COMMUNITY MINING AND RECOMMENDATION SYSTEM
BASED ON LOCATION, USER INTEREST, AND SPATIO-TEMPORAL TRAJECTORY

Zhao Yue 16307090084 **Feng Mengdi 16307100076** **Zhang Yuqing 16307130308**

Abstract

Location-based social network (LBSN) keeps a large number of spatial-temporal behavior track data of users. In our project, we choose to use foursquare Tokyo check-in dataset and make community discovery and division. Then we visualize the temporal and spatial trajectories of places of interest and make a deeper interpretation from the social perspective based on the analysis results. Also, we build a recommendation system and make an User Interface.

KEYWORD LSBN Community Discovery Recommendation System

1 INTRODUCTION

1.1 Introduction to Location-based Social Networking(LSBN)

LBSN, a social network based on location service, is a combination of location and traditional social network. Two core attributes of LSBN are location and user. With the development of the Internet and positioning technology, a large number of users generate a lot of location information through check-in, and the collection of these location information constitutes the spatial-temporal behavior track of users. LBSN keeps a large number of spatial-temporal behavior track data of users.
These data contain rich spatial structure information and user behavior information. Valuable knowledge in these data can be obtained by analyzing and utilizing data mining. In the traditional analysis of the characteristics of social network, it is mainly the non-geographic topological network based on graph theory, without considering the influence of geographic space. However, subsequent studies have shown that there is a strong correlation between the spatial attributes of user nodes and network performance, and a large number of users' social network relationships are related to their geographical location, indicating that geographic space has a restrictive effect on social networks.

1.2 Brief Introduction of Our Work

Through the analysis of the data of user's behavior trajectory, the methods of user activity prediction and user interest area discovery are obtained.

1. First,we preprocess the data and find the similarity matrix between users based on location, category and track of time and space.
2. Then we do the community division using *Walkstrap* algorithms with the tool *Gephi*.
3. After the community division, we visualize the track of users using packages *plotly* and *dash* in *python*.
4. Combined with the above factors, we have managed to conduct a recommendation system and make a User Interface using *PyQt5* in *Python*.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2 DATA

2.1 Introduction of Dataset

In our project, we choose to use foursquare Tokyo check-in dataset *dataset_TSMC2014_TKY.csv* provided in Kaggle website. It includes 2293 users and 537303 check-in records from April 12, 2012 to Feb 16,2013.
Foursquare is a location-based mobile service that encourages mobile users to share information about their current location with others. The dataset contains 8 columns of data:

* User ID
* Venue ID
* Venue category ID
* Venue category name
* Latitude
* Longitude
* Timezone offset in minutes
* UTC time

As shown in the figure below, the geographical location of the whole data set is circular. Therefore, it is inferred that this dataset is not a dataset that takes the geographical location of Tokyo as the check-in place alone, but a dataset that makes a circle centered around the center of Tokyo city and covers the data of Tokyo map as much as possible.

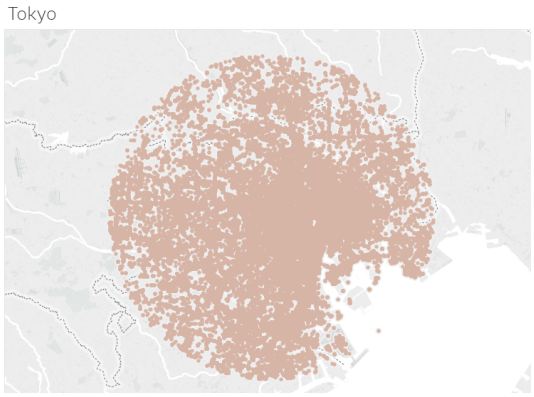


Figure 1: Tokyo check-in data on map

The below figure shows the distribution of check-in times. It can be seen from the figure that a large number of users check-in between 100-200, and only a few users check-in more than 1000 times in the past year.

