform. Naturally, we should attend to these blogs differently when inferring user interests. Therefore, we introduce a self-attention mechanism that automatically assigns different weights to each blog after training. The user context representation is given by the weighted summation of all blog vectors:

$$a_i = \frac{\exp(e_i)}{\sum_{j=1}^{T}\exp(e_j)} \quad \text{where } e_i = v^T \tanh(W_s s_i + U h_i)$$

$$c(u) = \sum_{i=1}^{T} a_i h_i$$

where a_i is the weight of the i-th blog, s_i is the one-hot source representation vector, v ^n, W ^{n' ×m}, U ^{n' ×n}, s_i ^m, h_i ^n, and m is the number of source platforms.

After obtaining the user context representation, we concatenate the keyword matrix of all blog keywords extracted in Task 1. These final features are fed to an ANN layer that trains user embeddings from the training set and predicts probability distributions over the 42 interest tags for validation and test sets.

3.3 User Growth Value Prediction

Based on Task 3' s description, growth value can be estimated as a measure of user activeness. Our approach incorporates user interaction information and behavior statistical features into a supervised learning model, as illustrated in Figure 3 [Figure 3: see original paper].

We employ a stacking framework [16] to enhance prediction accuracy. After basic behavior statistics analysis, selected original features serve as inputs to the stacking model, which consists of two layers: base and stacking. In the base layer, we select Passive Aggressive Regressor [17] and Gradient Boosting Regressor [18, 19] as basic regressors due to their excellent performance. In the stacking layer, we use a support vector machine model, specifically NuSVR, which can control its error rate. This yields the final user growth value predictions.

3.3.1 Original Feature Selection

Figure 4 [Figure 4: see original paper] shows an example of daily user behavior statistics, including posting, browsing, commenting, voting up, voting down, adding favorites, following, and sending private messages. For growth value prediction, dynamic behavior changes over time are more informative. To avoid sparse data issues, we adopt monthly rather than daily statistics.

We use correlation analysis to exclude "vote down" behavior due to its negative contribution to prediction. Through feature selection, we derive features for the stacking model using averages, logarithmic calculations, and growth rates of original data:

$$LOG(d) = \log(d + 1)$$

where LOG(d) represents the adjusted calculation result for data d, and GR(d_t) represents the growth rate calculation from data d_t in month t to d_{t+1} in month t+1:

$$GR(d_t) = \frac{d_{t+1} - d_t}{d_t + 1}$$

3.3.2 PAR/GDR-NuSVR-Stacking Model (PGNS)

After obtaining monthly statistics and derivative features, we feed them as inputs to Passive Aggressive Regressor and Gradient Boosting Regressor independently. By averaging predictions from these two base models, we create a new feature for the stacking model NuSVR. Due to inherent randomness in base models, we adopt a 10-fold cross-validation self-check mechanism.

If the trained model achieves a score higher than threshold S* under given scoring rules, we input validation or test set features to obtain a prediction value saved to a candidate set. Conversely, if the 10-fold cross-validation score falls below S*, the model is discarded and training restarts. To reduce single-round training errors, we set a minimum of R* training rounds, adding all predictions scoring above S* to the candidate set. Based on our experience, the candidate set size to R* ratio is approximately 0.45. After completing all training rounds, we average all predictions in the candidate set to generate final results.

4. Evaluation

Our model uses the Jieba[3] toolkit for Chinese word segmentation and trains 300-dimensional word embeddings [11].

Table 3 compares our Task 1 approach variations. The best results are achieved when using data from all three aspects to capture blog main ideas.

Table 4 tests our combined neural network with different embedding inputs. To obtain individual embedding results, we train a new CNN model for blog embedding and compute similarity between blog content and keywords in the embedding space. Results show blog content embedding proves more effective than keyword embedding alone, while their combination achieves the best performance.

Table 5 displays our system's best-run performance on each task, which achieved sixth place in the competition.

5. Conclusions and Future Work

This paper presents our system for the User Profiling Technology Evaluation Campaign of SMP CUP 2017. For Task 1, we extract keywords from three aspects of a user's blogs: the blog itself, blogs on the same topic, and other blogs by the same user. For Task 2, we construct a unified neural network model with a self-attention mechanism based on multi-scale convolutional neural networks to capture both local and global profile information. For Task 3, we adopt a stacking model for user growth value prediction. According to SMP CUP 2017

metrics, our model achieved final scores of 0.563, 0.378, and 0.751 on the three tasks, respectively.

Future work includes analyzing relationships between users and blogs. Currently, we only utilize user behavior in Task 2 while ignoring blog publication timing. We plan to incorporate network embedding into our model, collect more blogs with real-time information, and attempt to integrate temporal information into our weighting schema.

Author Contributions

B. Li (byli@uir.edu.cn) is the UIR-SIST system leader who designed the overall framework. J. Lu (lj1230@nyu.edu) built the keyword extraction model. L. Chen (lec@boyabigdata.cn) and K. Meng (kmmeng@uir.edu.cn) constructed the user interest tagging model. F. Wang (wangfengyi18@mails.ucas.ac.cn) summarized user growth value prediction. J. Xiang (xiang.j@husky.neu.edu) and N. Chen (nchen@uir.edu.cn) conducted evaluation and error analysis. X. Han (hanxu@cnu.edu.cn) drafted the paper. All authors revised and proofread the manuscript.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 61502115, No. 61602326, No. U1636103 and No. U1536207), and the Fundamental Research Fund for the Central Universities (No. 3262017T12, No. 3262017T18, No. 3262018T02 and No. 3262018T58).

