**Location prediction in location-based social**

Abstract

Depending on the developments in mobile devices and wireless networks, location-based social networks have been gaining popularity in recent years. These social networks allow their users to explore new places and share their location, videos and photos and make friends. Location-based social networks also help us to get huge amount of useful information about the mobility of users. This information can be used to improve location-based social networks by letting them make good assumptions about users’ behaviors.

This paper studies the problem of predicting the next check-in of users who are using location-based social networks. For making good predictions first we analyze the datasets we have which is based on the check-in data and place data. Datasets are obtained from two popular social networks, which are Foursquare and Gowalla. Gowalla dataset has about 10 million check-ins made by 100 thousand users on 1 million different places from all over the world while Foursquare dataset only covers London check-ins and has about 500 thousand check-ins made by 10 thousand users. After analyzing the datasets we obtain some features that we can use for prediction and acquire deep understanding of users’ behaviors. Features are place popularity, place popular time range, place distance to user’s home, users’ past visits, category preferences and friendships. First, we use features individually and make predictions. Then we propose a new method which uses all features together to make predictions. We use hit and precision as metrics. Hit can be described as whether one of our prediction in the prediction list is correct or not. Precision is defined as hit accuracy. When our correct prediction is in the first place in prediction list, precision is 1 and it decreases to 0 while our hit moves to last place in the prediction list. Finally, we compare all the results we acquire and we observe the improvement with new method.

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

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 the people’s needs. Recently a new type of social network has started to become popular, location-based social network. While users can share photos, videos and their opinions like 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 grow their popularity and public recognition. Location-based social networks contain huge amount of useful information about the mobility of users. This information can be used to improve location-based social networks by letting them make good predictions about users’ behaviors. These predictions help three sides of the networks. Social network providers can earn their users’ and commercial advertisers’ trust. While users’ loyalty and pleasure to these networks are growing, commercial advertisers’ investments are becoming strong on these networks.

1. Problem Definition

In this work the next check-in prediction problem is studied. We are trying to predict the next check-in of users with the help of features obtained from the database preprocessing. Check-in data contain 3 information, which are user, location and time. While database contains more than thousands of different places it is hard to guess check-in with just one prediction. To overcome this problem we make a list of predictions with different sizes and compare results. For every check-in we are trying to predict, we take the check-in from the database and we use rest for training and try to predict chosen check-in.

In previous works the problem is handled with different approaches. For location prediction Bayesian networks are used in [1] and collaborative filtering is used in [2, 3]. A study on user activity patterns is published in [4], which gives deep understanding of location-based services and user activities. Random walk algorithm is used in [5], which is studying new venue recommendation problem. User mobility, global mobility and temporal features are used for location prediction in [6]. New venue recommendation is also studied in [7] and specialized bookmark-coloring algorithm is used in the study. A study on location prediction offers two steps method [8], in the first step system tries to predict location category, then in the second step system tries to find location from the predicted category.

1. Data Analysis

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

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

The first information we obtain is users’ and places’ check-in numbers. Later we can use this information for place popularity and user mobility frequency. Users with less than 9 check-ins for the whole time are considered as inactive and these users are removed from the database. As far as we are trying to build a personalized recommendation system, users with few check-ins would be useless in this context. Places with only 1 check-in are also deleted. Then we label all places with city information using another database which contains cities’ latitude and longitude values. Separating database into different cities will help both reducing computation time and increasing prediction accuracy. While we don’t have any information about users’ home locations, we assume that most frequently visited place is the home location of the user. We are going to use this location as starting point and calculate every other places’ distances to this point. We split one day into four time range and calculate every places’ visiting frequency 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 5 different features that we can use for prediction. The features are users’ previous visiting places and categories, place popularity and time range frequencies and users’ distances to places.

Table 1. Total number of check-ins, total number of users, total number of places, average number of check-ins by user and average number of check-ins per place from 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 |

1. Check-in Prediction

This section formalizes the next check-in prediction problem. Given the current information about database, we aim to predict the next check-in place of users. We are calculating a score for every place in the system for specified user whose next check-in place is trying to predict and then we are ranking them according to the scores. In the end, we are getting the top most N scored places and we are checking if the next check-in place is in the list or not. We are using different list sizes from 10 to 100.

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

* 1. Problem Formulation

We have a set of users U and a set of places P. Each check-in c made by u ∊ U is defined as a tuple {p,t}, where p and t represent where and when check-in was made. C is defined as the total set of check-ins where Cu is the set of check-ins for a specific user u. Pu is the set of places and Cau is the set of categories which are visited by user u. And Fu represents the set of users, which are friend of user u. Given a user u, we a calculating the score for every p ∊ P and then we are ranking them according to these scores.