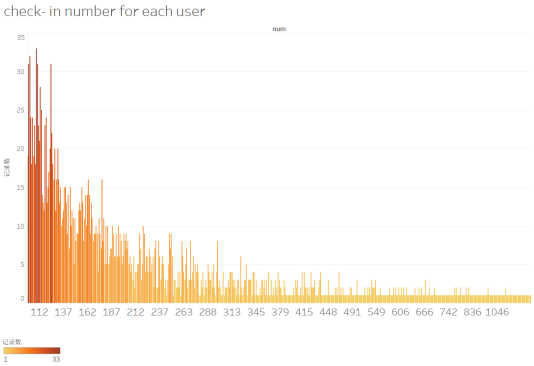


Figure 2: Check-in number for each user

Nearly 200,000 pieces of data are collected at train stations and 40,000 pieces of data are collected at subway stations, which is in line with the traffic habits of Tokyo residents, namely, extensive use of JR train and subway.

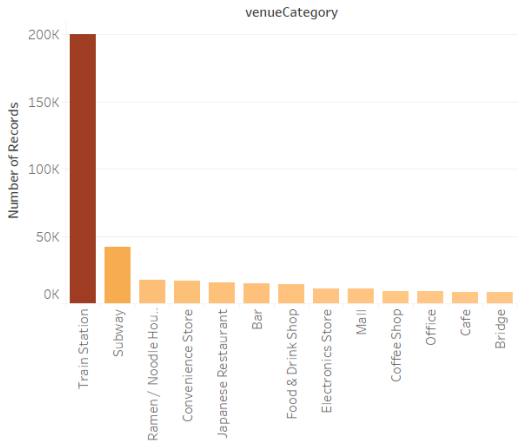
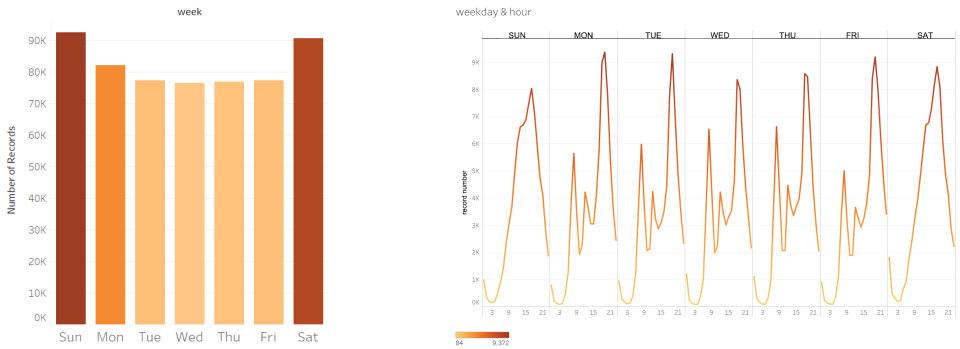


Figure 3: Venue category distribution

In the time series of check-in data, it can be observed that users have more check-in data on weekends than on weekdays. In weekdays, there is an obvious three-phase peak , which occurs in the morning, noon and evening. While on weekends, the data distribution is comparatively more average – less in the morning but grows steadily to the peak in the evening.



(a) changing rule in one week (b) changing rule in one day

Figure 4: changing rule in day and week

Additionally, from the classification of interests and the time of check-in, it can be concluded that there are a group of people in the data set of Tokyo who are very obvious in line with the characteristics of office workers, that is, they go to work in the morning, leave work in the evening, eat out at noon, and commute by train or subway.

2.2 Data Processing

A.Time

Since UTC Time is used in this dataset, standard Time needs to be converted to Time in zone of Japan firstly, and then we extract year, month, day, hour, and minute respectively from the correct time.

B.Center of Venues

Rank the check-in data records of each user according to the number of Venue ID, and take the

five sites with the most check-in data records of each user. We can obtain the center place of these five specific sites as the central location of each user.

C.Distance

Calculate the distance between the center point and all the places that the user goes to, and get the distance between a certain place that a user goes to and the center point of the user.

D.Average Distance

Sort the check-in data of a user by time, calculate the distance between each two adjacent locations, and divide by the number of locations to get the average distance of a user.
When calculating these distances, because the original data are the coordinates of the earth(Latitude and Longitude), the distance should be converted into the spherical distance to obtain more accurate values.

3 USER SIMILARITY AND COMMUNITY DISCOVERY

There are three dimensions of data in the data set, one is the location of longitude and latitude, one is the check-in time, and the other is the venue category.
Also, three computing dimensions of user similarity can be obtained. One is location-based user similarity, the other is location category-based user similarity, and the last one combines location and time to obtain location-based user similarity.

