A Label Propagation Based User Locations Prediction Algorithm in Social Network

Huan Ma¹ and Wei Wang²(✉)

¹ National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China

² Harbin Engineering University, No. 145, Nantong Street, Harbin, China
w_wei@hrbeu.edu.cn

Abstract. Network community detection is an important service provided by social networks, and social network user location can greatly improve the quality of community detection. Label propagation is one of the main methods to realize the user location prediction. The traditional label propagation algorithm has the problems including “location label countercurrent” and the update randomness of node location label, which seriously affects the accuracy of user location prediction. In this paper, a new location prediction algorithm for social networks based on improved label propagation algorithm is proposed. By computing the K-hop public neighbor of any two point in the social network graph, the nodes with the maximal similarity and their K-hopping neighbors are merged to constitute the initial label propagation set. The degree of nodes not in the initial set are calculated. The node location labels are updated asynchronously is adopted during the iterative process, and the node with the largest degree is selected to update the location label. The improvement proposed solves the “location label countercurrent” and reduces location label updating randomness. The experimental results show that the proposed algorithm improves the accuracy of position prediction and reduces the time cost compared with the traditional algorithms.

Keywords: Social network · Location prediction · Label propagation · Social relationships · User location probability

1 Introduction

As social networks with location-based information are increasingly popular, the users' location in social network attracts more attention than before. Location information can help to shorten the gap between the virtual and the real world, such as monitoring residents' public health problems through online network [1], recommending local activities or attractions to tourists [2, 3], determining the emergency situation and even the location of the disaster and so on [4–6]. In addition, users' offline activity area and trajectory can also be analyzed through their locations in social networks. Due to the increasing awareness of privacy protection, people will cautiously submit their personal location information or set the visibility of the location of the message in social networks,

which make it difficult to acquire their real location information. Therefore, how to accurately predict the actual location information of social network users is an important and meaningful research question.

This paper proposes location prediction algorithm for social network users based on label propagation, which solves the following two key problems:

- (1) The accuracy of the traditional label propagation algorithm is not high in the user location prediction, and “countercurrent” phenomenon will appear in the iterative process, which will lead to the increase of the time overhead.
- (2) Improve the accuracy of social network users’ location prediction by using their offline activity location.

2 Related Work

There are three scenarios for user location prediction in social networks, such as user’s frequent location prediction, prediction of the location of messages posted on the user’s social network, and forecasts of the locations mentioned in messages. The main methods of location pre-diction include location prediction based on the content of message published by users, user friend relationships, and so on.

Laere et al. chose two types of local vocabulary and extracted valid words to predict the location of users [7]. Ren [8] and Han et al. [9] were inspired by the frequency of reverse documents, using the reverse position frequency (ILF) and the reverse city frequency (ICF) to select the position of the vocabulary, they assumed that the location vocabulary should be distributed in fewer locations, but with large ILF and ICF values. Mahmud et al. [10] applied some column heuristics to select local vocabulary. Cheng [1] makes the position word distribution conform to the spatial change model proposed by the Backstorn [11], secondly they make local or non-local mark on 19,178 dictionary words, and use the Labeled Vocabulary Training classification model to discriminate all words in the tweet dataset.

Backstrom [12] established probability models through physical distances between users to express the possibility of relationships between users, which has no effect on the position prediction of friends considering different degrees of tightness. Kongl [13] on the basis of Backstorn work by adding the weight of the edge to predict the user’s position, where the weight of the edge is determined by a social tight coefficient. Li [14] considered the location of user neighbors, and captures the information of users’ neighbors that intuitively consider the location of users. The user location is allocated randomly, then the user’s location is iteratively updated from the user’s neighbors and the location name mentioned, and then the parameters in the update are improved by measuring the prediction error of the known location of the user. Davis Jr et al. [15] thought that the most frequent user’ locations that appear in the user’s social network as a basis for predicting their location. Jurgens et al. [16] extend the concept of location prediction into location label propagation, which is made by the location of the label space to explain the location of label propagation, they think that the position of the user through the iterative process that many times.

Li et al. [17] thought that the literature assume the user has only one home location is a defect, they think that users should have the relationship with a number of positions, so they have defined the location information of a user and user set as the set of locations, and these users about the system is not only a geographical location the range is not a point, is not a temporary and user related position, but a long-term position, so they set up a MLP in the paper (Multiple Location Profiling Model) to establish a model containing a plurality of position information of the position of archives to the user, and this model is to the location file according to the target user relationships and their tweets content released.

