Ecole Polytechnique Federale de Lausanne

MasterThesis

Thesis Title

Author:
Khalil Hajji

Supervisor:
Dr. Fabien Cardinaux

A thesis submitted in fulfilment of the requirements for the degree of Master of Science in Communication Systems

in the

Sony Technology Center
Communication Systems Faculty

August 2015

EPFL

Abstract

Faculty Name
Communication Systems Faculty

Master of Science in Communication Systems

Thesis Title

by Khalil HAJJI

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too…

Acknowledgements

The acknowledgements and the people to thank go here, don't forget to include your project advisor…

Contents

List of Figures

List of Tables

Symbols

a	distance	m
P	power	W (Js^{-1})
ω	angular frequency	$rads^{-1}$

For/Dedicated to/To my…

Chapter 1

Introduction

With the rapid and large deployment of the internet, the emergence of the cloud and the entrance to the internet of things area, the recent years were (and are still) marked by the exponential increase of the data stored and available of any kind. Nowadays, data streams for daily life; from computers, credit cards, TVs, trains, sensor equipped buildings and factories. The availability of this huge quantity of data is changing the way companies lead their business and are changing the rules of competitiveness. In a publication that zooms in the challenges and the power of big data, Mc Kinsey affirms that "the use of big data will become a key basis of competition and growth for individual firms" and they estimate that a retailer using big data to the full has the potential to increase its operating margin by more than 60 percent.

In this work, we are interested in the data collected from a gadget that shares the life of the user; his smartphone. Smartphone data contains the locations a user visits, the activities he does, the notification he receives, the applications he launches and many other information. This kind of data is unique in the sense that it represents a complete snapshot of a user's life. In a world driven by the power of data and the ability to anticipate and understand the needs of users, companies shows a big interest in studying this emerging kind of dataset. In the other hand, the richness and the diversity of this data attracts the curiosity of researchers.

In this context, many work have been done with this kind of logs, many paths have been explored and multiple questions answered. Those researches are discussed later with more details. In this work, we tackle an important question that escaped to the interest of previous researches: Having the smartphone logs of one user, can we find a model that exhibits his particular behaviors and habits? More practically, let's imagine the following example. Let's imagine that Bob has some particular habits: he does running while listening music on Saturdays, he visits his parents on Sundays, he reads news when he is in his office in the mornings, and he puts an alarm clock at 7am during his working days. The question we are answering in this work is the following: Can we find a model that discovers those particular behaviors by analyzing Bob smartphone's logs? How precise can this model be in doing this task? It is important to note that we are interested in discovering the individual behaviors of a user, and thus we aim to find a method that takes as input the logs of a unique user.

In the recent few years, key innovations allowed smartphones to drastically evolve from cell-phone devices used for calling to powerful ”pocket-computers” devices that can be used as cell-phone, camera, calendar, clock, game consol, web browser and many other roles at the same time. As an important company in the smartphone industry, Sony is one of the actors of the smartphones evolution and is aiming on keeping innovating this sector.
Our work is a part of this continuous research for innovation. Indeed, it aims to allow smartphones building a personal relationship with their owner by adapting to their specific needs and answering heir specific requests. Let's keep the parallel with Bob to understand what does building a personal relationship with a user concretely means. Let's suppose that Bob's smartphone is able to learn the specific habits of Bob. When Bob forgets to put his alarm clock on a working day, his smartphone can remind him to do it. When an unusual traffic congestion appears in Sunday in the rode that Bob use to take to reach his parents place, his smartphone can inform him. Finally, when his smartphone does not have enough power to play music on Saturday morning, Bob's smartphone can remind him to recharge it because he will probably need it for running. Our work is a start in making it possible for a smartphone to adapt to it's owner's behaviors and to react interactively in some contexts. In other words, it is a start is making smartphones behaves smarter.

Scientifically speaking, this problem can be seen as a clustering problem: our goal is to find different clusters of data points where each cluster represents a particular behavior of the user. Coming back to Bob, running while listening to music on Saturdays morning can be represented by a cluster that contains the running activity, the application launch music, the day Saturday and the time frame 8am-12am. Putting an alarm clock at 7 pm during the working days can be represented as a cluster containing all the days of the week, the notification alarm and the time frame 7pm-8am.
Clustering is a widely addressed problem in machine learning and data analysis, and it has been applied to many contexts and topics. It as been used for example in corpus-text modeling, recommender systems and image recognition. Different approaches have been developed to answer those challenges; the probabilistic latent topic modeling and the matrix factorization are examples of these different approaches. A parallel between our methods and these approaches is made multiple times in this thesis and it will be

shown that it is of a strong benefit.
Our problem sits at the interface between an emerging area of research that takes profit of the existence of a new kind of data and an area that constitutes one of the basis of the emergence of the machine learning and data analysis techniques. From a scientifically point fo view, addressing the clustering problem in a new emerging context makes our problem particularly challenging.

