

M. Sc. Engineering Thesis

**A Hierarchical Approach for Identifying Social Groups
from Mobile Phone Call Detail Records**

by

Fahim Hasan Khan

Submitted to

Department of Computer Science and Engineering

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Department of Computer Science and Engineering

BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Dhaka-1000, Bangladesh

December, 2015

**A HIERARCHICAL APPROACH FOR IDENTIFYING SOCIAL
GROUPS FROM MOBILE PHONE CALL DETAIL RECORDS**

By

Fahim Hasan Khan

Student ID: 0411052063P

Session: April, 2011

Supervised By

Dr. Mohammed Eunus Ali

Professor

Department of Computer Science and Engineering

MASTER OF SCIENCE IN COMPUTER SCIENCE AND
ENGINEERING

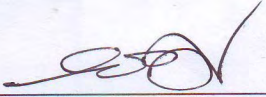
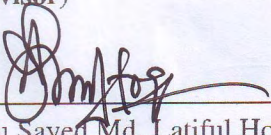
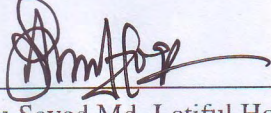
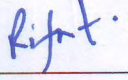

Department of Computer Science and Engineering

BANGLADESH UNIVERSITY OF ENGINEERING AND
TECHNOLOGY

December, 2015

The thesis titled "A HIERARCHICAL APPROACH FOR IDENTIFYING SOCIAL GROUPS FROM MOBILE PHONE CALL DETAIL RECORDS" submitted by Fahim Hasan Khan, Roll No.: 0411052063 P, Session: April/2011 has been accepted as satisfactory in fulfillment of the requirement for the degree of M.Sc. in CSE on 30th December, 2015.

Board of Examiners

- 
1. Dr. Mohammed Eunus Ali **Chairman**
Professor
Dept of CSE, BUET, Dhaka-1000
(Supervisor)
- 
2. Dr. Abu Sayed Md. Latiful Hoque **Member (Ex-Officio)**
Acting Head and Professor
Dept of CSE, BUET, Dhaka-1000
- 
3. Dr. Abu Sayed Md. Latiful Hoque **Member**
Professor
Dept of CSE, BUET, Dhaka-1000
- 
4. Dr. Rifat Shahriyar **Member**
Assistant Professor
Dept of CSE, BUET, Dhaka-1000
- 
4. Dr. Md. Mahbubur Rahman **Member(External)**
Professor
Dept of CSE, MIST, Dhaka-1216

CANDIDATE'S DECLARATION

It is hereby declared that neither this thesis paper nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.



Fahim Hasan Khan

DEDICATED TO PROPHET HAZRAT MUHAMMAD (SWM.)

Acknowledgement

Foremost, I am thankful to Almighty Allah for his blessings for the successful completion of my thesis. I would like to express my heartiest gratitude, profound indebtedness and deep respect to my supervisor, Dr. Mohammed Eunus Ali, Professor, Dept. of CSE, BUET, Dhaka, Bangladesh, for his constant supervision, affectionate guidance and great encouragement and motivation. His keen interest on the topic and valuable advices throughout the study was of great help in completing thesis.

I am especially grateful to Department of Computer Science and Engineering (CSE) of Bangladesh University of Engineering and Technology (BUET) for providing their all out support during the thesis work. My sincere thanks goes to CSE Office staffs for providing logistic support to me to successfully complete the thesis work.

Finally, I would like to thank my family: my parents Late Engr. Md Musa Khan and Mrs. Masuma Jamin and all of those who supported me for their appreciable assistance, patience and suggestions during the course of my thesis.

-Fahim Hasan Khan

Abstract

With the increasing use of mobile devices, now it is possible to collect different data about the day-to-day activities of personal life of the user. Call Detail Record (CDR) is the available mobile phone usage dataset at large-scale, as they are already constantly collected by the mobile operator mostly for billing purpose. By examining this data it is possible to analyze the activities of the people in urban areas and discover the human behavioral patterns of their daily life. These datasets can be used for many applications that vary from urban and transportation planning to predictive analytics of human behavior. In our research work, we have proposed a hierarchical analytical model for finding facts from CDR dataset for progressive exploration of facts on the day-to-day social activities of urban users in multiple layers. In our model, only the raw CDR data are used as the input in the initial layer and the outputs from each consecutive layer is used as new input combined with the original CDR data in the next layers to learn more detailed and deeper facts on social interaction, work and travel activity, friends, family and working relationship and predicting social groups based on these facts. Our proposed model starts with an aggregated overview of the activities of the users in their social life and allows us to gradually focus on smaller groups, using multiple layers of abstraction by applying clustering techniques and prediction classifiers. The uniqueness of our model is that the output in each layer is dependent on the results of the previous layers, thus, allow us to explore fact on social relationships and groups which can not be predicted in a single layered approach. This model utilized the CDR dataset of one month collected from the Dhaka city, which is one of the most densely populated cities of the world. So, our main focus of this research work is to explore the applications of CDR data containing spatio-temporal traces of the mobile phone users for progressive predicting of facts and features of social groups and relationships in a busy city.

Table of Contents

| | |
|--|------------|
| Title page | i |
| Board of Examiners | ii |
| Candidate's Declaration | iii |
| Dedication | iv |
| Acknowledgement | v |
| Abstract | vi |
| Contents | x |
| List of Figures | xii |
| List of Tables | xiv |
| 1 Introduction | 1 |
| 1.1 Overview of Problem Domain | 2 |
| 1.2 Limitations of Previous Works and Our Motivation | 4 |
| 1.3 Objectives and Scope of the Thesis | 5 |
| 1.4 Our Solution Overview | 7 |
| 1.5 Applications | 9 |
| 1.6 Outline of the thesis | 10 |
| 2 Related Works | 11 |

| | | |
|----------|---|-----------|
| 2.1 | City Status Analysis | 11 |
| 2.2 | Human Mobility and Activity Pattern Analysis | 13 |
| 2.3 | Traffic and Transportation Analysis | 14 |
| 2.4 | Other Works | 16 |
| 2.5 | Summary | 17 |
| 3 | Preliminaries | 18 |
| 3.1 | Preliminaries | 18 |
| 3.1.1 | Machine Learning for Spatio-temporal Prediction | 18 |
| 3.1.2 | Techniques for finding facts from Big Data | 19 |
| 3.1.3 | Classification | 20 |
| 3.1.4 | Naive Bayes classifier | 22 |
| 3.1.5 | Support Vector Machines | 22 |
| 3.1.6 | Clustering | 23 |
| 3.2 | Summary | 27 |
| 4 | Our Proposed Framework | 28 |
| 4.1 | The Hierarchical Exploration Model | 28 |
| 4.2 | The Layered Approach | 30 |
| 4.2.1 | Layer 1 | 30 |
| 4.2.2 | Layer 2 | 31 |
| 4.2.3 | Layer 3 | 32 |
| 4.2.4 | Layer 4 | 33 |
| 4.2.5 | Layer 5 | 34 |
| 4.2.6 | Beyond Layer 5 | 34 |

| | | |
|----------|--|-----------|
| 4.3 | Aggregated Social Closeness (ASC) Score | 35 |
| 4.4 | Summary | 35 |
| 5 | Methodology | 36 |
| 5.1 | Validation | 37 |
| 5.2 | Preprocessing | 39 |
| 5.3 | Layer 1 | 40 |
| 5.4 | Layer 2 | 46 |
| 5.4.1 | Classifying City blocks | 50 |
| 5.5 | Layer 3 | 51 |
| 5.5.1 | Tagging City blocks | 54 |
| 5.6 | Layer 4 | 56 |
| 5.7 | Layer 5 | 61 |
| 5.8 | ASC Score and Aggregated Social Group Prediction Model | 63 |
| 5.9 | Summary | 64 |
| 6 | Results and Analysis | 65 |
| 6.1 | Data Collection and Dataset | 65 |
| 6.2 | Experimental Setup | 66 |
| 6.3 | Results | 68 |
| 6.4 | Validation and Accuracy | 82 |
| 6.5 | Summary | 85 |
| 7 | Software and Visualization | 86 |
| 7.1 | Graphical User Interfaces | 86 |
| 7.2 | Outputs and Visualizations | 89 |

| | | |
|----------|---|-----------|
| 7.3 | Supplementary Visualization Tools | 91 |
| 7.3.1 | Weka | 91 |
| 7.3.2 | Google Map API | 92 |
| 7.4 | Summary | 93 |
| 8 | Conclusion and Future Works | 94 |
| 8.1 | Conclusion | 94 |
| 8.2 | Future Works | 95 |
| | Bibliography | 95 |

List of Figures

| | | |
|------|--|----|
| 1.1 | Examples of (a) user accessible CDR (b) encrypted CDR collected by a cell operator | 3 |
| 3.1 | Process of Knowledge Discovery in Databases (KDD) | 20 |
| 4.1 | The Layered Approach | 29 |
| 5.1 | The Hierarchical Progressive Exploration Model | 38 |
| 5.2 | Tree-based Hierarchical Presentation of Predicted Groups in Five Layers | 39 |
| 6.1 | Directory structure of the combined knowledge database | 67 |
| 6.2 | a) Number of calls per day b) Number of active user per day | 68 |
| 6.3 | User activity in different periods of the day | 69 |
| 6.4 | Initial voronoi diagram on the city map | 70 |
| 6.5 | Voronoi diagram on the city map after colorization based on call activity | 70 |
| 6.6 | Voronoi diagram showing call activity in different times of day | 71 |
| 6.7 | Comparing call traffic variations from usage score, μ in three consecutive weeks | 72 |
| 6.8 | Result of using EM clustering algorithm on the CDR to find POIs | 74 |
| 6.9 | Visualization of clusters as a result of EM indicating POIs | 74 |
| 6.10 | Finding Home of an User using a) EM Clustering Algorithm b)XMeans Clustering Algorithm | 74 |

| | |
|--|----|
| 6.11 Finding Workplace of an User using a) EM Clustering Algorithm b)XMeans Clustering Algorithm | 75 |
| 6.12 Visualization of home, workplace and other POIs of an user | 76 |
| 6.13 Working pattern of the users: Regular vs Irregular | 78 |
| 6.14 Mobility Pattern of Regular Working Users | 79 |
| 6.15 Traveling Route of an user from home to workplace | 79 |
| 6.16 Finding family members from home location and calling relation | 82 |
| 6.17 Part of data collected directly from users for validation | 83 |
| 7.1 Main menu of our software | 87 |
| 7.2 GUI for finding home and workplace for a single user | 87 |
| 7.3 Built-in Map Viewer for viewing home and workplace | 88 |
| 7.4 GUI for processing full CDR for home workplace info | 88 |
| 7.5 GUI showing output summary after successful completion of execution | 89 |
| 7.6 The generated output file showing the location of home and workplaces of users and traveling distance between two locations | 90 |
| 7.7 The generated output file showing the predicted traveling pattern of a user | 90 |
| 7.8 Preprocessing CDR for Clustering Using Weka | 91 |
| 7.9 Using Google Map API for location data visualization | 92 |

List of Tables

| | | |
|------|--|----|
| 5.1 | Notations used in this text and their definition | 40 |
| 6.1 | Sample data from CDR dataset | 66 |
| 6.2 | User activity in different periods of the day | 69 |
| 6.3 | Comparing call traffic variation in two consecutive weeks | 71 |
| 6.4 | Classifying user based on call activity | 72 |
| 6.5 | Detected calling relationship among users | 72 |
| 6.6 | POIs of a random user | 73 |
| 6.7 | Status of City Areas in Working Hour | 75 |
| 6.8 | Part of Home and Workplace information data of all the users | 77 |
| 6.9 | Working pattern of the Users: Regular vs Irregular | 77 |
| 6.10 | Traveling distance to workplace for regular workers | 78 |
| 6.11 | Traveling Route of a user | 80 |
| 6.12 | Part of list of users living in same neighbor-group | 81 |
| 6.13 | Tracefile generated by Transport Type classifier | 82 |
| 6.14 | Examples of user profile predicted upto Layer 3 | 82 |
| 6.15 | Accuracy of predicting home, workplace and other POIs | 84 |
| 6.16 | Accuracy of predicting working groups | 84 |
| 6.17 | Prediction of city area type | 84 |

| | |
|---|----|
| 6.18 Accuracy of detecting social groups in the final layer | 85 |
|---|----|

Chapter 1

Introduction

With the rapid developments in technologies involving hand-held devices and wireless communications people are increasingly using more mobile phones day by day. The ubiquitous mobile devices carried by people all over the world are regularly being used as a massive source of spatio-temporal data collected for different purposes [21]. Spatio-temporal data collected from any mobile device are a very convenient tool to understand city dynamics, environment, traveling and the behavioral patterns of the people within [39, 15, 28]. All the modern smartphones has a number of sensors (GPS, g-sensor, magnetic sensor, etc.) in-built which facilitate us to collect spatio-temporal data about user activity very easily. But, as not everyone uses smartphone and all the variations of smartphones do not have similar sets of sensors. Even, if they have, using the sensor is dependent on the decision of the users. So, collecting data in this way can be expensive and inconvenient at times.

Mobile phone service providers continuously collect data in trace files from all the active mobile phones using cellular service from the towers of their cell networks. Although this data is primarily collected for billing purposes and network traffic analysis [18], they contain spatio-temporal traces of the mobile phone users. This spatio-temporal trace proves to be a valuable source of inexpensive data about user activity, which is being collected continuously. Furthermore, in this method all types of mobile phones, ranging from the cheapest phone to most expensive smartphone functions as a similar type of spatial-temporal sensor. This kind of data can be very effective in understanding the relationship of the people and their daily activities in the

context of an urban setup.

By analyzing such data, it is possible to identify patterns and relations, revealing insightful information about the city, which in turn facilitates the authority, service providers and citizens with a better way of understanding, decision-making, discovery and exploration of the urban life [38]. Again, analysis of human behavior and mobility patterns is a very important research topic in various fields such as geography, urban and transportation planning, telecom sector, social science, and human psychology.

Accompanied by the usage of increasing number of mobile phone users, the mobile networks have become an enormous ubiquitous sensing system. These networks are constantly collecting call data logs from mobile phones of all types. So, mobile phone call data recorded by various cell towers all over the city are a cheap and easy-to-collect source of data contains a huge amount of spatio-temporal traces of the users. As mobile devices are regularly being carried by a huge part of the overall population of earth, those are potentially a very effective tool of pervasive sensing platform for collecting nearly real-time, fine-grained spatio-temporal data. This data is collected more effectively with finer granularity in highly populated urban areas and particularly in the developed and developing countries, where mobile phone penetration is almost hundred percent. Therefore, utilization of these spatio-temporal responses from mobile device users can be a prospective option for the researchers to formulate various methods for observing and sensing the city dynamics, identifying mobility patterns and predicting social features of the citizens in the urban area.

1.1 Overview of Problem Domain

Call Detail Record (CDR) is a data record generated by telecommunication equipment like a telephone exchange or cell tower. This data records are log files containing details of a single instance of communication activity, like voice calls, short messaging service (SMS) texts, and Internet and data services initiated by the phone user, which has been processed by specific telecommunication equipment. The mobile phone service providers keep records of all outgoing communication activity of mobile devices for billing and other purposes. Every single entry of CDR data includes the

| Call Date | Call time | Called Number | Duration | Charge | Call Type | FNF | Usage Type |
|-----------|-----------|---------------|----------|--------|-----------|-----|------------|
| 27-Sep-15 | 18:25:26 | 880[REDACTED] | 74 | 0.0000 | IN | N/A | VOICE |
| 26-Sep-15 | 16:24:32 | GPRS | 80 | 0.0186 | N/A | N/A | EDGE |
| 26-Sep-15 | 14:10:22 | GPRS | 748 | 0.1720 | N/A | N/A | EDGE |
| 26-Sep-15 | 8:17:38 | 880[REDACTED] | 38 | 0.0000 | IN | N/A | VOICE |
| 26-Sep-15 | 19:41:42 | 880[REDACTED] | 276 | 0.0000 | IN | N/A | VOICE |
| 26-Sep-15 | 21:09:02 | 880[REDACTED] | 96 | 2.1321 | OUT | N | VOICE |
| 26-Sep-15 | 13:15:03 | 880[REDACTED] | 16 | 0.4264 | OUT | N | VOICE |
| 26-Sep-15 | 19:19:11 | 880[REDACTED] | 0 | 0.0000 | IN | N/A | SMS |
| 26-Sep-15 | 8:11:05 | 880[REDACTED] | 64 | 0.0000 | IN | N/A | VOICE |
| 26-Sep-15 | 8:38:39 | 880[REDACTED] | 404 | 0.0000 | IN | N/A | VOICE |
| 25-Sep-15 | 9:53:31 | 880[REDACTED] | 0 | 0.0000 | IN | N/A | SMS |

a)

| | | | | | |
|----------------------|------------|------------|--------|-------------|-------------|
| "AAH03JAAQAAA09VAA+" | "20120714" | "10:44:04" | "68" | "23.758301" | "90.402199" |
| "AAH03JAAQAAA09VAA+" | "20120714" | "21:16:23" | "60" | "23.758301" | "90.402199" |
| "AAH03JAAQAAA09VAA/" | "20120708" | "22:10:37" | "1527" | "23.701700" | "90.429199" |
| "AAH03JAAQAAA09VAA/" | "20120711" | "12:12:54" | "1103" | "23.724199" | "90.405602" |
| "AAH03JAAQAAA09VAA/" | "20120711" | "10:26:33" | "304" | "23.722200" | "90.409203" |

b)

Figure 1.1: Examples of (a) user accessible CDR (b) encrypted CDR collected by a cell operator

following parameters along with others: a random ID number of the phone, independent of the user, device and phone number; the exact time and date; call duration and location in latitude and longitude of the cell tower that provided the network signal for the communication activity of the mobile device. The CDR data is stored in encrypted files for ensuring the anonymity and privacy of the mobile device users.

The two major limitations of the CDR data are that, a data entry is generated only when a communication activity is initiated and the location of that activity is recorded as the geographic position of the cell tower facilitating the activity. So, the availability of data is dependent on the frequency of the user activity in mobile devices. On the other hand, the spatial granularity of the location data depends on the granularity of a cell tower and the density of cell tower varies in different areas. According to the International Telecommunication Union (ITU), in 2011 there were more than 85 mobile phone users among every 100 residents of many major which has increased almost 5 times from 2001 to 2011 [39]. Because of the much greater number of cell towers and frequency of communication activity, the CDR data collected from a densely populated city can

largely overcome the aforementioned limitations and make urban activity analysis more effective.

Furthermore, massively collected spatio-temporal data or Big Data like CDR consists of billions of lines of data entries in log files, mostly in basic text format, which can take terabytes of space in the storage media. So, applying the proper techniques for preprocessing and finding facts from them is another challenge[18]. Processing these massive data demands a good amount of computational resources. Although, given the recent development of processing power and storage capability, it is becoming easier to handle these kinds of data. Yet, without well-handled techniques and optimized algorithms, processing and fact finding from big data can be a challenging task.

1.2 Limitations of Previous Works and Our Motivation

Many of the foundational concepts of sociology and social psychology have originated from the observation of the activities of people living in urban areas. A social group within social sciences has been defined as two or more people who interact with one another, share similar characteristics, and collectively have a sense of unity. Individuals in groups are connected to each other by social relationships. CDRs from mobile phones have become very popular in social and urban activity analysis over the last decade. Researchers from Sensible City Lab, MIT developed City Browser, a software tool to perform spatio-temporal analysis of CDR data to discover human activity, mobility behavior and flow of people across the city in a time window [1]. By presenting an insightful overview, Steenbruggen and others developed a typology of the various studies which utilize mobile phone data for spatial research in a smart city environment [39]. In a related research work, a strong correlation is established in daily activity patterns within the group of people who share a common work area's profile using CDR [28]. A few more works were done to establish correlations between CDR data and daily life activity patterns of the citizens of urban area [23, 8, 7]. Likewise, CDR data were used for estimating the presence of citizens and its spatio-temporal pattern in different urban regions [40].

Another promising research area is using CDR data to analyze travel activities like

origin-destination matrix, human trajectory and traffic flow analysis, etc. [15, 16, 46, 42]. In a research, information was found on how the usage of mobile phones correlates with individual travel behavior by exploring the correlation between mobile phone call frequencies and three indicators of travel behavior: radius, eccentricity, and entropy [46]. Another recent work on human activity-travel behavior (ATB) showed the monthly variability in human activity spaces and locations after analyzing one year's CDR data [17].

As we can see that, CDRs have been used extensively in a good number of researches on social and urban analysis. Although many researches has been done exploring different kinds of urban activity from CDR, no efforts have been made so far to identify different social groups, their relationships and investigate city area based on social group activities. The research works done so far mainly focused on analysis of various types of user activities, travel behaviors or city dynamics from a different perspective. Besides, the frameworks proposed in all the previous works used CDR data in a single-layered procedure to find facts, which has limited exploration potential. According to the best of our knowledge, no hierarchical exploration model with multiple layer is applied so far for analyzing spatio-temporal CDR data for finding social groups and relationships in urban areas.

The efficient processing and utilization of CDR data are also major challenges for the researchers, especially since there is a vast amount of information collected, which renders the problem as a Big Data problem. In order to address the large volume of the data, our proposed solution provides a unique and novel framework comprising layered approach which begins with an aggregated overview of the whole CDR and allows gradual focus on smaller sets of data.

1.3 Objectives and Scope of the Thesis

The objective of our thesis is to progressively explore social activities, social groups and relationships by applying fact finding techniques on CDR data in a hierarchical layer approach. Using our proposed framework we have predicted the social interactions between mobile phone users and social groups based on social interaction, work and

travel activity, friends, family and working relationship, and investigated status of the city areas based on social activities [20].

Our main objective is to discover social relationships and groups from CDR data gradually by using our proposed hierarchical exploration model with multiple layers. We have proposed a set of algorithms in different layers based on predictive classifiers and clustering techniques to investigate social groups and relationships by statistical analysis of the CDR. From our comprehensive analysis of the CDR we predict social groups such as extroverts and introverts, regular and irregular working people, frequent travelers, family, friends, co-workers, etc. in different layers progressively. The information obtained in different layers of our model led to the investigation of social lifestyles, like, working patterns, traveling patterns, patterns of social relations in a densely populated urban area.

Investigating some city area features related to social activities is another objective of our research. Using our framework on CDR data we have identified city areas like places of common interests, densely populated area, residential area, commercial area, etc. and predicted the status of social activities in these areas in different time periods. As the feature of any city is highly dependent on the underlying social structure and culture, exploration of these features is very much correlated with the investigation of social activities and relationships.

To accomplish the objective we formulate and devise a progressively exploring hierarchical prediction model with expandable numbers of layers utilizing fact finding algorithms for investigation of social groups and related facts from CDR data. The novelty of our work is that, the hierarchical framework provides flexibility of exploration and as the number of layers increase we can learn deeper facts on the city based on CDR data. In each of the layers we propose a set of prediction classifier and clustering techniques to find facts and predict social patterns.

The auxiliary objective of our work is to develop software for preprocessing the CDR data, applying fact finding algorithms and visualizing the different results obtained in progressive layers of our model in an automated or semi-automated way.

1.4 Our Solution Overview

In this thesis work, we have used CDR data to identify social groups and relationships in a hierarchically designed layered approach for progressive exploration by analyzing and correlating different facts including social interactions, work and travel activities. Our proposed model starts with an aggregated overview of the activities of the users in their social life and allows us to gradually focus on smaller groups, using multiple layers of abstraction. In each of the layers we have designed and employed a number of prediction classifiers and clustering techniques for finding facts on social activities and investigating social groups. As we proceed deeper within the layers, we obtain more detailed information using the facts obtained in the previous layers. Obtained facts from the modules in each layers are stored in a combined knowledge base along with the raw CDR utilized in the layers progressively.

In the very first layer, at first we introduce a number of feature extraction algorithms that addresses the challenges of extracting information from massive CDR data and extract features comprising calling pattern and unique locations visited by individual users. Then, we have used the raw CDR data to analyze the daily call activity and calling pattern of the citizens to predict a number of social groups including Heavy callers, Regular callers, Minimal callers, late-night callers, professional callers. We propose linear prediction classifier by using values for each feature of call activity to predict caller groups. Also, we examine the overall calling pattern of the city in different time periods and locations. Additionally, we generate a call graph to determine the possible relations among callers.