The label propagation algorithm can effectively deal with large data sets, so in this paper, we are in the position of the user prediction based on label propagation algorithm, but with the label propagation algorithm in-depth study, we found that the label propagation algorithm will position the label “countercurrent” label update and node location is random, this algorithm cannot guarantee the accuracy of prediction of the position of the user, in order to improve the accuracy of location prediction algorithm and reduce the time overhead, this paper pro-poses a label propagation based on user location prediction algorithm (Label Propagation Algorithm-Location Prediction, LPA-LP).

3 Related Concept and Problem Definition

Definition 1 Social Network. A social network can be represent by a graph $G = (V, E, A)$, where V represents the collection of the users who are in the social network, and $n = |V|$. E represents the collection of the relationship between users and $m = |E|$, and A represents the collection of the activities and $a = |A|$. Beyond that, L represents the set of locations, including users’ locations and activities’ locations, and $n_l = |L|$, U_o is the set of the users whose locations are known, on the contrary, U_n is the set of users whose locations are unknown.

Definition 2 Shortest Path Length. It refers to the shortest path between the two nodes i and j in the social network graph. It means the minimum number of paths through the node i to the node j . It can be used $d(i, j)$ to represent the shortest path length between two nodes.

Definition 3 K -Hopping Neighbors. It means that the user to its neighbor needs a k hopping to achieve, that is to say, the shortest path length of the two node is k .

Definition 4 K -Hopping Public Neighbors. $G = (V, E, A)$ is a social network diagram, where V represents the user set in the graph, $E = (v_i, v_j, w_{ij})$ represents the set of relations between the user nodes with weights, w_{ij} represents the weight of the edges between nodes. The k -hopping public neighbors set of the nodes is defined as follows:

$$\Gamma_k(v_i, v_j) = \{v | d(v_i, v) = d(v_j, v) = k\} = \Gamma(v_i, k) \cap \Gamma(v_j, k), k \geq 1 \quad (1)$$

In the formula (1), $\Gamma(v_i, k)$ represents the set of k -public neighbor of node v_i , and represents the set of node v_j , represents the set of k -public neighbor between v_i and v_j .

Definition 5. Similarity of k -hopping public neighbors. The value of k is determined by the network itself, it can be defined on formula (2).

$$\bar{k} = \frac{\sum_{i \neq j} k_{\max} |\Gamma(i) \cap \Gamma(j)|}{|V|} \quad (2)$$

In the formula (2), $k_{\max} |\Gamma(i) \cap \Gamma(j)|$ represents the max public neighbor hops between two nodes. The k value in the network refers to the average of any two nodes in the network. The similarity of the two node k -hopping public neighbors is defined by formula (3).

$$S(v_i, v_j) = \frac{|\Gamma(v_i, k) \cap \Gamma(v_j, k)|}{|\Gamma(v_i, k) \cup \Gamma(v_j, k)|}, k \geq 1 \quad (3)$$

Definition 6 Similarity of Nodes. It means denominator size of the similarity between the k -hopping public neighbors between nodes subtracts the two nodes. It can be defined by formula (4).

$$\gamma = \frac{|\Gamma(v_i, k) \cap \Gamma(v_j, k)|}{|\Gamma(v_i, k) \cup \Gamma(v_j, k)| - 2}, k \geq 1 \quad (4)$$

Definition 7. The max degree between nodes and users set. If the user is divided into different sets $L_1, L_2, \dots, L_e$ according to their locations, nodes are set up by users who are not labeled as location labels. The max degree of users divided into different sets according to their location is the degree and the maximum of some nodes in the nodes. It can be defined by formula (5).

$$d(v_i, L_i) = \max\{d(v_i, L_1), d(v_i, L_2), \dots, d(v_i, L_e)\} \quad (5)$$

Definition 8 K -Hopping Weight. We believe that the most important impact on user location is its 1 hop neighbors. Moreover, the offline location of users also has a great impact on user location, and its weight can also be set to 1. For $k > 1$, when setting the weight of the edge, it will be attenuated according to the speed of 1/5, that is, the weight of the edge of the 1 hop neighbor is 1, the weight of the 2 neighbors is 0.8, and so on.

Now given the location prediction problem definition: In the social network G , the unknown location information of the user u , according to the location information and the users of their k -hopping neighbors, to predict the unknown location information of the user u in the prediction of the probability of the position of L .