References

[1] M. Ciot, M. Sonderegger, & D. Ruths. Gender inference of twitter users in non-English contexts. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1136–1145. doi: 10.1126/science.328.5974.38.

[2] L. Wendy, & R. Derek. What' s in a name? Using first names as features for gender inference in Twitter. In: *Proceedings of the 2013 AAAI Spring Symposium: Analyzing Microtext*, 2013, pp. 10–16. doi: 10.1103/PhysRevB.76.054113.

[3] W. Liu, F.A. Zamal, & D. Ruths. Using social media to infer gender composition of commuter populations. In: *Proceedings of the International Conference on Weblogs and Social Media*. Available at: http://www.ruthsresearch.org/static/publication_{files}/LiuZamalRuths_{WCMCW}.pdf.

[4] D. Rao, & D. Yarowsky. Detecting latent user properties in social media. In: *Proceedings of the NIPS MLSN Workshop*, 2010, pp. 1–7. doi: 10.1007/s10618-010-0210-x.

[5] M. Pennacchiotti, & A.M. Popescu. A machine learning approach to Twitter user classification. In: *Proceedings of the Fifth International Conference on*

Weblogs and Social Media, 2011, 281–288. doi: 10.1145/2542214.2542215.

[6] M.D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, & F. Menczer. Political polarization on Twitter. In: *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 2011, 89–96. Available at: https://journalistsresource.org/wp-content/uploads/2014/10/2847-14211-1-PB.pdf?x12809.

[7] C. Tu, Z. Liu, & M. Sun. PRISM: Profession identification in social media with personal information and community structure. In: *Proceedings of Social Media Processing*, 2015, pp. 15–27. doi: 10.1007/978-981-10-0080-5_2.

[8] SMP CUP 2017. Available at: http://www.cips-smp.org/smp2017/.

[9] R. Mihalcea, & P. Tarau. TextRank: Bringing order into text. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004. Available at: http://www.aclweb.org/anthology/W04-3252.

[10] D.M. Blei, A.Y. Ng, & M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research* 3(2003), 993–1022. Available at: https://dl.acm.org/citation.cfm?id=944937]&preflayout=flat#source.

[11] T. Mikolov, K. Chen, G. Corrado, & J. Dean. Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at International Conference on Learning Representations (LCLR)*, 2013, pp. 1–13. Available at: https://www.researchgate.net/publication/234131319_{Efficient}*{Estimation}*{of}*{Word}*{Representations}{

[12] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, & L.D. Jackel. Handwritten digit recognition with a backpropagation network. In: *Proceedings of Advances in Neural Information Processing Systems*, 1990, pp. 396–404. Available at: https://dl.acm.org/citation.cfm?id=109279%22.

[13] A. Krizhevsky, I. Sutskever, & G. Hinton. ImageNet classification with deep convolutional neural networks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 1–9. doi: 10.1145/3065386.

[14] Y. Kim. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1746–1751.

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, & P. Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research* 12(2011), 2493–2537. doi: 10.1016/j.chemolab.2011.03.009.

[16] D.H. Wolpert. Original contribution: Stacked generalization. *Neural Netw* 5(2)(1992), 241–259. doi: 10.1016/S0893-6080(05)80023-1.

[17] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, & Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7(3)(2006), 551–585. Available at: http://www.jmlr.org/papers/v7/crammer06a.html.

[18] J.H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5)(2001), 1189–1232. doi: 10.1214/aos/1013203451.

[19] J. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38(4)(2002), 367–378. doi: /10.1016/S0167-9473(01)00065-2.

Author Biography

Junru Lu is currently a Master' s degree candidate at the Center of Urban Science and Progress, New York University. He received his Bachelor' s degree from the University of International Relations in 2018. His research interests include natural language processing, text mining, and social computing.

Le Chen received his Bachelor' s degree from the University of International Relations in 2018. He now works as a data analyst at Beijing Boya Bigdata Co., Ltd. His research interests include text mining and social computing.

Kongming Meng currently works as a data engineer at DeepBrain Company. He received his Bachelor' s degree from the University of International Relations in 2018. His research interests include data mining and data analysis.

Fengyi Wang is currently a graduate student at the University of Chinese Academy of Sciences (CAS). She received her Bachelor' s degree from the University of International Relations in 2018. Her research interests include natural language processing and social network analysis.

Jun Xiang is currently a graduate student in Computer Systems Engineering at Northeastern University. She received her Bachelor' s degree from the University of International Relations in 2018 and has published two papers in international conferences and Chinese journals during her undergraduate studies.

Nuo Chen obtained her Bachelor' s degree from the School of Information Science and Technology, University of International Relations in 2018. Her research interest is knowledge graphs.

Xu Han is an Assistant Professor at Capital Normal University. She received her PhD degree in 2011. Her research interests are artificial intelligence and mobile cloud computing. She has published over 30 research papers in major international journals and conferences.

Binyang Li is an Associate Professor at the School of Information Science and Technology, University of International Relations. He received his PhD degree from the Chinese University of Hong Kong in 2012. His research interests include natural language processing, sentiment analysis, and social computing. He has published over 50 research papers in major international journals and conferences.

---

[1] https://github.com/LuJunru/SMPCUP2017_{ELP}
[2] https://github.com/coderSkyChen/2016CCF_{BDCI}_{Sougou}

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

Note: Figure translations are in progress. See original paper for figures.