In the next layer, we propose algorithms based on two clustering techniques, X-means and EM to detect the home, workplace and other frequently visited places, collectively called Points of Interest (POIs) for the users. We have applied both centroid-based clustering, X-means and distribution based clustering, EM for finding POIs of the users and compared their performances. This is one of the critical information for our work, as most of the major social relations are based on home and workplace of the citizens. Also, we have developed prediction models to successfully predict working days and weekends of a user, rush hours and non-rush hours in different city areas in different times of the day.

Using the classifiers proposed in the third layer, we apply knowledge gathered so far to categorize the regular working people like professionals, businessmen, students, etc. and the home staying people like homemakers, retired and unemployed people, etc. Further, we have classified the working people on the basis of their mobility and traveling distance in the city and developed a prediction model to predict their regular traveling routes to workplace. Additionally, in this layer, our city area tagging classifier categorizes city areas as residential and commercial.

The fourth layer involves microlevel analysis of social groups and relationships to find smaller social circles, as now we have enough information in our combined knowledge database. At first we propose some hypotheses based on our observation of the social groups and then apply them to formulate a set of classifiers to predict social groups like, neighbors, co-workers, and special relationship groups like fellow travelers and people living in official accommodations.

In our final layer, we formulate and devise prediction classifiers to find the personal level social relationships and groups revolving around them based on social communications and interaction. The social groups we predicted here are family, friends, colleagues and closely acquainted people based on common relationship features.

Furthermore, we propose a statistical prediction model to find the closeness of two users based on the social interactions between them predicted in different layers. As a final result of our overall fact finding scheme, we calculated a score from all the combined facts to measure the probability of social closeness between two users, which we call the Aggregated Social Closeness (ASC) between them.

We have used real life CDR collected from the largest telecom operator of a busy city as a test dataset to design our progressive exploration model. We evaluated the accuracy of our methods using k-fold cross validation on our test CDR data. Also, as the users identification in our CDR is encrypted by the provider telecom operator to maintain anonymity, we have collected call record data from some volunteer users with known social relationships and group membership from the same city and validated our model further. From our validation we can see that in the initial layers, the accuracy of our results are almost 100 percent. But, as we explore deeper into the social groups and relationships in the later layers, the accuracy of our results diminishes.

1.5 Applications

The main application of our model is to formulate and devise methods and tools for the concerned people to help them visualizing and understanding the social activities, social groups and their relationship, as well as providing a better way of discovery and exploration of a busy city. Among many possibilities, a group of applications comprises those whose main beneficiary is the owner of a mobile phone. Even the simple phone acts as an intermediate tool to access data on its location and will position it on a map. Some of the possible applications for mobile phone users are as follows:

- interactive information service
- traffic services
- advertisements and news services
- recommender services
- social group and networking services

Another series of applications consists of those whose main beneficiary is not the owner of a mobile phone but rather other bodies, generally public authorities and private companies. This group of applications includes the services that this paper is concerned with, and some of them are listed here:

- mobile phone operators
- emergency services
- family safety services
- law-enforcement authorities
- real time traffic systems
- transport management authorities
- city planning authorities
- online shops and telemarketers

These applications represent one of the sectors with greatest market potential in the context of mobile telecommunications systems. The ongoing trends are already showing that in future, these services will become one of the principal sources of income for phone operators. In addition, non-operator corporations may well think up new applications and/or modify existing ones, thanks to the possibilities offered by

localization processes, which will help add value to their products, boost sales and open up fresh opportunities in new markets.

Furthermore, our work has applicability for research in the fields of geography, urban and transportation planning, telecom sector, business, social science, predictive analytics, etc. Also, this unique research work was done based on the CDR from a densely populated urban area of an underdeveloped country with many distinctive features. We have designed and developed a software tool and visualizer that will assist city managers and different service providers to render better service to the citizens.

1.6 Outline of the thesis

In Chapter 2, we have presented some of the recent research works with different objectives using CDR related to our topic.

Chapter 3 briefly explains some preliminary topics related to our problem and proposed framework. Then, we have discussed the framework of our proposed hierarchical exploration model in Chapter 4.

Chapter 5 illustrates the detail of different methods and algorithms we have used in different layers of our model. This chapter presents the overall technical detail of our methodology

Chapter 6 focuses on the experimental setups and results. It also illustrates the experimental data as well as environment of our research and examines the experimental results. Finally, the analysis of different experimental results are presented in this chapter.

We have presented the software in Chapter 7, which we developed as a part of our thesis work to process the CDR using our proposed model and visualize the results. Here we demonstrate all the features of our software. Additionally, the supplementary tools used in our work is discussed briefly in this chapter.

Finally, Chapter 8 concludes our thesis. This Chapter also includes the outlines of some future works related to this dissertation.

Chapter 2

Related Works

A rapidly increasing number of mobile phone users has motivated researchers from various fields to study its social and economic impact. With the extensive records of mobile phone data such as calling pattern and location of the mobile phone user, analyses have been performed on numerous aspects including behavioral routine, call prediction, and dynamics in human mobility. Over the years, diverse approaches are taken by researchers to exploit the applicability of simple mobile phone call logs, in several cases combining with GSM/WiFi/GPS traces or other types of additional data.

2.1 City Status Analysis

An insightful overview is given and a typology is developed of the various studies which utilize mobile phone data for spatial research in a smart city environment by Steenbruggen et al [39]. A research was done showing the relevance between real human trajectory and the one obtained through mobile phone data of real cellular network activity in the Boston metropolitan area using different interpolation methods and taking mobility parameters into consideration [15]. In a similar work based on a large mobile phone data of nearly one million records of the users in the central Metro-Boston area, a strong correlation is established in daily activity patterns within the group of people who share a common work areas profile [28]. In addition, within the group itself, the similarity in activity patterns decreases as their work places become apart. A software tool named City Browser was developed by researchers

from Sensible City Lab, MIT to perform spatio-temporal analysis of CDR data to discover human mobility behavior and flow of people across the city in given time windows [1].

An article [14] presents a field experiment nicknamed Mobile Century, which included 100 vehicles carrying a GPS-enabled Nokia N95 phone driving loops on a 10-mile stretch of I-880 near Union City, California, for 8 hours. Data were collected using virtual trip lines, which are geographical markers stored in the handset that probabilistically trigger position and speed updates when the handset crosses them. A survey was done in [21] on existing mobile phone sensing algorithms, applications, and systems used in many sectors, including business, healthcare, social networks, environmental monitoring, and transportation. A review is presented in [33] considering the state of practice in relation to using mobile phones as traffic probes, assesses the prospects for this data collection option and identifies unresolved issues that may have implications for obtaining realtime traffic information using mobile phones as probes.

A research group presented Nericell [25], a system that performs rich sensing by piggybacking on smartphones that user carry with them in normal course. In this work, they focus specifically on the sensing component, which uses the accelerometer, microphone, GSM radio, and/or GPS sensors in these phones to detect potholes, bumps, braking, and honking. Another paper [38] provides a systematic overview of the main studies and projects addressing the use of data derived from mobile phone networks to obtain location and traffic estimations of individuals, as a starting point for further research on incident and traffic management. In addition to a literature review, the main findings on a project called Current City project are presented, which is a test system in Amsterdam, Netherlands for the extraction of mobile phone data and for the analysis of the spatial network activity patterns. In [38], authors extracted GSM signaling data from a selected area around Munich, Germany for three months in order to detect road traffic congestion information directly from the mobile network. Another work is done in [24] which present a method to analyze the urban blocks' property and activity patterns based on real world cell phone data from Beijing.

2.2 Human Mobility and Activity Pattern Analysis

Human mobility pattern is highly predictable, especially, for urban citizens who tends to live a well-organized routine life. People tend to return a few frequent locations and follow simple repeated patterns despite the diversity of the their travel history. Gonzalez et al. [13] examine six-month trajectory of 100,000 mobile phone users and find a high regularity degree in human trajectories contrasting with estimation by Levy flight and random walk models. The most recent study in human mobility based on a large mobile phone data by Song et al. [37], whose result is consistent with Gonzalez et al.s [13] that human mobility is highly predictable. Based on data from 50,000 mobile phone users, they find that predictability in human mobility is independent of distance that each individual regularly travel and show that the predictability is stabled at 93 percent for all regular traveled distances of more than 10km. Using GPS data, Do et al. [10] found out that most people visit 2 - 4 places every day and calendar (day/time) has significant impact on peoples pattern of visiting places.

Azevedo et al. [2] study pedestrian mobility behavior using GPS traces captured at Quinta da Boa Vistas Park in Rio de Janeiro (Brazil). Movement elements are analyzed from data collected from 120 pedestrians. They find that the velocity and acceleration elements follow a normal distribution while the direction angle change and the pause time measure fit better to log normal distribution. Based on 226 daily GPS traces of 101 subjects, Lee et al. [23] develop a mobility model that captures the effect of human mobility patterns characterized by some fundamental statistical functions. With analytical and empirical evidence, they show that human movement can be expressed using gaps among fractal waypoints [30] It is also reflected from their work, that people are more attracted to more popular places. With a large set of mobile phone data, Candia et al. [8] study spatiotemporal human dynamics as well as social interactions. They investigate the patterns in anomalous events, which can be useful in real-time detection of emergency situation. At the individual level, they find that the interevent time of consecutive calls can be described by heavy-tailed distribution, which is consistent with the previous reports on other human related activities.

Research works cane out with a strong assumption that users move linearly over time. This hypothesis is in a high contrast with the results obtained in [12] that show the

tendency of users to stay in the vicinity of their call places. Authors in [12] propose a probabilistic inter-call mobility model, using a finite Gaussian mixture model to determine users position between their consecutive communication events (call or SMS) using Call Data Records. The model evaluates the density estimation of the spatio-temporal probability distribution of users position between calls, but it does not give an approximation of the fine-grained trajectory between calls. User displacements using GPS traces have been analyzed in [31]; the authors find the displacement behavior show Levy walk properties. Another research suggested that human interaction data, or human proximity, obtained by mobile phone Bluetooth sensor data, can be integrated with human location data, obtained by mobile cell tower connections, to mine meaningful details about human activities from large and noisy datasets [11]. A literature proposed NextMe a novel scheme to enhance the location prediction accuracy by leveraging the social interplay revealed in the cellular calls [47]. While very interesting in order to model inter-contact time distributions and general massive mobility, such random-based approaches cannot give precise approximations between given points on a per-user basis.

2.3 Traffic and Transportation Analysis

Spatio-temporal data collected from mobiles phones have been used extensively for traffic and transportation analysis in urban area for a long time. The paper in [4] provides a review about how to obtain parameters related to traffic from cellular-network-based data, describing methods used in existing simulation works as well as field tests in the academic and industrial field. Similarly, a technical note [22] was published on Collection Methods and Applications of Road Traffic Data. Another work in [9] presents two Bayesian framework based traffic estimation models by the measurement of cell handoff data of floating vehicles.

An analysis of mobile phone call intensity and taxi volume in Lisbon, Portugal was carried out in [29, 41], where, based on one month of observation, the authors found that the variation in the amount of mobile phone calls was strongly correlated with the taxi volume of the previous two hours. Hence taxi volume can be used to predict mobile phone call intensity of the next two hours. In addition, they found that the level

of inter-predictability varied across different time of the day; taxi was a predictor during PM hours while mobile phone call intensity became a predictor for taxi volume in AM hours. Another research provided a deeper understanding of how usage of mobile phones correlates with individual travel behavior by exploring the correlation between mobile phone call frequencies and three indicators of travel behavior: radius, eccentricity, and entropy [46]. The methodology is applied to a large dataset from Harbin city in China.

An approach for extracting origin destination information from mobile phone data was made in [45] and the work is updated in [44]. Origin-destination matrices was developed in another work using the same CDR data form Dhaka which we used for our work [16]. Traffic origin destination data is one of the most important pieces of information required for effective network management and strategic planning. Origin destination (OD) matrices provide an estimate of the number of vehicles traveling between points on a network over a given period of time. A similar effort is made in [6] using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. Another approach was proposed in [5], where the flow of mobile phones in a cell-phone network is measured and correlated to traffic flow. This methodology is based on the fact that a mobile phone moving on a specific route always tends to change the base station nearly at the same position.

Some work was done on Human activity-travel behaviour (ATB), which is a complex pattern of paths and activities in space and time and is the outcome of the interconnection between individual factors, interaction with other individuals, and external factors such as the surrounding environment and social structure [17]. It is reshaped by the socio-economic attributes, as well as the needs, life values, preferences, attitudes, prejudices and habits of individuals. The degree of variability clearly varies due to methodological differences in how human ATB and it was measured by some reserachers [35]. Variability in individual weekly ATB has been examined in some studies covering from one week up to a period of six weeks [34]. The results suggest a weekly pattern in human ATB that is spatially and temporally stable: individual ATB is more routine during the working week, while at weekends it is more dispersed with respect to activities and spatial extent. In contrast, Buliung et

al. [3] found the spatial extent of ATB to be larger during weekdays, whereas the first study also reported greater day-to-day variation in spatial behaviour during the week than on weekends. However, the latter study suggests that, towards the weekend, activities are more impulsive in terms of perceived flexibility in time and space. On the other hand, Schnfelder and Axhausen [36] argue that some activities are performed over different durations, which results in different patterns for certain activities, such as leisure activities or shopping in particular. Overall, Buliung et al. [3] note that while a set of various factors influencing individual ATB has been explored, less effort has been directed towards the study of temporal variation in ATB.

2.4 Other Works

Many other diversified research work was done using CDR and other spatio-temporal data. Authors in [43] infer the top-k routes traversing a given location sequence within a specified travel time from uncertain trajectories. Here, they use check-in datasets from mobile social applications. Their proposed methods permit to identify the most popular travel routes in a city, but they do not allow constructing time-sensitive routes. Authors in [26] propose a spacetime prism approach, where the prism represents reachable positions as a spacetime cube, given users origin and destination points i.e., the assumption of knowing the location of a user at one time and then again at another time fits well mobile phone data in which we only know users position during their communication events as well as time budget and maximum speed. Spatial prisms so allow evaluating of binary statements, such as the potential of encounter between two moving users. However, the maximum speed cannot be set for all users in general, which limits the model applicability. There is an US patent [19] for a method and gadget for providing vehicular traffic information using existing mobile phone network. There are two Google Patents, one on method and apparatus for collecting diagnostic messages and collating them into correlated groups to be matched to specific calls, to identify and diagnose issues with those calls [27] and another on method and apparatus for analyzing customer call data and related call information to determine call characteristics [32].

2.5 Summary

In this chapter have reviewed some of the research works done in the past decades related to our work. The purpose of the overall discussion in this chapter is to setup a baseline for the framework we are going to propose in the following chapter.

Chapter 3

Preliminaries

3.1 Preliminaries

3.1.1 Machine Learning for Spatio-temporal Prediction

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

Machine learning is closely related to computational statistics, which aims at the design of algorithm for implementing statistical methods on computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Example applications include spam filtering, optical character recognition (OCR), search engines and computer vision. Machine learning is sometimes conflated with data mining, although that focuses more on exploratory data analysis. Machine learning and pattern recognition "can be viewed as two facets of the same field." When

employed in industrial contexts, machine learning methods may be referred to as predictive analytics.

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are,

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.

Between supervised and unsupervised learning is semi-supervised learning, where the teacher gives an incomplete training signal: a training set with some (often many) of the target outputs missing. Transduction is a special case of this principle where the entire set of problem instances is known at learning time, except that part of the targets are missing.

3.1.2 Techniques for finding facts from Big Data

The process of finding and predicting facts or data mining is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Here, the goal is the extraction of patterns and knowledge from large amount of data, not the extraction (mining) of data itself. So, it is applied to any form of large-scale data or information processing like collection, extraction, warehousing, analysis, and statistics as well as any application of computer decision support system,

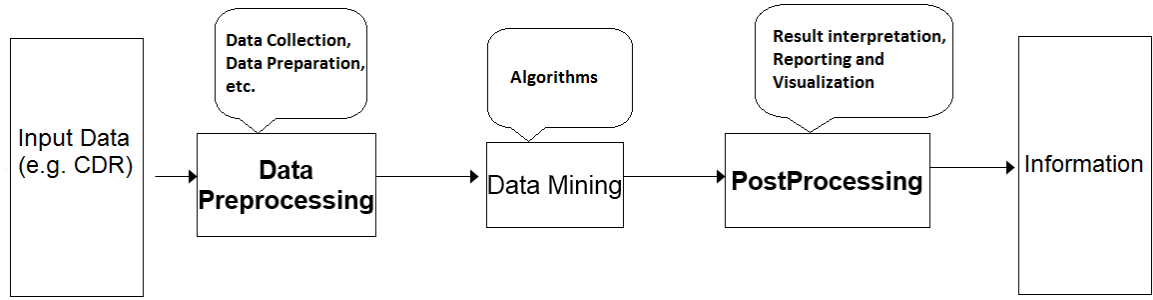


Figure 3.1: Process of Knowledge Discovery in Databases (KDD)

including artificial intelligence, machine learning, and business intelligence. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation and preprocessing, nor result interpretation, reporting and visualization is part of the data mining step, but do belong to the overall knowledge discovery in databases (KDD) process as additional steps.

Data mining involves six common classes of tasks. They are, Anomaly detection, Association rule learning, Clustering, Classification, Regression and Summarization. In our work, we have mainly used Classification and Clustering for finding social groups from the spatio-temporal information inside CDR.

3.1.3 Classification

In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations or instances whose category membership is known. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is

available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

Often, the individual observations are analyzed into a set of quantifiable properties, known variously as explanatory variables or features. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category. Terminology across fields is quite varied. In statistics, where classification is often done with logistic regression or a similar procedure, the properties of observations are termed explanatory variables and the categories to be predicted are known as outcomes, which are considered to be possible values of the dependent variable. In machine learning, the observations are often known as instances, the explanatory variables are termed features (grouped into a feature vector), and the possible categories to be predicted are classes.

Linear classifiers

A large number of algorithms for classification can be phrased in terms of a linear function that assigns a score to each possible category k by combining the feature vector of an instance with a vector of weights, using a dot product. The predicted category is the one with the highest score. This type of score function is known as a linear predictor function and has the following general form:

$$\text{score}(X_i, k) = \beta_k \cdot X_i \quad (3.1)$$

where X_i is the feature vector for instance i , β_k is the vector of weights corresponding to category k , and $\text{score}(X_i, k)$ is the score associated with assigning instance i to category k . In discrete choice theory, where instances represent people and categories represent choices, the score is considered the utility associated with person i choosing category k .

Algorithms with this basic setup are known as linear classifiers. What distinguishes them is the procedure for determining (training) the optimal weights/coefficients and the way that the score is interpreted.

3.1.4 Naive Bayes classifier

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $X = (x_1, \dots, x_n)$ representing some n features (independent variables), it assigns to this instance probabilities

$$p(C_k | x_1, \dots, x_n) \tag{3.2}$$

for each of K possible outcomes or classes.

3.1.5 Support Vector Machines

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear

gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When data is not labeled, a supervised learning is not possible, and an unsupervised learning is required, that would find natural clustering of the data to groups, and map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering is highly used in industrial applications either when data is not labeled or when only some data is labeled as a preprocessing for a classification pass; the clustering method was published.

3.1.6 Clustering

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to some predesignated criterion or criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated for example by internal compactness (similarity between members of the same cluster) and separation between different clusters. Other methods are based on estimated density and graph connectivity. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic

task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Here we have discussed some of the popular clustering algorithms which we have used in our work.

K-means clustering

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

X-means Clustering

Despite its popularity for general clustering, k-means suffers three major shortcomings; it scales poorly computationally, the number of clusters K has to be supplied by the user, and the search is prone to local minima. X-means provides solutions for the first two problems, and a partial remedy for the third. Building on prior work for algorithmic acceleration that is not based on approximation, it introduces a new algorithm that efficiently searches the space of cluster locations and number of clusters to optimize the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) measure. The innovations include two new ways of exploiting cached sufficient statistics and a new very efficient test that in one k-means

Algorithm 1 KMEANS(X, k)

Input:

```
1:  $X \leftarrow \{ x_1, x_2, \dots, x_n \}$ 
2:  $k \leftarrow$  (number of clusters)
3:  $MaxIters \leftarrow$  (limit of iterations)
4:  $S \leftarrow$  some initial candidate solution
5: for each  $c_i \in C$  do
6:    $c_i \leftarrow e_j \in E$  (e.g. random selection)
7: end for
8: for each  $e_i \in E$  do
9:    $l(e_i) \leftarrow \operatorname{argminDistance}(e_i, e_j)_{j \in 1 \dots k}$ 
10: end for
11:  $changed \leftarrow false$ 
12:  $iter \leftarrow 0$ 
13: repeat
14:   for each  $c_i \in C$  do
15:      $UpdateCluster(c_i)$ 
16:   end for
17:   for each  $e_i \in E$  do
18:      $minDist \leftarrow \operatorname{argminDistance}(e_i, e_j)_{j \in 1 \dots k}$ 
19:     if  $minDist \neq l(e_i)$  then
20:        $l(e_i) \leftarrow minDist$ 
21:        $changed \leftarrow true$ 
22:     end if
23:   end for
24: until  $changed = true$  and  $iter \leq maxIters$ 
25: return  $S$ 
```

sweep selects the most promising subset of classes for refinement. This gives rise to a fast, statistically founded algorithm that outputs both the number of classes and their parameters. Experiments show this technique reveals the true number of classes in the underlying distribution, and that it is much faster than repeatedly using accelerated k-means for different values of K .

Expectation Maximization (EM) algorithm

In statistics, an expectation maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

The EM algorithm is used to find (locally) maximum likelihood parameters of a statistical model in cases where the equations cannot be solved directly. Typically these models involve latent variables in addition to unknown parameters and known data observations. That is, either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component that each data point belongs to.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values viz. the parameters and the latent variables and simultaneously solving the resulting equations. In statistical models with latent variables, this usually is not possible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation.

The EM algorithm proceeds from the observation that the following is a way to solve these two sets of equations numerically. One can simply pick arbitrary values for one of the two sets of unknowns, use them to estimate the second set, then use these new values to find a better estimate of the first set, and then keep alternating between the two until the resulting values both converge to fixed points. It is not obvious that this will work at all, but in fact it can be proven that in this particular context it does, and that the derivative of the likelihood is (arbitrarily close to) zero at that point, which in turn means that the point is either a maximum or a saddle point. In general there may be multiple maxima, and there is no guarantee that the global maximum will be found. Some likelihoods also have singularities in them, i.e. nonsensical maxima. For example, one of the "solutions" that may be found by EM in a mixture model involves setting one of the components to have zero variance and the mean parameter for the same component to be equal to one of the data points.

Algorithm 2 ExpectationMaximization(X, Z)

Input:

- 1: Given observed variables X , unobserved Z
 - 2: Define $(\theta'|\theta) = E_{Z|X,\theta}[\log P(X, Z|\theta')]$
 - 3: Where $\theta = [\pi\mu_{ji}]$
 - 4: **E Step:**
 - 5: Calculate $P(Z(n)|X(n), \theta)$ for each example $X(n)$.
 - 6: Use this to construct $Q(\theta'|\theta)$
 - 7: **M Step:**
 - 8: Replace current θ by $\theta \leftarrow \operatorname{argmax} Q(\theta'|\theta)$
-

3.2 Summary

In the opening section of this chapter we have introduced some of the preliminary topics related to our work, which included, machine learning, data mining and clustering, the technique we have used after certain modification in the layers of our hierarchical exploration model.

Chapter 4

Our Proposed Framework

Our proposed framework consist of a hierarchical exploration model which contain multiple numbers of layers for finding facts from CDR to discover useful information on social activities and relationships in urban areas. Each of these layer has a number of module developed using automated or semi-automated supervised and unsupervised learning algorithms with modification to make them suitable for our intended fact finding procedures. Each modules works independently using the overall knowledge base, containing the basic CDR database and amassed information from previous layers. Thus, informations acquired from each modules of every layers contributes in the overall knowledge base which can be used for further progressive exploration of fact in the subsequent layers. Therefore, our proposed framework, modeled after the deep learning technique is capable of progressive exploration of deeper facts as the number of layers increases and develop a larger knowledge base.