4 Label Propagation Based User Location Prediction Algorithm

In this section, a correlation algorithm for location prediction for users of unknown location information in social networks is proposed. This paper proposed a location prediction algorithm based on label propagation (Label Propagation Algorithm-Location Prediction, LPA-LP), the algorithm is mainly divided into two parts, one part is to run before the label propagation algorithm of data preprocessing algorithm, the other part is the use of label propagation of location prediction algorithm.

Algorithm 1 Data Preprocessing

Input: G-dataset, L-location label set

Output: pre-processed user sets C

Initialization: set C is empty

while $S(v_i, v_j) \geq \gamma$ do

 for each $v \in U_n$ do

 if $\Gamma(v) \neq \emptyset$ do

 set $k = 1$

$\Gamma(v)_i \leftarrow BFS(v)$

$k = i$

 end for

 get $\Gamma_k(v_i, v_j)$ for every nodes based on formula(1)

 select all nodes with k_{max}

 calculate the similarity of k-neighbors based on formula(3)

 choose the k-neighbors as the begin set

 update the nodes' label in the begin set

 calculate node similarity based on formula(4)

 end while

Algorithm 1 is pretreated before running the label propagation algorithm to initialize the data set, according to the Definition 5, the node with its maximum similarity and the k hop neighbor as the set of starting processing for the user location prediction, and according to the known label to the data in the collection of the label, which is in order to be able to quickly and accurately using the label propagation algorithm for unknown location information in a social network user node location prediction. After preprocessing the data set, location prediction algorithm based on label propagation can be used to predict the location of users who have not tagged location labels in the processed data set. Algorithm 2 gives a description of the location prediction algorithm (LPA-LP) based on the label propagation.

Algorithm2LPA-LP algorithm description

```

Input: pre-processed user sets  $C_i$ 
Output: estimated user location,  $E$ 
Initialization: set  $E$  is empty
    choose the begin set  $C_i$ 
    while the nodes label change do
        for each node do
            sort the nodes as the  $k$  to every set  $C_i$  to made  $F$ 
            end for
            for  $v \in F$  do

$$F_v(t) = f(F_{v_{i1}}(t), \dots, F_{v_{im}}(t), F_{v_{im+1}}(t-1), \dots, F_{v_{in}}(t-1))$$

                update the nodes' label in the begin set build  $E'$ 
            end for
             $E = E'$ 
        end while
  
```

In Algorithm 2 location prediction algorithm based on label propagation in the iterative process of user location labels are updated, and the location information of the user location information of neighbors and user participation in the offline activities are taken into account, which significantly improves the prediction accuracy of the locations of users, and in the operation of label propagation algorithm for data sets are preliminary the treatment improve the performance of the label propagation algorithm of user location prediction algorithm, the following will be proved by experiments.

5 Experiment Result and Analysis

In this section, we will analyze the experimental results, the experimental results are divided into two parts, one part is the results of algorithm time overhead and the other is the accuracy of user locations prediction algorithm.

5.1 Data Set Description

In this paper, we use the dataset is NLPiR microblogging corpus. We extracted several datasets from the dataset. In order to compare the accuracy of the improved algorithm for user location prediction and improve the execution efficiency of the algorithm, we extract different scale datasets from the data set to compare the experimental results. The detail of our data sets are described in Table 1.

5.2 Experimental Results Analysis

The location prediction algorithm based on the label propagation (LPA-LP) is an improvement on the preprocessing of the data set and the selection strategy of the location label in the iterative process. It can avoid the “countercurrent” phenomenon of the position label and reduce the randomness to update the location tag, and improve the efficiency and the accuracy of the prediction. The whole experiment is divided into two

Table 1. Data sets description

Dataset	Users number	Relations number	Activities number
A	2748	12121	452
B	4025	61881	983
C	5194	77567	2392
D	9940	107979	4938

parts. The first part is using label propagation algorithm to predict user location on these four datasets of different sizes. The second part is using LPA-LP algorithm to predict location on four different scale datasets.

In the process of user location prediction, probabilistic LPA algorithm and LPA-LP algorithm with random or update the node label to a certain extent, the running times of the two algorithms may produce different results, so the choice between the four data sets of different size on the running times of experimental results for the 10, 30, 50, 70, 100 and mean value. The time required for the experiment to run on different scale data sets is shown in Fig. 1, 2, 3 and 4.

Fig. 1. Time overhead comparison with dataset A

Fig. 2. Time overhead comparison with dataset B

Fig. 3. Time overhead comparison with dataset C

Fig. 4. Time overhead comparison with dataset D

From these four figures, we can know that the running time of different dataset is similar between the improved algorithm LPA-LP and the algorithm LPA when the dataset have less than 5000 nodes, when the nodes are more than 9000 in dataset, we can see that the running time of the improved algorithm LPA-LP is obviously less than the algorithm LPA. It shows that the LPA-LP algorithm can be effectively applied to large-scale data sets.