* + 1. Individual Features Prediction
* 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 specified place.
* Place Distance: We defined home location of user as user’s most visited place. Then we calculate the distances to all places in the same city with the user. Then score is calculated inversely proportional to distance. uhome is the home location of user u.
* Close Popular Places: The first two individual features are combined and used as a single prediction feature. First, we found the top 1000 closest places to users, then we ranked them according to their popularity.
* Former Place Visits: We are using user’s former visited places for prediction problem. Score of every place is simply how many times the user visited that place.
* Friendship: We count total number of visits by every friend of user u on every single place and give this count number as a score of this place.
* Place Time Range Frequency: We divided the day into 4 different time ranges and we calculate the each frequency with how many times check-ins made in these time ranges. We didn’t use this feature individually for prediction, but use it later for the proposed prediction method along with the other features.
  + 1. Proposed Method Prediction

After testing the individual features we obtained the first results. Then we had a foresight to use and combine which 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 this combined feature and it gets better results than both individual features. We also wanted to use former visits because it always gets the best results among all individual features. Finally, we can use category prediction because it simply gives us an opinion about user’s preference.

For reducing candidate places and better simulating the real location-based social networks, we decided to use check-in location as starting point. Then we find the top 1000 closest locations to this starting point. We decided to use this because location-based social networks normally know the location of users and using this information for venue recommendation. We can cover a circle around the starting point with an average 2-4 miles radius for different cities. When we are starting to give scores to places, we are going to use this subset of places with 1000 locations.

Finally, we decided to use list as two equal pieces. First piece gets candidate places from user’s former visits. Second piece gets candidate places from the categories which user visited. For selecting from the categories, we used popularity in the time range which check-in was made. Then we normalized both pieces and merged them. We used the final list as a prediction list and tested with different list sizes on both datasets.

Let’s assume ListA and ListB are equal size and half of the prediction list. pt represent the place popularity at time range when specified check-in was made. Then we can formalize the proposed method as follows:

* 1. 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 selected check-in’s place. We use the same method for the entire dataset.

We use two different metrics which are hit and precision. Hit is counted as 1 when one of our places in the prediction list is selected check-in’s place. If none of our predictions is correct, then hit is counted as 0. Precision is defined as hit accuracy. Let’s assume the selected check-in’s place is p and our prediction list is L. Then formula can be generated as:

When our correct prediction is in the first place in prediction list, precision is 1 and it decreases to 0 while our hit moves to last place in the prediction list. Let’s assume k. prediction in the list is correct. Then formula can be generated as:

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

1. Results

We test both datasets with obtained features and proposed method and we get results. Figure 1 shows Gowalla results and Figure 2 shows Foursquare results. While prediction list size is increasing, every feature’s average hit value is getting better. On the other side average precision value is generally getting worse. Former place visits feature dominates all other features while getting best hit results from every list size. This approach works well because, if a user visits same place more than once it is very likely it would be hit on prediction. On Gowalla results, close popular and category-based hit results are close to each other, but close popular has better precision. On Foursquare closest feature gets second best results. Popularity has worse results on both 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 individual features. It gets best average hit and average precision result combination on every list size for both datasets.On Gowalla results, when list size is over 60 former visits gets better hit than proposed method but not better precision. When we evaluate results using both metrics, we can still say proposed method is better than former visits feature.

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

1. Conclusion and Feature Works

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

In 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 a much more stable and bigger datasets. Using 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.