4.1 The Hierarchical Exploration Model

The CDR data entries contain the date, time, duration and geographical location of the cell tower facilitating the communication activity. This information can be considered as digital footprints of user activities and can be utilized to detect facts like social activity patterns, social groups, their relationships and properties of city area based on social activities by applying a set of fact finding techniques if a reasonable amount of data is given. Dhaka is a densely populated city with frequent communication activity

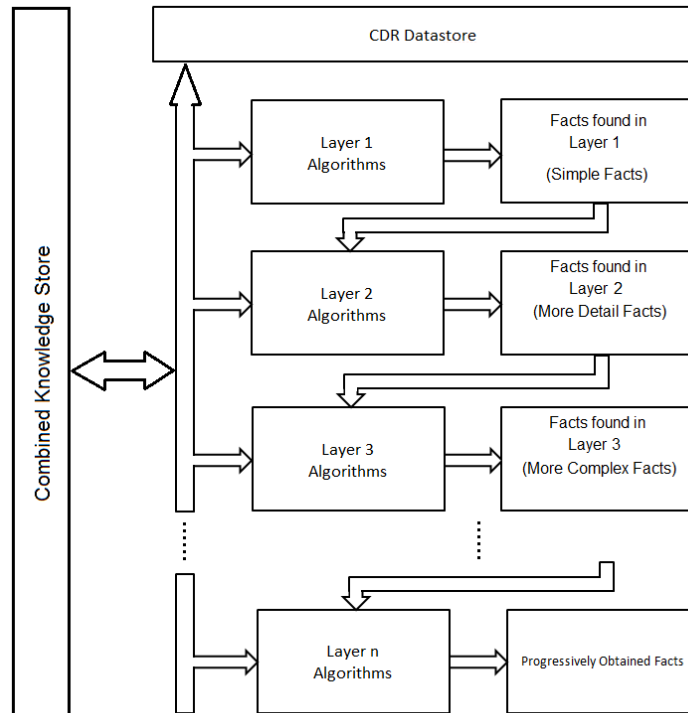


Figure 4.1: The Layered Approach

and has greater number of the cell towers generating fine grained CDR useful for effective fact finding.

In our work, the developed hierarchical exploration model with multiple layers can be expanded further for processing and analyzing large volume of CDR data collected from any large city to discover complex facts about the social groups and relationships in different layers. For each layer of our model we furnished a set of fact finding algorithms, including regressions, prediction classifiers and clustering techniques for analyzing spatio-temporal data to obtain information on social activities and relationships. The novelty of our proposed model is that, every new layer exploits the facts derived in the preceding layers for progressive exploration of new facts. By utilizing this multilayer hierarchical learning approach, as we explore deeper we discover more significant information on the smaller social groups with closer relationships, which can not be detectable in a single layered approach.

4.2 The Layered Approach

The hierarchical model starts with the basic CDR dataset, starting with a large-scale overview of the social groups and activities and gradually focus on more specific subsets of the large amount of data, which leads to the discovery of the smaller and closer groups. The first layer of our model uses the CDR data to find facts like places of interest, resident, workplace, etc. of the users and their basic social activity patterns, like, calling pattern, type of callers, frequencies of calls in different times of the day, etc. In the next layer we obtain informations on further social activity patterns like working patterns, traveling routes and patterns, etc. Using this information, in the subsequent layers, we find different groups of people based on their social activity patterns and find broader social groups, like, regular/irregular working groups, home-staying groups, frequent traveling groups, etc and the possible interactions and relationships inside and among these groups. As we go deeper with these layers, we progressively narrow down these groups and identify more distinguishable social groups like, family members, friends and acquaintances, neighbors, co-workers, etc. and their social activity patterns and relationships. We have developed our framework with the hierarchical model to go even deeper and narrow down professional groups, family and friend circles with similar social activity and traveling patterns. Furthermore, we classify city areas based on the activities and properties of social groups discovered in each layers from spatio-temporal perspectives. In this work, with the CDR data with limited attributes we have explored up to five layers using our framework which leads us to a bulk of useful information with the progressive exploration and expansion of the combined knowledge base. We have elaborated the frameworks of the layers in the next subsections. It is worth mentioning that, given a CDR dataset with a reasonably good number of attributes, our framework allows us to expand the hierarchical exploration model deeper up to as many layer as possible until there is no useful information is left to find.

4.2.1 Layer 1

The first layer mainly deals with the raw CDR dataset only. So, procedures in this layer focus on preparing the data for using in the subsequent layer, rather than finding

major facts. Here, the objective is to find simpler facts using statistical classifiers from the raw CDR data and integrating them to the knowledge base. We perform the basic analysis of our raw data in this layer. Here we obtain information about the concentration of user activities and calling pattern of users in different time periods of the day in the dates we have recorded in the CDR. For example, visualizing the overall calling pattern of our data we can clearly see that people make social interaction over phone more on holidays. The spatial variation of the call activity at any given time can be conveniently displayed by means of maps divided in Voronoi cells showing the service area of each cell tower. The information obtained from Voronoi diagram reflects the obvious and apparently known notions regarding the calling habits of the users, i.e. people talk more in the daytime to before midnight and less after the midnight till dawn, whereas a certain group talks more after midnight. However, this is considered as detection of minor social group.

By using our software, we isolate and summarize the call activity of every single user and are able to generate user based call log and location based call log in this layer. We call all the Unique Calling Locations of an user as UCL. From the concentration of call activity of users, we classify three groups, Minimal users, Regular users and Heavy users. From other calling patterns, we classify other types of users. Definitely, this three grouping gives us some idea about socially extroverts and introverts peoples.

One important and resource consuming task done in this layer is generation of a call graph. In our CDR, we do not have the information of the callee for call activities. So, we have furnished a way to develop a caller-callee relationship graph by considering calls made in same time with equal duration as an identical call and establish a calling relationship among the user. However, the original CDR data collected by cell operator contains callee information which would enable skipping this step and allows using our model more efficiently.

4.2.2 Layer 2

The major fact finding task done in this layer is identifying the locations of home, workplace and other frequently visited places for every single user. This information

is crucial for next layers as homes and workplaces have a pivotal role in the society. We call home, workplace and other frequently visited places commonly as Places of Interests or POI for individual users and store them in the main knowledge base. We use the call log of an user to determine the POI from where he makes most of the calls using our clustering algorithm. Using classifiers on the top POIs and the concentration of call activity in usual working hours and off-hours leads to the discovery of the home and workplace of an user. Other POIs are considered as locations where the user visited frequently. Working days and holidays of the regular working people is also identified in this layer.

Also, starting from this layer, we identify some of the city area features based on the knowledge of calling activity from previous layer. The CDR data have fixed number of unique locations, which are the locations of the cell towers providing service to the users. We consider every unique location as a zone of the city and use a classifier to find the concentration of call activities in different times of the day, which determines how busy or densely populated the zone is at a certain time of the day. From the concentration of calling activity in city areas in different time periods of day and days of week we identified the busy areas in working hours and holidays, even in different times of the year. These facts indicate a very important social feature of a city.

4.2.3 Layer 3

As we reach deeper upto Layer 3, we already have a good knowledge base consisting of the base CDR and information learned from the previous layer. Therefore, in this layer we explore some really useful information about social activities and social groups.

Using the home and workplace information, at first we classify two major social groups based on the working hour activity patterns, regular working people and irregular working people. The group of regular working people includes the users who spend their working hours and off-hours in different places, in their office and home respectively. People from this group regularly travel from their home to workplace and have a recognizable traveling pattern. Office workers, businessmen, students, etc. are part of this group. The other group comprises of people who spend both working hours and off-hours in same place. They can be retired people, homemakers or even people

who works in the same place as their home.

By considering the UCLs of an user who travel from his home to workplace, we have predicted his regular traveling route. It is done by applying a single source shortest path algorithm considering the UCLs as nodes, home location as source and workplace as destination. Finding traveling route is more effective when we have a good number call data of the user making calls en route to his workplace from home. Also, we have predicted the type of transport used from the distance crossed in a given time using this information.

Another social group detectable from this layer is based on the traveling distance of the users. Some people travel a very short distance to their workplace, where others have to pass a long way. We classify the people of the city in classes based on the range of their distance traveled from home to workplace.

In this layer we also identify and tag city blocks as Residential, Commercial or other miscellaneous types of area by considering the number of home, workplace or other types of POI located in a certain city block.

4.2.4 Layer 4

As we explore deeper into our CDR and knowledge base in Layer 4, we find closer social relationships and activity patterns, which leads to the discovery of smaller social groups consisting of smaller set of users. In this layer we propose a number of probabilistic prediction classifiers based on several hypotheses performing statistical analysis of the facts found in the preceding layers. We have discussed here some of the hypotheses used to design classifiers and applied on our data. The prediction classifiers designed based on the CDR are later validated with real call data we collected from volunteers.

Hypothesis 1. *A group of users have the same home locations means that they lives in the same neighborhood. So, we can classify these loosely connected social groups as neighbors and members have a high probability of interaction.*

Hypothesis 2. *If a group of users have the same workplace locations, they have a high probability of being acquaintances, even colleagues.*

Hypothesis 3. *The type of transport used by a user can be classified and grouped from his traveling route and information of time differences and distances among call activities made by him in the same day on that route.*

Hypothesis 4. *If a two or more persons have common home, workplace or more than one common POIs and same or overlapping regular traveling route, we can perform predictive analysis of their social relationship and group membership using available information.*

4.2.5 Layer 5

Layer 5 utilizes the knowledge of calling relationship either present in the CDR or derived from the call graph generated in the previous layer. To use this relationship information we propose some prediction models based on the following hypotheses to detect social groups, including family, friends, colleagues and closely acquainted people featuring more complex and deeper relationships. The prediction classifiers designed based on our CDR are later validated with real call data we collected from volunteers.

Hypothesis 5. *If two or more persons share the same home location and have a frequent calling relationship, we can predict the probability of them being family members or close acquaintances.*

Hypothesis 6. *If two or more persons share the same workplace location and have a frequent calling relationship, we can predict the probability of them being colleagues, co-workers or friends.*

Hypothesis 7. *If two or more persons visit or stay in same POIs multiple times in same time periods and have a calling relationship, we can predict the aggregated probability of them being member of a social group like family, friends or other types of close relationships.*

4.2.6 Beyond Layer 5

As we mentioned earlier, our framework allows us to explore the hierarchical exploration model deeper up to as many layer as possible, the feasibility of which is

limited by the attribute and size of the base dataset. Given our dataset, it is possible, but not feasible to explore beyond Layer 5. More mining opportunity can be explored within the limit of Layer 5. Our work focuses on limited scale to detect patterns of social group based activities and traveling pattern of social groups.

4.3 Aggregated Social Closeness (ASC) Score

After exploring all the layer, we have our combined knowledge database which contain the different group membership information of the users. Now, for utilizing this knowledge about all the user, we have developed a statistical prediction model which calculate the Aggregated Social Closeness (ASC) between two users. we consider all the social relation predicted in different layers and combining them using a statistical prediction model, we calculate ASC between two users, which predict the depth of their relationship and interactions. Evidently, a higher ASC value indicates that the two users are family members or close friends, on the other hand, a lower ASC means that they have no relation or interaction at all.

4.4 Summary

This chapter elaborates on the framework of our hierarchical exploration model for progressive discovery of social groups by detecting social activities from a CDR dataset. We have explained the layered approach of our model by over-viewing they layers and fact finding activities performed in them. The technical details of the fact finding techniques and algorithms applied in these layers are explored in the following chapter.

Chapter 5

Methodology

This chapter presents the algorithms used in the layers of our hierarchical exploration model and their explanations. Our proposed model works with multiple layers for processing and analyzing large volume of CDR data collected from any widespread urban area for progressive discovery of facts about the social groups and relationships. The proposed algorithms employs supervised and unsupervised learning methods mostly consists of regression, statistical classifiers and clustering techniques for predicting social groups and relationships. The algorithms are designed as single modules. In this hierarchical exploration model all the module in every layer works as an autonomous dataset processor capable of handling the expected input feature vector and produce output independently. Collectively, all the modules works as the complete hierarchical exploration model which produce output in different layers. In our work, we have used the prediction models in both automated and semi-automated ways. After finding every facts in different layers they are added in the combined knowledge which also contain the main CDR data.

The raw CDR data are contained in simple but very large text files containing spatio-temporal information on the call activity of users. In our model, this information is considered as digital footprints of user activities and utilized to detect facts like social activity patterns, social groups, their relationships and properties of city area based on social activities by applying our proposed statistical prediction algorithms on the available massive amount of data after necessary preprocessing. Dhaka is a densely populated city with frequent communication activity and has greater number of the cell

towers generating fine grained CDR useful for effective fact finding.

In every layers of our model, a set of preprocessing tasks are required to effectively use the original CDR data. Even some of the output files from different layers require preprocessing. Each layer employs a number of statistical classifier and clustering algorithms to predict facts. The initial layers of our model use the CDR data to find basic facts like places of interest, resident, workplace, etc. of the users and their social activities like working patterns, traveling routes and patterns, etc as well as some broader social groups. Using this information, in the next layers, we classify people based on their activity patterns and find smaller social groups, like, regular/irregular working groups, home-staying groups, frequent traveling groups, etc and the possible interactions and relationships among members of these groups. In the subsequent layers, we explore deeper to progressively narrow down these groups and identify more distinguishable social groups like, family members, friends and acquaintances, neighbors, colleagues and co-workers, etc. and their social activity patterns and relationships. Additionally, we classify and tag city areas based on social group activities discovered in each layers from spatio-temporal perspectives. Finally, a statistical prediction algorithm is proposed to find the possible closeness among two users based on our aggregated findings in the layered exploration model. Our experimental results are presented using tables, charts, maps and other necessary visualizations in the following chapter.

5.1 Validation

We evaluate the accuracy of our methods using k-fold cross validation. CDR Dataset of n call records is randomly divided into two parts- training set and validation set. Let, n_t be the number of training set and n_v be the number of validation set. Because we are using k-fold cross validation, at each fold number of training data is $n_t = \lceil \frac{(k-1)*n}{k} \rceil$ and number of validating data is $n_v = \lfloor \frac{n}{k} \rfloor$. We perform all the analysis on training dataset which contains n_t number of call records and propose our model based on this dataset. Afterwards, we verify our model using validation set which contains n_v number of call records. Further, we have validated our prediction results using some unencrypted call data collected from a number of volunteer users with known social relations and groups

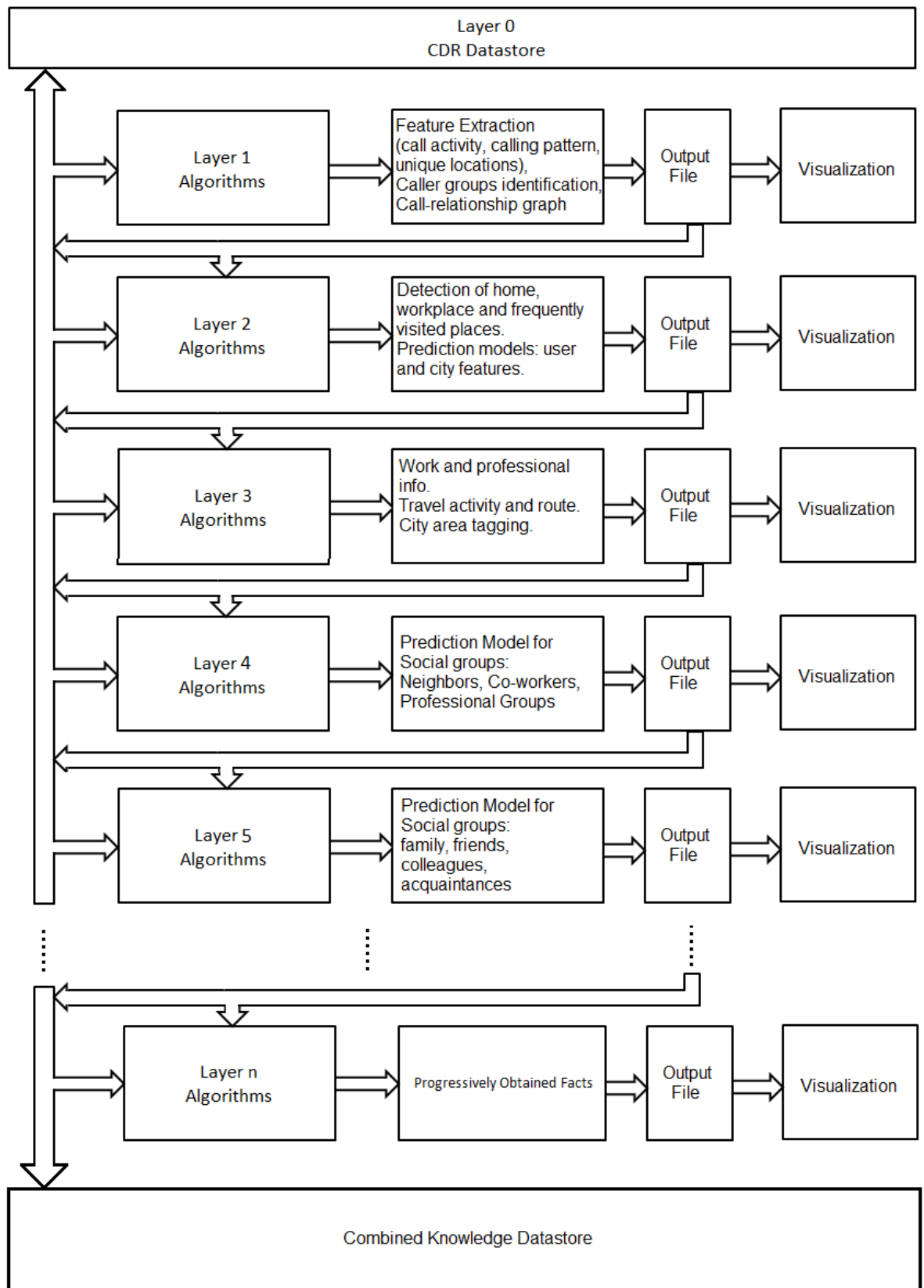


Figure 5.1: The Hierarchical Progressive Exploration Model

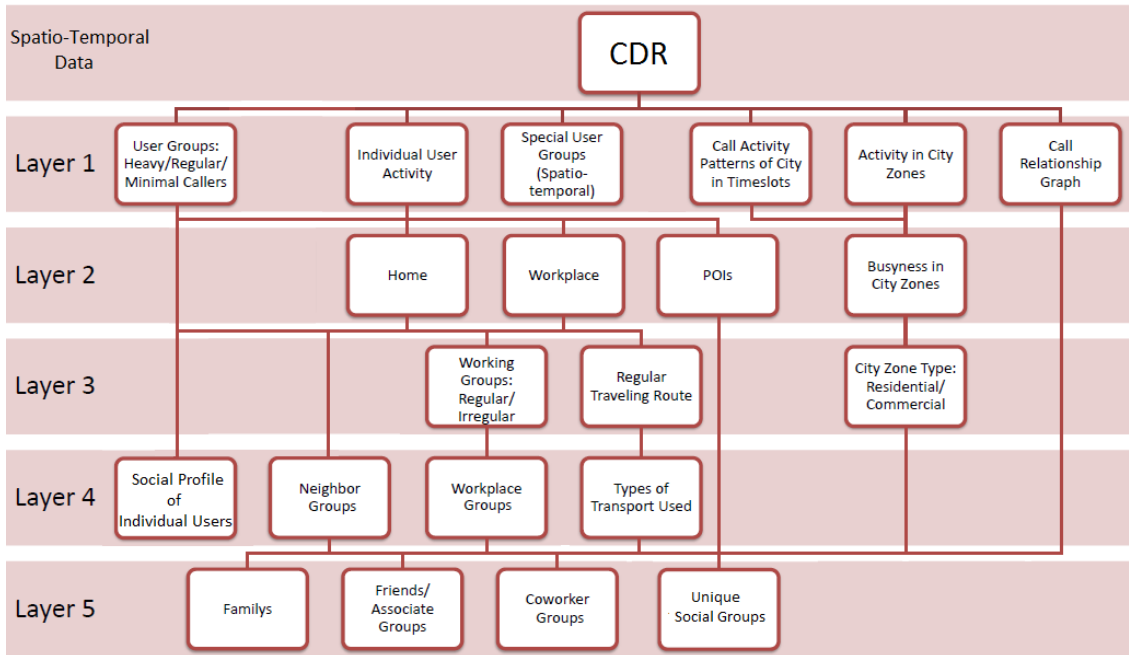


Figure 5.2: Tree-based Hierarchical Presentation of Predicted Groups in Five Layers

memberships.

5.2 Preprocessing

Preprocessing is necessary to effectively use the raw data for fact finding. Even some of the output files from different layers require preprocessing. As a matter of fact that, some of the output files from the first few layers itself are large files containing overview of the extracted facts about the whole city. This preprocessing techniques generates the appropriate feature vectors from the CDR data and knowledge base obtained from the layers to be given as input for every modules. The data entries from our original CDR contain the encrypted user id, date, time, duration and geographical location of the cell tower facilitating the communication activity. The preprocessing involves automatic or semi-automatic extraction of feature vectors for each modules. None of the single modules uses all the features available in the CDR data. For better representation of the features they are represented using some symbols and notations in this text. The definition of the symbols and notations used corresponded to the features of original CDR is summarized in 5.1.

The raw CDR data contains six features which we used in different layers of our

Table 5.1: Notations used in this text and their definition

| Notation/Symbol | Definition |
|-----------------|--|
| CDR | The full call detail record |
| u | Encrypted User ID assigned to every user |
| LUsr | List of all unique users in a full dataset |
| dt | Date of a call/communication activity |
| tm | Time of starting a call/communication activity |
| dur | Duration of a call |
| lat | Latitude of the cell tower providing the call |
| lon | Longitude of the cell tower providing the call |
| loc | A location consisting of a (lat, lon) pair |
| LLoc | List of all unique location in a full dataset |

framework to find fact. We have considered each entry of the raw CDR as a 6-dimensional feature vector x_i where $i = 1, 2, \dots, n$ and n is a positive integer indicating the number of entry in the available CDR. Each The features contained in each x_i are $(u, dt, tm, dur, loc(lat, lon))$. The set of feature vectors containing the full CDR data is X , where

$$X = \{x_i | \{u, dt, tm, dur, loc(lat, lon)\} \in x_i, i \in \mathbb{Z}^+\} \quad (5.1)$$

In every layer, necessary features are extracted from X using different types of preprocessing operations as necessary. Preprocessing is a computationally resource demanding process when applied to big dataset like our CDR, which is stored as a unstructured database in a text-based log file format.

In the following sections, we have explained the algorithms used in every layer for finding different facts on social groups and relationships.

5.3 Layer 1

At the first layer, using the raw CDR data we apply linear classifiers and predict comprehensive facts and identify patterns in it. As, the objectives is to find comprehensive facts from the raw CDR data, we have developed algorithms using simple linear classifiers to perform statistical analysis in this layer. Here, we propose algorithms for extracting information from massive CDR data and extract features

comprising calling pattern and unique locations visited by individual users. Then, we have used the raw CDR data to analyze the daily call activity and calling pattern of the citizens to predict social groups including heavy callers, regular callers, minimal callers, late-night callers and professional callers by using values for each features of call activity and propose prediction classifiers to predict caller groups. Also we examine the overall calling pattern of the city in different time periods and locations. Finally, We generate a call graph to determine the possible relations among callers.

To begin with, we isolate CDR logs for every single user and generate a log summarizing his call activities in all the unique locations he visited in the given time period of the available CDR using Algorithm 3. In this log we calculate the usage score ζ_{loc} of that user in every unique locations from no of call nc and total of call durations dur in that location using following equation,

$$\zeta_{loc} = nc_{loc} + \omega \sum dur_{loc} \quad (5.2)$$

Where, ω is a constant weight factor

These user based logs generated in this layer are inputs for the individual and aggregated facts prediction classifiers in the next layers.

