Global Journal of Information Technology: Emerging Technologies

Volume 7, Issue 3, (2017) 149-156

www.gjit.eu

Location prediction in location-based social networks

Mucahit Baydar*, Department of Computer Engineering, Yildiz Technical University, 34349 Istanbul, Turkey, and Department of Nanotechnology, Anadolu University, 26470 Eskişehir, Turkey.

Songul Albayrak, Department of Computer Engineering, Yildiz Technical University, 34220 İstanbul, Turkey.

Suggested Citation:

Baydar, M. & Albayrak, S. (2017). Location prediction in location-based social networks. *Global Journal of Information Technology: Emerging Technologies*. *7*(3), 149-156.

Received July 02, 2017; revised October 11, 2017; accepted November 29, 2017.

Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, North Cyprus.

Abstract

Developments in mobile devices and wireless networks have led to the increasing popularity of location-based social networks. These networks allow users to explore new places, share their location, videos and photos and make friends. They give information about the mobility of users, which can be used to improve the networks. This paper studies the problem of predicting the next check-in of users of location-based social networks. For an accurate prediction, we first analyse the datasets that are obtained from the social networks, Foursquare and Gowalla. Then we obtain some features like place popularity, place popular time range, place distance to user's home, user's past visits, category preferences and friendships, which are used for prediction and deeper understanding of the user behaviours. We use each feature individually, and then in combination, using the new method. Finally, we compare the acquired results and observe the improvement with the new method.

Keywords: Location prediction, location-based social network, check-in data.

* ADDRESS FOR CORRESPONDENCE: **Mucahit Baydar,** Department of Nanotechnology, Anadolu University (2017) Hocataskin mah. Incirli cad. No: 68/5 Yıldırım/Bursa, 16350 Eskisehir, Turkey.

E-mail address: mucahitbaydar@gmail.com / Tel.: +90-537-4585831.

1. Introduction

Social networks are now an inseparable part of human life. While people are using social networks more than ever, social networks are adapting and evolving according to people's needs. Recently, a new type of social network that has started to become popular is the location-based social network. While users can share photos, videos and their opinions like in any other social networks, location-based social networks' starting point and main focus is users' locations. These networks help their users to discover new places and venues, and also help venues to increase their popularity and public recognition. Location-based social networks contain a vast amount of useful information about the mobility of users, which helps them in making good assumptions about users' behaviours and improves location-based social networks. These predictions help the networks in three ways. Social network providers earn their users' and commercial advertisers' trust. While users' loyalty and pleasure to these networks are growing, the commercial advertisers' investments are becoming strong on these networks.

2. Problem Definition

This study deals with the next check-in prediction problem. We attempt to predict the next check-in of users with the help of features obtained from the data analysis phase. Check-in data contain three pieces of information: user, location and time. While the database contains more than thousands of different places, it is hard to guess the check-in with just one prediction. To overcome this problem, we make a list of predictions of different sizes and compare the results. For every check-in, we try to predict, take the check-in from the database and use the rest for training and try to predict the chosen check-in.

In previous works, the problem was handled with different approaches. For location prediction, Bayesian networks are used in Park, Hong and Cho (2007) and collaborative filtering is used Ye, Yin and Lee (2010) and Berjani and Strufe (2011). A study of the user activity patterns is published in Noulas, Scellato, Mascolo and Pontil (2011) which gives deep understanding of location-based services and user activities. The random walk algorithm is used in Noulas, Scellato, Lathia and Mascolo (2012], which studies the new venue recommendation problem. User mobility, global mobility and temporal features are used for location prediction in Noulas, Scellato, Lathia and Mascolo (2012). New venue recommendation is also studied in Wang, Terrovitis and Mamoulis (2013) and specialised bookmark-colouring algorithm is used in the study. A study on location prediction offers a two-step method: in the first step, the system tries to predict the location category (Ye, Zhu & Cheng, 2013) and in the second step, the system tries to find the location from the predicted category.

