Distributed systems seminar, fall 2015
Profiling mobile subscribers through cellplan enrichment with POI data
Elis Kõivumägi

Abstract

Applying semantic meaning to mobile subscribers' trajectories has been an interesting research topic during recent years. Mobile operators are beginning to understand the value in the vast amounts of subscribers' location data and are interested in enriching it with data from external data sources. It could provide them additional revenue sources as well as give them a better understanding about their customers desires and habits. There are several research questions that need to be solved to achieve the satisfactory result: how to detect placement episodes from location data that indicate when subscribers are conducting an activity at a certain place, how to define the physical boundaries for the episode that has been detected and how to choose the most probable semantic meaning from several possibilities? With the present work the author is proposing a framework/platform which outlines the algorithms and rules that ultimately enable profile creation from subscribers' movements that are registered by mobile networks.

Introduction

A person who is carrying a mobile phone can generate hundreds of location events per day depending on the usage. This data can be used to re-generate the trajectory and analyze the movements to understand what this person may have been doing at certain locations. Usually though, the location facts itself do not reveal much if the area is unknown to the person who conducts analysis. Even worse - it is impossible to analyze millions of subscribers' trajectories.

It is clear that another dataset is required that can be used by an algorithm trying to understand what activities could be done in different locations in limited timeframes. During this work a POI (point of interests) database is used which includes data about businesses, restaurants, activity parks etc.

Still, having the trajectories and POI data is not enough before profiles for subscribers could be calculated. The activities need include calculating the cell coverage areas for each cell which is used to identify subscribers' actual physical position, understand when subscribers are not moving and conducting an activity at a certain location and define the rules how semantic meaning is deduced based on the episode length and physical properties of the location that the activities is taking place in.

The following chapters propose (pre)processing rules and algorithms to solve the questions set.

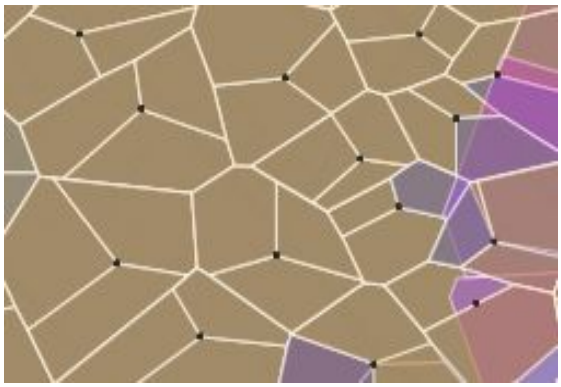
Cell coverage areas

The first challenge lies in mapping the mobile location event to physical world. Location events occur in mobile network cells that have finite coordinates - a point of the mobile site it is attached to. Describing the location events with these coordinates is not the best approach for solving the current problem in hand successfully as the mobile cell actually radiates coverage to an area in the physical world (similarly to WIFI). With location as an area we can estimate the location of the subscriber more accurately which can be improved further when they attach to several cells that cover the same area with each additional cell increasing the accuracy of the subscriber's' actual location.

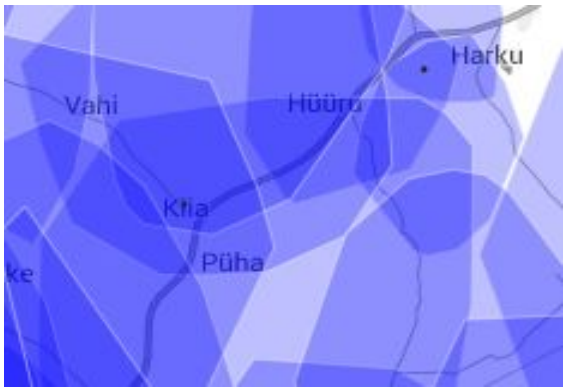
How the cells cover the physical world is considered a trade secret. The operators are only willing to share only the cell metadata for research purposes which leaves the researchers with the task to interpolate the coverage areas with available metadata which usually includes cell frequency and technology, radius, azimuth and others. But the ones given are most useful.

With given attributes, the most common approach for calculating coverage areas [1] is using a Voronoi algorithm [2]. There are several downsides to Voronoi with the biggest one being the lack of overlapping between different cells which operate at the same frequency. See figure 1.

Figure 1. Cell coverage area comparison



Cell coverage area calculation using Voronoi algorithm



Cell coverage area calculated by specialized planning software used by operators

Our experiments have shown that similar results to operators planning software can be achieved when using a set of rules widely accepted [3]. The more cell metadata available the better the results. Below there is a reference set of attributes for calculating cell coverage areas based on metadata:

Metadata type	**Value**	**Description**
Frequency	800MHz	These cells expand to up to 35km in urban and 10km in rural areas.