Next, we determine facts about the daily concentration of user activities and calling pattern of users. Later, we find the concentration of the call activities in different time periods of the day of the dates available in CDR. The call activity is measures by calculating busyness score β in a certain time period T . Busyness score, Number of calls made and active user in different dates is detected using Algorithm 4. The algorithm uses a straight forward counting approach to count the user and call number in a day, where $T = 24$ hours of a day and uses them to calculate β for everyday using the following equation.

$$\beta_{T,loc} = \omega_1 NU_{T,loc} + \omega_2 NC_{T,loc} \quad (5.3)$$

Where,

$NU_{T,loc}$ = total user in time period T in location loc

Algorithm 3 INDLUSER(X, U)

Input:

- 1: $X \leftarrow \{x_1, x_2, \dots, x_n\}$ (n entires in full CDR)
- 2: $U \leftarrow$ unique ID of user U

Output:

- 3: $CDR_U \leftarrow$ (call detail for user U)
 - 4: $nc(loc) \leftarrow$ (number of calls made in location loc)
 - 5: $dur(loc) \leftarrow$ (duration in location loc)
 - 6: $\mu(loc) \leftarrow$ (usage score in location loc)
 - 7: $uloclist \leftarrow \{loc, nc(loc), dur(loc), us(loc)\}$ (list for m unique loc for U)
 - 8: **BEGIN**
 - 9: **for all** $x_n \in X$ **do**
 - 10: **if** $uid_n \in x_n = u$ **then**
 - 11: $CDR_U \leftarrow$ ADD $x(n)$ to the list
 - 12: **end if**
 - 13: **if** $loc_n \notin uloclist$ **then**
 - 14: $uloclist \leftarrow$ ADD loc_n to the list
 - 15: $nc(loc_n) \leftarrow nc(loc_n) + 1$
 - 16: $dur(loc_n) \leftarrow dur(loc_n) + dur(n)$
 - 17: **else**
 - 18: $nc(loc_n) \leftarrow nc(loc_n) + 1$
 - 19: $dur(loc_n) \leftarrow dur(loc_n) + dur(n)$
 - 20: **end if**
 - 21: **end for**
 - 22: **for all** $loc_m \in uloclist$ **do**
 - 23: Calculate ζ_{loc_m}
 - 24: **end for**
 - 25: **END**
-

$NC_{T,loc}$ = total number of call in time period T in location loc

ω_1 and ω_2 are constant weight factors, whose values are tuned up using a linear classifier.

The spatial variation of the call activity at any given time can be conveniently displayed by means of maps divided in cells of Voronoi tessellation, which delimit the area of influence of each cell tower or antenna. The Voronoi tessellation partitions the plane into polygonal regions, associating each region with one cell tower. The partition is such that all points within a given Voronoi cell are closer to its corresponding tower than to any other tower in the map. These information reflect the obvious and apparently known notions regarding the calling habits of the users, i.e. people talk more in the daytime to before midnight and less after the midnight till dawn, where as a certain group talks more after midnight. Comparing the Voronoi diagram of different time periods leads

Algorithm 4 TCALLACTIVITY($X, LUsr$)

Input:

- 1: $X \leftarrow \{x_1, x_2, \dots, x_n\}$ (n entires in full CDR)
- 2: $LT \leftarrow \{T_1, T_2, \dots, T_m\}$ (list of defined time periods in CDR)
- 3: $LUsr \leftarrow \{u_1, u_2, \dots, u_i\}$ (list of i unique users)

Output:

- 4: $CallNo \leftarrow \{cn_1, cn_2, \dots, cn_j\}$ (number of calls in every T_j)
 - 5: $UserNo \leftarrow \{un_1, un_2, \dots, un_j\}$ (number of active users in T_j)
 - 6: **BEGIN**
 - 7: **for all** $T_j \in LT$ **do**
 - 8: $cn_j \leftarrow$ count number of calls
 - 9: $un_j \leftarrow$ count number of active users
 - 10: Calculate $\beta_{Tj} = \omega_1 NU_{Tj} + \omega_2 NC_{Tj}$
 - 11: **end for**
 - 12: **END**
-

the discovery of different, spatially distinct activity patterns. Besides different spatial patterns, each particular time of the day, as well as each day of the week, is characterized by a different overall level of activity.

The call volume shows strong variations with time and day of the week, but differences across subsequent weeks are generally mild, provided one considers call traffic in the same place, time and day of the week. To capture the weekly periodicity of the observed patterns, we define $nc_i(loc, t, T)$ as the number of calls recorded at location loc , which can either denote a single Voronoi cell or a group of neighboring cells during the i th week between times t and $t + T$. As we have access to continuous data for N weeks, the mean call activity is given by,

$$n(loc, t, T) = \frac{1}{N} \sum_{i=1}^N n_i(loc, t, T) \quad (5.4)$$

On the basis of concentration of monthly call activities, we classify three groups with the name of Minimal users, Regular users and Heavy users using a linear classification algorithm. First we calculate the usage score μ of every user using the following formula,

$$\mu = \omega_c NC + \omega_d \sum dur \quad (5.5)$$

Where, NC = number of calls made by the user in a month

$\sum dur$ = total duration (in second) of calls made by user in a month

ω_c and ω_d are constant weight factors, whose values are tuned up using a linear classifier.

Now we define the aforementioned three groups on the basis of μ tuned up based on our training data found from k-fold cross validation. We assign every user in the right group based on his μ . This classification shows us the overview of the mobile phone usage pattern of a city, which is closely related to it's social feature.

Algorithm 5 USEAGEGROUP($U, LUsr$)

Input:

- 1: $U \leftarrow$ unique ID of user U
- 2: $LUsr \leftarrow \{u_1, u_2, \dots, u_i\}$ (list of i unique users)
- 3: $\mu list \leftarrow \{\mu_1, \mu_2, \dots, \mu_i\}$ (usage score for each of i unique users)
- 4: $NC_{minM}, NC_{maxM}, NC_{maxR}, NC_{maxH} \leftarrow$ (Ranges for number of calls)

Output:

- 5: $LU_H \leftarrow$ (List of Heavy Users)
 - 6: $LU_R \leftarrow$ (List of Regular Users)
 - 7: $LU_M \leftarrow$ (List of Minimal User)
 - 8: **BEGIN**
 - 9: **for all** $n_i \in NoC$ **do**
 - 10: Calculate $\mu_i = \omega_c NC_i + \omega_d \sum dur_i$
 - 11: **if** $NC_{minM} \leq \mu_i \leq NC_{maxC}$ **then**
 - 12: ADD u_i to LU_M
 - 13: **else**
 - 14: **if** $NC_{maxC} \leq \mu_i \leq NC_{maxM}$ **then**
 - 15: ADD u_i to LU_R
 - 16: **else**
 - 17: **if** $NC_{maxM} \leq \mu_i \leq NC_{maxH}$ **then**
 - 18: ADD u_i to LU_H
 - 19: **end if**
 - 20: **end if**
 - 21: **end if**
 - 22: **end for**
 - 23: **END**
-

Using similar linear classifiers we find other minor social groups with known properties based on call activities. For example, a certain group of people talks more after midnight. Combined this with real life fact that a group of teenagers tends to have this types of calling pattern. Another example is that, we can see that a group of people talk noticeably more in business hours. We can deduce that these people are a group of

professionals whose business prerequisite is to communicating with people. These are very helpful information to narrow down the list of social and professional groups. To predict this group using a linear classifier we use the following equation based on Equation 5.3 and Equation 5.5.

$$\mu = \omega_c NC_T + \omega_d \sum dur_T \quad (5.6)$$

Where, T is the time period of activity for a certain user group.

Another important task done in this layer, which provides critical information in the next layers, is generating caller-callee relationship graph. We have the time and duration of every call activity of user. By cross referencing the whole database we detect the destination of the call activity for every user. That is done by using Algorithm 6 by considering the fact that if two call activity is taking place in the same time and have the same call duration, the two user is actually communicating with each other. This way we generate a calling relationship graph among the users in our CDR. The number of call made by two user is assigned as the weight of calling relationship edge of two user node.

Algorithm 6 USERRELATIONGRAPH($X, LUsr$)

Input:

- 1: $X \leftarrow \{x_1, x_2, \dots, x_n\}$ (n entires in full CDR)
- 2: $LUsr \leftarrow \{u_1, u_2, \dots, u_i\}$ (list of i unique users)

Output:

- 3: $r_{pq} \leftarrow$ number of calls made between random user p, u_p and user q, u_q , indicate calling relationship
 - 4: $R \leftarrow \{\text{set of } r_{pq}\}$
 - 5: **BEGIN**
 - 6: **for all** $u_i \in UID \subseteq X$ **do**
 - 7: **for all** $x_j \in X$ **do**
 - 8: **if** there is an entry where $dt_i = dt_j$ AND $tm_i = tm_j$ AND $dur_i = dur_j$ **then**
 - 9: $r_{ij} \in R$
 - 10: **end if**
 - 11: **end for**
 - 12: **end for**
 - 13: **END**
-

This step is simpler if we use the CDR which contains both the source and destination of all the call activities. In that case we only count the number of calls made between

users which denotes the weight of relational edge between them. But, as our main CDR does not contain this information, we had to acquire this information using algorithm 6

5.4 Layer 2

In this layer we predict the locations of home, workplace and other frequently visited places for every single user. We call home, workplace and other frequently visited places commonly as "Place of Interest" or POI. After detecting home, workplace and other POIs for individual users, we store them in the main knowledge base. We use the call log of an user to determine the top POI locations from where he makes most of the calls activity determined by his usage score μ for those locations. Then, from the maximum usage score u in usual working hours and off-hours leads to the discovery of the home and workplace respectively from the list of POIs. Other POIs are locations where the user visits frequently. It is a well-derived fact that all user have a home location, a good number of user, who are working people, have an workplace location and a small percentage of user have one or more regularly visited POIs.

We have used clustering based techniques in Algorithm 7 and Algorithm 8 to find the POIs of an user. Every sizable clusters from the list of cluster detected by a clustering algorithm is considered as a POI. From the various clustering algorithms, we have selected X-Means and EM clustering algorithms to find the POIs based on the performance of our experimental result.

Algorithm 7 use X-Means clustering to predict the POIs of an user. As we already discussed, X-Means is an extension of K-means with additional capabilities. In our work, our main limitation of using k-means is that, we don't know the number of POIs or clusters a user have, which needs to be supplied as value of k to use k-means. For X-means we don't have to provide the number of cluster as it is calculated by the algorithm it self. X-means provide us the number of POI i.e. clusters as well as the value of the clusters i.e. the locations of the POIs in a automated way. Besides X-means is computationally more feasible then K-means. So, we choose X-means to find the POIs from our user data.

The performance of X-Means on our datasets diverges on the activity pattern of every

Algorithm 7 POIXMEANS(X, U)

Input:

- 1: $U \leftarrow$ user id of the user whose POI is to be detected
- 2: $X \leftarrow \{ x_1, x_2, \dots, x_k \}$ (The full CDR)

Output:

- 3: $CDR_U \leftarrow \{ x_1, x_2, \dots, x_j \}$ (CDR for U)
 - 4: $POI_U \leftarrow$ List of POI for user U
 - 5: **BEGIN** Call $INDLUSER(X, U)$ and assign returned values to CDR_U
 - 6: $CList \leftarrow$ List of loc_i clustered locations with percentile value
 - 7: $Th \leftarrow$ Minimum percentile value of cluster to be POI
 - 8: Call $XMEANS(CDR_U)$ and assign returned values to $CList$
 - 9: **for all** $loc_i \in CList \subset LLoc$ **do**
 - 10: **if** loc_i is in a cluster with value $\geq Th$ **then**
 - 11: ADD loc_i to POI_U
 - 12: **end if**
 - 13: **end for**
 - 14: **return** POI_U
 - 15: **END**
-

Algorithm 8 POIEM(X, U)

Input:

- 1: $U \leftarrow$ user id of the user whose POI is to be detected
- 2: $X \leftarrow \{ x_1, x_2, \dots, x_k \}$ (The full CDR)

Output:

- 3: $CDR_U \leftarrow \{ x_1, x_2, \dots, x_j \}$ (CDR for U)
 - 4: $POI_U \leftarrow$ List of POI for user U
 - 5: **BEGIN** Call $INDLUSER(X, U)$ and assign returned values to CDR_U
 - 6: $CList \leftarrow$ List of loc_i clustered locations with percentile value
 - 7: $Th \leftarrow$ Minimum percentile value of cluster to be POI
 - 8: Call $EM(CDR_U)$ and assign returned values to $CList$
 - 9: **for all** $loc_i \in CList \subset LLoc$ **do**
 - 10: **if** loc_i is in a cluster with value $\geq Th$ **then**
 - 11: ADD loc_i to POI_U
 - 12: **end if**
 - 13: **end for**
 - 14: **return** POI_U
 - 15: **END**
-

individual user. As every user has unique mobility pattern, for most of the users X-Means perform well, while for some it fails to detect all the POIs. The reason is that as X-Means is a centroid-based clustering, it assumes the variance of the distribution of each attribute is spherical, which is not true for all the users. As a result, for some users a distribution-based clustering performs better for finding POIs. So, we also use a distribution-based clustering, expectation-maximization (EM) algorithm to predict the POIs too. Then we compare the list of POIs from both algorithms and select the optimal solution. Algorithm 8 use EM clustering to predict the POIs of an user.

After finding the list of POIs for a user, we can predict his home, regular workplace, if available and also other places of special interest. It is a known fact that every user has a home, a more or less permanent place to live regularly which is prominently visible in his daily activity pattern. Excluding a negligibly few exceptions, all the users spend his off-hours, which is in the night, in his home. From this knowledge we predict the home of an user by comparing his usage score μ at every POIs at off-hours in the full time period of available CDR data. Here, μ is calculated using the following formula,

$$\mu_{poi} = \omega_c NC_{T,poi} + \omega_d \sum dur_{T,poi} \quad (5.7)$$

Where,

NC = number of calls made by the user in T period from a POI

$\sum dur$ = total duration (in second) of calls made by user in T period from a POI

T = OFF-HOURS/WORKING-HOURS of the city

ω_c and ω_d are constant weight factors, whose values are tuned up using a linear classifier.

Obviously, the POI with highest μ in OFF-HOURS is the home of any user. Algorithm 9 predicts the home of any user.

We use a similar technique by finding μ of POIs in WORKING-HOURS to find workplace. WORKING-HOURS are quite similar in most of the urban areas of the world and for every city it is a known fact. The important thing we need to consider while predicting workplace is that all the user do not have a work places and a good

Algorithm 9 FINDHOME(U, CDR_U, POI_U)

Input:

- 1: $U \leftarrow$ user id of the user whose home is to be detected
- 2: $CDR_U \leftarrow \{x_1, x_2, \dots, x_j\}$ (CDR for U)
- 3: $POI_U \leftarrow \{loc_1, loc_2, \dots, loc_i\}$ (List of POI for user U)

Output:

- 4: $home(lat, long) \leftarrow$ Location of home for user U
 - 5: **BEGIN**
 - 6: Define $T \leftarrow$ OFF-HOURS for the city
 - 7: **for all** $loc_i \in POI_U$ **do**
 - 8: Calculate $\mu_{poi} = \omega_c NC_{T, poi} + \omega_d \sum dur_{T, poi}$ from CDR_U
 - 9: **end for**
 - 10: Find POI with Maximum Value of μ_{poi} and assign to $home(lat, long)$
 - 11: **return** $home(lat, long)$
 - 12: **END**
-

number of users stays at home. For them, the maximum valued POI detected during WORKING-HOURS also indicate home. But, we can easily identify this type of users by matching the location of maximum valued POI detected during WORKING-HOURS with the location of home. Algorithm 10 predicts if the user has a workplace and find it's location.

Algorithm 10 FINDWORKPLACE(U, CDR_U, POI_U)

Input:

- 1: $U \leftarrow$ user id of the user whose workplace is to be detected
- 2: $CDR_U \leftarrow \{x_1, x_2, \dots, x_j\}$ (CDR for U)
- 3: $POI_U \leftarrow \{loc_1, loc_2, \dots, loc_i\}$ (List of POI for user U)

Output:

- 4: $workplace(lat, long) \leftarrow$ Location of workplace for user U
 - 5: **BEGIN**
 - 6: Define $T \leftarrow$ WORKING-HOURS for the city
 - 7: **for all** $loc_i \in POI_U$ **do**
 - 8: Calculate $\mu_{poi} = \omega_c NC_{T, poi} + \omega_d \sum dur_{T, poi}$ from CDR_U
 - 9: **end for**
 - 10: Find POI with Maximum Value of μ_{poi} and assign to $workplace(lat, long)$
 - 11: **if** $workplace = FINDHOME(U, CDR_U, POI_U)$ **then**
 - 12: **return** "NO WORKPLACE"
 - 13: **else**
 - 14: **return** $workplace(lat, long)$
 - 15: **end if**
 - 16: **END**
-

5.4.1 Classifying City blocks

We classify city areas based on their level of activity (BUSY/IDLE) in different time periods. The CDR has a fixed number of unique locations and we consider them as a zone of the city. We find the concentration of call activities and no of active users in different times of the day, which determines how busy or densely populated the zone is at a certain time period of the day, including working hours and holidays. We have used Linear SVM to develop our prediction classifier to find the status of a location as BUSY, denoted by 1 or IDLE denoted by -1. Here, we use the feature vector set $L \subset X$ extracted in the first layer, where,

$$L = \{L_{loc} | \{NU_{T,loc}, NC_{T,loc}\} \in L_{loc}, loc \in Z+\} \quad (5.8)$$

$NU_{T,loc}$ = total user in time period T in location loc

$NC_{T,loc}$ = total number of call in time period T in location loc

Now, given some training data $D \in L$, a set of n points of the form

$$D = (L_{loc}, s_{loc}) | L_{loc} \in L, s_{loc} \in \{-1, 1\}_{loc=1}^n \quad (5.9)$$

where the s_{loc} is either 1 or -1, indicating the class to which the point L_{loc} belongs. Each L_{loc} is a p -dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $s_{loc} = 1$ from those having $s_{loc} = -1$. Any hyperplane can be written as the set of points L_{loc} satisfying

$$w \cdot L_{loc} - b = 0 \quad (5.10)$$

where \cdot (dot) denotes the dot product and w the normal vector to the hyperplane. The parameter $\frac{b}{w}$ determines the offset of the hyperplane from the origin along the normal vector w . As the training data are linearly separable for busyness of city area, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. These hyperplanes can

be described by the equations for two groups respectively,

$$w.L_{loc_i} - b \geq 1 \quad (5.11)$$

$$w.L_{loc_i} - b \leq -1 \quad (5.12)$$

We have used busyness score β from Equation 5.3 in our SVM and by doing so the final form of Equation would be,

$$w.\beta_{loc_i} - b = 0 \quad (5.13)$$

Algorithm 11 CLASSIFYCITYBLOCK($L, ULoc, T$)

Input:

- 1: $L \leftarrow \{ L_1, L_2, \dots, L_j \}$ (j entries in CDR)
- 2: $ULoc \leftarrow \{ loc_1, loc_2, \dots, loc_i \}$ (List of i unique locations in CDR)
- 3: $T \leftarrow$ Time Period in which level of activity to be detected

Output:

- 4: $loclists \leftarrow$ list of locations with status (BUSY/IDLE)
 - 5: **BEGIN**
 - 6: **for all** $loc_i \in ULoc$ **do**
 - 7: Calculate $\beta_{loc_i, T}$ for T from X
 - 8: **end for**
 - 9: **for all** $c_i \in C$ **do**
 - 10: **if** $w.\beta_{loc_i, T} - b \geq 1$ **then**
 - 11: $loc_i \leftarrow$ BUSY
 - 12: **else**
 - 13: $loc_i \leftarrow$ IDLE
 - 14: **end if**
 - 15: **end for**
 - 16: **END**
-

5.5 Layer 3

The first operation performed in this layer is to find out the home and workplace (if any) of all the users. Using the home and workplace information, at first we distinguish two major social groups based on the working activity patterns, the working people, who regularly goes to a certain workplace and irregular working

people, who either stays home or have some indistinguishable irregular working pattern. The group of regular working people spend their working-hours in their office and off-hours in their home. People from this group regularly travel from their home to workplace and have a recognizable traveling pattern. People like professionals, office workers, businessmen and students belongs to this group. The irregular working group includes home staying people like homemakers, retired and unemployed people. We have used a linear classifier based on SVM to predict these two major social groups.

Algorithm 12 WORKINGGROUP(X)

Input:

- 1: $X \leftarrow \{x_1, x_2, \dots, x_j\}$ (j entries in CDR)
- 2: $LU_{sr} \leftarrow \{u_1, u_2, \dots, u_i\}$ (list of i number of unique users)

Output:

- 3: $WGROUP, NGROUP$ (grouplist of working and non-working users respectively)
 - 4: $\Delta T_{u_i} \leftarrow$ Traveling distance of user u_i
 - 5: **BEGIN**
 - 6: **for all** $u_i \in LU_{sr} \subset X$ **do**
 - 7: $CDR_U \leftarrow$ Call $INDLUSER(X, u_i)$
 - 8: $POI_U \leftarrow$ Call $POIEM(X, u_i)$
 - 9: $home \leftarrow$ Call $FINDHOME(U, CDR_U, POI_U)$
 - 10: $workplace \leftarrow$ call $FINDWORKPLACE(U, CDR_U, POI_U)$
 - 11: **if** $FINDWORKPLACE(U, CDR_U, POI_U) \neq$ "NO WORKPLACE" **then**
 - 12: ADD u_i to $WGROUP$
 - 13: **else**
 - 14: ADD u_i to $NGROUP$
 - 15: **end if**
 - 16: $\Delta T_{u_i} \leftarrow HAVERSINE(home, workplace)$
 - 17: **end for**
 - 18: **END**
-

The other social groups discovered from this layer is based on the traveling distance ΔT of the users. While, finding home and workplaces of the users, we also calculated their traveling distance from home to workplace, if any. Some of the working people travel a very short distance to their workplace, while others have to pass a long way. We have created five groups based on the regular traveling distance of the working people. We have used the Haversine formula to calculate the great-circle (surface of Earth) distances between the home and workplace from their longitudes and latitudes. We calculate distance between two location using the Algorithm 13, where the coordinates of the two locations are $(lat1, lon1)$ and $(lat2, lon2)$.

We have predicted the regular traveling route of an user by considering his UCLs

Algorithm 13 HAVERSINE($lat1, lon1, lat2, lon2$)

Input: $lat1, lon1, lat2, lon2$

Output: $d \leftarrow$ (Distance between two places)

```
1: BEGIN
2:  $R \leftarrow$  (The radius of the Earth)
3:  $dlon \leftarrow lon2 - lon1$ 
4:  $dlat \leftarrow lat2 - lat1$ 
5:  $a \leftarrow (\sin(dlat/2))^2 + \cos(lat1) * \cos(lat2) * (\sin(dlon/2))^2$ 
6:  $c \leftarrow 2 * atan2(\sqrt{a}, \sqrt{1-a})$ 
7:  $d \leftarrow R * c$ 
8: return  $d$ 
9: END
```

(Unique Call Locations) during his travels between home to workplace. It is done by applying Dijkstra's algorithm for finding single-source shortest paths considering the UCLs as nodes, home location as source and workplace as destination. To do so first a graph $G(V, E)$ is created, where,

$V \leftarrow loc_1, loc_2, \dots, loc_i$, List of UCL of the user.

$E \leftarrow$ edges representing all possible paths between ever pairs of UCL of the user, created from the real map data and the distance between them are the values of edges.