3.1 User Similarity Based On Location Or Venue Category

First, we calculate the time of each user visiting each location and write it as *CheckInTimes.csv*. Then we calculate the similarity using Pearson coefficient:

$$sim(U_u, U_v) = \frac{\sum_i (r_{u,i} - \overline{r}_v)(r_{v,i} - \overline{r}_u)}{\sqrt{(r_{u,i} - \overline{r}_u)^2}\sqrt{(r_{v,i} - \overline{r}_v)^2}}$$

However, the location-based similarity matrix is too sparse, because the similarity between most users is 0.
Then we can use the exactly same method in similarity matrix on Venue Category, which is not so sparse as above.

3.2 User Similarity Based On Track of Time And Space

3.2.1 Algorithms

As mentioned above, the majority of users check in about 100-200 times of the total in a year, most of which among the data are in the train station. And it has obvious classification characteristics between workday and weekend.
Therefore on the time dimension, we choose one day of users' trajectory instead of one week. The main reason is that in seven days a week, most of the users with lower check-in frequency are likely to have all the records in seven days at the train station.
We divide the location visited by the user into 48 time dimension in accordance with the time, of which 24 0:00-1:00 to 23:00-0:00. As a result that the 24 dimension time division method has certain limitations, When a user's check-in time is in the critical area between two dimensions, it is easy to produce a fuzzy division. Therefore we introduced another 24 dimension for half hour, namely 0:30-1:30 to 23:30-0:30, which delay for 30 minutes correspondingly. This is equivalent to a sliding window to increase the flexibility of time matching.
We get the place where users go most in each time dimension as the feature location in this time dimension. If no place has been visited during this period, it will be empty for this period. Finally, the check-in records of each user for nearly a year are compressed into feature tracks in 48 time dimensions of a day.

For two users, calculate the distance of the characteristic location in each time dimension:

$$SeqSim(Seq) = \frac{1}{1 + distance(C_i, C_i')}$$

And calculate the average distance:

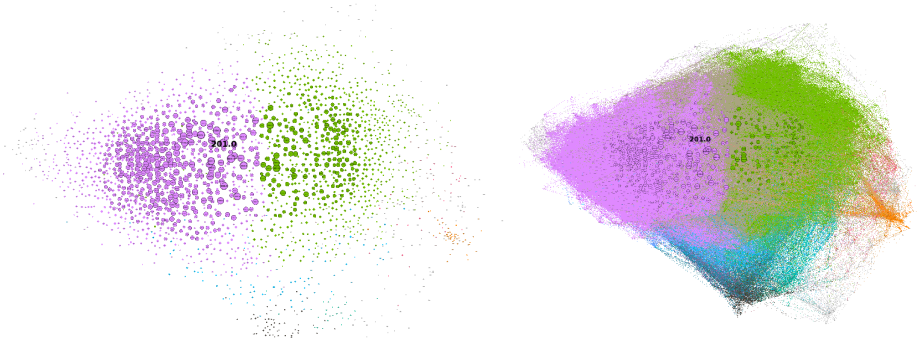
$$UserSim(LocTra1, LocTra2) = \frac{\sum_{j=1}^{m} SeqSim(Seq[j])}{m}$$

Finally, a matrix of 2293*2293 is obtained to represent the trajectory similarity between users.

Table 1: 6*6 similarity matrix

1	0.09488	0.055849	0.374954	0.531905	0.120881
0.092523	1	0.114348	0.114747	0.145113	0.13359
0.055829	0.119279	1	0.057623	0.080785	0.103615
0.367305	0.116135	0.057613	1	0.340955	0.177097
0.531559	0.149628	0.080522	0.34625	1	0.167533
0.119353	0.155191	0.109138	0.178677	0.16741	1

3.2.2 Classification Result



(a) classification result(only node) (b) classification result(nodes and edges)

Figure 5: changing rule in day and week

For classification, we used the obtained user similarity matrix as above and set 0.2 as the threshold: the edge less than 0.2 was set as 0.
Walkstrap and *spinglass* algorithms were used to obtain community classification respectively. Since this data has no criteria for community classification, it is an unsupervised learning with no real answers. Therefore, in the final evaluation, more attention is paid to the readability evaluation of the results than the clustering quality.
The final community adopted walkstrap algorithm with step number of 4, and the classification results were visualized as above (including edges in the figure(b). In this classification result, there were a total of 2293 nodes and 488,725 edges, which were divided into 26 categories, among which there are 10 clusters with more nodes.

