Location Prediction Modeling based on Mining Movement Pattern from Mobile Phone Data

Wang, Jingwei

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
9
(開始ページ / Start Page)
57
(終了ページ / End Page)
66
(発行年 / Year)
2014-03
(URL)
https://doi.org/10.15002/00010522

Location Prediction Modeling based on Mining Movement Pattern from Mobile Phone Data

Jingwei Wang
Graduate School of Computer and Information Sciences
HOSEI University
Tokyo, Japan
Jingwei.wang.3j@stu.hosei.ac.jp

***Abstract*—This paper proposes a location prediction model that can be used for predicting a moving route of any mobile phone user. It is said that one's moving route may reflect his/her hobby, regular moving pattern, or sociability. With the advancing network and communication technologies, mobile phones have become an indispensable device in our daily life. Data analysis from mobile phone users has become a hot topic. This paper describes how a mobile phone user's one year data is used to build the location prediction model by data partition, cleansing, and event sequence analysis and pattern extraction. With one's location prediction model, it is possible to predict a next place he/she will likely visit. This may be beneficial to some enterprise businesses. For example, if a company can predict someone will go to a certain location, it is possible to push some commercial advertisement related to sales in this location or near around to his/her mobile phone.**

Keywords—Data mining; prediction model; human behavior; mobile phone data; event sequence analysis.

I. INTRODUCTION

Making prediction from human activities has become a recently hot topic. No doubt it is possible but difficult. Song, et al. addressed the limits of such probability of human mobility in their article [1]. There have been many potential applications for researches on human mobility prediction, such as predicting how a disease virus is to be spread depending on human moving tendency. However, how precise we can predict and foresee the whereabouts and mobility of individuals is a fundamental question concerned. There are many existing studies and researches on this concerned issue.

This research is trying to explore the limits of predictability in human dynamics by studying the mobility patterns of anonymous mobile phone users. By measuring the entropy of each individual's trajectory, it is found that it is possible to have a 93% potential predictability in user mobility across the whole user base. Despite there are significant differences in one's travel patterns, it is still possible to find invariability by extracting regularity from his/her mobility data. There have been many ongoing researches in this area. For instance, A.-L. Barabasi who is a scientist in The United States Northeastern University with his colleague researches on the activity pattern [1] that comes from the anonymous mobile phone users. They find out that in general the activity of human is considered as unpredictable and free rein but in fact human activity is surprisingly abided by some regular. On the basis of Albert Laszlo's finding, this research starts to put efforts to create the predict model from anonymous mobile phone data and think about how to improve the accuracy of prediction. To achieve a better accuracy, an event sequence analysis is applied to the pre-processed a mobile phone user data. It is expected that a human mobility prediction model can be one of the elements in a human model [2].

II. RELATED WORK

Although this research topic has received increasing attention from many researches and scientists but it is still in its immature and infant stage. Currently, one often adopted approach is prior probability and normal distribution. Of course, some researcher use characteristic conditional probability distribution to create two parameters as current time variable t_i=t and last visited location variable $v_{i-1}=l_k$, to calculate the probability of the current location [3].

$$\begin{aligned} &p(v_i=l|t_i=t,v_{i-1}=l_k) \\ &=\frac{p(v_i=l,t_i=t|,v_{i-1}=l_k)}{P(t_i=t)}\propto p(v_i=l,\ t_i=t|v_{i-1}=l_k) \\ &=p(t_i=t|v_i=l,\ v_{i-1}=l_k)*p(v_i=l|v_{i-1}=l_k) \\ &=p(t_i=t|v_i=l)p(v_i=l|v_{i-1}=l) \end{aligned} \quad (1)$$

The above paper describes that the predict location only has relationship with the current visited time but not the last visited location. So the formula likes follow:

$$p(t_i=t|v_i=l,\ v_{i-1}=l_k)=p(t_i=t|v_i=l) \quad (2)$$

In here the time is separated into d and t. "d" means work day. "t" means hours. Then create the relationship with prior probability and normal distribution. Throw the normal distribution calculate the result:

$$\begin{aligned} &p(h|\mu_h,\ \delta_h^2,l) \\ &=\prod_{i=1}^{N_l} N_l(h_i|,\ \mu_h,\ \delta_h^2) \\ &=\prod_{i=1}^{N_1}\frac{1}{(2\pi\delta_h^2)^2}\exp\{-\frac{1}{2\delta_h^2}(h_i-\mu_h)^2\} \end{aligned} \quad (3)$$

From the above formula we get the relationship between prior probability and normal distribution. The author considers this method as HPHD algorithm.