Algorithm 14 TRAVELINGROUTE(U, CDR_U)

Input:

```
1:  $U \leftarrow$  for whom traveling route to be predicted
2:  $CDR_U \leftarrow$  CDR for user  $U$ 
3:  $home_U \leftarrow$  Home loaction for user  $U$ 
4:  $workplace_U \leftarrow$  Workplace location for user  $U$ 
```

Output:

```
5:  $TR \leftarrow \{ loc_1, loc_2, \dots, loc_n \}$  (Sequential list of all locations in the traveling route)
6: BEGIN
7:  $MAPDATA \leftarrow$  real world map data
8:  $V \leftarrow$  Find all unique calling locations  $\{ loc_1, loc_2, \dots, loc_m \}$  of  $U$ 
9: for all pairs of locations  $loc_p, loc_q$  pairs  $\in ULoc$  do
10:   if path exists between  $loc_p$  and  $loc_q$  as per  $MAPDATA$  then
11:      $E_{pq} \leftarrow$  Calculate HAVERSINE( $loc_p, loc_q$ )
12:   end if
13: end for
14: Create Graph  $G(V, E)$ 
15:  $TR \leftarrow DIJKSTRA(G(V, E), home_U, workplace_U)$ 
16: Verify  $TR$  with  $MAPDATA$  for a valid path
17: END
```

5.5.1 Tagging City blocks

Here we classify the type of city area from the types Residential, Commercial and Miscellaneous and tag them accordingly. We consider all the unique locations for tagging the city area. Here, we have used a linear SVM for developing the classifier to tag a location as residential (tagged as RES), denoted by 1 or commercial (tagged as COM) denoted by -1. Given some training data $D \in Z$, a set of n points of the form

$$D = (z_{loc}, s_{loc}) | z_{loc} \in Z, s_{loc} \in \{-1, 1\}_{loc=1}^n \quad (5.14)$$

where the s_{loc} is either 1 or -1, indicating the class to which the point z_{loc} belongs. Each z_{loc} is a p -dimensional real vector and member of the feature vector set Z . The full feature vector set Z can be defined as follows

$$Z = \{z_{loc} | \{nH_{loc}, nW_{loc}, \beta_{T1}, \beta_{T2}\} \in z_{loc}, loc \in LLoc\} \quad (5.15)$$

Where,

nH_{loc} = Number of home in loc

nW_{loc} = Number of workplace in loc

β_{T1} = Busyness Score of loc in WORKING-HOURS, $T1$

β_{T2} = Busyness Score of loc in OFF-HOURS, $T2$

Combining Equation 5.15 with Equation 5.3 we can say that every feature vector z_{loc} contains the features, $NU_{T,loc}, NC_{T,loc}, nH_{loc}$ and nW_{loc} . We want to find the maximum-margin hyperplane that divides the points having $s_{loc} = 1$ from those having $s_{loc} = -1$. Any hyperplane can be written as the set of points z_{loc} satisfying

$$w \cdot z_{loc} - b = 0 \quad (5.16)$$

Here, w is the normal vector to the hyperplane and the parameter $\frac{b}{w}$ determines the offset of the hyperplane from the origin along w . For this linearly separable data, we can select

two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. For two classes these hyperplanes can be described by the following equations,

$$w \cdot z_{loc_i} - b \geq 1 \quad (5.17)$$

$$w \cdot z_{loc_i} - b \leq -1 \quad (5.18)$$

We tag different location of the the city using the following algorithm based on the linear SVM discussed in this section.

Algorithm 15 CITYAREATAG($X, LUsr, LLoc$)

Input:

- 1: $LUsr \leftarrow \{ u_1, u_2, \dots, u_i \}$ (list of i number of unique users)
- 2: $X \leftarrow \{ x_1, x_2, \dots, x_j \}$ (j entries in CDR)
- 3: $LLoc \leftarrow \{ loc_1, loc_2, \dots, loc_k \}$ List of all unique location

Output:

- 4: $LLtag \leftarrow$ List of type (RES/COM) of all area $\in LLoc$
 - 5: **BEGIN**
 - 6: $nH_{loc} \leftarrow$ Number of home in loc
 - 7: $nW_{loc} \leftarrow$ Number of workplace in loc
 - 8: **for all** $u_i \in LUsr$ **do**
 - 9: **if** FINDHOME(u_i) = loc **then**
 - 10: $nH_{loc} \leftarrow nH_{loc} + 1$
 - 11: **end if**
 - 12: **if** FINDWORKPLACE(u_i) = loc **then**
 - 13: $nW_{loc} \leftarrow nW_{loc} + 1$
 - 14: **end if**
 - 15: **end for**
 - 16: **for all** $loc_k \in LLoc$ **do**
 - 17: Calculate β_{T1,loc_k} and β_{T2,loc_k} using Equation 5.3
 - 18: SELECT $\{ nH_{loc_k}, nW_{loc_k}, \beta_{T1}, \beta_{T2} \} \in z_{loc_k}$
 - 19: **if** $w \cdot z_{loc_k} - b \geq 1$ **then**
 - 20: $Tag_{loc_k} \leftarrow$ RES
 - 21: **else**
 - 22: $Tag_{loc_k} \leftarrow$ COM
 - 23: **end if**
 - 24: **end for**
 - 25: **return** $LLtag$
 - 26: **END**
-

5.6 Layer 4

In Layer 4, we explore more closer social relationships and activity patterns, which leads to the discovery of smaller social groups. We have proposed some probabilistic prediction models using statistical classifier based on the hypotheses introduced in the previous chapter.

Hypothesis 1. *A group of users have the same home locations means that they lives in the same neighborhood. So, we can classify these loosely connected social groups as neighbors and members of same group have a high probability of interaction.*

The first hypothesis explains that, the people who live in the same area, which is under the same cell tower can be predicted as neighbors, which is an important social group in almost all culture of the world. Also, the probability of two random people living in the same area and having any kind of social relationship is much higher than two random people living in two different area. Let us consider there is n people, and the probability that two people, A and B, chosen at random know each other or have a social relationship, r_{AB} is,

$$P(r_{AB}) \in (0, 1) \quad (5.19)$$

Now, if we randomly choose three people, A,B and C, so that A and B live in same neighborhood and C lives in different one. In that case, the fact predicted from previous layers of our model would be,

$$FINDHOME(A) = FINDHOME(B)$$

and

$$FINDHOME(A) \neq FINDHOME(C)$$

In that case, using Equation 5.19 we can explain Hypothesis 1 as follows,

$$P(r_{AB}) > P(r_{AC}) \quad (5.20)$$

The calculation of $P(r_{AB})$ and $P(r_{AC})$ can be calculated through a chain of $\leq k$ people,

where $k \leq n$. But, this calculation is out of our context.

Here, according to Equation 5.20 we consider that people living in same neighborhood belongs to a social group and has a good probability of knowing each other. Also, this probability is inversely proportional to, n , the number of people living in that neighborhood as represented below,

$$P(r_{AB}) \propto \frac{1}{n} \quad (5.21)$$

The number of people living in an area, n , can be estimated by multiplying the number of active user in that area and percentage of mobile phone user in that city. So, we propose Algorithm 16 to predict the social groups of people who live in the same neighborhood according to Hypothesis 1. This algorithm finds the location of home information of all the user, check in which neighborhood the home location belongs and add that user to that neighbor group. Besides, it keeps a counter for the number of member in each group, which helps us to measure the probability of social relation among members. Thus, we predict the group membership of all the users and total number of members in each group.

Algorithm 16 NEIGHBORGROUPS($LU_{sr}, LLoc$)

Input:

- 1: $LU_{sr} \leftarrow \{ u_1, u_2, \dots, u_j \}$ (list of i number of unique users)
- 2: $LLoc \leftarrow \{ loc_1, loc_2, \dots, loc_i \}$ (List of i unique locations in CDR)

Output:

- 3: $NeighborGroups \leftarrow \{ ng_1, ng_2, \dots, ng_i \}$ (List of all neighbor groups and their members)
 - 4: **BEGIN**
 - 5: $NeighborGroups \leftarrow ULoc$ (Assign every unique location to a group's location, i.e. $loc_{ng_i} \leftarrow loc_i$)
 - 6: **for all** $u_j \in LU_{sr}$ **do**
 - 7: $home_{u_j} \leftarrow FINDHOME(u_j)$
 - 8: **for all** $loc_{ng_i} \in NeighborGroups$ **do**
 - 9: **if** $loc_{ng_i} = home_{u_j}$ **then**
 - 10: ADD u_j to ng_i
 - 11: **end if**
 - 12: Count number of u_j added to hg_i
 - 13: **end for**
 - 14: **end for**
 - 15: **END**
-

Hypothesis 2. *If a group of users have the same workplace locations, they have a high probability of being acquaintances, even colleagues.*

Hypothesis 2 works in a similar fashion as Hypothesis 1 for predicting relationship among people who have workplace in the same city area. From real life experience we know that, people working in the same area may know each other, may even work in the same office making a group of office workers. Also, people who have business or shops in the same area usually have a kind of mutual relationship which indicates a social group of businessman. The probabilistic calculations are same as we did for people live in same neighborhood and we can apply all the equation for working groups too.

The proposed algorithm finds the workplace location information of all the user, check in which workplace group the location belongs and add that user to that group. As previous algorithm, we keep a counter for number of member for each of the groups. Thus, we predict the group membership of all the users and total number of members in each working group.

Algorithm 17 WORKGROUPS($LU_{sr}, LLoc$)

Input:

- 1: $LU_{sr} \leftarrow \{ u_1, u_2, \dots, u_j \}$ (list of i number of unique users)
- 2: $LLoc \leftarrow \{ loc_1, loc_2, \dots, loc_i \}$ (List of i unique locations in CDR)

Output:

- 3: $WorkGroups \leftarrow \{ wg_1, wg_2, \dots, wg_i \}$ (List of all work groups and their members)
 - 4: **BEGIN**
 - 5: $WorkGroups \leftarrow ULoc$ (Assign every unique location to a group's location, i.e. $loc_{ng_i} \leftarrow loc_i$)
 - 6: **for all** $u_j \in LU_{sr}$ **do**
 - 7: $workplace_{u_j} \leftarrow WORKPLACE(u_j)$
 - 8: **for all** $loc_{ng_i} \in WorkGroups$ **do**
 - 9: **if** $loc_{ng_i} = workplace_{u_j}$ **then**
 - 10: ADD u_j to wg_i
 - 11: **end if**
 - 12: Count number of u_j added to wg_i
 - 13: **end for**
 - 14: **end for**
 - 15: **END**
-

Hypothesis 3. *The type of transport used by an user can be classified and grouped from his traveling route and information of time differences and distances among call activities made by him in the same day on that route.*

Using the Algorithm 18 based on Hypothesis 3 we have predicted the class of transport used by users in their trips through traveling routes. To do so, we calculated the speed of transport by calculating distance between two locations and time difference from consecutive CDR entries. Then, we calculated the average speed and used these features in a SVM to detect the class of a transport. We have used Linear SVM to develop our prediction classifier to find the type of transport, Tpt for each trips $T = (T_1, T_2, \dots, T_k)$. Here the Tpt_k is either MANUAL, denoted by 1 or MOTORIZED denoted by -1, indicating the class to which the transport type belongs. Given some training data $D \in X$, a set of n points of the form

$$D = (T_k, Tpt_k) | T_k \in T, Tpt_k \in \{-1, 1\}_{k=1}^n \quad (5.22)$$

Algorithm 18 TRANSPORTCLASS(U, CDR_U)

Input:

- 1: $TR \leftarrow (T_1, T_2, \dots, T_k)$ Trips through traveling route TR for user U
- 2: $T_k \leftarrow (x_1, x_2, \dots, x_n) \in CDR_U$ CDR entries in trip T_k for user U

Output:

- 3: $Tpt \leftarrow$ class of the transport (MANUAL, MOTORIZED)
 - 4: **BEGIN**
 - 5: **for all** x_i, x_j pair in TR **do**
 - 6: $Distance \leftarrow$ HAVERSINE(loc_i, loc_j);
 - 7: $TimeDifference \leftarrow |tm_i - tm_j|$;
 - 8: Calculate $Speed$ from $Distance$ and $TimeDifference$;
 - 9: **end for**
 - 10: Calculate Average $AVG(Speed)$
 - 11: **for all** $T_k \in TR$ **do**
 - 12: **if** $w.T_k - b \geq 1$ **then**
 - 13: $Tpt_k \leftarrow$ MANUAL
 - 14: **else**
 - 15: $Tpt_k \leftarrow$ MOTORIZED
 - 16: **end if**
 - 17: **end for**
 - 18: **END**
-

We want to find the maximum-margin hyperplane that divides the points having $Tpt_k = 1$ from those having $Tpt_k = -1$, which can be written as the set of points T_k satisfying

$$w.T_k - b = 0 \quad (5.23)$$

We select two hyperplanes in a way that they separate the data and there are no points between them, and then maximize their distance. These hyperplanes can be described by the equations for two groups respectively,

$$w.T_k - b \geq 1 \quad (5.24)$$

$$w.T_k - b \leq -1 \quad (5.25)$$

Using these equations in Algorithm 18 we predict the class of transport user by a user in a trip.

Hypothesis 4. *If a two or more persons have common home, workplace or more than one common POIs and same or overlapping regular traveling route, we can perform predictive analysis of their social relationship and group membership using available information.*

We have applied this hypothesis to predict social relationship based on probability of interactions, working and traveling pattern. For, example If a group of people share the same home and workplace, we can predict that they are co-workers and live in same residential facility, which may be provided by the employer. This prediction would be more established is we find a similar traveling pattern, which we detect by predicting this traveling route and time period of traveling.

Any traveling route, R is a sequential set of i number of location coordinates as below,

$$R = loc_1, loc_2, \dots, loc_i$$

So, if R_A and R_B is the traveling route of user A and B respectively, we predict their common or intersecting traveling route R_{AB} in T time period, where,

$$R_{AB,T} = R_{A,T} \cup R_{B,T} \quad (5.26)$$

Thus we predict the common or intersecting traveling route for two random users.

In this layer, we predict individual user's social profile based on the predictions about

him up to this layer. We examine the time period and location based social activities using a modified version of Equation 5.6 which is as follows,

$$\mu_u = \omega_c NC_{T,loc} + \omega_d \sum dur_{T,loc} \quad (5.27)$$

Where,

T is the time period of call activity for user u .

loc is the location of call activity for user u .

Here, we can change the parameters for location and time period of the user and find his group membership. For example, the users with high usage score in the time period 12 AM to 4 AM are late night callers. Similarly, working user with high usage score during working hours belongs to special group of professionals who highly focus on communication. Also, location based prediction is an important feature of our exploration model. For example, if an user has a working place in an university area, he has a high probability of being a student or teacher of that university. Now, if he lives in the residential area for students and a late night caller, it is more likely that he is a student. Similarly, user with workplace in a large shopping is likely a shopkeeper, user with workplace in a cantonment is a prospective member of military and user with workplace in a hospital area is probably a doctor, nurse or medical personnel. We predict the working days and offdays of a working user from his days staying in workplace during working hours.

5.7 Layer 5

In the fifth layer we consider the knowledge of calling relationships among the members of the groups detected earlier and design algorithms to utilize them for exploring deeper into more intimate social relationships and groups. In this layer we use the hypotheses proposed in the previous chapter. We detect social relationships and groups including family, friends, colleagues and closely acquainted people by combining the relationship knowledge predicted in previous layer and calling relationship graph. the calling relationship graph itself is a huge social network.

Hypothesis 5. *If two or more people share the same home location and have a frequent calling relationship, we can predict the probability of them being family members or close acquaintances.*

According to Hypothesis 5, we use calling relationship to find the family members of the users. As calling relationship is the strongest indication of any social relationship, we predict that the users who have calling relationship and live in the same place have a good probability of being a family member or at least close acquaintance. The number of calls made between two users is another important factor for predicting closeness of social relation. The closeness of this type of relationship can be further investigated by examining the time and duration of calls and frequency of call made. We propose Algorithm 19 to predict the family members of a user considering this facts.

Algorithm 19 FAMILYMEMBERS(LU_{sr}, R)

Input:

- 1: $LU_{sr} \leftarrow \{ u_1, u_2, \dots, u_n \}$ (list of i number of unique users)
- 2: $R \leftarrow$ Calling Relationship graph

Output:

- 3: $RL_{u_n} \leftarrow$ List of users having social relation with u_n
 - 4: **BEGIN**
 - 5: $r_{th} \leftarrow$ minimum number of call made between two users in a certain time period to consider a calling relation as social relation
 - 6: **for all** u_j, u_i pair $\in LU_{sr}$ **do**
 - 7: **if** $FINDHOME(u_j) = FINDHOME(u_i)$ AND $u_j \neq u_i$ **then**
 - 8: **if** $r_{u_i, u_j} \in R$ and $r_{u_i, u_j} > r_{th}$ **then**
 - 9: ADD u_j to RL_{u_i} as Familymember
 - 10: ADD u_i to RL_{u_j} as Familymember
 - 11: **end if**
 - 12: **end if**
 - 13: **end for**
 - 14: **END**
-

Hypothesis 6. *If two or more people share the same workplace location and have a frequent calling relationship, we can predict the probability of them being colleagues, co-workers or friends.*

Similarly, Hypothesis 6 explore the workplace based relationships among users. All the people work in the same area can be considered as colleagues or friends based on their calling relationship pattern. Calling relationship is analyzed to predict the type of this relationship. The features from calling relationship we considered are the time of call,

call duration and frequency of call made. If two users from same workplace call each other in the off hours and even after midnight frequently, it indicates a closer social relationship. Also, call duration and frequency of call made is proportional to closeness of social relationship. That means, socially close people call more often and talk for longer time.

5.8 ASC Score and Aggregated Social Group Prediction Model

Hypothesis 7. *If two or more persons visit or stay in same POIs multiple times in same time periods and have a calling relationship, we can predict the aggregated probability of them being member of a social group like family, friends or other types of close relationships.*

Using the combined knowledge database about all the user, we have developed a statistical prediction model which calculate the Aggregated Social Closeness (ASC) score between two users and predict aggregated Social Group. we consider all the results of the classifiers in different layers and combining them using a statistical prediction model, we calculate ASC between two users, which predict the probability of closeness. Evidently, a higher ASC value indicate that the two users are family members or close friends, on the other hand, a lower ASC means that they have no relation or interaction at all.

ASC is calculated using the chain rule of probability. To calculate the ACR between two user A and B, We consider the following probability found in different layers of our model.

$P_{AB}(H)$ = Probability of A and B's home in same location. $P_{AB}(W)$ = Probability of A and B's workplace in same location. $P_{AB}(cr)$ = Probability of A and B's calling relation. $P_{AB}(tr)$ = Probability of A and B's overlapping traveling route.

$$ACR = P_{AB}(H)P_{AB}(W)P_{AB}(cr)P_{AB}(tr) \quad (5.28)$$

We have developed our aggregated social group prediction model based on naive Bayes probabilistic model. All our findings in the previous layers are represented by a

feature vector, $R = (r_1, \dots, r_n)$ representing n types of social relationship, it assigns to this instance probabilities

$$p(C_k|x_1, \dots, x_n) \quad (5.29)$$

for each of K possible outcomes or classes. Its' equivalent joint probability model is,

$$p(C_k, x_1, \dots, x_n) \quad (5.30)$$

Which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$p(C_k)p(x_1, \dots, x_n|C_k) \quad (5.31)$$

Using the naive Bayes probability model we can propose a naive Bayes classifier. It is the following function that assigns a class label $y = C_k$ for some k as follows:

$$y = \underset{k \in 1, \dots, k}{\operatorname{argmax}} p(C_k) \prod_{i=1}^n p(x_i|C_k) \quad (5.32)$$

This naive Bayes classifier give us a social group prediction model based on our finding in the previous layers. We calculate ASC between two users, which predict the depth of their relationship and interactions. Clearly, a higher ASC value indicates that the two users are family members or close friends, on the other hand, a lower ASC means that they have no relation or interaction at all.

5.9 Summary

In this chapter we discuss the detailed methodology of our work base on the framework of our hierarchical exploration model explained in the previous chapter. So, we presented the algorithms we developed for applying in different layers to explore progressively the fact on social groups and their relationship.

Chapter 6

Results and Analysis

We implement our proposed algorithms, executed them on our CDR and perform validation test with some real life data collected from some people around us, who volunteered to provide us their personal call data. This chapter contains the detail of the dataset, experimental settings, results and analysis of our thesis work. In Section 6.1, we explain our datasets. Afterwards, in Section 6.2, we explain our experimental environment and settings for our implementation. Rest of this chapter contains the experimental results and analysis of the results.

6.1 Data Collection and Dataset

This research work is done using the data of the people who live in Dhaka, the capital city of Bangladesh. Our datasets are CDR data obtained from Grameenphone Ltd, a major telecom operator of Bangladesh. This comprises of 971.33 million anonymous call records made by 6.9 million users, which are more than 55 percent of the total population of the study area. It was collected for a duration of one month, from June 19, 2012 to July 18, 2012. As we have already mentioned, for each record the CDR Dataset has the following parameters: a random ID number of the phone; the exact time and date; call duration and location (latitude and longitude) of the cell tower that provided the network signal for the mobile device activities. The random ID number is generated by the operator for every mobile phone, independent to the phone or the SIM card number, which allows us to link each phone owner with their mobile device activities.

Data were recorded following the legislations of Bangladesh for billing purposes by the operator and not for our study. Geographical information was obtained from the geographical coordinates (latitude and longitude) of network antennas. The precision of the spatial accuracy of the mobile device activity corresponds to the coverage of a network antenna. The coverage area is not spatially fixed and varies according to population density. As Dhaka is one of the most densely populated cities of the world, the analysis on the sample collected from here gives us some unique results because of the uniqueness of the city.

Table 6.1: Sample data from CDR dataset

| User ID | Date | Time | Duration | Latitude | Longitude |
|--------------------|----------|----------|----------|-----------|-----------|
| AAH03JAAQAAAO9VAA+ | 20120714 | 10:44:04 | 68 | 23.758301 | 90.402199 |
| AAH03JAAQAAAO9VAA+ | 20120714 | 21:16:23 | 60 | 23.758301 | 90.402199 |
| AAH03JAAQAAAO9VAA/ | 20120708 | 22:10:37 | 1527 | 23.701700 | 90.429199 |
| AAH03JAAQAAAO9VAA/ | 20120711 | 12:12:54 | 1103 | 23.724199 | 90.405602 |
| AAH03JAAQAAAO9VAA/ | 20120711 | 10:26:33 | 304 | 23.722200 | 90.409203 |

Another smaller set of data we have used in our work is the personal CDR form a few volunteers to validate the results found from our hierarchical exploration model using the massive anonymous CDR data. Now, most of the mobile operators allow the users to retrieve their personal CDR data but using online user account. We took the benefit of this feature and collected the CDR data form the online account of a few users. this data is more comprehensive in feature and they contain the information about the Call Date, Time, Called Number, Actual, Duration(Sec), Charges in BDT, Call Type (IN, OUT), FNF, Usage Type (VOICE, SMS, DATA) for every call activity. We have collected this type of personal CDR in a limited scale and validated the results obtained from the original massive CDR collected from the phone operator.

6.2 Experimental Setup

In our thesis work, the proposed hierarchical exploration model with multiple layers for processing the CDR data for identifying the user activities and mobility patterns. According to the framework of our model, in each layer we have used a set of fact finding and prediction algorithms to find out different information, detect patterns about the social activity, relationships and group belongingness of the users. To do so we

started with the raw CDR data in the initial layer and later the information obtained in each layers is added in the combined knowledge base alongside the original CDR. The novelty of our proposed model is that, the algorithms in every layer use the original CDR data and also the information and facts derived in the preceding layers collectively present in the combined knowledge base to discover new facts and patterns. Thus, as further we go on with the layers, gradually we focus on more detailed information about smaller groups and closer relationships, which were not identifiable using the original CDR data directly in a single layered approach. Therefore, algorithms in each layer is dependent on the outputs of the previous layers justifying the requisite of the hierarchical approach.

The raw CDR data are contained in simple but very large text files. For processing this data using our algorithm we developed a programs based on our proposed prediction classifiers and clustering algorithms in the layers of our hierarchical framework using JAVA. We also utilize WEKA data mining tool to implement all these classifier and clustering techniques in JAVA. We have obtained most of the the outputs in two types, output files and visualizations, while some of the summarized results are simply shown in a GUI. For better understanding of the results discovered in different layers, the visualization of the location data is necessary. For different visualizations related to map and location data, we have used Google Maps API in our JAVA programs.

The experiments on Big Data like CDR is resource demanding and time consuming. So, all the experimental implementations of this thesis are done on a number of personal computer parallely equipped with Intel Core i5 CPUs running at 2.0 GHz or more and equipped with 4 to 6 GB RAM.