4 VISIUALIZATION RESULT AND ITS INTERPRETION

4.1 Visualization display of user track

We use packages *plotly* and *dash* in python to make an interactive visual tool.

1.run the code trackofmap.py and then copy the website URL to browser

2.after reloading the page, we can see the webpage below. In the legend on the right side of the page, the color indicates the user's feature location within 24 hours, as well as its central location(the big yellow dot).

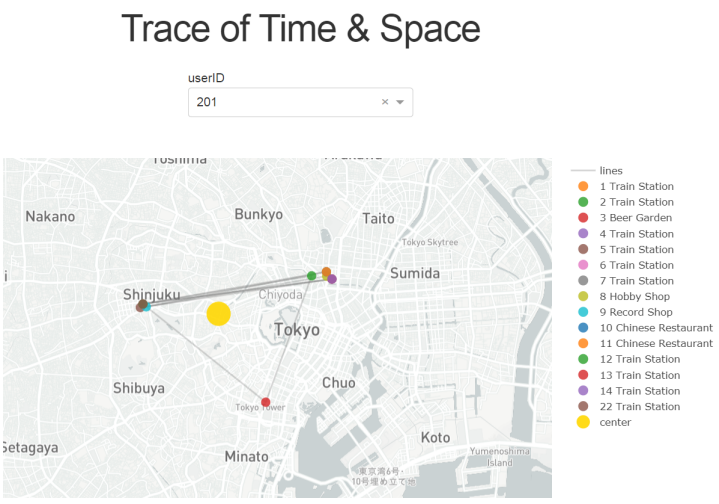


Figure 6: UserId:201

4.2 Track interpretation

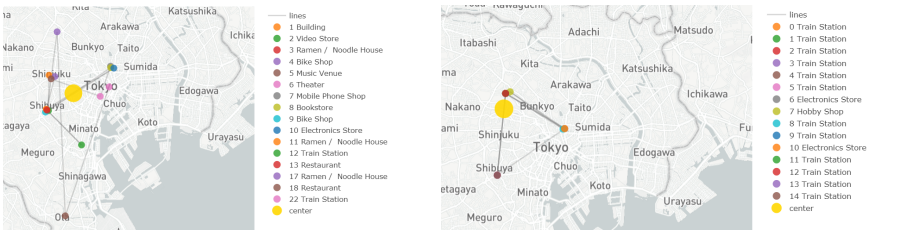
Previously, through the trajectory map of time and space, it can be found that there are two categories of groups, namely the groups marked with pink and green. By looking at the characteristic trajectory map, the trajectories of these two types of users are relatively similar.

1) The first class
   The typical user in the first class is 201he realy has "enormous" degree. We can find that the whole track map is a standard triangle, and the contents are mostly railway stations, one of the three types of points is a bar at 3 am, and the other two are transportation hubs of Shinjuku and Akihabara.
   Additionally,Users 947, 2071, 395, and 303 in the first category are similar as well.One spatio-temporal feature is around 6 to 12 o 'clock in akihabara, and there is few evening data.Comparing 395 users with 303 users, the two users are concentrated in two areas, one is akihabara area, the content is mainly university (Kyoto university, Meiji university), akihabara (video store, etc.), the other is in shinjuku.User 395 frequently enjoys himself in bars and concert halls, while user 303 frequently visit coffee shops, restaurants and department stores.

(a) UserId:947 (b) UserId:395 (c) UserId:303

Figure 7: The first class

2) The second class
   Where users in category 2 are quite close to category 1 users is that they all incline to appear in akihabara around 8 to 12 o'clock. Their unique characteristics are that around 2 PM they like to be in shibuya nearby, and often appear in shinjuku to Toshima among this area (Toshima region: one district in Tokyo).



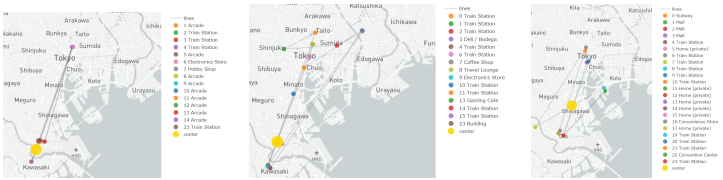
(a) UserId:651 (b) UserId:741

(c) UserId:1878 (d) UserId:1537

Figure 8: The second class

To summarize, the first type of feature is concentrated in from akihabara to shinjuku, then the second type of feature is akihabara - shibuya - shinjuku - Toshima.
Apart from these two categories, the remaining categories are very small.

