Master Project in Sony : Mining smartphone data

Report 1 : 16/02/2015-13/03/2015

Khalil Hajji

EPFL supervisor : Patrick Thiran

Sony supervisor : Fabien Cardinaux

16 March 2015

1. PROJECT GOAL

Mobile and wearable devices have become very popular; they have the advantage that they are carried with the user during all the day. Using the different sensors of the smartphone (camera, microphone, GPS, accelerometers, etc) and the users' actions on the smartphone, Sony was able to build a dataset that contains a lot of information about some users´ behaviors and actions during the day observed during several months. It is for example possible to know if a user is walking, running or driving a car during one particular moment of the day. It is also possible to know which places a user visited, which applications he opened or what notifications he received.

The goal of the project is to take profit from the existence of such data to learn the user´s behaviors and actions. The long term goal would be to make these smart devices smarter by automatically taking some actions or making some personalized recommendations.

At this stage, the project subject is still open in the sense that for now the goal is to investigate the dataset and see what meaningful information can be extracted from it and what kind of impacting applications it can have.

For privacy reasons, the objective is to use this data in an application where the data of a user never leaves his smartphone. For that reason, the analysis that we will do and the models that we will develop do not mix information of multiple users but treat each user´s data separately.

To have a better insight about the possible applications and the possible work, we discuss in the next section the dataset that is available, the different features available and the quantity of data that we have.

2. DATASET

2.1- General statistics

The dataset contains the activity log of 6 Sony employees recorded using internal Sony software installed in their smartphone. Below we present the period of observation of those users and the number of records that we collected for each user:

	User1	User2	User3	User4	User5	User6
#days_of_observation	300	231	89	249	229	224
#samples	56940	56777	28259	43326	48874	108445
#samples/day	246	189	317	174	213	484

The number of users that we have is really small. However, as we want to do the learning by user (separately), the amount of data that we have per user is more important than the number of users. In our case, we have a big amount of data per user observed during several months.

2.2- Features

The data is collected when one of the following events E occurs; notification, application launch, screen on, screen off, launcher on or launcher off. When one of those events occurs, a data vector containing multiple features is recorded. This makes a lot of information about the user´s daily life and activities available. Some of the features recorded are the following:

- Location: the place of the user at the moment of the record.
- Time: the exact time of the record.
- Notification: if the event is a notification, information about the name of the application from which the notification is coming from and its priority level.
- Application launch: if the event is an application launch, information about the name of the launched application.
- Activity: The activity that the user is doing. It could for example be: running, in a car, in a bug, sitting, standing, on a bicycle, not carrying the phone.
- Bluetooth: if the Bluetooth is activated, information about the detected devices and about the devices that are paired (or on pairing) with the user´s smartphone.
- Battery: information about the battery level of the phone and about its current state (not charging, charging with an usb, charging with a wireless device).
- Wi-Fi routers: If the Wi-Fi is on, information about the Wi-Fi routers currently detected and the Wi-Fi on which the device is connected.
- Telephony: information about the location of the base station to which the smartphone is connected and whether cellular data is used or not.
- Headset plug: information about if there is a headset plugged on the smartphone or not.
- Usb plug: information if there an usb device plugged into the smartphone or not.

Not all the features are present in all the records but only the available ones. For example if the GPS is not activated, we will not have information about the location.

To have complete information about all the features available, the different attributes for each feature and a description of the different values that can be taken by each attribute, please have a look to the file smartphone_data_documentation.pdf (attached).

3. WORK DONE

The software that logs the different users´ data was developed in Sony Japan. In the same way, the data collection, storage and structure is being handled in Japan. Thus the data we are using is unknown to everybody. For that reason, we started getting work with the data from scratch.

3.1- Discovering the data content

As the content of the data was unknown and as no documentation was created to explain the content and the structure of the data, the first step was to understand how was the data structured and what it was representing.

Once the structure was understood, the second step was to discover the different features contained in the records, the meaning of their different attributes and the meaning of the values of that attributes.

