First place solution for NLPCC 2017 shared task Social Media User Modeling

Lingfei Qian, Anran Wang, Yan Wang, Yuhang Huang, Jian Wang*, Hongfei Lin

Dalian University of technology, Dalian Liaoning 116023, China
*wangjian@dlut.edu.cn

Abstract. With the popularity of mobile Internet, many social networking applications provide users with the function to share their personal information. It is of high commercial value to leverage the users' personal information such as tweets, preferences and locations for user profiling. There are two subtasks working in user profiling. Subtask one is to predict the Point-of-Interest (POI) a user will check in at. We adopted a combination of multiple approach results, including user-based collaborative filtering (CF) and social-based CF to predict the locations. Subtask two is to predict the users' gender. We divided the users into two groups, depending on whether the user has posted or not. We treat this task subtask as a classification task. Our results achieved first place in both subtasks.

Keywords: location prediction, gender prediction, user modeling, collaborative filtering, classification algorithm.

1 Introduction

User modeling on social media is critical both in recommendation system and precise advertisement to get the target users [1]. With the rapid development of mobile devices and popularity of social applications, users also like to share their location by logging their point-of-interests (POIs) while posting both text and photos. The (POIs) previously logged are used to recommend and predict new places users may be interested in. Meanwhile, people are really active posting blogs on social media and user profiling is getting more attention as precise advertisement is now very essential. Therefore it's possible and necessary to extract information from social media to build user profiles.

This shared task contains two subtasks. Subtask one is check-in location prediction for users, which can be considered as a point-of-interest (POI) recommendation problem in location-based social networks (LBSNs). Subtask two is user's attributes prediction, which can be considered as a classification problem. We will introduce the two subtasks respectively in the following sections.

1.1 Subtask one

There are some common approaches for POI recommendation. The classic approach to the recommendation problem is collaborative filtering (CF) [2], which is comprised of two main methods, memory-based CF and model-based CF [3]. The former includes user-based CF [4] and item-based CF [5] or both of them together [6], the latter is mainly based on Matrix Factorization (MF) [7].

Similar to traditional recommendation systems, users' location check-in data can be processed by CF approaches to calculate the similarities of one user to candidate POIs (items) in POI recommendation. M et al [3] adopted user-based CF to obtain POI interest score for a user according the user's most similar neighbors. As for the MF method, by mapping users and POIs to a lower dimensionality joint latent factor space, the algorithm learns the interest score for POIs the users haven't visit [8].

On the other hand, POI recommendation has some unique characteristics compared to other recommendation systems. The first is a geographical clustering phenomenon [9]. Users prefer and are more likely to visit POIs nearby. Several studies show that the geographical information can contribute to the POI recommendation result. Ye M et al [3] quantified the geographical influence by applying a power-law distribution to model the relationship between distance and the possibility of user's visiting. Lian D [10] proposed a method called GeoMF, which combines geographic information with MF by augmenting the activity areas matrix into users' latent feature matrix and influence area matrix into POIs'. The second characteristic is social relationship influence. Experiment result shows that adding friend-based CF [3, 11] or integrating social influence with probabilistic matrix factorization [11] can improve the POI recommendation quality.

For subtask one, we proposed a fusion method on memory-based CF methods and rules to realize users' POI prediction. There are three parts in our method, we will elaborate in the next section.

1.2 Subtask Two

There are many works in the area of gender prediction with textual information and other data from social media. Some work has been done to analysis the writing style and preference of words of authors to infer the latent attributes of authors such as gender [12]. Some work paid attention to the POS (part of speech) of words and select features to improve the accuracy of gender classification [13].

Other works take author's affective factors into consideration [14]. However, these works tend to focus on collections of lengthy text posts and extract features from text. There are also works that focus on short text, combining textual features with other information. Burger [15] et al extract n-gram features from users' microblog, users' personal description and user names. They combined the improved balance Windows algorithm [16] to do the gender prediction. Some researchers have proved that social networks can also be united with logistic regression to improve the accuracy rate of user region [17]. Another related work has taken social tags of users into consideration to build user model and proves this is a helpful approach [18]. Ma et al.

[19] have used factorization of matrix by analyzing locations and interaction records of users to classify the users.