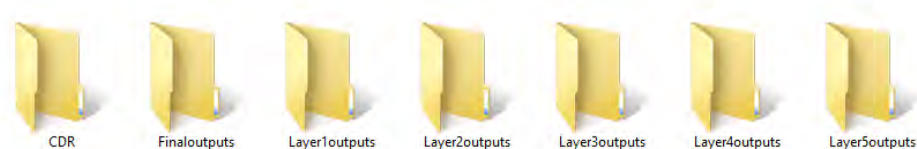


Figure 6.1: Directory structure of the combined knowledge database

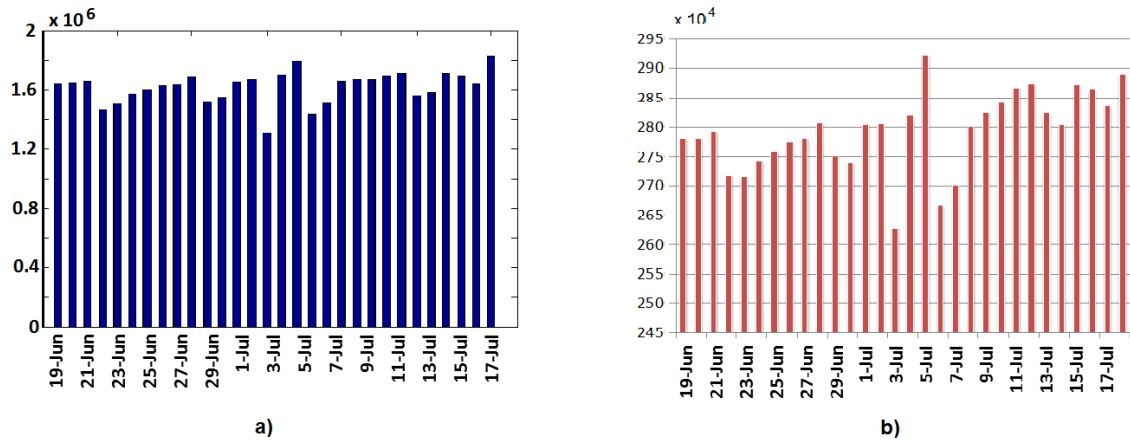


Figure 6.2: a) Number of calls per day b) Number of active user per day

6.3 Results

In this section, the results are presented and discussed as they are obtained in different layers of our hierarchical exploration model.

Layer 1

In the first layer, we perform some analysis of our raw CDR data. For understanding the city dynamics, we can detect the call activities of all users in any given time slot of the day. It gave us an idea about intensity of activity in the city during different periods of the day, as phone calls are directly related to other regular activities of urban life. Figure 6.2 shows the number of calls made per day and number of active user per day in the whole city, which presents us with a pattern representing the city status in days of the week. Here we can see that, maximum number of active user and call is on 5th July, 2012, which was a national holiday in Dhaka city. So, we can clearly see that people make social interaction over phone more on holidays. On the other hand it is also clear that people make less call on weekends. These two fact are significant information on social activity.

Also, we have detected the call activity in different times of the day, which reflect the overall activity and dynamics of the city (Table IV and Figure V). The time slots can be selected as per the requirement of analysis. We divided the day in four slots and identified the user activity accordingly.

Table 6.2: User activity in different periods of the day

| Timeslot | Number of Calls | Number of Active Users |
|---------------|-----------------|------------------------|
| 12 AM to 6 AM | 370311 | 192928 |
| 6 AM to 10 AM | 1256755 | 578975 |
| 10 AM to 5 PM | 3442933 | 960704 |
| 5 PM to 12 AM | 3187238 | 968891 |

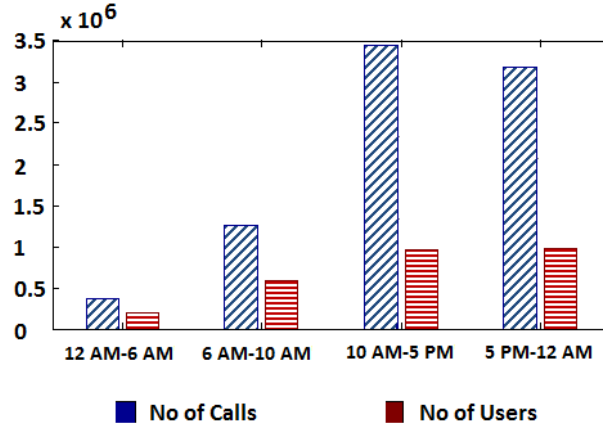


Figure 6.3: User activity in different periods of the day

Voronoi tessellation is used to conveniently display the spatial variation of the call activity at any given time. Maps divided in Voronoi cells delimiting the area of influence of each cell tower or antenna. The Voronoi tessellation partitions the plane into polygonal regions, associating each region with one cell tower. The partition is such that all points within a given Voronoi cell are closer to its corresponding tower than to any other tower in the map. These information reflect the obvious and apparently known notions regarding the calling habits of the users, i.e. people talk more in the daytime to before midnight and less after the midnight till dawn, where as a certain group talks more after midnight.

Figure 6.6 shows activity maps for aggregated data corresponding to a 1-hour interval. The left panel shows the activity pattern for a peak hour, while the right panel shows the same neighborhood of Dhaka city during an off-peak hour . The differences between both panels reflect the intrinsic rhythm and pulse of the city. We can expect call patterns during peak hours to be dominated by the hectic activity around business and office areas, whereas other, presumably residential and leisure areas can show increased activity during off-peak times, thus leading to different, spatially distinct activity patterns. Besides different spatial patterns, each particular time of the day, as

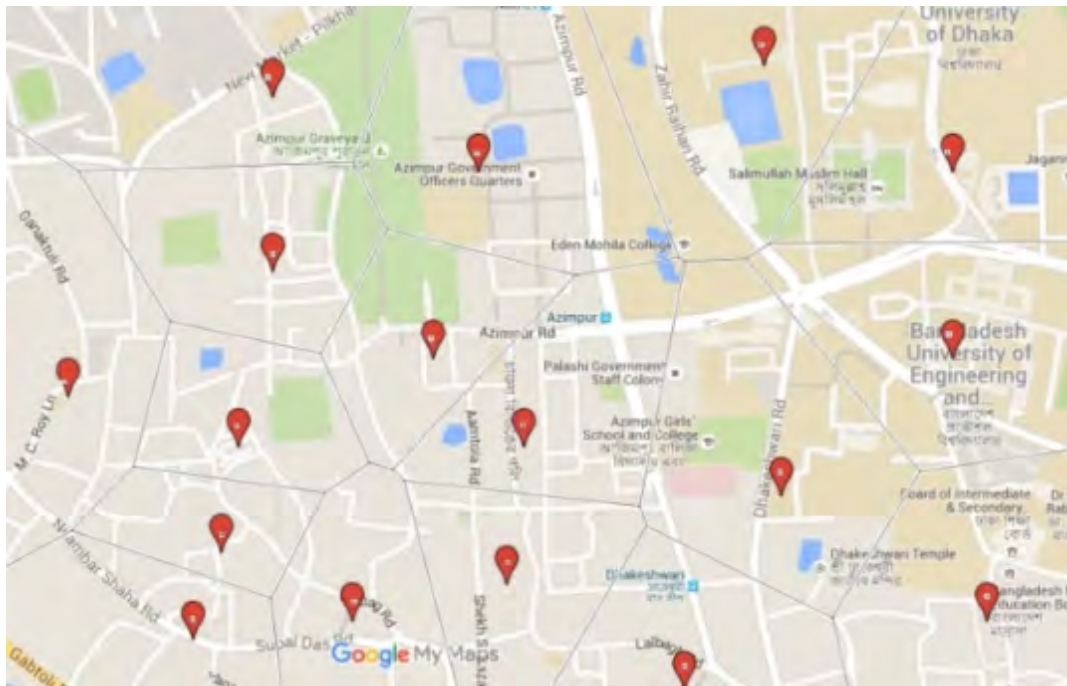


Figure 6.4: Initial voronoi diagram on the city map



Figure 6.5: Voronoi diagram on the city map after colorization based on call activity

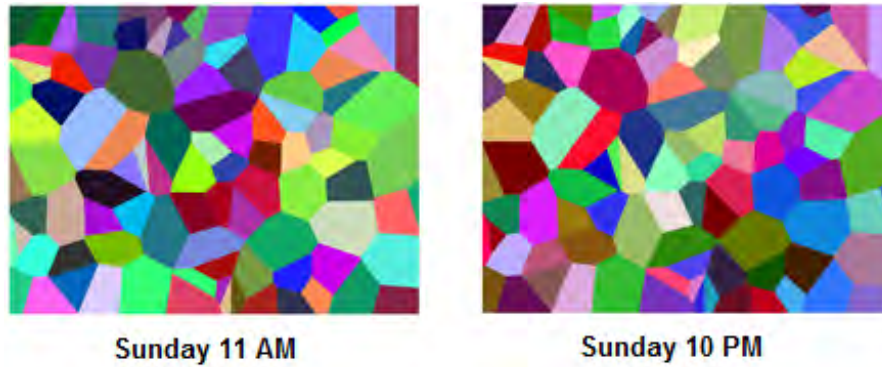


Figure 6.6: Voronoi diagram showing call activity in different times of day

Table 6.3: Comparing call traffic variation in two consecutive weeks

| Day | Date | No of users | No of calls | Date | No of users | No of calls |
|-----------|--------|-------------|-------------|--------|-------------|-------------|
| Tuesday | 19-Jun | 2780393 | 16376825 | 26-Jun | 2774048 | 16259300 |
| Wednesday | 20-Jun | 2782646 | 16437123 | 27-Jun | 2780499 | 16359352 |
| Thursday | 21-Jun | 2793478 | 16589529 | 28-Jun | 2807500 | 16878637 |
| Friday | 22-Jun | 2717526 | 14660436 | 29-Jun | 2751396 | 15178572 |
| Saturday | 23-Jun | 2714540 | 15030609 | 30-Jun | 2741296 | 15477377 |
| Sunday | 24-Jun | 2744098 | 15682032 | 1-Jul | 2805570 | 16514478 |
| Monday | 25-Jun | 2758351 | 16005153 | 2-Jul | 2805998 | 16673047 |

well as each day of the week, is characterized by a different overall level of activity.

The call volume shows strong variations with time and day of the week, as shown in Figure 6.6, but differences across subsequent weeks are generally mild, even if we consider call traffic of the whole city as shown in Table 6.3 and . It is more prominent if we consider call traffic in the same place, time and day of the week.

From the usage score US , which signifies call activity of users, we classify three groups, Minimal users, Regular users and Heavy users. To perform all kinds of calling pattern analysis we only considered the active users from our whole data sample. Active users are the people who made 10 or more calls in our one month window. The users who made less than 10 calls are inactive users and can't provide sufficient data to analysis. We assumed that these user are redundant users who have another primary connection. It is a common scenario due to the inexpensiveness and availability of mobile phone connecting SIM cards. According to our analysis, more than 20 percent of the total users are inactive. Among the active user, we have shown the percentage of above three group memberships of users in 6.4. This classification has a prominent social implication. It can be safely inferred that, the heavy users are socially more active,

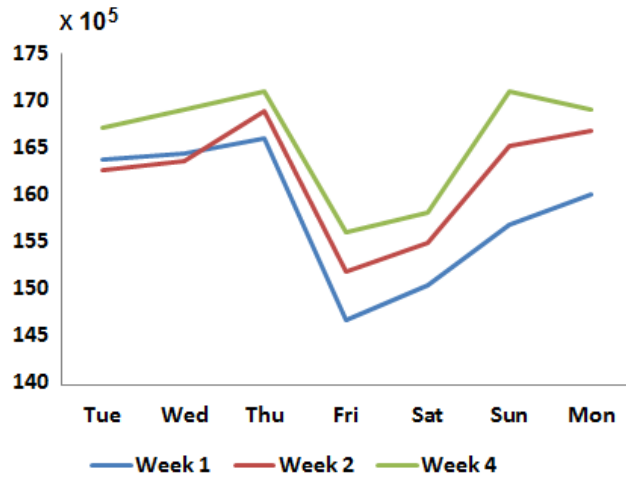


Figure 6.7: Comparing call traffic variations from usage score, μ in three consecutive weeks

while other two groups are less active proportionately.

Table 6.4: Classifying user based on call activity

| Group | No of Users | Percentage |
|---------------|-------------|------------|
| Minimal users | 2478480 | 36 |
| Regular users | 4036447 | 58 |
| Heavy users | 412046 | 6 |

We generate caller-callee relationship graph by cross referencing the call activities taking place in same time with same duration. Thus we detected the both participant call activity for every user. 6.5 shows a few lines of this relationships detected by our algorithms. Every relationship is given a score based on number of calls and call duration between two users.

Table 6.5: Detected calling relationship among users

| USER A | Call Time | Call Duration | USER B |
|--------------------|-----------|---------------|--------------------|
| AAH03JAAQAAA09VAAR | 75209 | 145 | AAH03JAARAAACttAjh |
| AAH03JAAQAAA09VAAq | 43246 | 32 | AAH03JAAbAAH86JAAN |
| AAH03JAAQAAA09VABK | 75633 | 33 | AAH03JAAbAAH86JAZ8 |
| AAH03JAAQAAA09VAAj | 67063 | 54 | AAH03JAAbAAH86KAOE |
| AAH03JAAQAAA09VAA1 | 71834 | 12 | AAH03JAB+AAAgO5AUv |

Layer 2

In this layer, we have used the CDR data to find the POIs of the users. For every user, in the CDR data, we have their locations with the date and times of the call activity made. We have used clustering algorithms to find clusters from user log, which are the POIs. Every cluster found from the algorithms is a POI. As we want the automated detection of number of POIs or clusters, we have used X-Means and EM clustering algorithms to find the POIs. Due to much better performance, we selected EM and used it for next algorithms to find home and workplace. For this method to work effectively, we need a reasonably good number of call records for each user. As we have a massive amount of CDR data, which is practically achievable in minimal efforts. Also, in Dhaka city the density of the cell tower is reasonably high we have considered the location of the cell tower providing the signal for the call activity as the approximate location of the users. 6.6 is showing the POIs of a random user predicted using clustering algorithm. Also the following figures illustrate the clusters and visualization of POIs.

Table 6.6: POIs of a random user

| Location | No of Total Calls | No of Calls in Working Hours | No of Calls in Off Hours |
|-------------------------|-------------------|------------------------------|--------------------------|
| "23.856100","90.402802" | 95 | 88 | 7 |
| "23.858900","90.408302" | 66 | 23 | 43 |
| "23.819201","90.417198" | 26 | 26 | 0 |
| "23.783300","90.395302" | 12 | 12 | 0 |
| "23.928600","90.300301" | 11 | 11 | 0 |

Later, we predicted the home and workplace of the same user using the same clustering technique. We know the usual working hour of Dhaka city, which is 9 AM to 5 PM or a slight variation of this time-slot. Now, for each of the POIs, if a user made most of the calls in the working hours, we consider it as his workplace. On the other hand, if the user made most of the calls from a stay location during the off-hours or in the usual holidays, we considered it as his home. By applying this fact with our clustering algorithm, while predicting home, we take the top cluster of the calls made in off-hours, and, for predicting workplace we consider the calls made in working hours only.

On the map of Figure 6.12, we can see the visualization of the POIs of a single user. A map marker is placed in the locations from where he has made one or more calls.

EM
 ==

Number of clusters selected by cross validation: 8

| Attribute | Cluster | | | | | | | |
|-----------|---------|---------|---------|---------|---------|---------|---------|---------|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| | (0.31) | (0.26) | (0.16) | (0.03) | (0.11) | (0.03) | (0.02) | (0.09) |
| ===== | | | | | | | | |
| 24.000601 | | | | | | | | |
| mean | 23.8777 | 24.0189 | 23.8405 | 23.9057 | 23.9461 | 23.8692 | 23.8475 | 23.8695 |
| std. dev. | 0.0015 | 0.0107 | 0.0027 | 0.0064 | 0.0177 | 0.0126 | 0.0041 | 0.0121 |
| 90.250298 | | | | | | | | |
| mean | 90.3897 | 90.2457 | 90.3867 | 90.3223 | 90.2812 | 90.3005 | 90.2712 | 90.3963 |
| std. dev. | 0.0018 | 0.0027 | 0.0024 | 0.0031 | 0.0151 | 0.0199 | 0.0084 | 0.0089 |

Figure 6.8: Result of using EM clustering algorithm on the CDR to find POIs

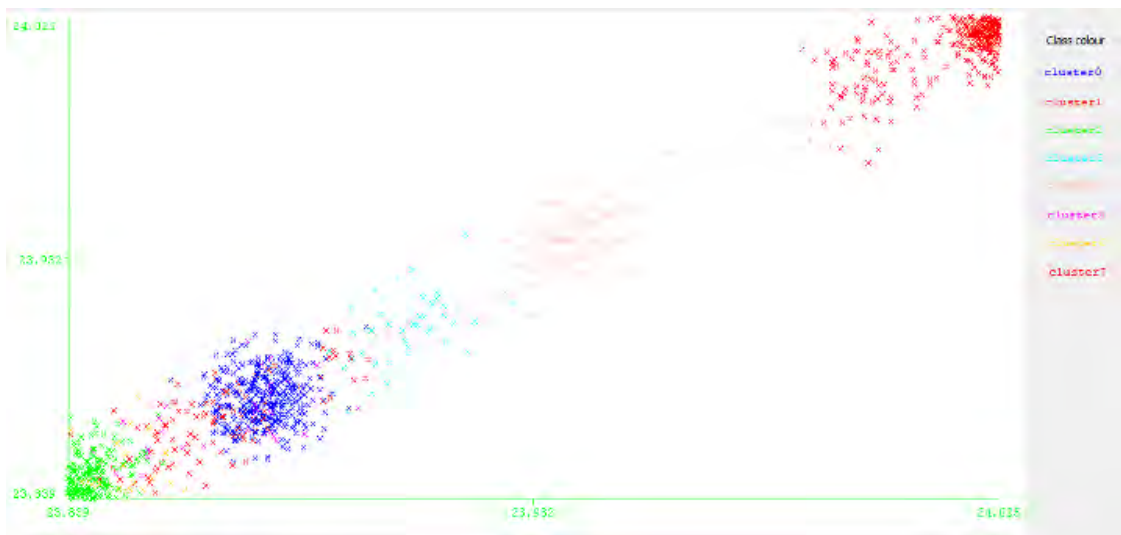


Figure 6.9: Visualization of clusters as a result of EM indicating POIs

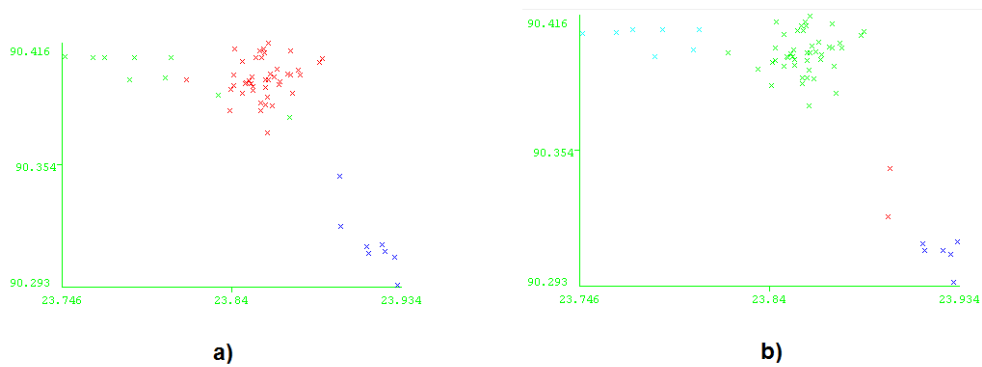


Figure 6.10: Finding Home of an User using a) EM Clustering Algorithm b)XMeans Clustering Algorithm

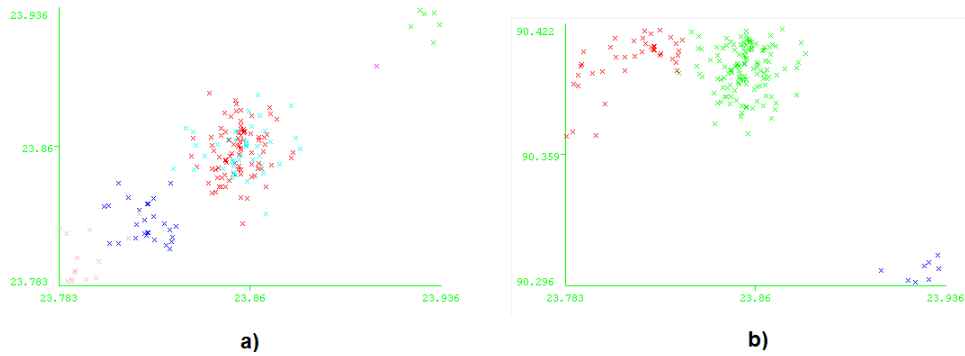


Figure 6.11: Finding Workplace of an User using a) EM Clustering Algorithm
b) XMeans Clustering Algorithm

The home and workplace are marked with different symbols. Beside every marker the number of calls made from that location is mentioned. From this visualization, we get the complete picture of the mobility and city area covered by a single user. This information is used to find UCLs, which later utilized to predict the usual routes of traveling in the city.

We classifying city areas in two classes BUSY and IDLE in different time periods of the day using a linear classifier based on SVM. We have 1360 unique locations in our CDR and we consider them as a zone of the city. We find the concentration of call activities in different times of the day, which determines how busy or densely populated the zone is at a certain time of the day. The Voronoi tessellation created in layer one is useful to visualize this information.

