Predicting next location and Recommend service using Geo-social data

Smriti Nigam,
Student
[1]Computer Science,
[1]PICT, Pune, INDIA.

Abstract : Location prediction is the key to many applications like traffic planning and controlling, weather forecasting, recommendation services according to the location like the hotel, food, homeland security, and travel recommendation.
Previously location prediction was done with the help of history or past moving pattern of an individual but it may fail sometimes because exact prediction doesn't reflect only past data. It divides into two phases, first phase analyses the individual preferences and the next phase is to find the social group in which an individual belongs.

In this dissertation work, the focus is to solve the location prediction problem using a semi-supervised approach. In this approach, the first phase identifies the social groups of a user by using HDBSCAN clustering algorithm with a geosocial dataset which has latitude and longitude information of a user. Haversine distance is used for finding the distance between geosocial points. The next phase applies Random forest classification in identified POIs to predict the correct location of the user. It also identifies preferred POI and what are the services available on that POI.

Impact of the system is to recommend location-based services to the user. Location-based services used in many ways, it can help to understand user mindset, with the help of rating and preference, it recommends the services to a user which enhances the business of the service providers.

It also provide ease to the user so that they can easily access their service which the user need.

IndexTerms - ***HDBSCAN clustering, Haversine Distance, Geosocial data, Location Prediction, Recommendation of Services.***

I. Introduction

Rapid growth in the acquisition of mobility data of the user. Collection of data from the online social network, mobility log, traffic data, wifi signal, GPS data, credit card transactions, and check-in data are possible.
\newline Previously we use historical data for human moving patterns and forecasting future locations, but now for predicting the future location, we can use their preferences and social group which they belong. Geo-tags and check-in data set are used for prediction their preferences and external groups.
\par The Geosocial dataset also provides latitude, longitude and time stamp information which can help for predicting the next location of the user.

According to existing studies, person moving patterns are highly regular and periodic. Person moving pattern is many times limited to several recurrent locations such as eatery places, offices, and house, but the movement of some depends upon external factor like the location of a taxi driver. Taxi driver location depends upon the POI where many people are going. For predicting correct future location human behavior is also an important factor. We can understand human behavior from two feature: a person's liking and his or her exterior social influences, so the location of the user depends upon the combination of internal factor and external factor~\cite{bm1}.
According to behavior theory, the location of the user depends upon a combination of internal factor and external factor. A person's decisions are generally driven by an individual's preferences, work category, routines, social contacts, professional contacts and advice from friends and family.

Location prediction is a combination of four steps in the first step we analysis the POI(point of interest) area by given dataset. A second step we analyze the user preference location by using clustering. The third step we identify the social group in which user interacts and its group movement or the preferred location. Fourth step we identify the combination of all three by using the content we identify the next location of the user.
Recommendation of services is the second aspect of the proposed work in which we identify preferences of a user which is going to a particular location. In which we will identify the services which provided at the particular location and what are the preferences of the user and find out the matched preferences of user and location by association rule ~\cite{bm7} and then send the request to the user that they can avail those services.
This application is useful for service provider companies which can easily access their customer which prefer their services.
There are many challenges in exact location prediction :

1. Many users cannot disclose their exact location in the social network and provide general information. This problem can be resolved by using network-based estimation.
2. Many user tags in multiple locations by uncleaned check-in data of social platforms, it increases the cost of data process and analysis.
3. Use of raw location data without any semantic information makes it hard to the personal purpose of daily route.
4. For recommendation if billion of product and million of a user so time complexity of the system is increased.
5. Some very similar items can have different names or contexts but recommendation system can not recognize it, for resolving this problem Collaborative Filtering (CF).
6. New user information not available to the recommendation is not appropriate.

II.RELATED WORK

Location prediction and service recommendation are done in two-phase in the first phase predict the correct location and second phase recommend the services.

A. Location Prediction

Mobility log and GPS trajectories take and find out stay point [1] and this stay point detects spatial feature. It gathers the Point of interest data, check-in data of user matches with a spatial feature, and fetch it a temporal and sequential feature which applied on decision tree modal and predict next location of a user.