For subtask two, we divided the users into two groups based on whether the user has post information or not. We propose different feature combinations for the two groups. These combinations will be described in the following sections.

2 Method

2.1 Subtask one

User-based CF

The key to get the POI interest score by the user-based CF approach is to find users similar to the target user. In this subtask, we adopted check-in data to build the similarity between users. The user set denoted as $\mathrm{U} = \{u_1, u_2, \cdots, u_M, \}$, and location (POI) set denoted as $\mathrm{L} = \{l_1, l_2, \cdots, l_N, \}$. We also built user-location check-in matrix $C \in \mathcal{R}^{\mathrm{M}\times\mathrm{N}}$, and each element $c_{i,j} = 1$ or 0 depends on if the user i has visited location j.To simplify the calculations, we utilized matrix multiplication to get the user similarity matrix M_{user_user} after L2 normalized matrix C by each line.

$$M_{user\_user} = \mathrm{Norm}(C) \cdot \mathrm{Norm}(C)^T \tag{1}$$

Where, $M_{user\_user} \in \mathcal{R}^{\mathrm{M}\times\mathrm{M}}$ and each element $a_{i,k}$ represents the similarity between user i and user k.With the user similarity, we can gain the score of new POIS for each user using another matrix multiplication as follow:

$$P^{user\_cf} = M_{user\_user} \cdot C - C \tag{2}$$

Where, $P^{user\_cf} \in \mathcal{R}^{\mathrm{M}\times\mathrm{N}}$ and each element $p_{i,j}^{user\_cf}$ represents the score for user i and location j.

Social friend (SF) based CF.

Similar to user-based CF, we gained user similarity matrix through social networks check-in data. The friend set denoted as $\mathrm{F} = \{f_1, f_2, \cdots, f_Z, \}$ and the social network Matrix$S \in \mathcal{R}^{\mathrm{M}\times\mathrm{F}}$, where$S_{i,j} = 1$ or 0 depends on if the user i has followed user j. Same steps as user-based CF, finally we can get the score matrix $P^{sf\_cf} \in \mathcal{R}^{\mathrm{M}\times\mathrm{N}}$, where each element $p_{i,j}^{sf\_cf}$ represents the score for user i to location j.

Social location (SL) based CF.

Using social networks to measure user similarity is helpful, but the problem is some people might be irrelevant when they just follow popular people like stars. Therefore, we adopted another way to get the similarity $b_{i,\mathrm{k}}$ between users by their POI coincidence instead of friend coincidence. We adopted the Jaccard similarity coefficient to measure the similarity between a user and his friends as follows:

$$b_{i,k} = \frac{|L_i \cap L_k|}{|L_i \cup L_k|} \quad (3)$$

Where L_i and L_k denote the location set for user i and user k who is watched by user i. And the score for SL-based CF is:

$$p_{i,j}^{sl\_cf} = \sum_{k \in F_i} (b_{i,k} \cdot c_{k,j}) \quad (4)$$

Where $c_{k,j}$ means if user k has gone to location j.

Social location-item (SLI) based CF.

In the consideration of the large set of POIs, item-based CFs have a great amount of computing cost. Therefore, we chose the POIs that occur in close friends' check-in history. The close friends for user i, are those who have $b_{i,k} > 0$, and their POI list will be the mini-candidate POIs for user i, except the POIs user i has already gone to. First we measure the new POI score by using item-based CF on the mini-candidate POI set as follows:

$$p_{i,j}^{item\_cf} = \frac{\sum_{q \in L_i} w_{j,q}}{\sum_{q \in L_i} c_{i,q}} \quad (5)$$

Where $w_{j,q}$ denotes the similarity between location j and location q, which is also calculated in the Jaccard similarity coefficient as follow:

$$w_{j,q} = \frac{|U_j \cap U_q|}{|U_j \cup U_q|} \quad (6)$$

Then we combine the item-CF on the mini-candidate POI set with SL-CF as follows, to rearrange the candidate POIs after item-based CF by how many friends of user i have been to location j.

