Master Project in Sony : Mining smartphone data

Report 2 : 16/03/2015-30/04/2015

Khalil Hajji

EPFL supervisor : Patrick Thiran

Sony supervisor : Fabien Cardinaux

4 Mai 2015

REMINDER

We have a dataset of various types of users' traces collected from their smartphones during several months. Those traces include the GPS traces, the Wi-Fi names connected to the phone, the applications launched ect… This data is collected each time that one of the following events E occurs: [notification received, application launched, screen is on, screen is off, launcher is on, launcher is off]

The goal of the project is for now to explore this dataset, see how far we can go with this kind of traces and what kind of applications we could use exploiting the presence of such information.

Highlights

During the first month, we mainly focused on discovering the content of the dataset and exploring the quality and the accuracy of the traces we have.

We discovered a rich and abundant dataset; it contains various types of information concerning each user and a lot of samples per day were recorded for each of them.

However, we figured out that it contains some imperfections and needed to be cleaned. For example, we discovered that the presence of notifications coming from the system caused the data to be unequally recorded. At some periods of time a lot of samples are present whereas at some other periods there in not at all.

Planned work

Considering the outcome of the first data exploration we planned to do the following:

1- Solving the oversampling data problem caused by the system's notifications.
2- Cleaning the data and combining the traces to end up with more accurate and rigorous information
3- Starting the first learning algorithms to analyze and discover the data

REPORT PLAN

In this report, we first mention the similar work that has been already done with similar datasets. We mainly describe how their data traces were cleaned and what the goal of their work was.

Second, we describe our approach of reformatting the data to decrease the oversampling problem and to have a more accurate representation of the data.

Then, we explain how we combine the data traces to have a clean, more accurate and more often information about the locations visited by the user. We also explain the reason that led us to consider this feature in particular.

Finally, we will talk about the goal of the first analysis that we want to run in our dataset and describe the different methods we are considering to that end.

As usual, we end the report by presenting the things we are currently doing and the next steps we are planning to do.

1. STATE OF THE ART AND SIMILAR WORK

A lot of works have been done with datasets that combine a lot of users' mobile traces from smartphone and a various type of problems were addressed with these datasets. Thus, we want to exploit the abundance of such a work for three main purposes: As we are looking on how we can make use of this information, we are interested in seeing the applications and the problems that have been already tackled. In response to the fact that our data is noisy and needs some preprocessing, we are interested to see how previous research works proceeded to clean and formatted the dataset. Finally for our analysis part, we will also pay attention to the methods that were used to analyze and extract information from these traces.

1.1- Problems addressed

Cao et al. [1] addressed in 2010 the following problem: having a dataset of Nokia smartphone traces, they tried to discover the context that causes a user to have a certain interaction with the phone. An interaction could be "listening to music", "reading news", "having a message session". A context is a set composed by the features that are not phone interactions. For example a Context is C= {"Is holiday=Yes", "Day period=morning", "Phone profile= silent", "speed =High"}. They were for example interested in learning that the context C usually implies the interaction "reading news".

In extension to this work, Ma et al. [2] tried to detect similarities between users based on their habits (2012). Based on the same dataset and taking the result of [1] as input, they tried to cluster the users by behavior similarities. This work could be used for context-aware recommendation systems. Context aware recommendation systems take into account the context of the user to give him the right recommendation at the right time.

More work have been done with this kind of dataset as [6], [7] and [8], however we limit our description to the researches described below because it is the most relevant to our work.

1.2- Dataset Cleaning

In [1], the results presented show that the dataset is categorized. The interactions are categorized into "listening to music", "listening to visible radio", ect… Different categories of time ranges are selected and superposed in the same context as for example {"Is not holiday", "Day period = Noon", "Time range = PM3:00-4:00", "Day name"=Thursday}. The location of the user is decided by the cell to which he is connected to. However, no information about the way how this categorization is done is given. Nevertheless, it gives us insight about how our data should be categorized in order to be able to see meaningful results.