Table 6.7: Status of City Areas in Working Hour

| Location | Status |
|-------------------------|--------|
| "23.856100","90.402802" | BUSY |
| "23.858900","90.408302" | BUSY |
| "23.819201","90.417198" | IDLE |
| "23.783300","90.395302" | BUSY |
| "23.928600","90.300301" | BUSY |

Layer 3

In this layer we process the whole CDR database to find the home and workplace of all the users using our classifier iteratively. The result is a list containing the home, workplace and additional information of all the 6.9 million users, which is another Big

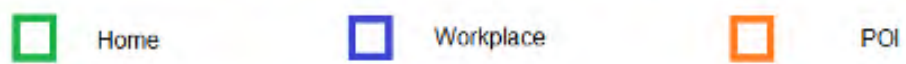
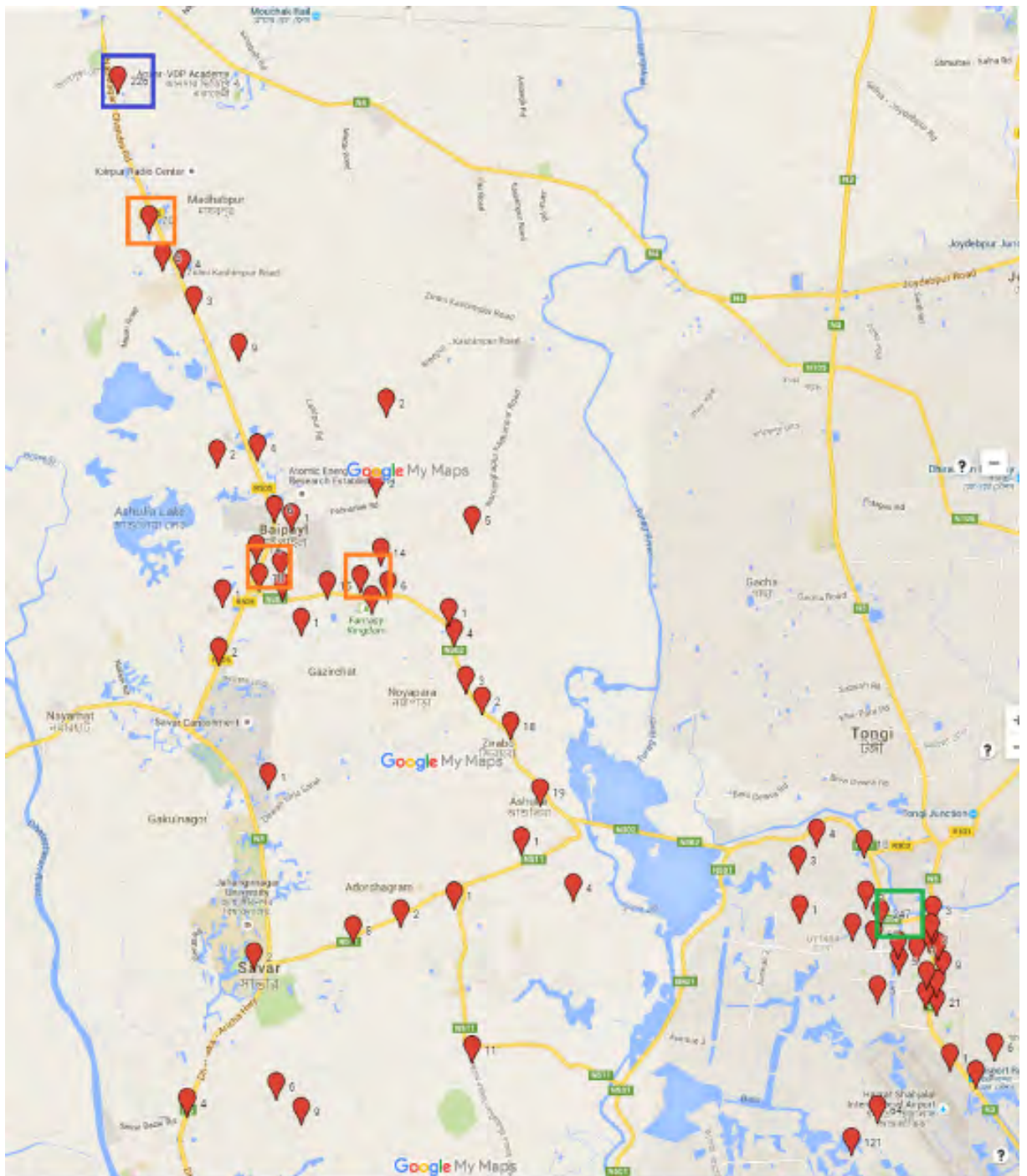


Figure 6.12: Visualization of home, workplace and other POIs of an user

Table 6.8: Part of Home and Workplace information data of all the users

| User ID | Home | | Workplace | | Distance | NCall | Dur Call | US |
|-----------|--------|--------|-----------|--------|----------|-------|----------|------|
| AAH86JAa9 | 23.789 | 90.408 | 23.787 | 90.415 | 0.72 | 22 | 3469 | 80 |
| AAH86JAa7 | 23.796 | 90.364 | NA | NA | NA | 966 | 78429 | 2273 |
| AAH86JAA8 | 23.707 | 90.410 | NA | NA | NA | 165 | 18731 | 477 |
| AAH86JAA9 | 23.846 | 90.421 | 23.793 | 90.402 | 6.20 | 50 | 6262 | 154 |
| AAH86JAAA | 23.710 | 90.404 | 23.812 | 90.255 | 18.93 | 26 | 4201 | 96 |
| AAH86JAa8 | 23.723 | 90.384 | NA | NA | NA | 101 | 23433 | 492 |

Data file. A sample of complete predicted data is shown in Table 6.8

After prediction of the two most important social hubs, home and workplaces of all the users in our CDR, we have used this information to investigate closer social relations and smaller social groups. When, we applied our SVM based classifier on our home and workplace database of all users, we found two groups of people based on their working patterns. One group, the regular workers, has a certain call activity pattern which enabled us to distinguish their home and workplace. Another group has no regular working pattern for distinguishing home and workplace and they are the irregular workers.

Subsequently, when the home and workplace of the regular working group is found, we have used this information to calculate the regular distance traveled between their home and working place. We calculated it from the coordinates of their home and working applying Haversine algorithm.

Table 6.9: Working pattern of the Users: Regular vs Irregular

| | | |
|------------------|---------|--------------|
| Irregular worker | 5163239 | 74.5 Percent |
| Regular Worker | 1763734 | 25.5 Percent |

Using our method we have seen that 1.8 million (25.5 percent) of the 6.9 million users of our CDR data have a consistent working schedule and the home and workplace of these people is clearly detected. The other 74.5 percent people have irregular patterns of home and workplace (Table III). According to our rational hypothesis, people like housewives, retired people, part-time and irregular workers, etc. or people who work and live in the same places belong to this group. This indicates that we have worked with a sample of around 12 percent of the total 15.4 million people [29] of Dhaka City.

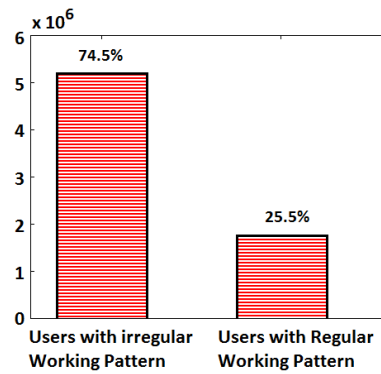


Figure 6.13: Working pattern of the users: Regular vs Irregular

By applying this method on all regular workers in our CDR data, we have found their distances traveled for attending workplace from home. Based on this information, we have classified these people in some groups for having some idea about the traveling pattern of the working people.

Table 6.10: Traveling distance to workplace for regular workers

| Traveling Distance | No of User | Percentage |
|--------------------|------------|------------|
| 0-2 km | 951217 | 53.93 |
| 2-5 km | 368484 | 20.89 |
| 5-10 km | 258620 | 14.66 |
| 10-20 km | 149168 | 8.46 |
| 20-100 km | 36245 | 2.06 |

From our findings presented in Table 6.10 we can see that 53.93 percent of the regularly working people live within two kilometers of their workplace and only 2.06 percent people live more than 20 km away from their workplace. By analyzing these groups, we can see that most of the people try to stay near their workplace and they try to travel less for going to the workplace as much as possible. Consequently, they usually select their home near their workplace or select their workplace near their home, as traveling in a densely populated city like Dhaka is difficult.

We have predicted the regular traveling route of an user by considering his UCLs during his travels between home to workplace. It is done by applying Dijkstra’s algorithm for finding single-source shortest paths considering the UCLs as nodes, home location as source and workplace as destination.

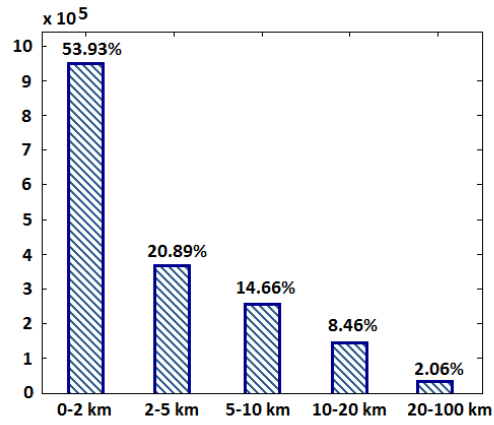


Figure 6.14: Mobility Pattern of Regular Working Users

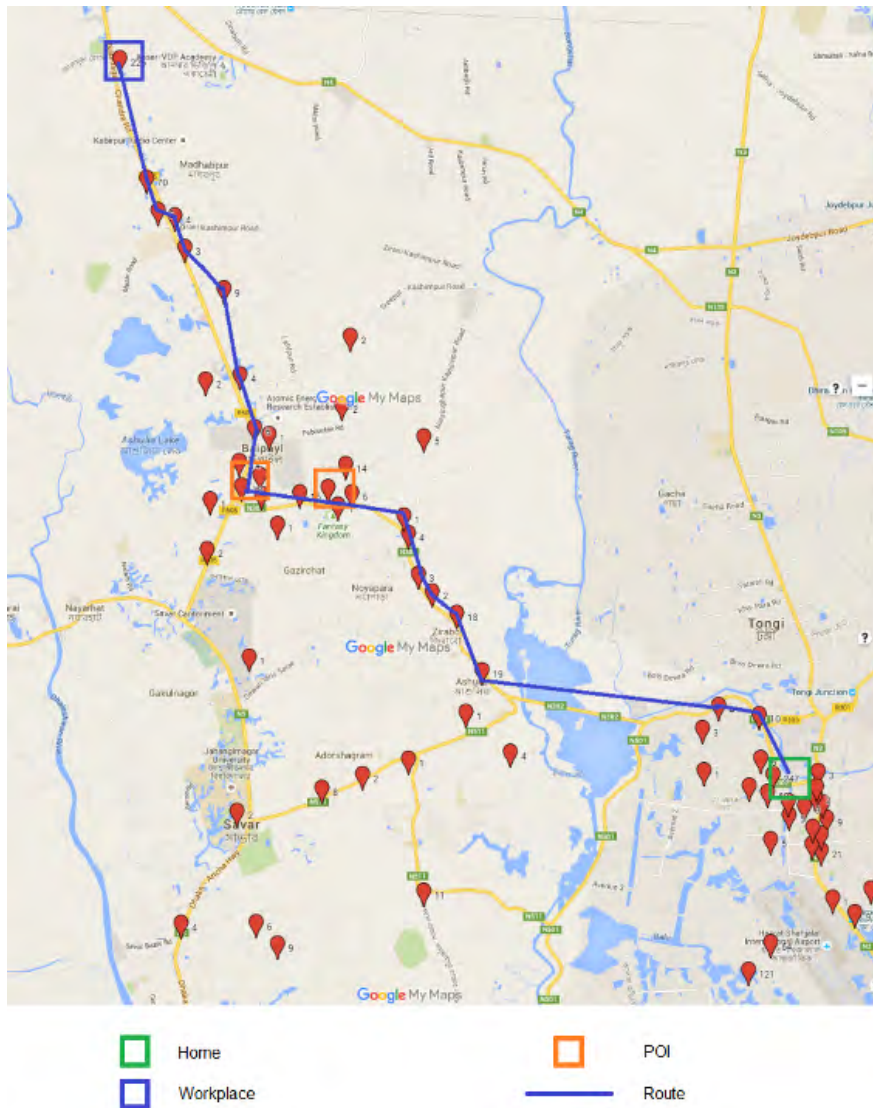


Figure 6.15: Traveling Route of an user from home to workplace

Table 6.11: Traveling Route of a user

| LAT | LONG | Remarks | |
|-----------|-----------|-----------------|-----------|
| 23.878599 | 90.390602 | Home | |
| 23.843901 | 90.279404 | Traveling Route | |
| 23.848301 | 90.274696 | | |
| 23.8547 | 90.312202 | | |
| 23.855801 | 90.266701 | | |
| 23.864401 | 90.3992 | | |
| 23.869699 | 90.402496 | | |
| 23.875 | 90.389397 | | |
| 23.875799 | 90.289398 | | |
| 23.8792 | 90.400597 | | |
| 23.883101 | 90.331703 | | |
| 23.8908 | 90.387497 | | |
| 23.937799 | 90.2714 | | |
| 23.942801 | 90.270798 | | |
| 23.948299 | 90.277496 | | |
| 23.949699 | 90.274399 | | |
| 23.9606 | 90.271103 | | |
| 23.9781 | 90.267502 | | |
| 23.9928 | 90.256699 | | |
| 24.025 | 90.244202 | | Workplace |

The traveling route is actually a list of sequential coordinates predicted from the list of UCL of the user. The traveling route of a random user is shown in Table 6.11. The traveling route can be visualized using map as presented in 6.15.

Layer 4

In layer 4, we find closer social relationships and activity patterns, which leads to the discovery of smaller social groups consisting of lesser set of users. The results predicted from layer 4 is presented below.

The type of transport used by an user can be classified and grouped from his traveling route and information of time differences and distances among call activities made by him in the same day on that route. For example, our prediction model can predict a fact which means,

”User X travel from Location P to Location Q using traveling route R using transport type T_1 and T_2 everyday”. The following Table 6.13 is showing the part of the trace file generated by the transport type detection classifier.

Table 6.12: Part of list of users living in same neighbor-group

| User ID | Home Lat | Home Long |
|--------------------|----------|-----------|
| AAH03JAAbAAH86JACK | 23.5394 | 90.1717 |
| AAH03JAAQAAA09VAWj | 23.5394 | 90.1717 |
| AAH03JAAQAAA09WAau | 23.5394 | 90.1717 |
| AAH03JAARAAACttAW5 | 23.5394 | 90.1717 |
| AAH03JAARAAACtuAJF | 23.5394 | 90.1717 |
| AAH03JAASAAAF6rAOi | 23.5394 | 90.1717 |
| AAH03JAAUAAABBDAAs | 23.5394 | 90.1717 |
| AAH03JAAUAAABBEAlt | 23.5394 | 90.1717 |
| AAH03JAAVAAAFwhATb | 23.5394 | 90.1717 |
| AAH03JAAVAAAFwiAnN | 23.5394 | 90.1717 |
| AAH03JAB/AAAQwpALF | 23.5394 | 90.1717 |

If a two or more people have common or overlapping regular traveling route from home to workplace, we can predict the probability of their interaction. Also, If two or more people have same home and workplace and use same traveling route in same time, we can predict the probability of closeness and using the same transport.

We predict social profile of users based on the facts detected in the previous layer about of home, workplace, POIs, regular traveling routes and working pattern. Some examples of predicted user profiles is shown in Table 6.14.

Layer 5

Using the calling relationship information with our prediction models we detect closer social groups, including family, friends, colleagues and closely acquainted people featuring deeper relationships. Then we validate the prediction classifiers designed based on our CDR with real call data from volunteers.

If two or more people share the same home location and have a frequent calling relationship, we can predict the probability of them being family members or close acquaintances. We can find the relationship from our predicted knowledge of home location and calling relation from previous layers. 6.16 illustrates an example of such relationship. We can strongly infer that that two highlighted users are closely related.

Table 6.13: Tracefile generated by Transport Type classifier

| CallTime | Lat | Long | Dis. Tra | Mob.Status | Time diff | Speed | Avg Sp. | Tpt Type |
|----------|--------|--------|----------|------------|-----------|-------|---------|-----------|
| 8:37:11 | 23.879 | 90.391 | 0 | Home | | | | |
| 8:53:34 | 23.879 | 90.391 | 0 | Home | | | | |
| 9:13:41 | 23.891 | 90.387 | 1.393 | Traveling | | | | |
| 9:51:04 | 23.938 | 90.271 | 12.91 | Traveling | 0:37:23 | 20.38 | | |
| 10:33:24 | 23.978 | 90.268 | 4.499 | Traveling | 0:22:38 | 11.74 | | |
| 10:46:03 | 23.993 | 90.257 | 1.969 | Traveling | 0:12:39 | 9.09 | 13.74 | Manual |
| 10:59:41 | 24.025 | 90.244 | 3.799 | Workplace | | | | |
| 11:51:31 | 24.025 | 90.244 | 0 | Workplace | | | | |
| 13:25:18 | 24.025 | 90.244 | 0 | Workplace | | | | |
| 16:35:41 | 24.001 | 90.250 | 2.783 | POI | | | | |
| 16:36:09 | 24.001 | 90.250 | 0 | POI | | | | |
| 18:39:41 | 24.001 | 90.250 | 0 | POI | | | | |
| 21:39:07 | 23.891 | 90.387 | 18.53 | Traveling | | | | |
| 21:49:16 | 23.879 | 90.401 | 1.854 | Traveling | 0:10:09 | 11.12 | | |
| 21:50:35 | 23.870 | 90.402 | 1.074 | Traveling | 0:01:19 | 53.70 | | |
| 21:51:59 | 23.864 | 90.399 | 0.6778 | Traveling | 0:01:24 | 33.89 | 32.90 | Motorized |
| 22:13:11 | 23.879 | 90.391 | 1.805 | Home | | | | |
| 0:17:19 | 23.879 | 90.391 | 0 | Home | | | | |
| 8:35:49 | 23.879 | 90.391 | 0 | Home | | | | |

Table 6.14: Examples of user profile predicted upto Layer 3

| User ID | P4EA _{cw} | BBDAYO | Pv/ADI |
|-------------------------|----------------------|--------------------|----------------------|
| User type | Regular Worker | Irregular Worker | Regular Worker |
| Home | 23.703501, 90.456299 | 23.8106, 90.371399 | 23.881701, 90.308899 |
| Workplace | 23.9508,90.2714 | NA | 23.751101,90.426399 |
| Traveling distance | 33.31 | 0 | 18.81 |
| Predicted Working Hours | 10 AM to 6 PM | NA | 8 AM to 4 PM |
| Predicted Off days | Friday | NA | Friday, Saturday |
| Predicted Social Group | Service Holder | Homemaker | Student |

| User ID | Home Lat | Home Long |
|---------------------|----------|-----------|
| AAH03JAAQAAA09WAau | 23.5394 | 90.1717 |
| AAH03JAARAAActtAW5 | 23.5394 | 90.1717 |
| AAH03JAARAAActuAJF | 23.5394 | 90.1717 |
| AAH03JAASAAAF6rAOI | 23.5394 | 90.1717 |
| AAH03JAAUAAAABDDAAs | 23.5394 | 90.1717 |
| AAH03JAAUAAAABBEAlt | 23.5394 | 90.1717 |
| AAH03JAAVAAAFwhATb | 23.5394 | 90.1717 |

| User 1 | User 2 | Rel Score |
|---------------------|---------------------|-----------|
| AAH03JAARAAActtAW5 | AAH03JAAUAAAABDDAAs | 32 |
| AAH03JAAUAAAABBEAlt | AAH03JAAVAAAFwiAnN | 3 |
| AAH03JAAVAAAFwhATb | AAH03JAB/AAAQwpALF | 11 |

Figure 6.16: Finding family members from home location and calling relation

6.4 Validation and Accuracy

Beside using k-fold cross validation, we have validated our results using some unencrypted call data collected from a number of volunteer users. In this dataset the

| | A | B | C | D | E |
|----|--------|--------|----------|----------|---------|
| 1 | USER1 | USER2 | Time | Duration | Locatio |
| 14 | CLR001 | CLE008 | 11:03:16 | 19 | LOC2 |
| 15 | CLR001 | CLE019 | 11:09:47 | 43 | LOC2 |
| 16 | CLR001 | CLE021 | 12:42:26 | 16 | LOC2 |
| 17 | CLR001 | CLE001 | 12:43:08 | 1 | LOC2 |
| 18 | CLR001 | CLE004 | 14:20:15 | 13 | LOC1 |
| 19 | CLR001 | CLE009 | 15:01:37 | 50 | LOC1 |
| 20 | CLR001 | CLE005 | 15:06:03 | 38 | LOC1 |
| 21 | CLR001 | CLE018 | 15:14:35 | 15 | LOC1 |
| 22 | CLR001 | CLE011 | 15:14:58 | 43 | LOC1 |
| 23 | CLR001 | CLE018 | 16:39:13 | 20 | LOC3 |
| 24 | CLR001 | CLE014 | 16:40:18 | 6 | LOC3 |
| 25 | CLR001 | CLE005 | 18:01:25 | 48 | LOC3 |
| 26 | CLR001 | CLE013 | 18:05:51 | 1 | LOC3 |
| 27 | CLR001 | CLE006 | 18:13:33 | 38 | LOC3 |
| 28 | CLR001 | CLE003 | 19:18:51 | 125 | LOC4 |

| A | B | C | D | E | F | G | H |
|-----------|-----------|---------------|----------|--------|-----------|-----|------------|
| Call Date | Call time | Called Number | Duration | Charge | Call Type | FNF | Usage Type |
| 23-Sep | 13:31:17 | 8801711578673 | 150 | 3.20 | OUT | N | VOICE |
| 23-Sep | 19:43:22 | 8801711578673 | 389 | 0.00 | IN | N/A | VOICE |
| 23-Sep | 14:53:51 | 8801712080983 | 99 | 2.13 | OUT | N | VOICE |
| 23-Sep | 9:13:25 | 8801717406219 | 171 | 0.00 | IN | N/A | VOICE |
| 23-Sep | 15:28:59 | 8801718128326 | 19 | 0.00 | IN | N/A | VOICE |
| 23-Sep | 18:19:17 | 8801718128326 | 25 | 0.00 | IN | N/A | VOICE |
| 23-Sep | 20:05:13 | 8801718128326 | 55 | 1.28 | OUT | N | VOICE |
| 23-Sep | 9:06:21 | 8801718562994 | 112 | 0.00 | IN | N/A | VOICE |
| 23-Sep | 16:02:52 | 8801764052331 | 5 | 0.21 | OUT | N | VOICE |

Figure 6.17: Part of data collected directly from users for validation

users and their social relations and group memberships are known. This data was collected from mobile phones of the users and by downloading operator provided personal CDR data which is available from web. We applied our techniques on this data and compared the results with the results found using the original CDR. Also, some of the city area related validation is done with the help of Google Map and available area information.

The results obtains in the first layer is considered hundred percent accurate as they are facts found from statistical analysis from actual data collected from real life. Also, the results as not related to the identity of the users which is the only unknown factor in our main CDR, thus, discards the requirement of validations. In the second layer,

the accuracy of predicting home, workplace and other POIs is validated using our data collected directly from the users. The accuracy of our method found is summarized below.

Table 6.15: Accuracy of predicting home, workplace and other POIs

| Type of place | Accuracy |
|----------------------|-----------------|
| Home | 100% |
| Workplace | 90% |
| Other POIs | 70% |

In Layer 3, our prediction model detects regular working people more accurately than users with irregular working pattern. The reason for that is, many seemingly irregular worker has a hidden regular working pattern which is not visible from their calling activity. The accuracy found on our data is as below,

Table 6.16: Accuracy of predicting working groups

| Type of place | Accuracy |
|----------------------|-----------------|
| Regular Worker | 100% |
| Irregular Worker | 60% |

Our model can tag city areas very accurately. Some example of city area predicted by our model and the result later validated from Google Map is presented in the table below.

Table 6.17: Prediction of city area type

| Location | Our Prediction | Validated from Map |
|----------------------|-----------------------|-------------------------------|
| 23.750299,90.358597 | Residential Area | Kaderabad Housing |
| 23.729401,90.383904 | Residential Area | Azimpur Govt Officers Quarter |
| 23.7075, 90.438599 | Commercial Area | Jatrabari Bazar |
| 23.740299,90.372803 | Residential Area | Dhanmondi Residential Area |
| 23.755301, 90.389198 | Commercial Area | Farmgate Intersection |
| 23.8333,90.415298 | Residential Area | Nikunja-2 Residential Area |
| 23.8717,90.390099 | Residential Area | Uttara Residential Area |
| 23.7817, 90.4058 | Commercial Area | Mohakhali Commercial Area |

The accuracy of predicting social groups declines as we explore deeper with the layers. The accuracy of predicting social groups found using our validation data in the final layer is mentioned in the table below,

Table 6.18: Accuracy of detecting social groups in the final layer

| Social Groups | Accuracy |
|----------------------|-----------------|
| Family/Friend | 75% |
| Coworker/Colleague | 45% |

6.5 Summary

In this chapter at first we have discussed the datasets used in our experiments and how we obtained it. Later we have explained our experimental setup and their parameters. Finally, we presented the findings and results of our work in the form of facts and figures.

Chapter 7

Software and Visualization

We have designed a software using JAVA for preprocessing and executing the modified machine learning algorithms and provides us with output and report files. The software also present visualization of the outputs.

7.1 Graphical User Interfaces

The GUI of our software provides the basic functionality to use our hierarchical model for exploration of informations by using our modified data mining algorithms. The software has a main menu for running the data processing operations using different modules in the layers of our model.

Figure 7.2 shows the screen shot of the Home and Workplace Finder Module of our software. Here, the user select the file containing the CDR of an individual user. As output the GUI display the location of the Home and Workplace of that user followed by a summary of calls made by him from different locations in working hours and off-hours.

Our software has a built-in map viewer which utilizes Google Map API to show different POI of the users. Figure 7.3 shows the screenshot of a map view of the location of home of a user using our integrated Mapviewer.