The above method is very ingenious but they did not realize that human behavior has strong relationship with the last place and next place. They did not consider the last time visited location effect the next visit place. Infect the active of the human has obvious sequence regular. For example people always go to "location-3" first and then go to "location-4" and third go to "location-8" and last go to "location-9". The purpose of my research is to find out the sequence regular among human activity.

In my research the model uses event sequence to predict the next place. Event sequence is a basic method in Data Mining. The idea of the method is that something always happens in some special order.

III. DATA PRE-PROCESSING

The data is cut by day and week. The paper cleanses the data to eliminate the noise. Currently the data is collected from the mobile phone users for one year. The format of the data is ".csv". The model focuses on the GPS information.

A. Change the GPS Information into Location Name

The purpose of research is to predict the user's moving route. So the paper uses Google location API. Throwing the Google location API I can get the location name. The GPS information is considered as input and the output is location name. Using the Google location API the model also can get the type of the location and during the location the moving speed of the user.

B. Place Threshold Initialization

I set the initialize threshold of the place. If the initialize threshold of the location is set very high, the location is very difficult considered as important place and very easy to be ignored. In this case the location cannot be put into the sequence. The speed is divided into two branches. Speed is faster than 7 and speed is slower than 7. If the speed is faster than 7, the user does not motionlessly stay in the area, it is possible to explain that the user is not interested in these areas. "s_l" is the support of the location. "c_l" is to calculate the number how many times the location exists in the history data."c_{all}" is the total number of the locations in the history data.

$$s_l=1-\frac{c_l}{c_{all}} \tag{4}$$

If the speed is smaller than 7, the user is on walking state. If these places' frequency are very high in the historical data, that proves that the user interested in these areas. These sites for modeling contribution are high, so the initial threshold of these sites should be set very low. In this way these locations will be taken into the sequence. "s_l" is the support of the location. "$\overline{v}$" is the average speed of this location when the user visited this location. "v_{lv}" means to calculate the variance of the speed of this location when the user visited the location.

$$s_l=\overline{v}*v_{lv} \tag{5}$$

To sum up the initialization formula of location support as follow:

$$s_l=\left(1-\frac{c_l}{c_{all}}*\frac{1}{\overline{v}}\right)*v_{lv} \tag{6}$$

IV. EVENT SEQUENCE ANALYSIS

Order in real life is very important, for example in shopping goods tend to buy beds, and over a period of time to buy sheet. In this paper, the user is usually the first to go to the station and then go to work place, and then go to the station to go to the supermarket and then go home. The purpose of this research is to achieve the basic user access location sequence, basing on the location sequence to achieve forecasting.

Filter out the location whose support is less than threshold, assuming the threshold as 5, and the sites: $<l_2>$ and $<l_3>$ are filtered out. Then get the frequent sequence pattern whose length is 1, such as $\{<l_1>, <l_2>, \ldots, <l_n>\}$. So that each place alone formed frequent sequential pattern with length is 1. Then element in the frequent sequence of length 1 cooperate one by one to generate new frequent sequence with length is 2. Such as follow:

$$\begin{aligned} C2=\{&<l_1, l_1>,<l_1, l_2>,<l_1, l_3>\ldots\ldots\ldots<l_1, l_n>\\ &<l_2, l_1>,<l_2, l_2>,\ldots\ldots\quad <l_2, l_n>\\ &\vdots\\ &<l_n, l_1>,<l_n, l_2>,\ldots\ldots\quad <l_n, l_n>\} \end{aligned}$$

Formation of the multiple locations sequence candidate is through the elements of the son generation. Then the judge each with a length of 2 sequences could meet the minimum support degree, if it is not satisfaction, it will be filtered out. The last generation of a frequent site sequence pattern. This method is very simple, but when the place of very large amounts of data, efficiency will become the bottleneck. Through the above description I believe you have found that, you need to scan the database when you want to generate the candidate item, which is affecting the efficiency. So here I want to introduce the professor that Jiawei Han's method [5].