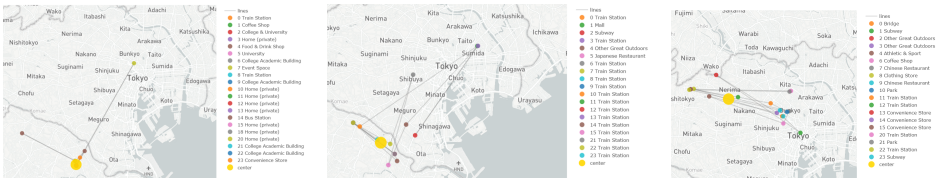
3) The third and fourth class
The third category is dominated by user 318, user 319 and user 1617, with obvious characteristics. According to the similarity of space and time, these users are obviously close to each other. However, if we look closely at the places they go to, we will find that they are almost completely different users.



(a) UserId:318 (b) UserId:319 (c) UserId:1617

Figure 9: The third class

Similarly, 1317 and 503 users are mostly located in Keio University, while 503 users are located in the area next to Keio University. They are similar in space-time trajectory, but their places and identities are obviously different.



(a) UserId:503 (b) UserId:1317 (c) UserId:2290(A separate class)

Figure 10: The fourth class and a separate class

These gaps indicate the deficiencies in the community division based on the track alone. The results of the division are better in time and track. However, from the point of view of location classification (that is, interest), some of them are wrong classification, which will lead to errors in the recommendation system later.
There are also some users in a separate category, such as user 2290. According to the track map, shopping and convenience store records are often found in remote horse training area, while restaurants, subway stations and other records are on the other side, so it is inferred that it is a person who lives in the suburbs and works near Shinjuku.

5 LOCATION RECOMMENDATION SYSTEM

5.1 Introduction of recommendation system

In LBSN (Location-based Social Networks), Users and POIs(point of interest) are two essential types of entities. In our project, we tend to explore three concepts of influence to design a recommendation system:

A. geographic location
B. user interest
C. track of time and space

Designing a good recommendation system needs to combine these factors. But it requires a huge data set to train to know how much weight to assigne to these factors is proper, and our data set can not complete this task. Therefore, we try to integrate these factors in recommendation processes rather than algorithms.

5.2 Idea of recommendation

As we learned in the lectures, user-based or item-based collaborative filtering techniques may be applicable to POI recommendation system. Since the data set doees not offer us detailed information of items(or to say the information of the places), so we mainly based on user-orientation collaborative filtering.

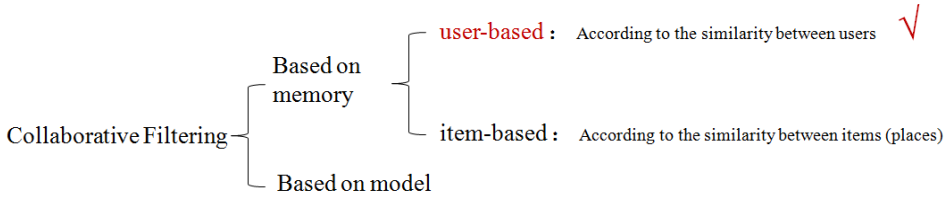


Figure 11: Classification of Collaborative Filtering

According to *Toblers First Law of Geography* that 'everything is related to everything else, but near things are more related than distant things', we can draw the conclusions that people tend to visit POIs close to their homes or offices. So the geographic features will be highlighted in our recommendation system.

Based on the law and the data and information we obtain, our idea of recommendation is below:

Step 1 Firstly, we first use the similarity matrix to find his or her similar users.

Step 2 Sum of users check-in times at each point and get the most frequented locations for similar users.

Step 3 Based on geographical preferences, we can filter the result above. There are two parts of geographical preferences: a.Geographic center b.Average travel distance.

Step 4 Take the above factors into consideration, we can get the recommendation more accurately.