3. Data Analysis

Datasets we used were obtained from two popular location-based social networks: Foursquare and Gowalla. Both the networks were launched in 2009 and became popular quickly. While Gowalla was shut down in 2012, Foursquare is still online and is one of the most popular location-based social networks. The Gowalla dataset has about 10 million check-ins made by 100,000 users on 1.5 million different places for 18 months from all over the world. The Foursquare dataset covers only London check-ins and has about 500 thousand check-ins made by 10,000 users for 9 months. Unlike Foursquare, Gowalla also has friendship information.

Location-based social networks contain a lot of information like place ratings, comments, price range, popularity and user ratings, preferences, friends etc., yet very little information is available. We have check-in data, which contains user, place and time information, and place data which contains the place name, latitude, longitude and category information. We also have friendship information only for the Gowalla dataset. We use these datasets and try to infer new useful information.

The first information we obtain is the check-in numbers of the users and places. Later, we can use this information for place popularity and user mobility frequency. Users with less than nine check-ins for the whole time are considered as inactive and they are removed from the database. So far as we are try to build a personalised recommendation system, users with a few check-ins would be useless in this context. Places with only one check-in are also deleted. Then, we label all places with city information, using another database that contains the city latitude and longitude values. Separating database into different cities helps reducing computation time and increasing prediction accuracy. While we don't have any information about users' home locations, we assume that the most frequently visited place is the home location of the user. We use this location as the starting point and calculate the distances of all other places from this point. We split 1 day into four time ranges and calculate the visiting frequency of the other places in these time ranges. Time ranges are 00.00–06.00, 06.00–12.00, 12.00–18.00 and 18.00–00.00. Finally, we gather six different features to use for prediction. The features are users' previous visiting places and categories, friendship, place popularity, time range frequencies, and users' distances between home and all other places.

Table 1. Total number of check-ins, users and places, average number of check-ins by user and average number of check-ins per place from the top 5 most popular cities from Gowalla and London from Foursquare dataset

Cities	Check-ins	Users	Places	Avg. check-in by user	Avg. check-in per place
Austin	497,935	9,954	20,733	51.9	24.01
San Francisco	362,547	9,458	19,516	38.33	18.57
Dallas	308,182	7,821	22,361	39.4	13.78
Stockholm	225,009	8,102	11,374	27.77	19.78
New York City	211,511	7,755	17,852	27.27	11.84
London	460,034	10,630	28,529	43.27	16.12

4. Check-in Prediction

This section formalises the next check-in prediction problem. Given the current information about database, we aim to predict the next check-in place of users. We calculate a score for every place in the system for a specified user whose next check-in place has to be predicted and then we rank them according to the scores. Finally, we get the top N highest scored places and we check if the next check-in place is in the list or not. We use different list sizes from 10 to 100.

We use the features obtained from the data analysis section for the prediction individually. As the Gowalla dataset was separated into different parts by cities, we use them separately for prediction. As the Foursquare dataset contains only London check-ins, we use it without any prior process.

4.1. Problem Formulation

We have a set of users U and a set of places P. Each check-in c is defined as a tuple $\{u, p\}$, where u and p represent are those who made that check-in and where the check-in was made. C is defined as the total set of check-ins where C_u is the set of check-ins for a specific user u. Ca_u is the set of categories which are visited by user u and F_u represents the set of users who are friends of user u. Given a user u, we calculate the score for every $p \in P$ and then we rank them based on these scores.

4.1.1. Prediction Using Individual Features

- Place popularity: The first feature we use for prediction is place popularity. Place popularity is simply how many check-ins were made by users on a specified place.

$$Score_p = \Sigma\{u, p\} \mid \forall u \in U$$

- Distance to home location: We define the home location of the user as the user's most visited place. Then we calculate the distance between home location and all places in the same city for the user. Then score calculated is inversely proportional to the distance. u_{home} is the home location of user u.

$$Score_{u,p} = \frac{1}{\text{dist}\{u_{\text{home}}, \text{p}\}}$$

- Close popular places: The first two individual features are combined and used as a single prediction feature. First, we found the top 1,000 closest places to each user's home location, then we ranked them according to their popularity.

$$Score_{u,p} = \left(\Sigma\{u, p\} \mid \forall u \in U\right) \wedge p \in close\left(u_{\text{home}}, 1000\right)$$

- Former place visits: We use the user's former visited places for prediction problem. The score of every place is simply how many times the user visited that place.