$$p_{i,j}^{sl\_item\_cf} = p_{i,j}^{item\_cf} \cdot p_{i,j}^{sl\_cf} \quad (7)$$

In the end, we integrated the ranking lists we discussed above. Let P denotes the results set $P = \{P^{user\_cf}, P^{sf\_cf}, P^{sli\_cf}\}$ and $S_{i,j}$ denotes the fused score for or user i to location j.We fused our results in linearly with hyper-parameter α for each result as follow:

$$S_{i,j} = \sum_{p \in P} \alpha_p \cdot p'_{i,j} \quad (8)$$

Where α_p represent the hyper-parameter for result p and $\sum_{p \in P} \alpha_p = 1$, $p'_{i,j}$ is the normalized score for each ranking set. The normalization method is shown as follows:

$$p'_{i,j} = \frac{p_{i,j} - min}{\max - \min} \quad (9)$$

Where max or min is the max or min in $\{p_{i,j}\}_{j \in L_i}$. We get the top 10 POIs for each user as our POI prediction results.

2.2 Subtask two

Subtask two is user profiling, here we just have to do the gender prediction. Gender prediction can be seen as a binary classification. With data of users' tweets, tags, social connections, and names of the places users had visited, predicting users' gender.

In this paper, we screen the users with tweet information and divide these as a group. Then all of the users are treated as a training set to predict the gender of users without tweet information. As is shown in Fig. 1, we extract textual information from users' tweets and this textual information is trained to predict genders of users with tweet information. On the other hand, almost all of the users have data about their social connections, tags, and names of check-in locations. These features are utilized to predict the genders of users without tweet information.

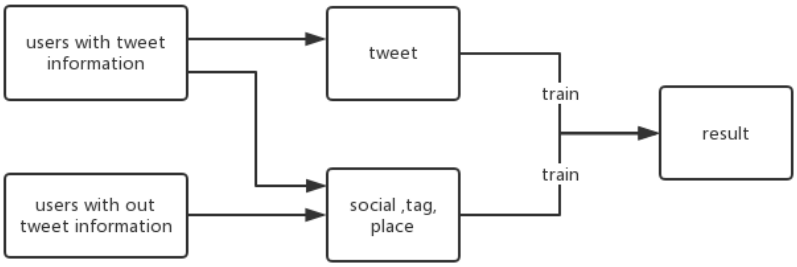

Fig. 1. The workflow for subtask two

Textual information

In this paper, all the tweets of the same user were connected together to get the user's textual information. Then TF-IDF was used to generate a sparse matrix, for every word in each user's document, first to calculate the TF (Term Frequency), and then to get the IDF (Inverse Document Frequency) of the word. TF represents the frequency of the word in this exact user's document. IDF counts how many documents contain this word, and take the reciprocal of it. TF then is multiplied by IDF to get the vector of the user. We use different models for comparison in the further experiments.

Social, tag, place information.

Social and tag information is made up of a serious of numbers, but the size of the number is meaningless, so we treat numbers as characters. We cut the names of the place into words. Then we extract statistical features of users' social network and tags to get the vector representation respectively. As for location information, we treat them as textual information. We also try different models with these features.

3 Experiments and Results Analysis

3.1 Subtask one

Dataset and evaluation.

We used check-in data and social network data for subtask one. We split check in data to form an offline training set and test set by POI groups. For each POI, split out 0.75 percent of the check-in data (rounded up) to get the training set, so that all the POIs already appear in training set. Then we filter out the users who don't exist in the training set from the rest of check-in data to become test data. Our offline training set has 7,017,632 check-in actions with 276,442 users and 620,195 POIs, and 1,699,732 check-in actions with 119,765 users and 173,397 POIs for the test set. Also in the test set, we found that less than half the users have more than 10 POIs recorded and those users' check-in data makeup 89% of all the data gathered. We removed the users who have less than 10 POIs record to form another test set, which contains 58,058 users which can be considered active users.