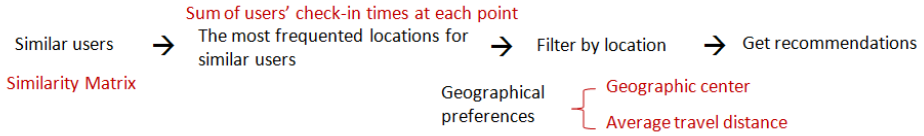


Figure 12: Recommandation Idea

5.3 Result of recommendation

	venueId	venueCategory	latitude	longitude	distance
1	4c787720dbaa76b0144f154b	Train Station	35.666963	139.758378	0.953398
2	4efd56c06c25c5ce7ee7f759	Hotel	35.698355	139.749599	0.919288
3	4b582e2bf964a5201a4d28e3	Park	35.683832	139.744592	0.822094
4	4b7bee49f964a5205c742fe3	Historic Site	35.683750	139.745604	0.713395
5	4c505692d49d76b0d4b2187b	Road	35.689666	139.751442	0.496750

(a) Based on venue category(major classes)

	venueId	venueCategory	latitude	longitude	distance
1	4b5e638ef964a520a58c29e3	Park	35.666383	139.752533	0.637988
2	4b600990f964a520e3d329e3	Train Station	35.666692	139.758133	0.941349
3	4c787720dbaa76b0144f154b	Train Station	35.666963	139.758378	0.953398
4	4b6a5612f964a52088d22be3	Fast Food Restaurant	35.667543	139.757322	0.845500
5	4de0b8a552b1c1b3ce39b9dc	Japanese Restaurant	35.664970	139.755471	0.817484

(b) Based on venue category(small classes)

	venueId	venueCategory	latitude	longitude	distance
1	4ba9be42f964a5207e363ae3	Office	35.700362	139.750428	0.987617
2	4b600990f964a520e3d329e3	Train Station	35.666692	139.758133	0.941349
3	4b835fdff964a5204e0431e3	Convenience Store	35.700734	139.750776	0.998990
4	4b46d0ebf964a520982826e3	Subway	35.695504	139.751705	0.761419
5	4e0333e4e4cdbd9a517691ce	Community College	35.676925	139.761125	1.049582

(c) Based on track similarity

	venueId	venueCategory	latitude	longitude	distance
1	4b600990f964a520e3d329e3	Train Station	35.666692	139.758133	0.941349
2	4b5ab0ddf964a520f5d028e3	Stadium	35.693204	139.749827	0.690288
3	4b443147f964a5208bf225e3	Train Station	35.695966	139.758067	1.052735
4	4d5216eec5ff6ea834feaa07	Ramen / Noodle House	35.701185	139.750235	1.026571
5	4b05879cf964a520549c22e3	Hotel	35.666867	139.743948	1.054748

(d) Based on location visited

Figure 13: Result of recommendation

Through the method above, we can use 4 kind of data to conduct our recommendation. Here we list first 5 recommendations:

A. venue category(major classes)

The result is with various category which cater for the users' taste, such as: hotel, park, historic site and so forth.

B. venue category(small classes)

These recommendations are based on small categories of which different restourants are in different categories.

C. track similarity

As we have calculated track similarity on the above section, we can find some interesting places for the users.

D. location visited

When based on location users have visited, the result may be more correct but a little bit boring

for the user.

We can see that only the recommendations based on category can find out places like 'Park' and 'Historic Site' that is worth going. Since 70% checkin points of our data set are places like 'Train Station' and 'Subway', it is not easy to pick out valuable places.

5.4 User interface

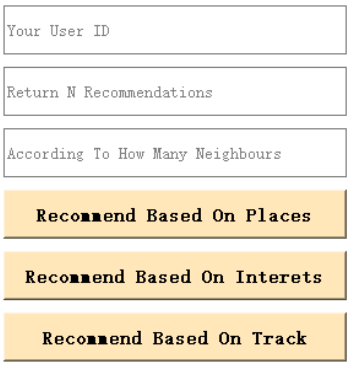


Figure 14: User Interface

We design the User Interface(UI) using *PyQt5* in Python.
There are three int figures you need to input:

* User Id
* Return N recommendations
* According to how many neighbors

Also, there are three recommendation options you can choose from:

* recommend based on places
* recommend based on Internets
* recommend based on track