In [2], one of the main challenges is to overcome the sparsity of the data. In fact, the individual users' behaviors are very different from one user to the other even if they represent the same semantic. Thud, a big challenge is to be able to categorize and discretize the data so that similar behaviors between users could be caught. First of all, to reduce the sparsity of the locations, they decide to label each user's locations into one of the three categories: "Home", "Work Place" and "Others". For that they use an approach proposed by Yang [3], which discovers automatically those three clusters for each user. Second, to reduce the sparsity of the user-phone interactions they use a set of 13 predefined interactions defined by Nokia Ovi store (www.ovi.com). To map the initial interactions into those categories, they use the work introduced in [4] which only needs a small seed of labeled interactions to be able to automatically classify the interactions into the target categories. Related to our work, we especially keep in mind [3] and [4] which can be useful if we need to reduce the sparsity of our data.

1.3- Analysis Methods used

In [1], they use association rules to learn the contextual information that leads to a phone interaction. The idea is to look through all the contextual features, build contexts and count how

many times a context appears with a certain type of phone interaction. As checking all the combinations of the contextual features exponentially explode, thresholds min_support and min_confidence are used to only go through the promising contexts.
The main critic we have on this work is that by dividing initially the features into interaction and contextual features, we can learn only context that lead to an interaction with a phone. Moreover, interaction features at some time can be contextual at another time. In fact, a present user interaction can be caused by a past interaction. For example, when the user is in the train, he always opens his mail (interaction 1) then reads the news (interaction 2). In this case the context {"In the train = Yes", "mail"} leads to the interaction "reading news". In this work, the interaction features are never considered as potential contextual features and a property as the one described above can never be detected.

In [2], they use a constrained based matrix factorization model described in [5] to extract super-behaviors (which are clusters of behaviors). This method is a matrix factorization method that allows using some prior constrains and that gives as an output a distribution of factor matrixes. We keep this method in mind and we may want to understand it better when we will begin the analysis part. Especially, we want to know how this method differs from a usual probabilistic model and what are its weaknesses and strengths with respect to such a model.

2. DATA FORMATTING

In this section we describe how we change the representation of the data to overcome the oversampling problem and to have a more accurate representation of the data.

As described in the Figure 1 and Figure 2 the idea is to fill in the gaps. We assume that two consecutive and equal realizations that occurs at time t1 and t2 imply that this realization always holds in the interval [t1, t2] if |t1-t2| is small enough (i.e smaller than a threshold). We represent our features separately with a time interval representation that indicates the start time and the end time of each realization.