The quality of subtask one is evaluated by F1@K (K=10), and we also list P@K R@K for analysis. The calculation formula is as follows:

$$P_i@K = \frac{|H_i|}{K} \tag{10}$$

$$R_i@K = \frac{|H_i|}{|V_i|} \tag{11}$$

$$F1_i@K = \frac{2 \times P_i@K \times R_i@K}{P_i@K + R_i@K} \tag{12}$$

$$\mathrm{P}@K = \frac{1}{N}\sum_{i=1}^{N} P_i@K \tag{13}$$

$$\mathrm{R}@K = \frac{1}{N}\sum_{i=1}^{N} R_i@K \tag{14}$$

$$\mathrm{F1}@K = \frac{1}{N}\sum_{i=1}^{N} F1_i@K \tag{15}$$

Where $|H_i|$ is the correctly predicted locations for user i's top K prediction, $|V_i|$ is the correct locations for user i. $P_i@K$, $R_i@K$, and $\mathrm{F1}_i@K$ is the precision, recall and F1 for a user i. N is the user count.

Experimental Results and Analysis.

We evaluate our method in both a normal test named test set 1, and a test with more than 10 POIs record for each user, named test set 2. First we evaluate every single method on both test sets and the results are shown in Table 1. Then we integrate the results by different combination and the results are shown in Table 2.

Table 1. The results of each CF method

Method	Test set 1			Test set 2		
	F1@K	P@K	R@K	F1@K	P@K	R@K
User-based CF	1.250%	1.683%	1.422%	1.661%	2.762%	1.291%
SF-based CF	1.999%	3.099%	1.932%	3.221%	5.682%	2.433%
SL-based CF	0.147%	0.188%	0.229%	0.163%	0.287%	0.126%
SLI-based CF	0.208%	0.267%	0.300%	0.243%	0.415%	0.188%

Based on the results in Table 1, we find that the SF-based CF has the best performance. This demonstrates that social network information is effective in POI recommendation compared to the normal user-based CF method. However, the performance of SL-based CF and its related methods are much worse than other CF methods. We believe the reason is that only 3.6% of friends in social networks have check-in data, this makes it hard to utilize users' check-in coincidence to measure user similarity. Also adding item-based CF and SL-based CF can achieve a slight improvement.

Table 2. The results of fusion work

Method	α_U	α_{SF}	α_{SLI}	Test set 1			Test set 2		
				F1@K	P@K	R@K	F1@K	P@K	R@K
U + SF	0.2	0.8	0	2.120%	3.227%	2.098%	3.334%	5.839%	2.528%
U + SLI	0.4	0	0.6	1.291%	1.727%	1.507%	1.695%	2.820 %	1.317%
SF + SLI	0	0.7	0.3	2.018%	3.111%	1.965%	3.230%	5.684%	2.444%
U+SF+SLI	0.1	0.6	0.3	2.122%	3.232%	2.078%	3.348%	5.862%	2.539%

In the Table 2, we list out the fusion result and the best hyper-parameters. Based on the table, we find that fusion work did improve the final result especially for user-based CF and SF-based CF. We adopted the U+ SLI method in the shared task, but the result shows SF-based CF is much more effective in this task. Also the α parameters indicate the significance of social network information.

3.2 Subtask two

Dataset and evaluation.

In this task, there are several data sets given, including users' check-in information, users' tag information, user's social information, users' tweets and users' gender and their picture information. The check-in information includes POI, different categories

of location, latitude, longitude, and names of the places. There are about 300,000 users in total, and about 75,000 users have tweet information. The quality of the User Profiling subtask is evaluated by accuracy, where δ is the indicator function where Label_i and Predict_i is the same.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \delta(\text{Label}_i, \text{Predict}_i) \quad (16)$$

Experimental Results and Analysis.

Some results from the experiments which use Logistic Regression (LR [20]) to train different features are shown in Table 3. Different models were used to train the tweet features and the features combination of tag, place and social. The results are displayed in Table 4. We attempted on LR, NB (Naïve Bayes), RF (Radom Forest) [21], XGBoost (eXtreme Gradient Boosting) [22]. We used two LR model where the penalty item parameter penalty of LR1 is L1 normalization, penalty of LR2 is L2 normalization.