In PSI approach [2] it find POI by clustering method. According to this algorithm user moving preference depends on its social groups. So it finds out each moving preference and its group moving pattern and combined with prefix span pattern and detect next location.

In using Ensemble method [3] use offline predict location which improving privacy and reducing power consumption through network usage.it breaks the user area into the region and finds the probability distribution for the likelihood of the trip. It uses the Markovian model for the non-region. GALLOP(global feature fused location prediction) [5] model it analyses the characteristic of all check-in and fetches the feature which finds the similarity by density estimation. It improves the robustness and more generality of the prediction method.

Probability-based location prediction [6] using cloaking to make user anonymous. For prediction uses historical user. ST-RNN [7] uses Times New Romanrecurrent neural network each layer of RNN is using with time-specific transition matrices and distance specific for different geographical distance and find more specific prediction by linear interpolation.This work uses Twitter API and finds the user information count of a follower and after applying Geo-coding API [10] find the location of the follower and apply k-means clustering to form cluster user based preferred location. Predict location by using Clustering.

B. Recommendation Services

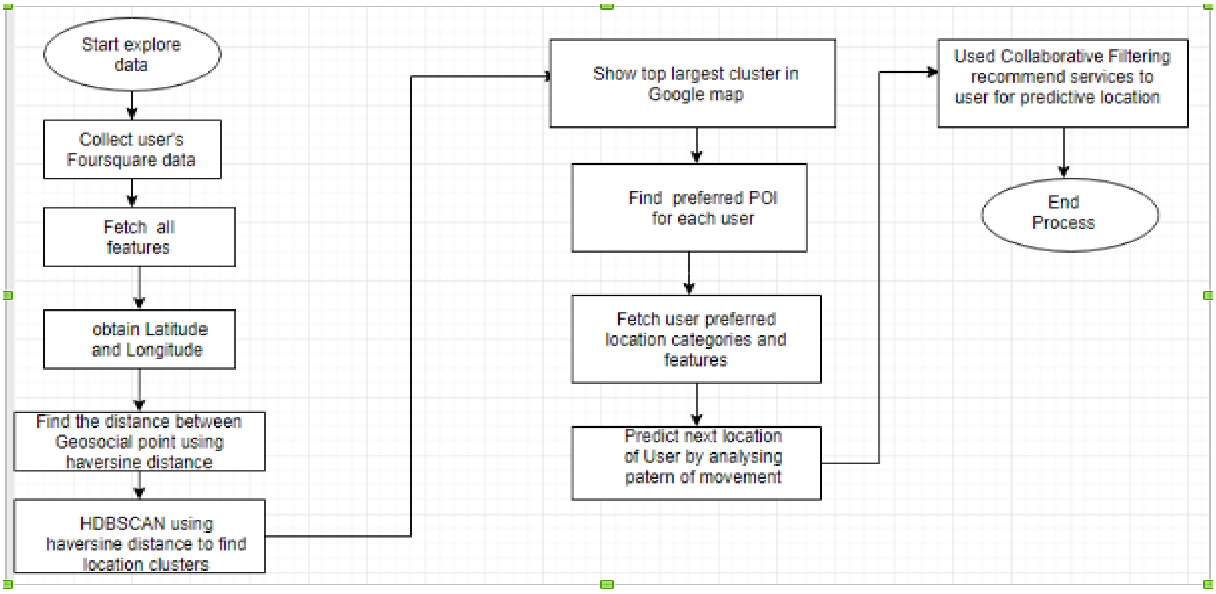
Web service recommendation aims to predict missing QoS (Quality-of-Service) values of Web services by utilizing the personalized influence of users [4] and services when measuring the similarity between users and between services along with Web service QoS factors, such as response time and throughput, usually depends on the locations of Web services and users. It uses a decision tree for analysis.The device has a unified rating prediction model by combine user and item geographical relationship, [8] user and user geographical relationship, and social interest similarity from social networks have location information of person using it which called as location depended on social networks. Recurrent model is personalized travel recommendation by analyzing social network photos and identify travel group. User profile information capture by Flickr and detect spatial and temporal feature and predict the interest by using probabilistic Bayesian learning framework [9].Mining people attribute by travel group algorithm [11] improved the user and item similarity approach by extracting the item feature and applying various item features weight to the item to confirm different item features. It uses collaborative filtering algorithm and detects the best match. It test performance with the help of Mean Absolute Error (MAE)

III. PROBLEM DEFINITION

To predict the future location of the individual with previous historical mobile datalog, and recommend the services related to his/her preference along with other user interest in the predictive geographical area.