At the end of this step, a complete and exhaustive file named smartphone_data_documentation.pdf (attached) was created to describe all the features present in the dataset, their attributes and the meanings of the values taken.

3.2- First statistics

The second step was to discover the dataset characteristics. For that purpose a lot of preliminary statistics were computed as the observation period, the number of samples per user, the number of features by user…

Moreover, we plotted the time repartition of the samples to see how samplings vary over the time period and we did the same for the features. The Figure 1 and the Figure 2 represent the number of records per day and the number of features per day for user 1.

After looking at those plots we noticed that the number of samples per day highly varies from one day to the other and that in some days the number of samples is anormally high for some days (>5000 samples in one day! Unrealistic!!). This is the case for all the users. We also concluded from the figure 2 that the notification feature is the feature that causes the number of samples to explode in certain days.

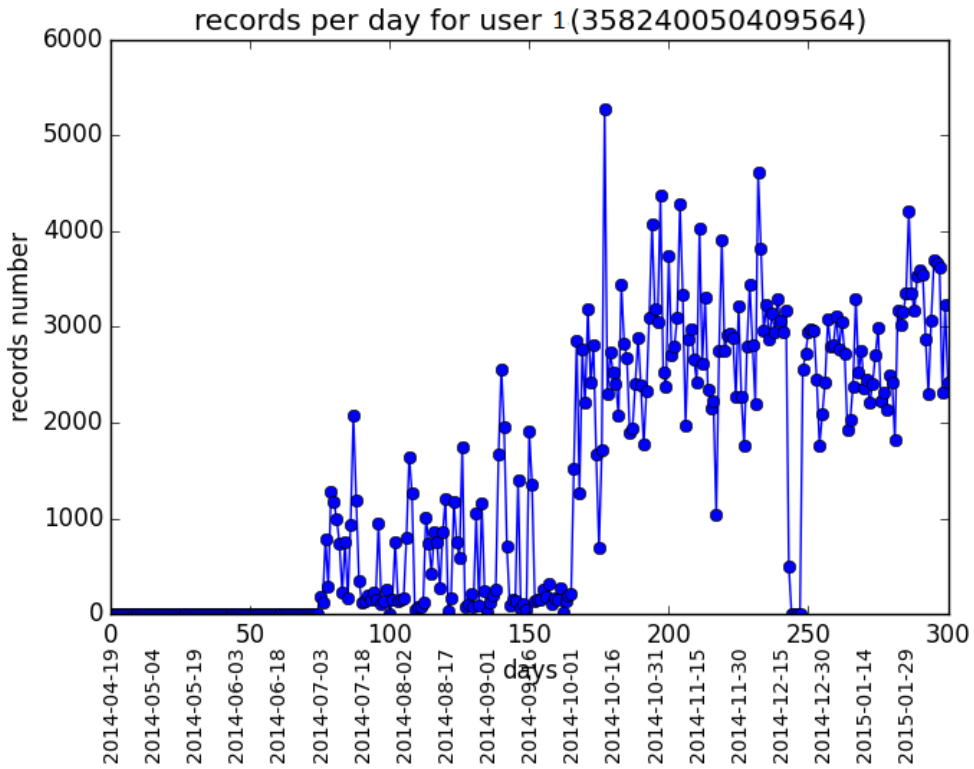

Figure 1: number of samples by day of user 1

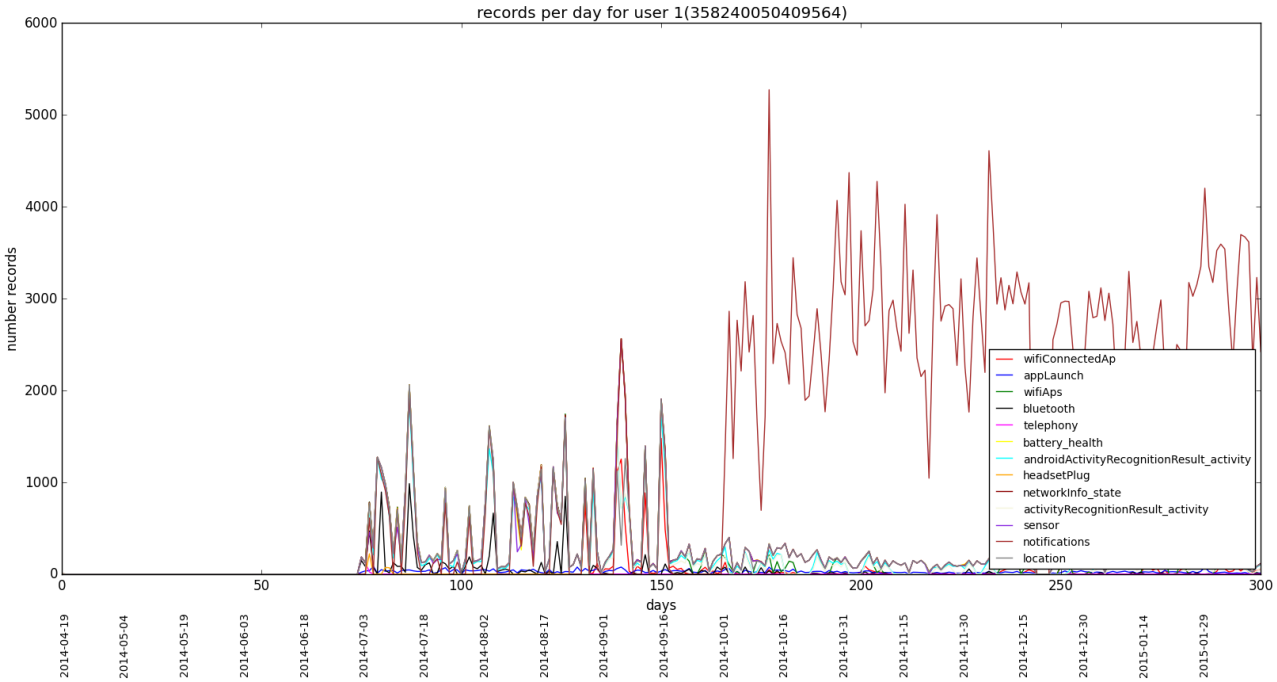

Figure 2: number of features by day of user 1

At the end if this step, we concluded that the dataset contains a lot of noise that is probably caused by the notification feature.

3.3- Cleaning the data

After having a deeper look into the notifications feature and after having observed a lot of data samples, we realized that the logger (the software responsible for collecting the data) was logging notifications that were internal to android (notifications that are not visible by the user). Those notifications are not meaningful to us because they do not have any influence on the behavior of the user. For that reason, they were removed.

Moreover, sometimes the logger was recording the same notification more than once (duplicate notification). Those notifications were also removed.

After cleaning the notifications feature, we ended up with a much more realistic number of samples as shown by the graph in the figure 3.

4. NEXT STEPS

The first step is to have a final cleaned version of the data that is represented in an easy and accurate way. This step is nearly finished. In fact, in the way the data is currently represented, each of the events belonging to E causes the recorder to extract all the possible information. For example if a user receives 10 notifications in one minute, then the software will do 10 records and each of them may contain the same redundant information about the place, the activity, the network and all the features described above. The goal is to get rid of this redundancy. The idea is to represent each feature as an event that occurred at a certain time. An event is represented only if it has a different value than the previous one. For example, for the location feature, if we have n successive times the information that user u was in place x1 and then moved to x2, we represent x1 only once. Thus the representation will be; user u was in x1 at time t1, in x2 at time t2 and so on. This will remove the redundant information in the dataset.. From now on, we will only talk about events. Events could be

launching application x, receiving notification y, visiting place z, detecting the Bluetooth device m or doing the activity n…