Table 3. The results of different features combination trained by LR

Feature combination	Accuracy
tweet	94.08%
tweet + place	93.45%
tweet + tag	93.97%
tweet + social	94.62%
tag	76.09%
place	73.27%
social	74.20%
tag + social + place	82.34%

Table 4. The performance of various models

Model	Accuracy(tweet)	Accuracy(tag+ place+ social)
LR1	93.07%	81.78%
NB	81.12%	70.89%
RF	80.24%	72.65%
XGBoost	94.16%	82.15%
LR2	94.08%	82.34%
LR1+XGBoost+LR2	94.08%	82.54%

As we can see from the table, the place and tag information did not help to improve the performance, but social connection does improve the results. However place and tag are useful without tweet. Naïve Bayes did not perform well in this dataset, probably because the attributes are not totally independent of each other. Tree models were not suitable for the dataset, probably due to over fitting. We use simple voting to

combine the results of LR1, LR2 and XGBoost. The voting algorithm doesn't work because the differences between these models are not large enough. We use tweet features for users who have posted tweets, and tag, place, social for users without tweets, and both of them are trained by LR2. The accuracy of offline tests was 85.25%, and the final result was 85.64%.

4 Conclusion and future work

In this paper, we elaborate our methods and ideas on user modeling shared tasks. For subtask one, we take it as a recommendation problem and adopted a memory-based collaborative filtering method. Our online submission results are based on user-based CF and SLI-based CF result. However, we found that SF-based CF is much more effective than SL-based CF in this subtask. We believe the reason for the undesirable performance for SL-based CF is due to the lack of check-in data for social networks. In the future, we are going to study how to utilize the geographical information and categories of locations in this subtask.

For subtask two, we treat the task as a classification problem. We focused more on data analysis and the combination of features. Our online submission divided the users into two groups and construct features respectively. However we found some features from other groups can also help to improve the result. In the future, ensemble learning should be investigated deeply. Also the information of categories for locations was not used in our method, this data could contribute to the results.

5 Acknowledgments

This research is supported by the National Key Research Development Program of China (No. 2016YFB1001103) and Natural Science Foundation of China (No. 61572098, 61572102）

References

1. Farseev A, Nie L, Akbari M et al. (2015) Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, p 235-242
2. Goldberg D, Nichols D, Oki BM et al. (1992) Using collaborative filtering to weave an information tapestry. Communications of the ACM 35:61-70
3. Ye M, Yin P, Lee W-C et al. (2011) Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, p 325-334
4. Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., p 43-52

5. Sarwar B, Karypis G, Konstan J et al. (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web. ACM, p 285-295
6. Wang J, De Vries AP, Reinders MJ (2006) Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, p 501-508
7. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8) p30–37
8. Berjani B, Strufe T (2011) A recommendation system for spots in location-based online social networks. In: Proceedings of the 4th Workshop on Social Network Systems. ACM, p 4
9. Cao X, Cong G, Jensen CS (2010) Mining significant semantic locations from GPS data. Proceedings of the VLDB Endowment 3:1009-1020
10. Lian D, Zhao C, Xie X et al. (2014) GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, p 831-840
11. Ye M, Yin P, Lee W-C (2010) Location recommendation for location-based social networks. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems. ACM, p 458-461
12. Schler J, Koppel M, Argamon S et al. (2006) Effects of Age and Gender on Blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. p 199-205
13. Mukherjee A, Liu B (2010) Improving gender classification of blog authors. In: Proceedings of the 2010 conference on Empirical Methods in natural Language Processing. Association for Computational Linguistics, p 207-217
14. Rangel F, Rosso P (2016) On the impact of emotions on author profiling. Information processing & management 52:73-92
15. Burger JD, Henderson J, Kim G et al. (2011) Discriminating gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, p 1301-1309
16. Littlestone N (1988) Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine learning 2:285-318
17. Rahimi A, Vu D, Cohn T et al. (2015) Exploiting text and network context for geolocation of social media users. arXiv preprint arXiv:1506.04803
18. Carmagnola F, Cena F, Cortassa O et al. (2007) Towards a tag-based user model: How can user model benefit from tags? User Modeling 2007:445-449
19. Ma H, Cao H, Yang Q et al. (2012) A habit mining approach for discovering similar mobile users. In: Proceedings of the 21st international conference on World Wide Web. ACM, p 231-240
20. Kurt I, Ture M, Kurum AT (2008) Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert systems with applications 34:366-374
21. Breiman L (2001) Random forests. Machine learning 45:5-32
22. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, p 785-794