IV. SYSTEM ARCHITECTURE

Spatial and temporal data of Foursquare is used for detecting the top POI which mostly preferred by the user. In Foursquare data has geographical distance.For finding distance between two points we have to find the length of the shortest curve between two points along the surface of the earth. Both DBSCAN and HDBSCAN algorithm is density-based spatial clustering method that group all point that is closely based on distance and also marks outlier which has low density and for distance measure using haversine distance which measures the great-circle distance between two points on a sphere given their longitudes and latitudes.Data is fetched from kaggle which is Foursquare data. It has temporal and geospatial features, users location information and time information. It fetches all feature which is used for predicting the location.


Start explore data
Collect user's Foursquare data
Fetch all features
obtain Latitude and Longitude
Find the distance between Geosocial point using haversine distance
HDBSCAN using haversine distance to find location clusters
Show top largest cluster in Google map
Find preferred POI for each user
Fetch user preferred location categories and features
Predict next location of User by analysing patern of movement
Used Collaborative Filtering recommend services to user for predictive location
End Process

A. Find POI Cluster

The POI cluster can be found by identifying the distance between geosocial points by using haversine distance and then using HDBSCAN clustering which is a hierarchical density- based clustering method that finds the cluster, based on the density. If the cluster size is larger, then we denoted the location as POI. For this system, the top 5 clusters are preferred as POI.

B. User Preferred POI

After fetching top POI identify each user-preferred POI, and POI which most users preferred. For a particular user, preference is identified by its preferred POI.

C. Predict Location

For finding user preference to time and location so we gather the time-series information of a user and identify its preferences over time using the random forest and predict its location.

D. Recommend Services:

To recommend the services to the user, it collects the preference of the user, category of locations where users have traveled in the past, and other location-based features. It applies collaborative filtering to find out the services in which the user might be interested in and recommends it.

V. SYSTEM ANALYSIS

To comprehensively study the performance of the algorithm,we conducted experiments on two real-world datasets: Portugal taxi GPS trajectory data and Foursquare dataset. All experiment was performed on a personal computer which has 3.5 GHz CPU and 8 GB of RAM using pycharm platform and python with machine learning libraries like pandas, numpy, sklearn, pickle, bokeh, etc.

VI. MATHMATICAL MODEL

The proposed system can be shown mathematically with components

S= (s,e,X,Y,FP,Ss,Sf)

s = start state of system

e = end state of system

Input X= (lat, lon, ta, tl,λ)

lat=latitudeoflocation

lon = longitude of location

ta = time to arrive

tl = time to leave

Output Y = Predicted Location

[FP]is Random forest classification Model ForPOI, Time series Function.

[Ss]is Success rate of system when it predict correct location and suggest correct services.

[Sf]is Failure accrue when correct location is predict

Objective Function

$V\mu(u,v)dudv=Y$

u = User Information.

v = Time

Y = Predicted Location at given time

Dataset: The portugal taxi dataset contains the trajectories of taxis from July 2013 to June 2014 in the city of Porto, Portugal.Foursquare dataset includes long-term (about 10 months) check-in data in New York City and Tokyo collected from Foursquare from 12 April 2012 to 16 February 2013 Dataset contains 227428 check-ins in New York City and 537703 check-ins in Tokyo. Each file contains 8 columns, which are:User ID (anonymized), Venue ID (Foursquare), Venue category ID (Foursquare), Venue category, name(Foursquare),Latitude, Longitude, Timezone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC), UTC time. There are many POI type find out with the help of categories.There are many types but in given system consider top 50 POI's. some types of POI are:

1)Home Location:

Home location is store as users private location. This location may be part of many users. It can not divert the POI. It's mainly used for prediction of the location not part of POI.