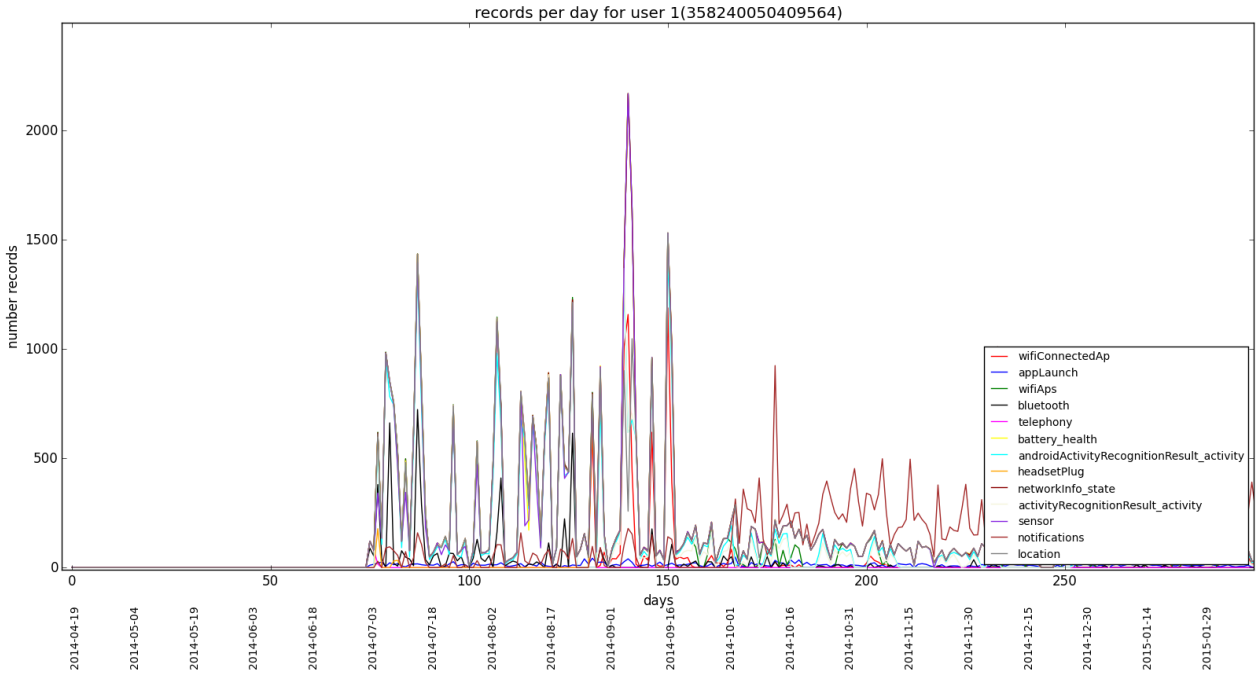

Figure 3: number of features by day of user 1 after filtering the notifications

Once a clean and well-structured data is obtained, the analysis part can begin. For that part, we plan as a first approach to try to find clusters of events that would describe the behavior of the user. This means that we would like to find the events that use to occur together. We hope that this could help us to understand the behavior of the user. For example we could have a cluster: Activity running, headset plugged, application launched music, time Saturday 10. This means for example that the user use to do running on Saturdays morning and use to listen to music while running.

We plan to start this step by the documentation step: we'll try to see if similar work or similar problems were already tackled and see how it was done.

5. SUGGESTIONS

Before ending this report, I would like to invite you for your feedback, remarks, possible suggestions or whatever you think is useful to say and especially for the following points:

- As you may have noticed, the project subject is still open. My feeling is that the dataset we have is quite powerful in the sense that we have a lot of different information often sampled concerning the users' habits. For that reason, I think that there is the possibility to really do something interesting and useful using this data. At this stage, I would be interested to know if you have a suggestion about possible applications and ideas that could be realized thanks to this dataset.
- I would also be interested in having your opinion on how the first step should be tackled; Clustering events close in time.

Finally, if you want to have more elements or more clarifications about certain points, please let me know.