In addition to comparing the running time of the algorithm, it is necessary to compare the accuracy of the algorithm. The results of the experiment are shown in Table 2.

Table 2. Algorithm accuracy comparison

Dataset	Accuracy	
	LPA	LPA-LP
A	59.3%	64.4%
B	62.2%	67.3%
C	66.5%	69.5%
D	70.7%	78.4%

6 Conclusion

This paper proposes a location prediction algorithm for social network users based on label propagation. The algorithm first obtains k -hop public neighbors at any two points in the social network graph, and uses the node with the largest similarity and its k -hop neighbors as the initial set of label propagation, and calculates the degree of the node to these sets. In each iteration, the node adopts the strategy of asynchronous update, and selects the node with the highest degree to update the position label, so as to avoid the “countercurrent” phenomenon of the position label and reduce the possibility of randomly updating the position label. Relevant experiments show that the algorithm proposed in this paper improves the accuracy of user location prediction and reduces the time cost of the algorithm.

References

1. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating Twitter users. In: The 19th ACM Conference on Information and Knowledge Management, pp. 759–768. ACM, Toronto (2010)
2. Yuan, Q., Cong, G., Ma, Z., et al.: Who, where, when and what: discover spatio-temporal topics for Twitter users. In: The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 605–613. ACM, Chicago (2013)
3. Noulas, A., Scellato, S., Lathia, N., et al.: Mining user mobility features for next place prediction in location-based services. In: 13th Industrial Conference on Data Mining, pp. 1038–1043, IEEE, New York (2013)
4. Rakesh, V., Reddy, C.K., Singh, D., et al.: Location-specific tweet detection and topic summarization in Twitter. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1441–1444. ACM, Niagara (2013)
5. Ao, J., Zhang, P., Cao, Y.: Estimating the locations of emergency events from Twitter streams. *Procedia Comput. Sci.* **31**, 731–739 (2014)
6. Lingad, J., Karimi, S., Yin, J.: Location extraction from disaster-related microblogs. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1017–1020. ACM, Rio de Janeiro (2013)
7. Van Laere, O., Quinn, J., Schockaert, S., et al.: Spatially aware term selection for geotagging. *IEEE Trans. Knowl. Data Eng.* **26**(1), 221–234 (2014)
8. Ren, K., Zhang, S., Lin, H.: Where are you settling down: geo-locating Twitter users based on tweets and social networks. In: Hou, Y., Nie, J.-Y., Sun, L., Wang, B., Zhang, P. (eds.) AIRS 2012. LNCS, vol. 7675, pp. 150–161. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35341-3_13
9. Han, B., Cook, P., Baldwin, T.: Geolocation prediction in social media data by finding location indicative words. In: 24th International Conference on Computational Linguistics, pp. 1045–1062. ACM, Mumbai (2012)
10. Mahmud, J., Nichols, J., Drews, C.: Where is this tweet from? Inferring home locations of Twitter users. In: Sixth International AAAI Conference on Weblogs and Social Media, pp. 73–77. AAAI, Dublin (2012)
11. Backstrom, L., Kleinberg, J., Kumar, R., et al.: Spatial variation in search engine queries. In: Proceedings of the 17th International Conference on World Wide Web, pp. 357–366. ACM, Beijing (2008)

12. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International Conference on World Wide Web, pp. 61–70. ACM, North Carolina (2010)
13. Kong, L., Liu, Z., Huang, Y.: SPOT: locating social media users based on social network context. *Proc. VLDB Endow.* **7**(13), 1681–1684 (2014)
14. Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.: Towards social user profiling: unified and discriminative influence model for inferring home locations. In: The 18th International ACM SIGKDD Conference, pp. 1023–1031. ACM, Beijing (2012)
15. Davis, Jr C., Pappa, G., de Oliveira, D., de L Arcanjo, F.: Inferring the location of twitter messages based on user relationships. *Trans. GIS* **15**(6), 735–751 (2011)
16. Jurgens, D.: That’s what friends are for: inferring location in online social media platforms based on social relationships. In: Seventh International AAAI Conference on Weblogs and Social Media, pp. 237–240. AAAI, Massachusetts (2013)
17. Li, R., Wang, S., Chang, C.: Multiple location profiling for users and relationships from social network and content. *Proc. VLDB Endow.* **5**(11), 1603–1614 (2012)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.