Frequency	1900MHz	7km in urban, 3-4 km in rural
Frequency	2600MHz	2-3 in urban, 1km in rural
Cell type	Micro	Used usually in metro stations or indoors. Usually coverage is limited to several hundred of meters
Radius	N meters	Sometimes operators give out the radius which also takes into account the tilt of the cell
Azimuth	Degrees	In which direction the cell is directed
Start/end angle	Degrees	How wide is the coverage area of the cell. In Voronoi 360 degrees is divided with the number of cells in the mobile site which is not correct. If angle is not given apply 120/160 degrees.

Generation of the cell coverage areas involves lot of trial-and-error as the quality of the data varies a lot. If possible, conduct tests to verify whether or not the cell is covering the area the calculated cell coverage areas indicate - measure the current location with GPS and at the same time identify the connected cell.

Detecting placement episodes

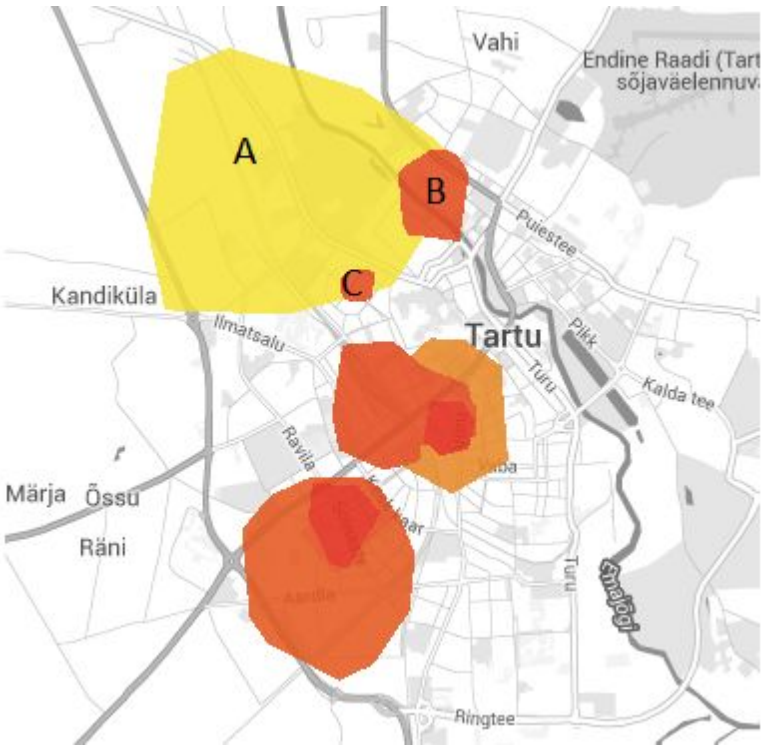
There are many articles published that propose how to detect anchors from mobile location data [4] [5]. Anchors indicate which are the most significant locations to the subscribers and are identified as home and work locations. The objective of this work is to find additional anchors which would enable the profiling.

Placement episode detection is heavily depending on the density of the location data. The more events (the more complete the trajectory) the better the results. When the data is sparse, as it is in the current work, data from several days have to be merged. From the merged data it is possible to detect additional hotspots which repeat weekly.

Also the quality of cellplan affects the outcome. Current algorithm is relying on overlapping among different cells and therefore would not work with Voronoi type of cell plan without adding buffers to cell areas to detect overlapping (and neighbouring cells). When considering overlapping threshold value the cell sizes have to be considered as well. In Figure 2 one orange cell overlaps with the big yellow one, but does not overlap with the other orange one which overlaps with the yellow one as well.

The test subject was actually not moving around at that moment. This indicates that cell areas are wrong because one of them is not covering test subject’s physical location. But due to the algorithm setup, this is considered as one episode because both orange cells overlap with yellow one. This helps to improve episode detection.

Figure 2. Anchors for test subject



From the Figure 2 it is clear that this test subject has 3 distinct anchors. The next step is assigning temporal attributes to these anchors - when and for how long usually the test subjects spends time in each of those locations and whether it is in the mornings, daytime, evenings etc.

The results for the three anchors are given in Table 1.

Anchor	**Time**	**Duration**
Top anchor in Figure 2	Mornings and afternoons	1-2 hours at a time
Middle anchor in Figure 2	Daytime	Several hours
Bottom anchor in Figure 2	Mornings, evening, nights	Several hours

The episodes that last several hours are clearly related to home and work locations. The one that appears during daytime is work and the other one is home. The interesting episode is the top one. Due to the short period of time that the test subject spends there, it is very difficult to understand the semantics and it cannot be deduced from only temporal parameters.

To understand the semantics a secondary layer of information is need that describes the physical world in terms of interesting places that people are visiting.