$$Score_{u,p} = \Sigma\{u, p\} \in C_u$$

- Friendship: We count the total number of visits by every friend of user u on every single place and give this count as the score of this place.

$$Score_{u,p} = \Sigma\{f, p\} \mid \forall f \in F_u$$

- Category preference: We use the user's category preference for prediction. We count the total number of visits by user u for every category, then calculate the ratios of category visits. Finally, we choose the most popular places from these categories by these ratios. $r_{u,ca}$ is a visiting ratio of user u for category ca.

$$Score_{u,p} = \sum_{i=1}^{n}\{u, p\} \mid \forall u \in U \wedge p \in List_i \wedge List_i \subset ca_i$$

$$size\left(List_i\right) \propto r_{u,ca_i}$$

- Place time range frequency: We divided the day into four different time ranges and we calculated each frequency with how many check-ins were made in these time ranges. We did not use this feature individually for prediction, but used it later for the proposed prediction method along with the other features.

4.1.2. Prediction Using Proposed Method

After testing the individual features, we obtained the first results. Then, we had the foresight to use and combine these features. We tried to use place popularity and time range frequency together. In this way, we can observe the most popular places in different time ranges. We tested the combined features and it gave better results than using both the individual features. We also wanted to use former visits because it always gets the best results among all the individual features. Finally, we can use category prediction because it simply gives us an opinion about the user's preference.

For reducing candidate places and better simulating of the real location-based social networks, we decided to use check-in location as the starting point. Then, we found the top 1,000 closest locations to this starting point. In this way, we can cover a circle around the starting point with an average of 2–

4 miles of radii for different cities. We made a decision to use this because location-based social networks normally know the location of users and use this information for venue recommendation. When we start to give scores to places, we use this subset of places with 1,000 locations.

Finally, we decided to use the list as two equal pieces. The first half gets candidate places from the user's former visits, while the second half gets candidate places from the categories which the user visited. For selecting places from the categories, we used popularity in the time range in which check-in was made. Then we normalised both halves and merged them. We used the final list as a prediction list and tested with different list sizes on both the datasets.

Let us assume that $List_A$ and $List_B$ are of equal size and are half the prediction list. p_t represents the place popularity at the time range when specified check-in was made. p_c is the place where check-in was made. Then we can formalise the proposed method as follows:

$$Score_{u,p1} = \sum_{i=1}^{n} \{u, p_t\} \mid \forall u \in U \wedge p \in List_{A_i} \wedge List_{A_i} \subset ca_i \wedge p \in close(p_c, 1000)$$

$$size(List_{A_i}) \propto r_{u,ca_i}$$

$$Score_{u,p2} = \Sigma\{u, p\} \mid \exists u \in U \wedge p \in close(p_c, 1000)$$

$$List_A = rank(norm(Score_{u,p1}))$$

$$List_B = rank(norm(Score_{u,p2}))$$

$$List = List_A + List_B$$

$$rank(List)$$

4.2. Methodology and Metrics

We use leave-one-out cross-validation for testing the system. We remove the selected check-in from the dataset and use the rest for the training. Then we try to predict the place of the selected check-in. We use the same method for the entire dataset.

We use two different metrics, hit and precision. Hit is counted as 1, when one of places in the prediction list is the selected check-in's place. If none of our predictions is correct, then hit is counted as 0. If we assume that the selected place of check-in is p and our prediction list is L, then the formula can be generated as

$$Hit = \begin{cases} 1, & p \in L \\ 0, & p \notin L \end{cases}$$

Precision is defined as hit accuracy. When our correct prediction is in the first place in the prediction list, precision is 1 and it decreases to 0 while our hit moves to the last place in the prediction list. If k prediction in the list is correct, then the formula can be generated as

$$Precision = \frac{|N| - k + 1}{|N|}$$

We calculate hit and precision for every check-in in the datasets for testing. Then we average the results and obtain the average hit and average precision. We use these average values for comparing the features of prediction performance.