2)Work Location:

Work location is the office location of users. Work location is treated as POI but it depends on time. Work location reflects many users POI who work on that location and also who worked for users like- taxi driver, street food vendors, and many more.

3)Entertainment:

Entertainment area is like a movie theater, park, museum, and many more which is used for entertainment purpose.

4)Institute:

Institute is the POI for the student, workers, and faculty and its time is also fixed. It's used for predicting the location of student, faculty, and worker.

5)Departmental Store: Departmental store is the location of grocery place. Its opening hour is fixed so we can easily predict for that location.

6)

Other: There many categories which are the not part of top 50 that are defined in other section.

B. Methodology

1) Pre-processing: In this pre-processing step, the location data is captured and identified more influencing features that train the model. In pre-processing removes all the missing values and add the types of the categories in the table. It also divides the data into two parts train and test set. With the train data set, the train data set is used to train the model and testdata is used to test the system.

2) Clustering:The clustering method is finding the types of POI. For clustering identify the distance between two points.

Haversine distance is used for finding the distance between two geo-location points. HDBSCAN clustering method used for finding the cluster in the system. HDBSCAN has used density method which finds the cluster by the density of the cluster size.

3) Random forest: Random forest classification method used to identify the location of the user. Random forest is in sklearn,sci-kit-learn python tool kit which is applied on a data frame. Random forest gives train accuracy 89and test the accuracy of 50

4) Prediction:Random forest used for predicting the next location. In this method, we create multiple trees accordingto timestamp. Combining all trees using the boosting method and choose the best prediction.

5) System result evaluation: System is tested by some of the evaluation methods which are Accuracy, ROC curve, MAE, and R2 method.

C. Modules

This section describes the implementation details of modules in the system.

1) Pre-processing: It takes the input one or more data log, performs pre-processing functions on them and stores pre-processed results in a temporary directory.Pre-processing functions are: -Tokenizing, null values removal or find POI, etc.

2) Processing includes functions like feature extraction,Feature selection. In feature selection procedure is the selection of some subset of a learning algorithms input variables upon which it should focus attention while ignoring the rest. It is also called dimension reduction. It captures the relevant feature which is more influence the prediction result and discards the other features.

3) Clustering: Form cluster from the feature vectors and type of feature used as POI.

4) Concating: Concrete the cluster output to feature vector.

5) Classification: Classifies the feature vector based on training.

6) Calculate Accuracy: Accuracy calculation is done by equating the predicted label with true labels.

D. Proposed Algorithm

1) Load the data set Foursquare which has user basic information, latitude, longitude, and timestamp.

2) In the first phase fetch the features which are used for illustration the user preferences in a given dataset, like timestamp,user-id, and geo-location which is in the form of latitude and longitude. For finding POI we find top most visited place clusters which are identified by using the HDBSCAN clustering algorithm and Haversine Distance.

D. Proposed Algorithm

1) Load the data set Foursquare which has user basic information, latitude, longitude, and timestamp.

2) In the first phase fetch the features which are used for illustration the user preferences in a given dataset, like timestamp,user-id, and geo-location which is in the form of latitude and longitude. For finding POI we find top most visited place clusters which are identified by using the HDBSCAN clustering algorithm and Haversine Distance.

3) In HBDSCAN collect the top 5 clusters which are known as 5 POIs of a given location which shown in figure.size of cluster denoted the density of users.

4) To finding POIs forgiven data by using foursquare data set at the particular timestamp. It denoted the top-visited places of the given time, those are the location of the office, railway station, and airport location, etc.

5) To find the number of users in each POI and find the most preferred POI. That POI is shown with the help of the bar chart.

6) Find out each user’s preferred POI. It considered that individual moving preferences change dynamically. So we capture an individual moving pattern over time.

7) In the next step, we identify each user moving pattern concerning time and use random forest method for identifying its next location prediction

8) After predicting the particular location of a person, then we identify its preferences about services and apply collaborative filtering for recommending the services to the user.