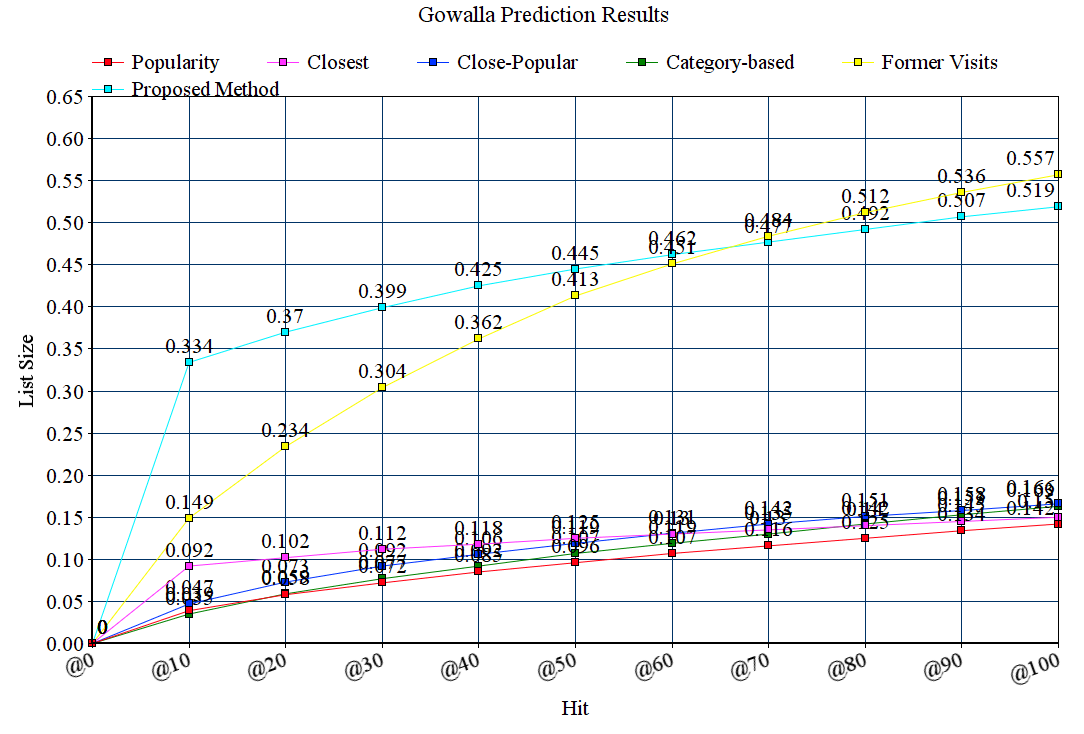


Figure 1. Gowalla prediction results

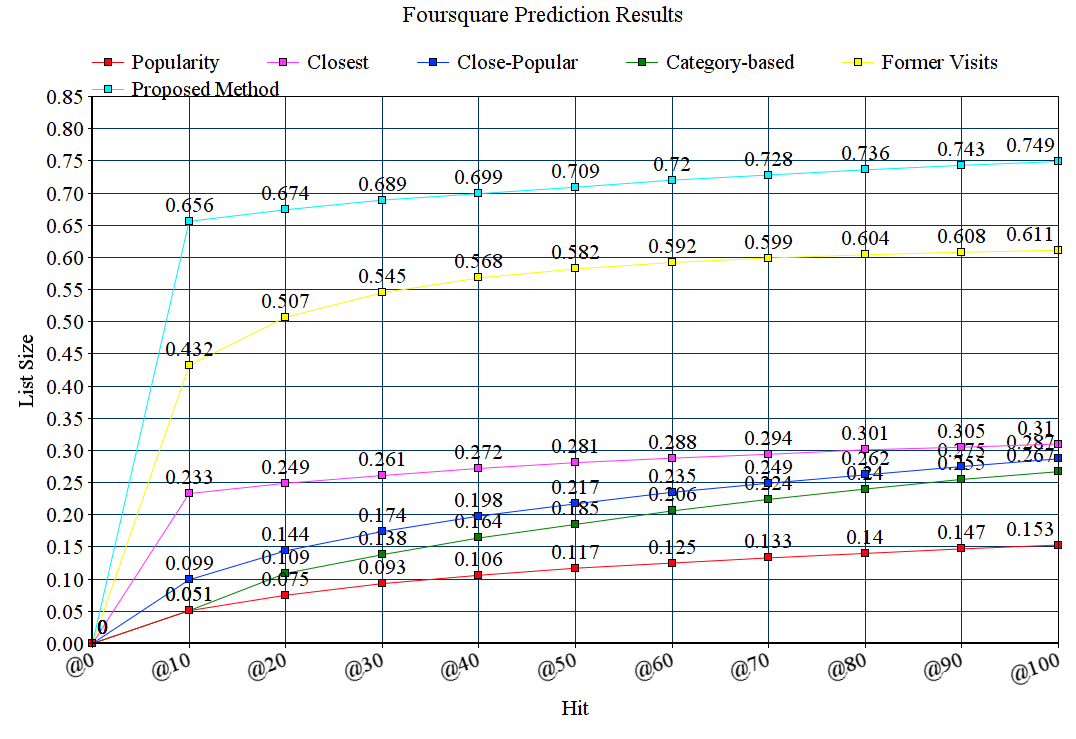


Figure 2. Foursquare prediction results

References

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

Ye, M., Yin, P., & Lee, W. C. (2010, November). Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 458-461). ACM.

Berjani, B., & Strufe, T. (2011, April). A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems* (p. 4). ACM.

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An Empirical Study of Geographic User Activity Patterns in Foursquare. *ICwSM*, *11*, 70-573.

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012, September). 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 Confernece on Social Computing (SocialCom)* (pp. 144-153). IEEE.

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

Wang, H., Terrovitis, M., & Mamoulis, N. (2013, November). 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). ACM.

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