Figure 7.4 represents the interface of the Home and Workplace Analyzer of our visualization software. This module works in the second and third layers of our

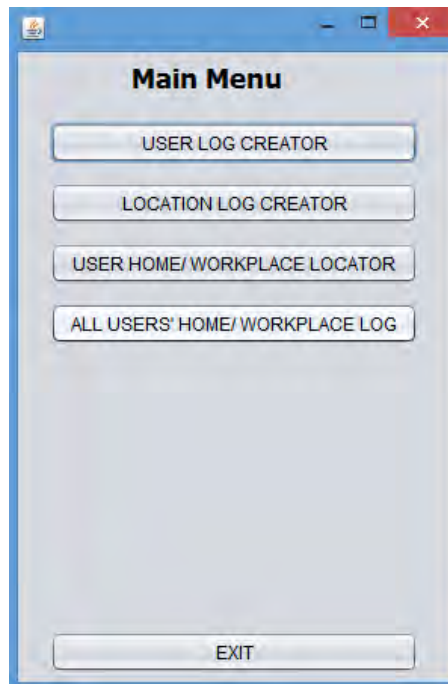


Figure 7.1: Main menu of our software

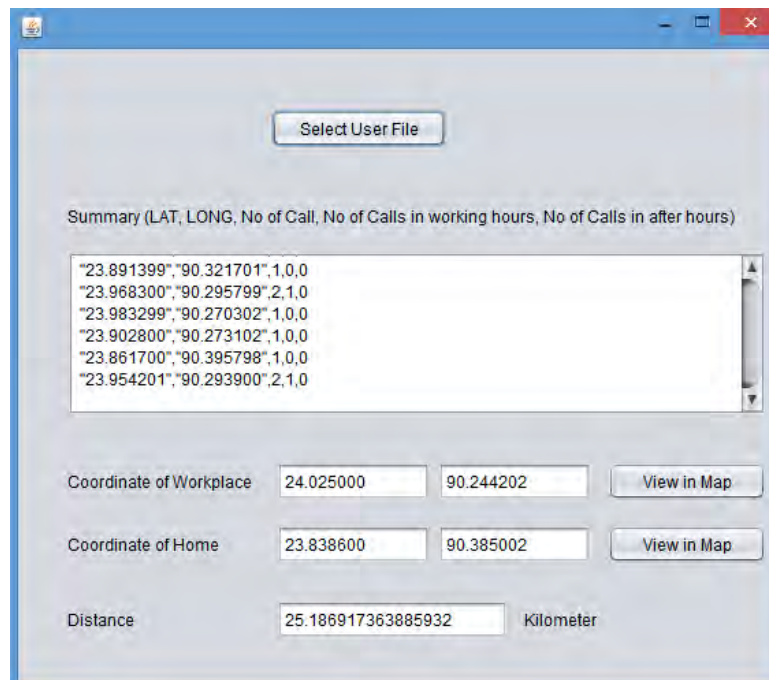


Figure 7.2: GUI for finding home and workplace for a single user

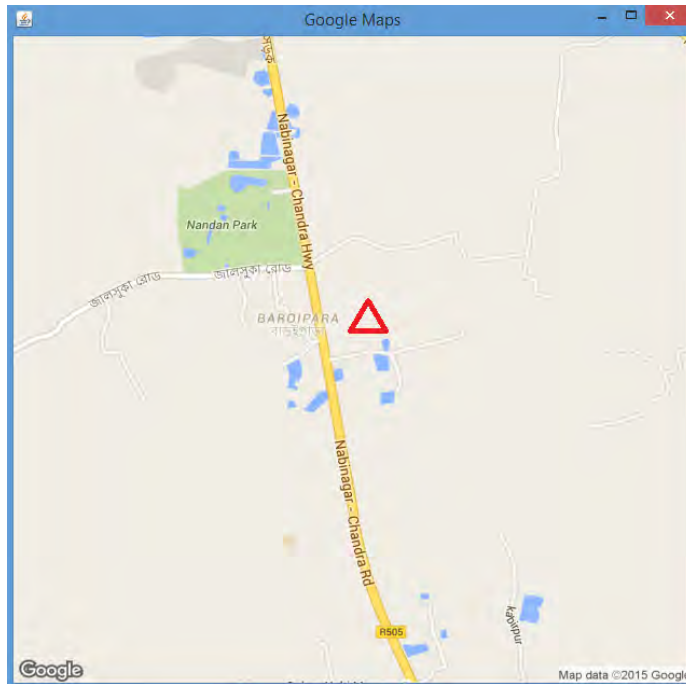


Figure 7.3: Built-in Map Viewer for viewing home and workplace

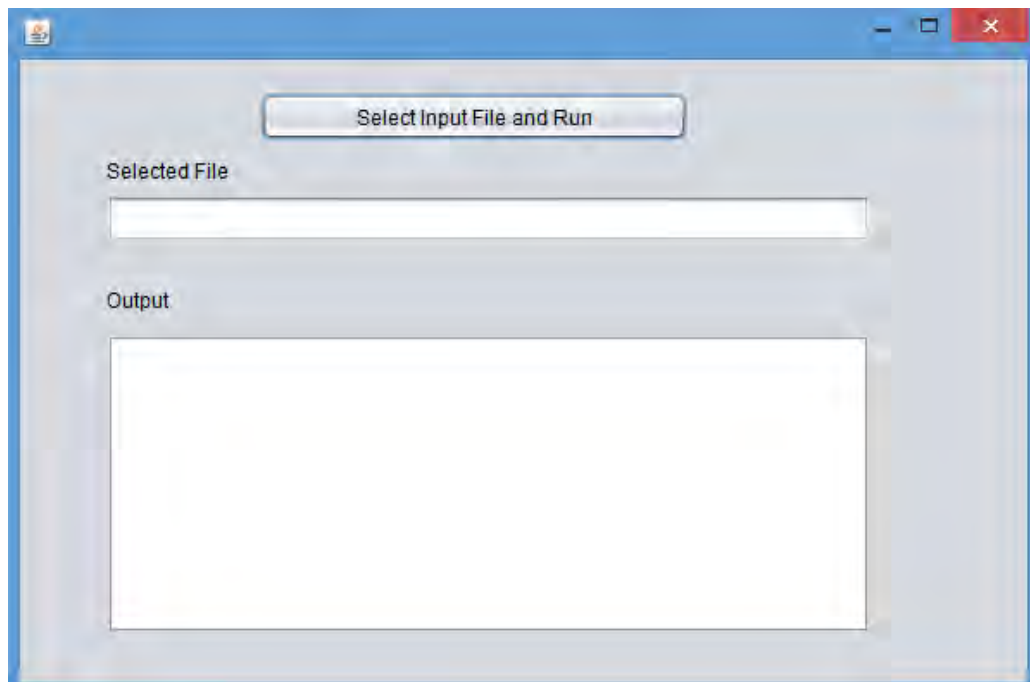


Figure 7.4: GUI for processing full CDR for home workplace info

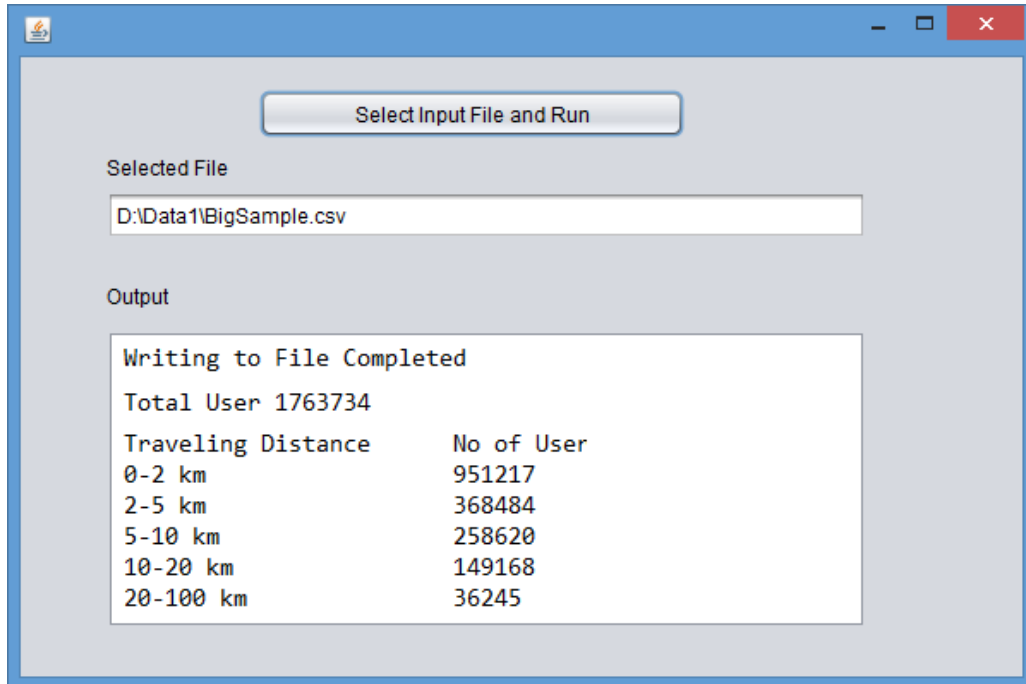


Figure 7.5: GUI showing output summary after successful completion of execution

software. Here, we provide the CDR datafile as input for processing. Our software can handle files in both text (.txt) and comma-separated values (.csv) formats as input. After processing the input file a summary is generated and the details information found from the data mining process is saved in a CSV datafile.

The summary of output generated from our Home and Workplace Analyzer can be seen Figure in the 7.5. The GUI mainly show the message regarding the successful generation of the output file. Also a summarized overview of the distance traveled by the working people of the city is displayed here, which is a output from the third layer of our model.

7.2 Outputs and Visualizations

The output file generated after the successful completion of the processing of the dataset is shown in 7.6. It is a CSV file storing the home and workplace information of all the users present in our main CDR file. Mentionable, there are many users who do not travel to a workplace from home. Our algorithm assigns zero values in places of the location of their workplaces. Obviously, the traveling distances are zero in those cases.

| | A | B | C | D | E | F |
|----|--------------------|-----------------|------------------|-----------------|------------------|-----------------|
| 1 | User ID | Home Lat | Home Long | Work Lat | Work Long | Distance |
| 2 | AAH03JAAQAAA09VAA+ | 23.758301 | 90.402199 | 23.758301 | 90.402199 | 0 |
| 3 | AAH03JAAQAAA09VAA/ | 23.7017 | 90.429199 | 23.724199 | 90.405602 | 3.468444156 |
| 4 | AAH03JAAQAAA09VAA0 | 23.9317 | 90.3078 | 23.9422 | 90.294701 | 1.770717899 |
| 5 | AAH03JAAQAAA09VAA1 | 23.937799 | 90.2714 | 23.913979 | 90.309616 | 4.701384154 |
| 6 | AAH03JAAQAAA09VAA2 | 23.7106 | 90.420303 | 23.7106 | 90.420303 | 0 |
| 7 | AAH03JAAQAAA09VAA3 | 23.718599 | 90.391899 | 23.718599 | 90.391899 | 0 |
| 8 | AAH03JAAQAAA09VAA4 | 23.805 | 90.419403 | 23.8353 | 90.416702 | 3.380390618 |
| 9 | AAH03JAAQAAA09VAA5 | 23.870001 | 90.428902 | 23.870001 | 90.428902 | 0 |
| 10 | AAH03JAAQAAA09VAA6 | 23.732201 | 90.431396 | 23.732201 | 90.431396 | 0 |
| 11 | AAH03JAAQAAA09VAAA | 23.8589 | 90.408302 | 23.8561 | 90.402802 | 0.640133124 |
| 12 | AAH03JAAQAAA09VAAB | 23.875299 | 90.397797 | 23.875299 | 90.397797 | 0 |

Figure 7.6: The generated output file showing the location of home and workplaces of users and traveling distance between two locations

| USER | TIME | LAT | LONG | DISTANCE TRAVELED | STATUS | TRAVELING SPEED | VEHICLE |
|---------|----------|-----------|-----------|-------------------|-----------|-----------------|----------|
| AFwhAI2 | 8:37:11 | 23.878599 | 90.390602 | 0 | Home | | |
| AFwhAI2 | 8:53:34 | 23.878599 | 90.390602 | 0 | Home | | |
| AFwhAI2 | 9:13:41 | 23.8908 | 90.387497 | 1.393 | Traveling | | |
| AFwhAI2 | 9:51:04 | 23.937799 | 90.2714 | 12.91 | Traveling | | |
| AFwhAI2 | 9:55:57 | 23.937799 | 90.2714 | 0 | Traveling | | |
| AFwhAI2 | 10:04:54 | 23.937799 | 90.2714 | 0 | Traveling | | |
| AFwhAI2 | 10:10:46 | 23.937799 | 90.2714 | 0 | Traveling | | |
| AFwhAI2 | 10:33:24 | 23.9781 | 90.267502 | 4.499 | Traveling | | |
| AFwhAI2 | 10:46:03 | 23.9928 | 90.256699 | 1.969 | Traveling | 13.73614152 | Rickshaw |
| AFwhAI2 | 10:59:41 | 24.025 | 90.244202 | 3.799 | Office | | |
| AFwhAI2 | 11:51:31 | 24.025 | 90.244202 | 0 | Office | | |
| AFwhAI2 | 12:56:10 | 24.025 | 90.244202 | 0 | Office | | |
| AFwhAI2 | 13:25:18 | 24.025 | 90.244202 | 0 | Office | | |

Figure 7.7: The generated output file showing the predicted traveling pattern of a user

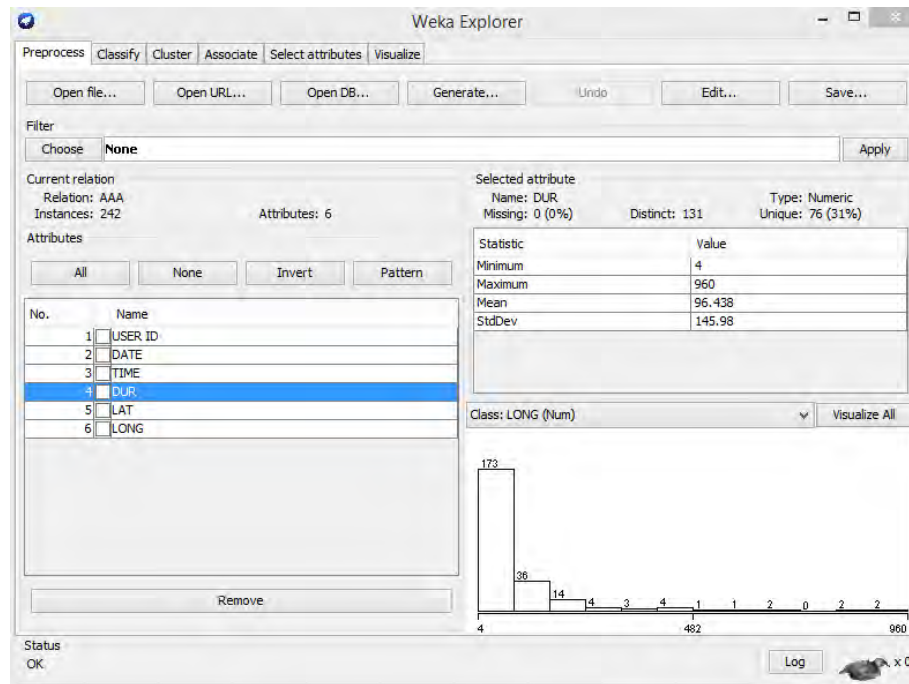


Figure 7.8: Preprocessing CDR for Clustering Using Weka

7.7 shows output file of the predicted traveling pattern of a user which includes the traveling route from home to work place and type of vehicle used for traveling.

7.3 Supplementary Visualization Tools

7.3.1 Weka

Besides the software tool we developed, we have used Weka as an additional tool to perform some of the clustering tasks required for our work. Weka is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.

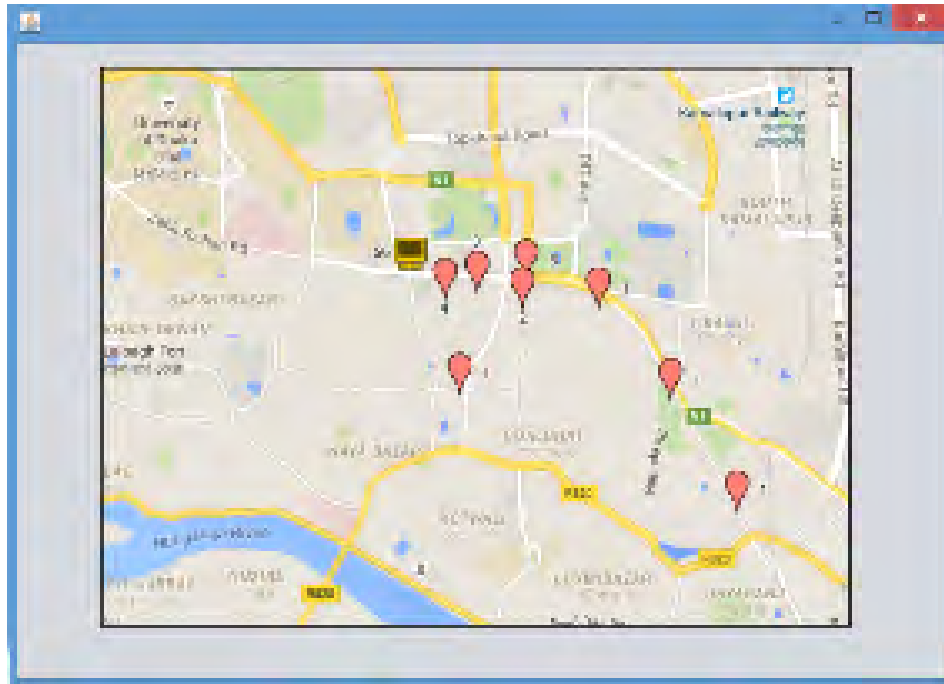


Figure 7.9: Using Google Map API for location data visualization

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes. Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query.

7.3.2 Google Map API

Also, for visualizing map data, we have used Google Map API. We have embedded Google Map API into our software to develop our built-in map viewer. By using the Google Maps API, it is possible to embed Google Maps site into an external website or application, on to which specific data can be overlaid. Although initially only a JavaScript API, the Maps API was expanded to include an API for Adobe Flash applications, a service for retrieving static map images, and web services for performing geocoding, generating driving directions, and obtaining elevation profiles. We have presented our spatial map data like home, workplace, traveling routes etc using Google Map API.

7.4 Summary

This chapter explains the features and functionalities of the software we developed for performing the fact finding task we have done in different layers of our framework. Also, we have discussed briefly about the additional tools we have used for our work.

Chapter 8

Conclusion and Future Works

8.1 Conclusion

In our thesis, we have used the spatio-temporal data extracted from CDR to identify and analyze user activity and mobility pattern by applying our proposed hierarchical model. We have been motivated by the advantage of using the CDR data for urban analysis, which is easy to collect in massive scale from a densely populated city to overcome its limitations. From the initial implementation of our model, in the first layer we have found various facts for individual users, like the locations of his home, workplace, frequently visited places, etc. Also, as a part of the bigger picture of the city we have detected the types of people in the city on the basis of their working pattern. In the next layer, with the support of the data found in the first layer we have identified facts like distance traveled by people to reach their workplaces, the city area covered in a certain time frame, mobility patterns, etc. Information obtains from different layers of our model has application in the investigation of the dynamics of a densely populated city by analyzing human activity and mobility pattern. As we continue deeper with more layers of our model, they enable us to identify further detailed activities including crowd and traffic density in different areas of city, social gathering, traveling routes of citizens, utilizing the facts found in the previous layers to identify and understand the city dynamics in a higher degree.

8.2 Future Works

We have developed and proposed the hierarchical exploration model and applied that on CDR for finding social activities and relationships from the perspective of the cellphone users. As a future work, our hierarchical model can be used effectively as a generalized model for applying on any types of spatio-temporal big datasets for prediction and progressive exploration of information from different depths of the layers.

The CDR we used has limited attributes which constrained us to discover informations from a narrower perspective. The full CDR data from the cell operators contains more than hundreds of attributes. The availability of CDR data is limited by the cellphone operator for maintain the anonymity and protect the personal information of their users. Using our model on the full CDR data will unlock a good number of useful information on the bigger picture of social characteristics and other features of a busy city. Also, the amount of data we have initially worked with is limited to one month. By using more data from a longer period of time, the efficiency of this model can be enhanced greatly.

Bibliography

- [1] Fahad Alhasoun, Abdullah Almaatouq, Kael Greco, Riccardo Campari, Anas Alfaris, and Carlo Ratti. The city browser: Utilizing massive call data to infer city mobility dynamics.
- [2] Tiago S Azevedo, Rafael L Bezerra, Carlos AV Campos, and Luís FM de Moraes. An analysis of human mobility using real traces. In *Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE*, pages 1–6. IEEE, 2009.
- [3] Ron N Buliung, Matthew J Roorda, and Tarmo K Remmel. Exploring spatial variety in patterns of activity-travel behaviour: initial results from the toronto travel-activity panel survey (ttaps). *Transportation*, 35(6):697–722, 2008.
- [4] N Caceres, JP Wideberg, and F García Benitez. Review of traffic data estimations extracted from cellular networks. *IET Intelligent Transport Systems*, 2(3):179–192, 2008.
- [5] N Caceres, JP Wideberg, and FG Benitez. Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, 1(1):15–26, 2007.
- [6] Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):0036–44, 2011.
- [7] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26:301–313, 2013.

- [8] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.
- [9] Peng Cheng, Zhijun Qiu, and Bin Ran. Particle filter based traffic state estimation using cell phone network data. In *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*, pages 1047–1052. IEEE, 2006.
- [10] Trinh Minh Tri Do and Daniel Gatica-Perez. The places of our lives: Visiting patterns and automatic labeling from longitudinal smartphone data. *Mobile Computing, IEEE Transactions on*, 13(3):638–648, 2014.
- [11] Katayoun Farrahi and Daniel Gatica-Perez. Probabilistic mining of socio-geographic routines from mobile phone data. *Selected Topics in Signal Processing, IEEE Journal of*, 4(4):746–755, 2010.
- [12] Michal Ficek and Lukas Kencl. Inter-call mobility model: A spatio-temporal refinement of call data records using a gaussian mixture model. In *INFOCOM, 2012 Proceedings IEEE*, pages 469–477. IEEE, 2012.
- [13] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [14] Juan C Herrera, Daniel B Work, Ryan Herring, Xuegang Jeff Ban, Quinn Jacobson, and Alexandre M Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568–583, 2010.
- [15] Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Guy Pujolle, and Carlo Ratti. Estimating real human trajectories through mobile phone data. In *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on*, volume 2, pages 148–153. IEEE, 2013.
- [16] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

- [17] Olle Järv, Rein Ahas, and Frank Witlox. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38:122–135, 2014.
- [18] Ilias Kalamaras, Stavros Papadopoulos, Anastasios Drosou, and Dimitrios Tzovaras. Mova: A visual analytics tool providing insight in the big mobile network data. In *Artificial Intelligence Applications and Innovations*, pages 383–396. Springer, 2015.
- [19] Joseph P Kennedy Jr. Cellular based traffic sensor system, September 24 1996. US Patent 5,559,864.
- [20] Fahim Hasan Khan, Mohammed Eunus Ali, and Himel Dev. A hierarchical approach for identifying user activity patterns from mobile phone call detail records. In *Networking Systems and Security (NSysS), 2015 International Conference on*, pages 1–6. IEEE, 2015.
- [21] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.
- [22] Guillaume Leduc. Road traffic data: Collection methods and applications. *Working Papers on Energy, Transport and Climate Change*, 1:55, 2008.
- [23] Kyunghan Lee, Seongik Hong, Seong Joon Kim, Injong Rhee, and Song Chong. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, pages 855–863. IEEE, 2009.
- [24] Weifeng Lv, Kan Liu, and Tongyu Zhu. Analyzing city blocks’ properties based on mobile data. In *Broadband, Wireless Computing, Communication and Applications (BWCCA), 2012 Seventh International Conference on*, pages 260–263. IEEE, 2012.
- [25] Prashanth Mohan, Venkata N Padmanabhan, and Ramachandran Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones.

- In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 323–336. ACM, 2008.
- [26] Tijds Neutens, Nico Van de Weghe, Frank Witlox, and Philippe De Maeyer. A three-dimensional network-based space–time prism. *Journal of Geographical Systems*, 10(1):89–107, 2008.
- [27] Mark Edward Overton. Call data correlation, August 22 2014. US Patent App. 14/466,453.
- [28] Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*, pages 14–25. Springer, 2010.
- [29] Santi Phithakkitnukoon, Marco Veloso, Carlos Bento, