



# Master Project in Sony : Zoom in Next steps

22/06/2015-15/08/2015

Khalil Hajji

EPFL supervisor : Patrick Thiran

Sony supervisor : Fabien Cardinaux

**22 June 2015**

# 1. EVALUATION METHOD

As explained in the report 3, one of the biggest challenges to which we are confronted is to find a way to evaluate the accuracy of our results. This is the task we want to solve in priority because just looking to the results will not enable us to decide which method works better than the others.

The literature and the similar work that has done in this context all evaluated their work in the same way: they asked directly the users to evaluate the correctness of their results. An example of this is the paper [1]. However, in our case, accessing personally to the users we gathered the data from will be complicated (according to my supervisor). Hopefully, we wish to have access to them for the final evaluation of our best model but in the meantime we need to set up a method that will indicate to us the performance of our methods.

We had a meeting today morning to that end and the best solution we could come up with for now is the following: If we hide intentionally some features, the best model that regroups well the activities and the habits of the users is the model that predicts the best those hidden features given the others. More explicitly, we hide for example the day feature, we train our model, and then use a test set where a classifier tries to classify each entry in {“Week_day”, “Week_end”}. The model with the best classification score indicates the model that separates well the data into representative habits of the user. Using a similar idea, we could hide the location and then try to guess if the user is at home, at work or at another place.

This solution is not optimal but it is at least the best one we could think of.

# 2. MODELS FOR OUR DATA

The second concern of the planned work is to set up and select models and algorithms that will enable us to achieve our goal: discovering the habits and the behaviors of the users. The idea is that we evaluate the different possibilities and alternatives, and according to the remaining time, select the best strategy to follow.

For now, the methods that we are considering are the following:

## 2.1- Bayesian Matrix Factorization

Constrained based matrix factorization model was used in [2]. It was used to extract super-behaviors (which are clusters of behaviors) from behaviors (This was discussed in the report 2). This method is a matrix factorization method that allows the use of some prior constrains and that gives as an output a distribution of factor matrixes. We need to explore and to document deeper about this method to understand in details what it is doing and what it is bringing comparing to the other methods. Depending on this evaluation, we can decide if we implement it or not. This method is discussed in [3].

## 2.2- Latent Dirichlet Allocation (LDA) or Probabilistic Latent Semantic Indexing (pLSI)

As explained in the last report (report 3), our data can be seen as a corpus of documents where the documents are the records and the words are the features. Thus, we can apply the LDA or(/and) PLSI algorithm and see if it enables us to extract the behavior of the users. We call a behavior a cluster of features.

### 2.3- Explicit Generative Model using LDA or pLSI

This is a model that we tuned ourselves believing that it may describe well our data and the goal we want to achieve.
Let us suppose that the features available we have are:

- Day = {Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}
- Hour = {Morning, Afternoon, Night}
- Location = {Home, Work, Others}
- Activity = {Still, In_Vehicle, Running}

The generative model works as follows (Let us consider the generative model using pLSI):

- To generate a record (doc), first we select a behavior $Z_k$
- Then we select:
    - A Day with probability $p(d_i|Z_k) s.t \sum_d \Pr(d|\ Z_k) = 1$
    - An Hour with probability $p(h_i|Z_k) s.t \sum_h \Pr(h|\ Z_k) = 1$
    - A Location with probability $p(l_i|Z_k) s.t \sum_h \Pr(l|\ Z_k) = 1$
    - An Activity with probability $p(a_i|Z_k) s.t \sum_h \Pr(a|\ Z_k) = 1$

Here we specify explicitly that a behavior is composed of one day, one hour, one activity and one location and each of these features has its own distribution (that's why named Explicit Generative Model).
As our records represent a time period of 1 hour, a record can contain more than one behavior of a user. For this reason, we tend to think that using LDA should be better.
This is just an idea we had; we have not derived the distributions details and the inference equations yet.

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