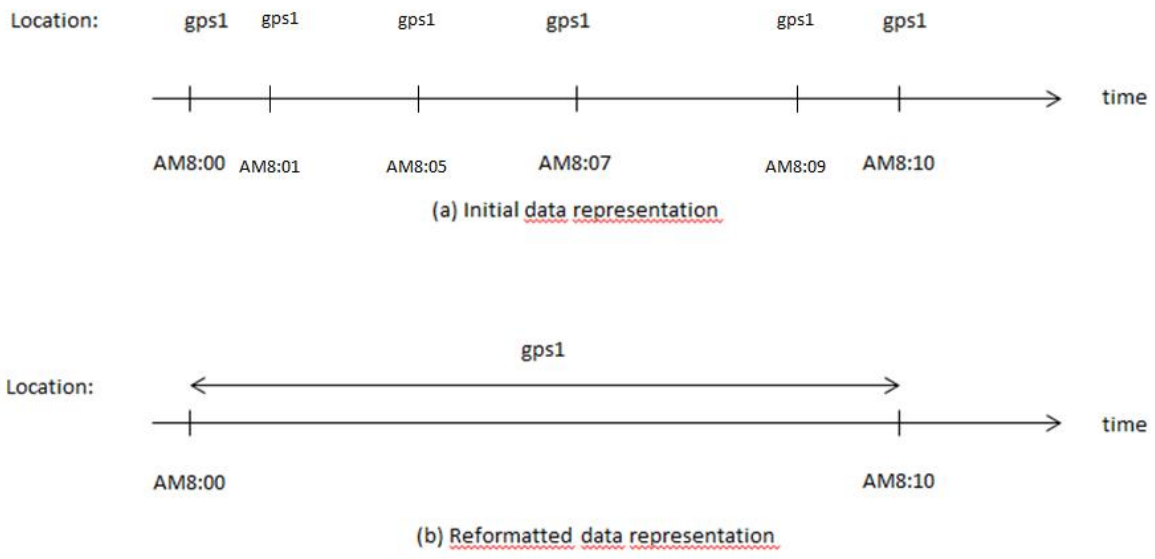

Figure 1: data representation of an oversampled period

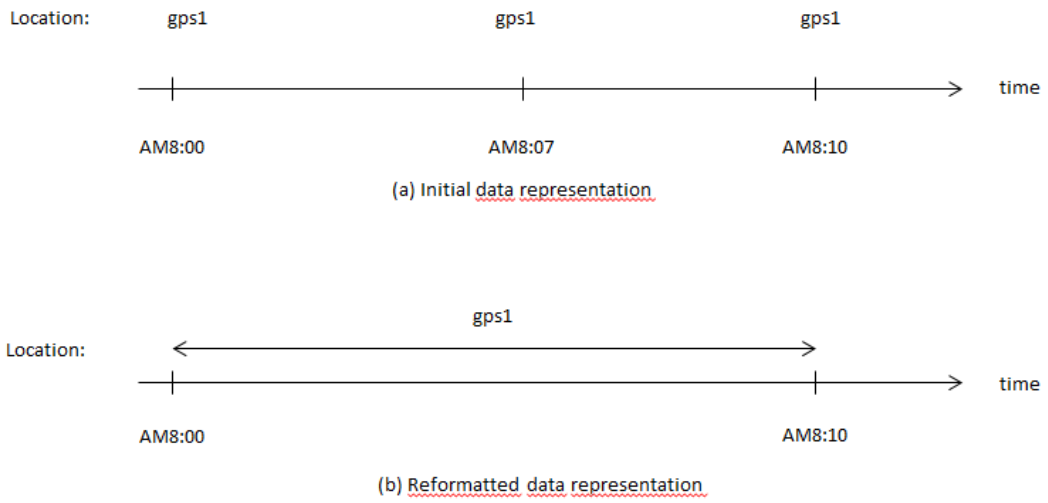

Figure 1: data representation of a downsampled period

By using this method we decrease the effect of the inequality of the data sampling. As shown in figure 1 and 2 the two periods result in the same reformatted data even if the sampling period of figure 2 is smaller than the one of Figure 1.

Moreover, this method allows us to remove the redundancy of the data and to have a more compact representation. The formatted dataset is 50% smaller in size than the original.

Last but not least, the formatted traces representation increases the amount of information we have about the user. It fills the initially unknown gaps with information.

Our data contains some punctual events as the notifications and applications launches. For those features we do not change the representation of the data.

3. LOCATION EXTRACTION

Having obtained a version of the traces where we increased the amount of information available and decreased the unequal sampling problem, our goal is to combine features and integrate different pieces of information to increase the accuracy and the quality of our data.

Most of the similar work already done invested a lot of effort in extracting accurate users' locations. This is explained by the fact that the location of the user is a factor that influences his behavior and often describes his habits. Thus, we invest in our turn a lot of effort to have accurate and consistent information about the location of the user.

This operation aims to achieve two goals. First, we would like to regroup individual coordinates in meaningful locations for the user. For example we would like that all the individual coordinates of the user that were took at his university be regrouped in the same cluster. We call this criterion the *location consistency*. Defining the clusters by a distance metric or using the cell id as in [2] will be weak with respect to this point; a university is an area that spreads out into hundreds of meters squares while a shop in a shopping street spreads out only to few dozens of meters squares. Thus we want our clusters represent different area coverage depending on the users' visits.

Second, we are interested in having information about the location visited by the user as often as possible. The more we have the more the location metric is accurate and the more patterns might be discovered. We call this criterion the *location accuracy.*

3.1- Combining features

We start from the following observation: If a smartphone detects the same WI-FI at two different moments t1 and t2, then it means that the phone is at the same place at t1 and at t2.

Extending this observation, we argue the following:

- All WI-FIs that are detected at the same time represent the same place
- All GPS coordinates observed at the same time than a WIFI belong to the same place than this WI-FI
- All cells observed at the same time that a WIFI belong to the same place than this WI-FI

Using these observations we construct clusters of locations where each cluster contains a set of WI-FI names, a set of GPS coordinates and a set of Cell ids. Then, by observing a GPS coordinate, a cell ID or a WI-FI, we can guess the location visited by the user.
To check that the clusters are realistic and meaningful we do the following coherence test: we compute the maximum distance that separate two GPS points in each cluster and verify that this distance do not exceed a threshold D.