5. Results

We test both the datasets with the obtained features and the proposed method and get the results. Figure 1 shows the Gowalla prediction results and Figure 2 shows the Foursquare prediction results. While the prediction list size increases, every feature's average hit value gets better. On the other hand, the average precision value generally gets worse. Former place visits feature dominates all the other features while getting the best hit results from every list size. This approach works well because, if a user visits the same place more than once, it is very likely it would be a hit on the prediction. On the Gowalla results, close popular and category-based hit results are close to each other, but close popular has better precision. On Foursquare, the closest feature gets the second best results. Popularity has the worse results on both the datasets. Using just popularity for prediction is obviously a bad idea while there are thousands of different places in datasets.

Our proposed method gets usually the best results over all of the individual features. It gets the best average hit and average precision result combination on every list size for both the datasets. On Gowalla results, when the list size is over 60, former visits get better hit than the proposed method, but not better precision. When we evaluate the results using both the metrics, we can still say that the proposed method is better than the former visits feature.

We can also infer that almost all the prediction techniques obtain better results on the Foursquare dataset. This could be explained by the fact that the Foursquare dataset is more recent than Gowalla and it has more check-ins that we can use for training the system.

6. Conclusion and Future Works

This work has studied the next check-in prediction problem in location-based social networks. We used large amounts of data from the Foursquare and Gowalla location-based social networks. First, we analysed the datasets and obtained some features for prediction. Then we proposed a new method, which combines the individual features and compared the results. We can infer from these results that using features individually is not a good strategy, while it only covers only one side of the problem. Our proposed method uses features together; this covers the problem more completely and thus obtains the best results.

In the future work, we can obtain much more information from location-based social networks and use them for better predictions. While these networks are growing, we can also have much more stable and bigger datasets. Using the user's comments and ratings, we can learn the user's preferences more accurately. With different techniques and better datasets, the problem would be much easier to solve.