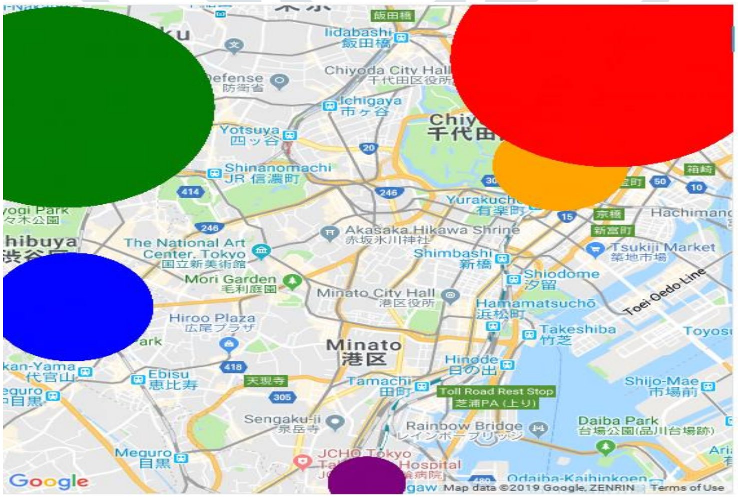
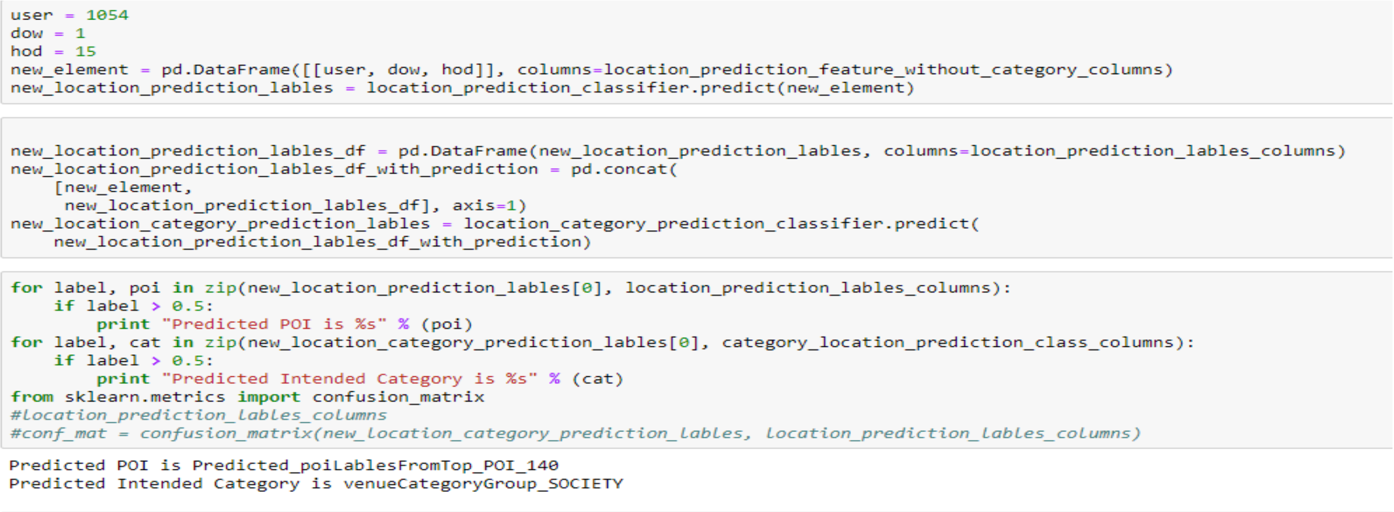


Fig. 2. Top most visited location of Tokyo

In Fig.3 is the result obtain by analysing the historical data by using random forest algorithm.It find out the POI as well as type of POI which can provide help in recommending the services to the user.