If a user is in the same place, he is more likely to detect the same WI-FI. If a user is at the university, he will always detect the WIFI of the university even if he visited two points far by hundreds of meters. If he goes into a restaurant in a shopping street he detects only the WI-FI of the restaurant. If he goes to a shop located at some dozen of meters, he detects only the WI-FI of the shop. This solution comes with a guaranty of location consistency.

Moreover, this method brings a huge gain: it increases a lot the location accuracy. In fact, our data does not contain all the time all the features. It happens very often that at some time t we only have information about the WIFI (and not about GPS location and the cell), or only about the GPS (and not about the WIFI and the cell). Thus, by using this method, we only need to know one of the three features to be able to know the location of the user.

3.2- Removing the mobile WI-FI routers

The coherence test pointed out that some WI-FIs appear in GPS locations that are very far (sometimes hundreds of kilometers). This problem is caused by the WI-FIs that are mobile in the space. Those WI-FIs are for example the WI-FIs in the train, the smartphones' hotspot, the private LANs ect…

Thus, we decided to keep only the WI-FIs for which we have the certitude that they are stationary. This implies that we remove both the mobile WI-FIs and the WI-FIs for which we have uncertainty.

We keep a WI-FI if the following two conditions hold:

- The maximum range of the distance in which it was observed is less than max_wifi_range.
- The WI-FI was observed enough (more than d different days)

This method allows us to discard all the mobile WI-FIs. The other WI-FIs that are discarded are the ones for which we do not have enough information. This means that they are not often visited by the user. Thus, they are not important to describe his behavior.

To conclude this section, we want to note that we plan to output more objective statistics and results that allow us to quantify the gain resulted from this method. For example we are interested in

comparing the time covered by the new location metric and the original one (considering only the GPS coordinates). We are also interested in seeing the maximum distance distribution of the clusters and the number of visit by cluster distribution to validate the consistency of our method.

Finally, as done in [2], we do not forget the possibility of keeping the two most visited clusters (which should represent home and work) and regroup all the others in a common cluster depending on the results we will have in our first analysis.

4. ANALYSIS GOAL

The first question on which we are trying to answer is the following: Can we discover user habits and routines using this data? We are interested in learning clusters that represents usual activities or recurrent behaviors of the user. Contrary at [1], we do not want learn the context that leads to a set of predefined events (phone interactions). We want to be able to learn any kind of habit and behavior that the user uses to have. For example, we would be interested in knowing each Saturday morning, the user plays some sport, or in the free days the user uses to visit his hometown.

5. NEXT STEPS

Currently we are transforming our data into a matrix representation and we are evaluating our location metric.

After that, to address our problem of discovering users' habits, we are considering some alternatives as baseline methods:

- Soft clustering our Data using SVD
- Hard clustering our Data using K-means. The distance between vectors taken into account in the K-means could be a tuned distance function and not an Euclidian distance function.
- Considering an approach inspired from [2], using entropy. The idea would be to iteratively select features that increase the predictability of the Data.

Concerning the latter point, your intuition on which method you think is the best could be a valuable input for us.

6. REFERENCES

[1] H. Cao, T. Bao, Q. Yang, E. Chen and J. Tian An Effective Approach for Mining Mobile User Habits (2010)

[2] H. Ma, H. Cao, Q. Yang, E. Chen and J. Tian A Habit Mining Approach for Discovering Similar Mobile Users (2012)

[3] G. Yang Discovering significant places from mobile phones (2009)

[4] H. Cao, D. H. Hu, D. Shen, D. Jiang, J. Sun, E. Chen and Q.Yang context-aware query classification (2009)

[5] M. Schmidt Linearly constrained Bayesian matrix factorization for blind source separation (2009)

[6] H. Zhu, K. Yu, H. Cao, E. Chen, H. Xiong and J. Tian Mining Personal Context-Aware Preferences for Mobile Users (2012)

[7] H. Zhu, H. Cao, E. Chen, H. Xiong and J. Tian Exploiting Enriched Contextual Information for Mobile App Classification(2013)

[8] V. Etter, M. Kafsi, E. Kazemi, M. Grossglauser, P. Thiran Where to go from here? Mobility prediction from instantaneous information (2013)