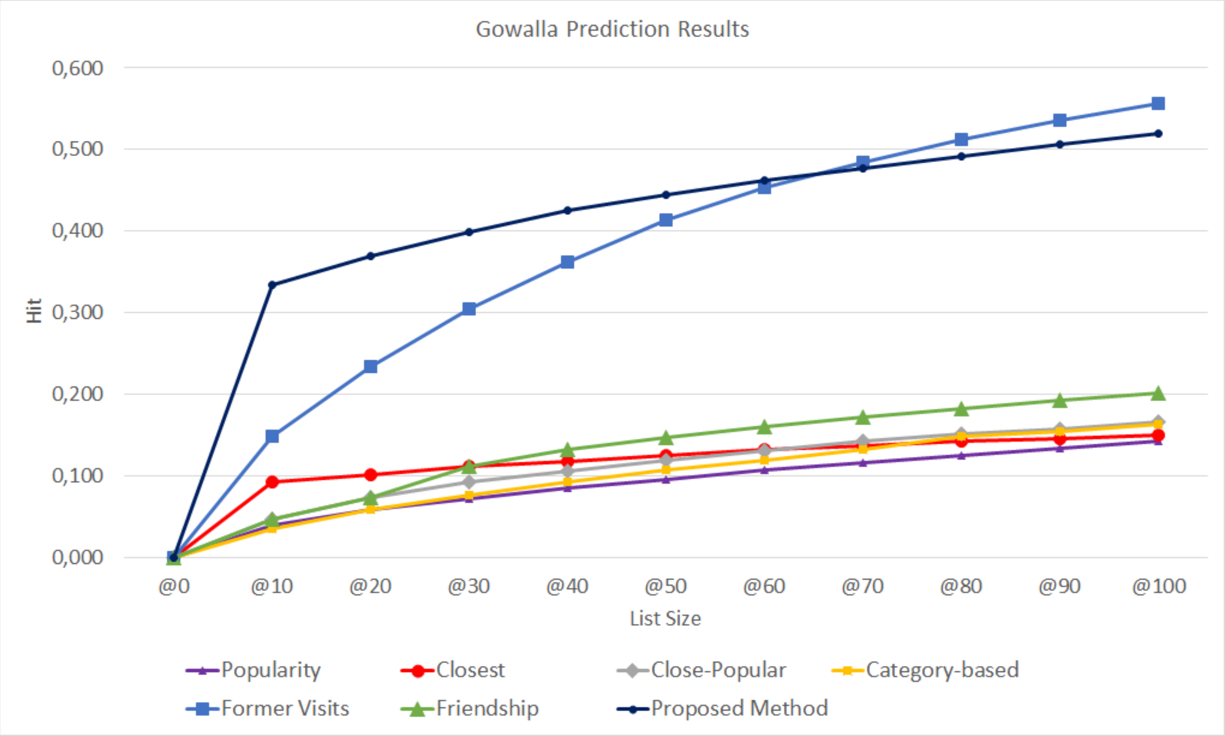

Figure 1. Gowalla prediction results

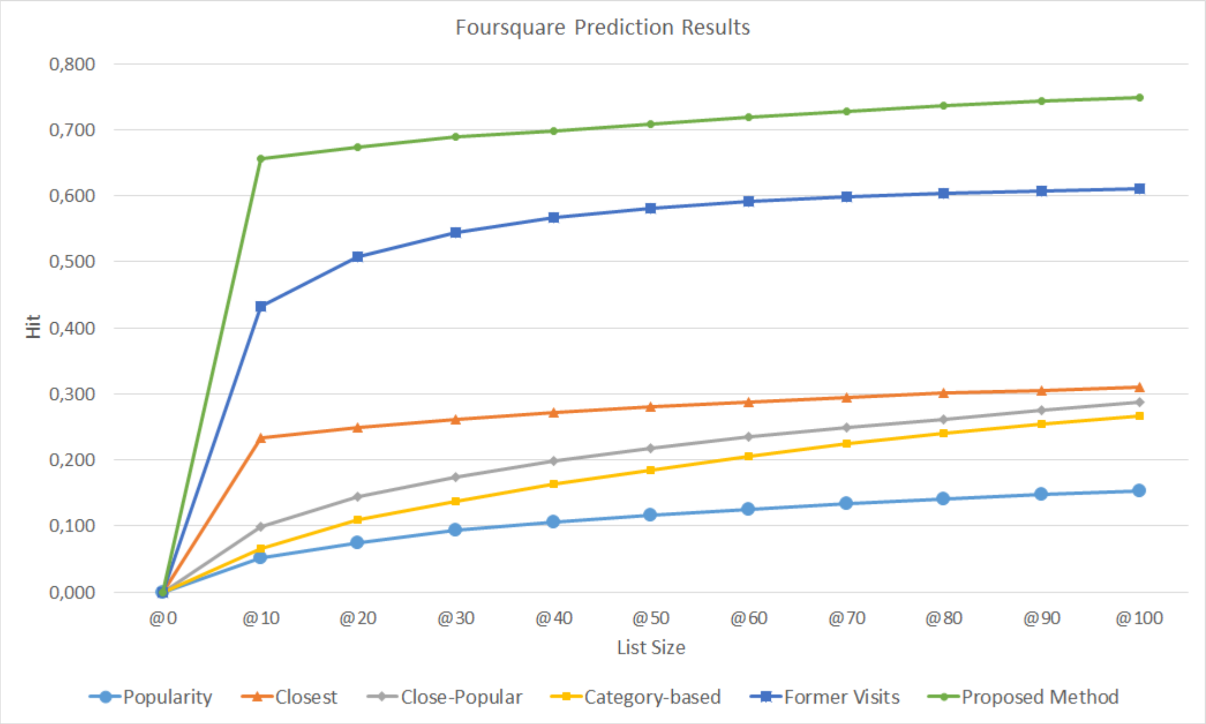

Figure 2. Foursquare prediction results

References

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.

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international conference on social computing* (socialcom) (pp. 144-153).

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th International Conference* (pp. 1038-1043).

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. *ICwSM, 11*, 70-573.

Park, M. H., Hong, J. H., & Cho, S. B. (2007). Location-based recommendation system using bayesian user's preference model in mobile devices. In *International Conference on Ubiquitous Intelligence and Computing* (pp. 1130-1139). Springer, Berlin, Heidelberg.

Wang, H., Terrovitis, M., & Mamoulis, N. (2013). Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 374-383).

Ye, J., Zhu, Z., & Cheng, H. (2013). What's your next move: User activity prediction in location-based social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 171-179).

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* (pp. 458-461).