The thesis is organized as follows: In chapter 2, we introduce some notations and definitions, state the problem in a mathematical way and go through the researches done in this field.
In chapter 3, we describe in details the Generative Hidden Class Model for Mixed Data Types (GHCM-MDT). It is model that we developed specifically to answer our needs and that shows to perform better than the other existing methods.
In chapter 4, we introduce other known and widely used models that has been shown to perform very well in doing tasks similar to ours. We use those models as baselines to evaluate the performances of GHCM-MDT. To have a complete overview, we both use some models based on the matrix factorization approach using some sophisticated techniques and others based on advanced methods of probabilistic latent topic modeling.
In chapter 5, we detail the metrics used to test the performance of the different models in performing the task needed.
In chapter 6, we present the results obtained with the different models and compare the performances of the different models to GHCM-MDT.
Finally, chapter 7 presents our conclusions.

Chapter 2

Preliminaries

2.1 Definitions and notations

2.2 Problem Statement

2.3 Related work

je [1]

2.3.1 Problems addressed

2.3.2 Methods used

2.3.3 Evaluation methods

2.3.4 Critics

Chapter 3

Generative Hidden Class Model for Mixed Data Types (GHCM-MDT)

We recall that our task is the following: by observing user logs, we want to discover the behaviors and habits that describe his life.
To that end, we use a usual and common practice when trying to extract some hidden properties (behaviors) from an observable structure (logs): We assume that the logs we are observing are generated by behaviors. Then our task is to find the behaviors that generated the data we are observing. This practice drives the models we are going to discuss in the next sections.

3.1 Hidden Class Model for Mixed Data Types (HCM-MDT)

3.1.1 HCM-MDT model

Let's consider the corpus representation of the smartphone logs. Smartphone logs are represented by a corpus containing $\boldsymbol{R}$

3.1.2 Relationship between HCM-MDT and Probabilistic Latent Semantic Indexing (pLSI)

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

3.2 Generative HCM-MDT (GHCM-MDT) model

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

3.3 GHCM-MDT inference and parameter estimation

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

3.3.1 Gibbs sampling

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

3.3.2 Hyperparameters estimation

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

3.4 Handling unseen data with GHCM-MDT

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

3.5 Relationship between GHCM-MDT and Latent Dirichlet Allocation (LDA)

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

Chapter 4

Baseline Models

4.1 Singular Value Decomposition (SVD)

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

4.1.1 matrix representation of the smartphone logs

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

4.1.2 TF-IDF transformation

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

4.1.3 Predictions with SVD

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

4.2 Linearly Constrained Bayesian Matrix Factorization (LCBMF)

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

4.2.1 The LCBMF model

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

4.2.2 Linear constrains for smartphone logs matrix

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

4.2.3 Predictions with LCBMF

Nunc posuere quam at lectus tristique eu ultrices augue venenatis. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam erat volutpat. Vivamus sodales tortor eget quam adipiscing in vulputate ante ullamcorper. Sed eros ante, lacinia et sollicitudin et, aliquam sit amet augue. In hac habitasse platea dictumst.

4.3 Latent Dirichlet Allocation (LDA)

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

Chapter 5

Evaluation metrics

5.1 Perplexity of unseen data

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

5.2 Missing Features prediction

Sed ullamcorper quam eu nisl interdum at interdum enim egestas. Aliquam placerat justo sed lectus lobortis ut porta nisl porttitor. Vestibulum mi dolor, lacinia molestie gravida at, tempus vitae ligula. Donec eget quam sapien, in viverra eros. Donec pellentesque justo a massa fringilla non vestibulum metus vestibulum. Vestibulum in orci quis felis tempor lacinia. Vivamus ornare ultrices facilisis. Ut hendrerit volutpat vulputate. Morbi condimentum venenatis augue, id porta ipsum vulputate in. Curabitur luctus tempus justo. Vestibulum risus lectus, adipiscing nec condimentum quis, condimentum nec nisl. Aliquam dictum sagittis velit sed iaculis. Morbi tristique augue sit amet nulla pulvinar id facilisis ligula mollis. Nam elit libero, tincidunt ut aliquam at, molestie in quam. Aenean rhoncus vehicula hendrerit.

Chapter 6

Results and Discussion

6.1 Presenting the Dataset

6.1.1 Subsection 1

6.1.2 Subsection 2

6.2 Experimental results

Chapter 7

Conclusion

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

[1] C. J. Hawthorn. Littrow configuration tunable external cavity diode laser with fixed direction output beam. *Review of Scientific Instruments*, 72(12):4477–4479, December 2001. URL http://link.aip.org/link/?RSI/72/4477/1.