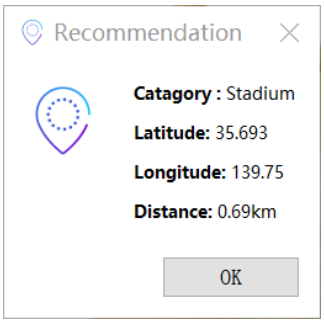


Figure 15: User Interface Result

5.5 Improvement of recommendation

There are two parts of improvement we currently come up with.
The first one focuses on similarity matrix: We can make full use of the characteristics of similarity

matrix, and make up for the unused factors in similarity in the recommendation algorithm.
To be specific, for similarity based on location, we may calculated the location type preferences of the weighted users in the recommendation ranking. We may give the users a weight w based on their similarity $sim(u, v)$.
And for track similarity, we may take the factor of time into account – At what point does the customer request a recommendation?

The second one focuses on filtering enhancement. To elaborate, we may choose not to recommend specific categories (like: transport). Also, for some categories (like: religion), we need to be more cautious during the filtering process.

And there are also some algoritms we can refer to.
1) The contact ratio of location and time
Suppose that the trajectory of user u can be represented by a combination of timestamp and location ID, like $< t1, l1 >, < t2, l2 >, , < tn, ln >$. Geographical distribution $GP(u, r)$ of user u can be calculated as:

$$GP(u, r) = \sum_{i=1}^{n} \frac{\delta(r, l_i(u))}{n(u)}$$

Combining the time factor, δT is the time precision (generally set at 1 hour), reflecting the proportion of all users in the same geographical location in adjacent time. At the same time, different weights of θ are set in working time and non-working time, taking into account the influencing factors of working time and non-working time. Then we can get the more accurate similarity between users:

$$CoL(u, r) = \frac{\sum_{i=1}^{n(u)} \sum_{j=1}^{n(v)} \theta(\Delta T - |T_i(u) - T_j(v)|)(\delta(l_i(u), l_i(v)))}{\sum_{i=1}^{n(u)} \sum_{j=1}^{n(v)} \theta(\Delta T - |T_i(u) - T_j(v)|)}$$

2) User relationship prediction model based on user social network and geographical location

$$MR(u, v) = \lambda \sum_{t \in F(u) \cup F(v)} \frac{1}{log(d_t} + (1 - \lambda) \sum_{r \in Loc} \frac{GP(u, r) \times GP(v, r)}{\|GP(u, r)\| \times \|GP(v, r)\|}$$

The first part is based on user social network, while in the rest, cosine similarity is obtained by "spatio-temporal slice" of the user.

6 SUMMARY

The final question is : **how to design a good recommendation system?**

First, we need an excellent dataset to support our algorithms.It may need:

* More complete check-in information
* Plentiful layers of venue categories
* User rating of location
* Average cost per location
* Social information between users
* Feedback on recommendations

Some of them, especially rating,categories and average cost, can be used in item-based collaborative filtering.

Secondly, after having a good dataset, we need to know how to deal with the information effectively. That is to say, we should try to know what customers are wanting and what customers will want in their mind.
Besides their behavior pattern, we need to guess their aims. For example, one may go to a restaurant for fill his stomach, but sometimes he may choose an expensive and decent one for enjoying and

feasting. So, time, place, consumption level, interests of users are all factors we need to take into consideration.

7 REFERENCE

1 W. R. Tobler. A Computer Movie Simulating UrbanGrowth in the Detroit Region. Economic Geography,46:234240, 1970.

2 Adamic L A, Adar E. Friends and neighbors on the Web[J]. Social Networks, 2003,25(3): 211-230.

3 Mao Ye1, Peifeng Yin1, Wang-Chien Lee1 Department of Computer Science and Engineering, The Pennsylvania State University, PA, USA. Department of Computer Science and Engineering, HKUST, Hong Kong Exploiting Geographical Influence for Collaborative Point-of-Interest Recommendation

8 APPENDIX

8.1 Task Of Group Members

- Zhao Yue 16307090084: community discovery and division; visual trajectory display
- Feng Mengdi 16307100076: recommendation system; bulding the User Interface
- Zhang Yuqing 16307130308: data processing; part of the algorithms; PPT & report writing

8.2 Acknowledgement

Wholeheartedly grateful to the instructor and TA's help, guidance and devotion in this semester.