Fig. 3. Prediction Result



```
user = 1054
dow = 1
hod = 15
new_element = pd.DataFrame([[user, dow, hod]], columns=location_prediction_feature_without_category_columns)
new_location_prediction_lables = location_prediction_classifier.predict(new_element)
```

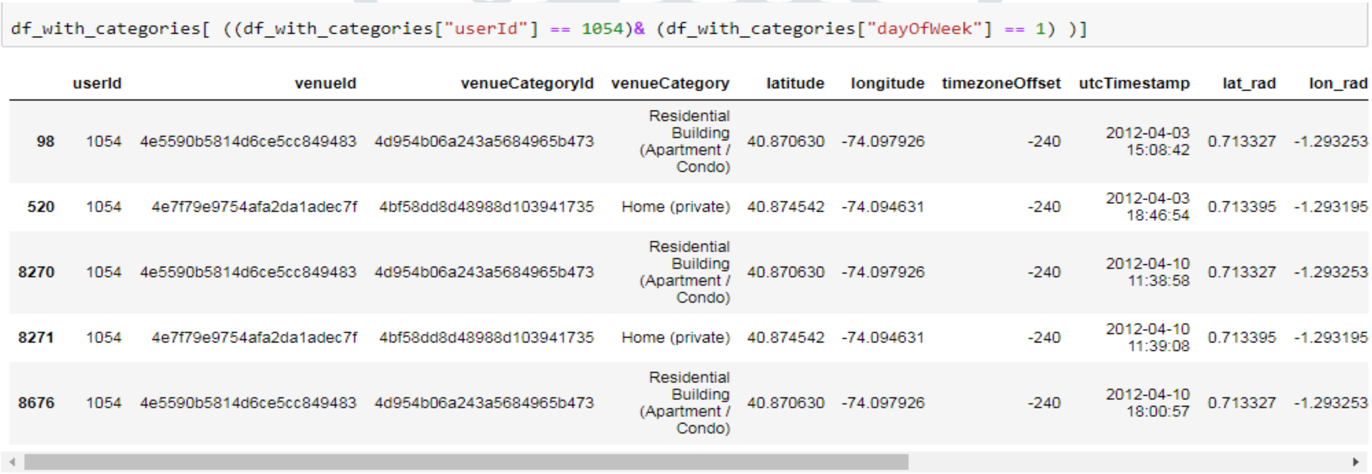
```
new_location_prediction_lables_df = pd.DataFrame(new_location_prediction_lables, columns=location_prediction_lables_columns)
new_location_prediction_lables_df_with_prediction = pd.concat(
    [new_element,
     new_location_prediction_lables_df], axis=1)
new_location_category_prediction_lables = location_category_prediction_classifier.predict(
    new_location_prediction_lables_df_with_prediction)
```

```
for label, poi in zip(new_location_prediction_lables[0], location_prediction_lables_columns):
    if label > 0.5:
        print "Predicted POI is %s" % (poi)
for label, cat in zip(new_location_category_prediction_lables[0], category_location_prediction_class_columns):
    if label > 0.5:
        print "Predicted Intended Category is %s" % (cat)
from sklearn.metrics import confusion_matrix
#location_prediction_lables_columns
#conf_mat = confusion_matrix(new_location_category_prediction_lables, location_prediction_lables_columns)
```

```
Predicted POI is Predicted_poiLablesFromTop_POI_140
Predicted Intended Category is venueCategoryGroup_SOCIETY
```

In Fig.4 define exact weekly prediction of a user by analyzing the result define its POI, type of POI it is used for checking the previous result is correct or not. It defines a test case result.

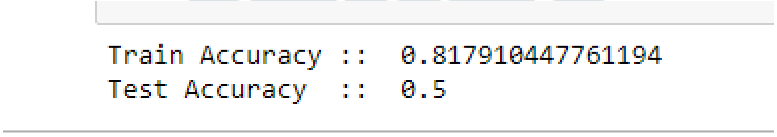
In the next step, we identify system accuracy by using MAE,R2,AUCandMSEmethodAfter predicting particular location of person,then we identify its preferences about services and apply collaborative filtering for recommending the services to the user



```
df_with_categories[ ((df_with_categories["userId"] == 1054)& (df_with_categories["dayOfWeek"] == 1) )]
```