A. PreFixSpan:

Its major idea is that, instead of projecting sequence databases by considering all the possible occurrences of frequent subsequences, the projection is based only on frequent prefixed because any frequent subsequence can always be found by growing a frequent prefix. The kernel of PreFixSpan algorithm is Depth-first traversal. PreFixSpan is similar to GSP algorithm. The model needs to scan the data base to generate the frequent sequence with length 1. And then use the elements of the frequent sequence as the start location to generate the sequence, such as $\{<l_1>:5\ <l_2>:7\ .<l_n>:6\}$ that is the frequent location set and use $<l_1>$ as the first location to generate the sequence, in order to get the result the algorithm need to delete all the elements that is present in front of the location $<l_1>$. The model will use "-" instead of $<l_1>$. I just persist to get the locations that appear on the back of $<l_1>$. The order in the history like this $\{<l_2>, <l_7, l_8, l_8, l_4>, <l_4>, <l_5, l_7>\}$, the result of projection is $\{<-, l_4>, <l_4>, <l_5, l_7>\}$. If in the history record there is only one item including $<l_1>$, delete the item directly and use the rest data as project data base. If the history record like this

Supervisor: Runhe Huang, Professor

{<l_1> , <l_7 , l_8, l_9 ,l_1 , l_4> , <l_4> , <l_5 , l_7>}, after the calculating the project data base likes this {< l_7 , l_8, l_9 , l_1 , l_4 >,<l_4>,<l_5 , l_7>} after deleting item <l_1>. After we get the <l_1>-projected data base, we will traversal every elements in the <l_1>-projected data base to calculate the support of them and filter out the location with low level support and generate the candidate element. And then to generate <l_1>-projected data base but the length is 2. At last the model will get all the sequence with the start <l_1>.

V. EXPERIMENTS ANS RESULTS ANALYSIS

On the original data, there is only GPS information. Original data do not have the location name information. To deal with this problem, the paper uses Google Service API to get the location name throw the GPS information.

A. *Loop through the file name*

The paper uses the file "getFilesName.rb" to get all the file names into the array "$list []". The paper loops the array "$list []" to show all the name of files.

B. *Generate the format of data*

The paper use file "locationSupportDataStore.rb" to generate the format of the data. Here every 7 days we make these data to be one group as an array. After that we put one week's data to be one big array. The graph as follow:

```
w1=[['a'],['a','b','c'],['a','c'],['d'],['c','f']]
w2=[['a','d'],['c'],['b','c'],['a','e']]
w3=[['e','f'],['a','b'],['d','f'],['c'],['b']]
w4=[['e'],['g'],['a','f'],['c'],['b'],['c']]
```

Fig.1. Data store format.

We use "w_i[]" to store the data where the user have visited separated by day. Base on this idea at last the paper generates the sequence model is about the weekly moving behavior. Every 7 days to store the data once, that because the regular of the person's activity is 7 days. During the file "LocationsupportDataStore.rb" the contents is that:

Class	Class	LSDSN
Method	def	looplist
Method	def	generateData

Fig.2. Structure of the LocationsupportDataStore.rb.

C. *Generation expansion frequent site*

File "generateFrequenceItems.rb" will make the above frequent location into the format of ["-", "locationName"]. And store the result into array "$itemsStore". The result as follow

```
irb(main):328:0>  $itemsStore
=> [["Mrs._Jeannine_Gilliard_Beer-Gabel"], ["Chemin_du_Crêt"], [
ue_du_Tir-Fédéral"], ["Ecublens"], ["Route_du_Bois"], ["Chemin_d
y"], ["E23"], ["Place_de_la_Gare"], ["Rue_de_Lausanne"], ["Renen
_de_Marcolet"], ["Pont_de_l'Avenir"], ["Route_du_Pont-Bleu"], ["
A9"], ["Chemin_du_Stand"], ["Route_de_Bussigny"], ["Route_de_Cri
, ["Rue_du_Villars"], ["Rue_de_Bassenges"], ["Route_de_la_Pierre
], ["Chavannes-près-Renens"], ["Lausanne"], ["Route_de_Lausanne"
```

Fig.3. Extend frequent locations.

D. *Generate Project Data Base*

Firstly chose the first element from array "$itemsStore []", we will generate the projected data base with the element. File "GPD.rb" is a special file that is used for generate project data base. During the file it includes method "def generated ("String")",and the parameter is "String" type. We will see that

"$itemsStore[0][0]="Mrs._Jeannine_Gilliard_Beer-Gabel". Then generate the project data base that is started with "Mrs._Jeannine_Gilliard_Beer-Gabel". Fig.3 as follow:

"-"means$itemsStore[0][0]="Mrs._Jeannine_Gilliard_Beer-Gabel".

```
irb(main):225:0> $pDB
=> [["-", "Chemin_du_Crêt", "Avenue_Piccard"], ["Avenue_Piccard", "Route_de_la_Sorge", "Avenue_du_Tir-Fédér
sanne", "Rue_de_Verdeaux", "Rue_du_Bugnon", "Chemin_de_Jouxtens", "Renens", "Sur_la_Croix", "Keni_Service_Sà
issier", "Bussigny-près-Lausanne", "Villars-Sainte-Croix", "E62", "A9", "Route_de_Berne", "Route_de_Bossière
onne", "E27", "Chemin_de_la_Tuilière", "Chemin_de_Pangires", "Route_de_Châtel-Saint-Denis", "Route_de_Montre
z", "Rue_du_Jura", "Route_de_Bussigny", "Route_de_Crissier", "Route_de_Buyère", "Route_de_Sullens", "Chemin_
lle-Sullaz", "Chemin_de_Cretenet", "Désir_Vert_Sàrl", "Daniel_Cruchon", "Route_de_Condémine", "Route_de_Rene
n_du_Coteau", "Route_de_la_Maladière", "Allée_du_Tilleul", "Chavannes-près-Renens", "Route_de_Chavannes", "L
"Epesses", "Puidoux", "Saint-Saphorin", "Corseaux", "Route_Cantonale_Vevey-Forel_Lavaux", "Chemin_de_Burign
s", "Grand'Rue", "Route_de_Forel", "Route_de_Vevey", "Rte_des_Prés_de_Banps", "Route_de_Lausanne", "Route_de
ugy", "Chemin_de_la_Dôle", "Chemin_de_la_Forêt"], ["Ecublens", "Galerie_de_Marcolet", "9", "Rue_du_Tissonet"
âteau", "Daillens", "Orchidée_Realty_SA", "A1", "E62", "Route_de_Crissier", "Bussigny-près-Lausanne", "Chemi
ois", "Chemin_du_Croset", "Mrs._Jeannine_Gilliard_Beer-Gabel"]], [["Renens", "Rue_de_Lausanne", "Avenue_du_1
que", "Rue_de_la_Savonnerie", "Rue_Neuve", "Rue_du_Simplon", "Avenue_du_Censuy", "Avenue_du_Silo", "Rue_du_C
e_de_Prilly", "Avenue_d'Echallens", "Avenue_Recordon", "Chemin_des_Clochetons", "Place_Chauderon", "Lausanne
contemporain", "Voie_du_Chariot", "Rue_du_Port-Franc", "Ecublens", "Au_Record_Muret", "Sentier_de_Denges", "
```

Fig.4. Projected data base.

Keep on using the file "GPD.rb", basing on the lasted projected data base to generate <$itemssStore [i][0]>-projected data base. On this situation the algorithm gets the sequence like this [location1] [location2]. Keep on looping. At last get the whole location sequence

The paper uses the data from the user 02. The data is from 1month 15 day 2012 year. The data is used as experiment data. Delete the repeat location. The result is given as Fig.4.

From Fig.5 the result shows that the first place the user visited is "Ecublens". And then the model finds the same place during the location sequence. All the sequence starting with the location "Ecublens" will be shown.

After the model generates the projected data base, uses the above method to calculate the support count of locations during the projected data base. At this situation the paper set a second level threshold for filtering out the infrequent locations. Here the paper sets the values as 2. The model will store the new frequent locations into array, and pick up the elements from the array one by one. The model will use the frequent element connect with the <$itemssStore [i][0]> to generate a sequence that the length is 2. When the 2 lengthy sequence is generated, the model will base on the 2 lengthy sequence to continue generate projected data base. At this case the result of the projected data base likes this <location-1, location-2>-projected data base. The predict model will keep on looping, until there is on location exists in the array. That means all the locations cannot satisfy the threshold value, then the loop will over.