Mapping POI data to locations

In order to give meaning to the unknown episode it has to be compared with the POI database - which POI-s exist in that area and are open during the hour the test subject is identified there as we concluded in the previous chapter.

As a next step it is necessary to identify POI's which lie in the coverage areas of the episode cells.

Cell plan enrichment is an area of little research as operator's cell plans are not widely available for this purpose. There exist open databases [9], but they rarely consist of the attributes required to model the cell coverage area correctly.

The most common approach to apply semantics to a certain physical location is using a purpose built mobile application that supports location labeling. This ensures that labels are 100% correct, but gathering such dataset is complicated and time consuming.

In [7] the authors tried to investigate how the behavioral routine of the mobile phone users can reveal their performed activities at specific locations. The data collected was from natural mobile phone users. The approach adopted is organized into four steps as follows:

- Characterizing comprehensive temporal variables for each call location.
- Selecting the most effective variables using features selection techniques.
- Building a classification models using decision tree and random forest based on machine learning.
- Enhancement of the inference performance by applying a post-processing algorithm

The final results showed that the system has 69.7% prediction accuracy.

During the current the following approach was chosen. For each cell in the episode POI's and weights will be added which represent the occurrence of the cell in the episode. The data can be represented as depicted in Table 2.

Cell	**POI type**	**Weight** (# of events)
A	Sport	8
A	Activities	8
A	Entertainment	8
B	Sport	2
C	-	4

In cell A, there exists three types of POI, in cell B there's only one type and cell C does not cover any interesting POI's.

To select the correct semantics for the episodes, a lookup table is also needed to identify times that apply to each POI category. For example, people usually go out for sports after work or even later in the evening. Usually entertainment activities start from 6-7 on weekdays and last for couple of hours.

Table 3 describes the temporal rules for POI selection.

POI type	**Mon-Fri**	**Sat-Sun**
Sport	7:00-9:00 or 18:00-21:00	14:00-18:00
Activities	-	10:00-20:00
Entertainment	19:00-02:00	18:00-02:00
Shopping	17:00-20:00	10:00-18:00
Food	12:00-14:00	18:00-20:00
Education	9:00-15:00	-

As a last step episode time will be compared against specified times for each POI type. The bigger the overlap, the higher the probability which means that results have to be normalized to be able to represent it as probability. Therefore, the maximum overlapping times will be assigned the probability of 1. So if a test subject was conducting an activity in the area of cell a on Tuesday from 18-20. The probability for sports will be set to 1 for sports. And 0.5 for entertainment because only one hour overlapped with the predefined activity time. For cell B, the probability will be set to 1 as well for sports. The weights are ignored in current case because cell C only includes sports and does not introduce any new POI types.

If cell C included a shopping POI type with the probability of 1 as well. Then the probabilities would have to be multiplied with the weights as follows:

Total events: 14
Sports = (8+2)/14 * 1 =0,71
Entertainment = 8/14 * 0,5 = 0,28
Shopping = 4/14*1 = 0,28

Experiment, results, conclusions and future work

For the experiment data was gathered during a period of one month. Data consists of location updates gathered in mobile network. Location updates are usually generated when subscribers use their mobile devices For example, make a call, send a SMS and initiate data session. This results in dense trajectories for those who use their mobile devices frequently and sparse for those who don’t.

Additionally a POI database is used which includes the POI categories given in Table 3. In total there is more than 8600 POI in the database.

Unfortunately the code was not finished and the results of the experiment cannot be presented during the fall seminar. The time spent manually analyzing the movements and episodes to deduce the initial rules for code exceeded expectations. Therefore, the author would like to continue to work on this topic during next semester's' seminar and finish the work then. This would include finalizing the code on current experiment and also tweaking the rules based on the results.

The results will be also used to publish a journal article on the same subject and the author did not want to produce something that was yielding the quality needed to accomplish it.

References

[1] A Boukerche, X Fei - 2007. A Voronoi Approach for Coverage Protocols in Wireless Sensor Networks

[2] A Okabe, B Boots, K Sugihara, SN Chiu - 2009. Spatial tessellations: concepts and applications of Voronoi diagrams

[3] 1996. Wireless communications: principles and practice

[4] R Ahas, A Aasa, A Roose, Ü Mark, S Silm - 2008. Evaluating passive mobile positioning data for tourism surveys: An Estonian case study

[5] E Kõivumägi, M Vait, A Hadachi, G Singer - 2015. Real time movement labelling of mobile event data

[6] Abson Sae-Tang, Michele Catasta, Luke K. McDowell, Karl Aberer. Semantic Place Prediction using Mobile Data

[7] Hui Wang, Zhisheng Huang, Ning Zhong, Jiajin Huang. Semantically Modeling Mobile Phone Data for Urban Computing

[8] F Li, N Clarke, M Papadaki - 2010. Behaviour profiling on mobile devices

[9] www.opencellid.org