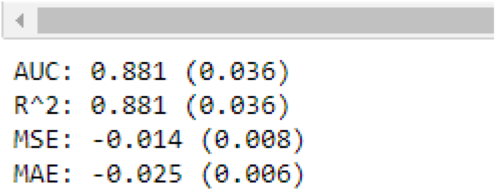
	userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp	lat_rad	lon_rad
98	1054	4e5590b5814d6ce5cc849483	4d954b06a243a5684965b473	Residential Building (Apartment / Condo)	40.870630	-74.097926	-240	2012-04-03 15:08:42	0.713327	-1.293253
520	1054	4e7f79e9754afa2da1adec7f	4bf58dd8d48988d103941735	Home (private)	40.874542	-74.094631	-240	2012-04-03 18:46:54	0.713395	-1.293195
8270	1054	4e5590b5814d6ce5cc849483	4d954b06a243a5684965b473	Residential Building (Apartment / Condo)	40.870630	-74.097926	-240	2012-04-10 11:38:58	0.713327	-1.293253
8271	1054	4e7f79e9754afa2da1adec7f	4bf58dd8d48988d103941735	Home (private)	40.874542	-74.094631	-240	2012-04-10 11:39:08	0.713395	-1.293195
8676	1054	4e5590b5814d6ce5cc849483	4d954b06a243a5684965b473	Residential Building (Apartment / Condo)	40.870630	-74.097926	-240	2012-04-10 18:00:57	0.713327	-1.293253

Fig. 4. Prediction of user's location given wee



```
Train Accuracy ::  0.817910447761194
Test Accuracy  ::  0.5
```

Fig. 5. Accuracy of location prediction system



```
AUC: 0.881 (0.036)
R^2: 0.881 (0.036)
MSE: -0.014 (0.008)
MAE: -0.025 (0.006)
```

Fig. 6. System evaluation metho

Table1 shows each user involvement, in particular, POI and it displays the particular user preference. So this table used for recommending the service to the user.

E. Evaluation Matrices

The accuracy score is not enough to measure the performance of overall system. Other parameters also contribute in calculation of evaluating the performance. There is various methods for performance evaluation one of those is classification report which specifies Area Under ROC Curve, Confusion Matrix, Classification Report Predicted location match with test data set and match it. we are calculating false positive rate because it impacts a lot while recommending wrong services to user which were not present service provider location.

VII. CONCLUSION

The location prediction is very essential in case of recommending the service, traffic control, find the interest of consumers and alerting for disaster.Location prediction system uses the HDBSCAN for finding the users influential point of the area.HDBSCAN is a clustering algorithm which is using haversine distance for finding the distance between two points. Random forest method is used for finding the location POI and also the type of POI. This method recommends the services given POI by using location-based services. In the future system can be done to examine the incorporation of morphological features of the urban environment extracted from street-level imagery and its potential to enhance our understanding of places and latent structures in the urban fabric, as well as the corresponding interactions of people with them

V.REFERENCE

[1] Linyuan Xia, Qiumei Huang , Dongjin Wu, "Decision Tree-Based Contex-tual Location Prediction from Mobile Device Logs, School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China Correspondence should be addressed to Qiumei Huang;hqium@mail2.sysu.edu.cn.

[2] Ruizhi wu, Guangchun luo , Qinli yang,and Junming shao, "learning Individual Moving Preference and Social Interaction for Location Pre-diction",Received December 20, 2017, accepted February 6, 2018,date of publication February 13, 2018, date of current version March 15, 2018.

[3] Region Formation for Efficient Offline Location Prediction.

[4] Yuxing Han, Junjie Yao, Xuemin Lin, and Liping Wang, GlobAL Feature Fused Location Prediction for Different Check-in Scenarios, ieee transac-tions on knowledge and data engineering, vol. 29,no. 9, September 2017

[5] Yushuang Yan, Qingqi Pei, Xiang Wang,Yong Wang Probability-based Location Prediction Algorithm, (Invited paper)

[6] Jianxun Liu, Mingdong Tang, Location-Aware and Personalized Collabo-rative Filtering for Web Service Recommendation, IEEE transactions on services computing, vol. 9, no. 5, september/october 2016.