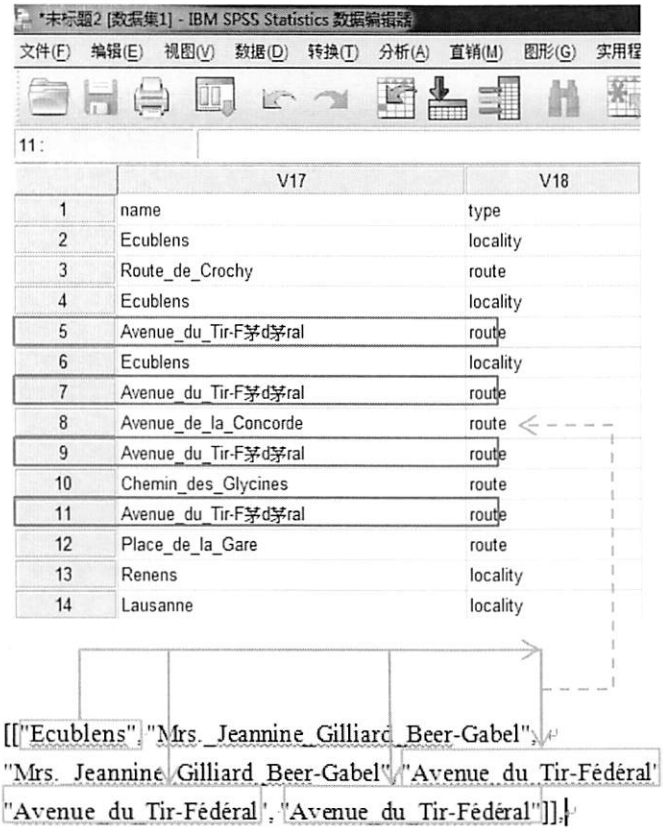

Fig.5. Location sequence start with "Ecublens".

Use the last half of the user's history data to do the test I find that base on de sequence model to predict the use's moving behavior the accuracy can get 60%.

VI. CONCLUSION AND FUTTURE WORK

This paper just separate the data into two parts, one is about work day another is weekday. The paper should keep on separating the data into more details. The model can extract the holiday from the weekday. On holiday people have a long time to travel so people will enjoy going on a long journey. So the travel trajectory will appear larger fluctuation. Many people will be out of Tokyo during holidays, traveling to Kyoto or Osaka. Others may go to farther in Hokkaido for example. In this case the prediction method of analysis will be changed, as if still using the frequency to measure cannot make a good result. Generally if some places the user just visited one time or it is a long time for the user to visit the place again, in this case it is not easy for the model continue to use the prediction sequence. At this time the model should get connect with user's FaceBook or internet access record together. Mostly people will search some information form the internet before they go out one month ago. So from the history record we can predict where the users want to go. So if the various channels of information together, accuracy of prediction will be greatly improved. If the user will go to some place but the location do not exist in the history data. The prediction model cannot be used in this situation. Because the method mentioned in this paper is based on the historical records. Present in our lab someone focuses on semantic analysis. Their purpose is to base on the twitter to analysis the emotion of the user to prevent suicide or some bad things. They throw the daily log and the history of internet to analysis the users. If I put prediction model and their research together, the accuracy of the prediction model will be improved.

In the future we can use the data to create a human model. The predicted model is just one part of the human model. Human model is complex model, it includes many sub-models. Such as the prediction model is just one part of the human model. During the human model it is also include emotion model. Now in our lab someone focuses on creating emotion model. Base on the event the emotion can simulate human's emotion. In the future I want to improve the accuracy of the prediction model by time sequence. The behavior of the human has strong relationship with time. Sometimes on some extract time the human will appear on some place. For example form the data we find the speed and latitude and longitude have some special relationship. When the speed has some change, such as from the high speed to low speed. The point when the speed changes, is the point of inflection of longitude and latitude. That means between the two that kind of points, the longitude and latitude show linear feature. Basing on this characteristic we can use *threshold autoregressive model* that is supported by H.Tong to predict the next location where the mobile phone user will visit.

ACKNOWLEDGEMENT

First and foremost, I would like to extend my sincere gratitude to my supervisor, Professor Runhe Huang, for her valuable instructions and suggestions on my research and this thesis. Without her patient help and illuminating instruction, it would be impossible for me to complete this thesis on time. My sincere thanks also go to Professor Bin Guo, for his kindly providing of the Mobile data and precious suggestions on the thesis. Last my thanks would go to the friends in Huang Lab, for their kind help and support.

REFERENCES

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi, "Limits of Predictability in Human Mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[2] Seungho Lee, Young-Jun Son "Dynamic Learning in Human Decision Behavior for Evacuation Scenarios under BDI Framework,"Proceedings of the 2009 INFORMS Simulation Society Research Workshop L.H. Lee, M.E.Kuhl, J.W.Fowler, and S. Robinson.

[3] Huiji Gao, Jiliang Tang, Huan Liu. "Mobile Location Prediction in Spatio-Temporal Context"
http://research.nokia.com/files/public/mdc-final72-gao.pdf

[4] M. De Domenico, A. Lima, and M. Musolesi, "Interdependence and Predictability of Human Mobility and Social Interactions," arXiv e-print 1210.2376, Oct. 2012.

[5] J. Pei, J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," 2001, pp. 215–224.

[6] R. Huang, X. Zhao and J. Ma, "The Contours of a Human Individual Model based Empathetic U-pillbox System for Humanistic Geriatric Healthcare", Journal of Future Generation Computer System (submitted), 15th, Jan, 2013