[7] Guoshuai Zhao, Xueming Qian, Chen Kang, Service Rating Prediction by Exploring Social Mobile Users Geographical Locations, IEEE trans-actions on big data, vol. 3, no. 1, january-march 2017.

[8] Liang Hu, Guohang Song, Zhenzhen Xie, and Kuo Zhao "Personalized Recommendation Algorithm Based on Preference Features",TSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214ll08/11llpp293-299 Volume 19, Number 3, June 2014.

[9] Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan" Predicting the Next Loca-tion: A Recurrent Model with Spatial and Temporal Contexts" Center for Research on Intelligent Perception and Computing National Laboratory of Pattern Recognition Institute of Automation, Chinese Academy of Sciences qiang.liu, shu.wu, wangliang, tnt@nlpr.ia.ac.cn.

[10] S. Hemamalini , K. Kannan and S. Pradeepa "Location Prediction of Twitter User based on Friends and Followers"International Journal of Pure and Applied Mathematics.

[11] Yan-Ying Chen, An-Jung Cheng, and Winston H. Hsu"Travel Recom-mendation by Mining People Attributes and Travel Group Types From Community-Contributed Photos"IEEE TRANSACTIONS ON MULTI-MEDIA, VOL. 15, NO. 6, OCTOBER 2013.

[12] Yantao Jia , Yuanzhuo Wang , Xiaolong Jin and Xueqi Cheng "TSBM: The Temporal-Spatial Bayesian Model for Location Prediction in Social Networks" 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

[13] Josh Jia-Ching,Ying Wang-Chien Lee ,Tz-Chiao Weng"Semantic Trajec-tory Mining for Location Prediction"National Science Council, Taiwan, R.O.C. under grant no. NSC100-2631-H-006-002 and NSC100- 2218-E-006-00

[14] Alisha M Baji, Salini M, Sandra S, Sreya R Thomas, Aswathy Raviku-mar "TRACKGO A Location Prediction Web Application" 2017 Interna-tional Conference on Networks Advances in Computational Technologies (NetACT) —20-22 July 2017— Trivandrum

[15] Chen Cheng , Haiqin Yang , Michael R. Lyu , Irwin King "Where You Like to Go Next: Successive Point-of-Interest Recommenda-tion"Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

[16] Gorkhan Yavas , Dimitrios Katsarosbis Manolopoulos"A data mining approach for location prediction in mobile environments"Received 3 May 2004; accepted 30 September 2004 Available online 30 October 2004

[17] Taehwan Kim,Yisong Yue,Sarah Taylor,Iain Matthews"A Decision Tree Framework for Spatiotemporal Sequence Prediction"KDD15, Au-gust 10-13, 2015, Sydney, NSW, Australia. Copyright is held by the owner/author(s). Publication rights licensed to ACM.

[18] Josh Jia-Ching Ying, Wang-Chien Lee, and Vincent S. Tseng, "Min-ing Geographic-Temporal-Semantic Patterns in Trajectories for Location Prediction"2013. Mining Geographic-Temporal- Semantic Patterns in Trajectories for Location Prediction. ACM Trans. Intell. Syst. Technol.

[19] Anastasios Noulas, Salvatore Scellato, Neal Lathia, Cecilia Mascolo "Mining User Mobility Features for Next Place Prediction in Location-based Services"2012 IEEE 12th International Conference on Data Mining

[20] Hsing-Kuo Pao , Junaidillah Fadlil , Hong-Yi Lin , Kuan-Ta Chen"Trajectory analysis for user verification and recognition" Knowledge-Based Systems 34 (2012) 8190

[21] Han Su Kai Zheng Jiamin Huang Haozhou Wang Xiaofang Zhou "Cal-ibrating trajectory data for spatio-temporal similarity analysis"Received: 18 November 2013 / Revised: 10 June 2014 / Accepted: 12 June 2014 Springer-Verlag Berlin Heidelberg 2014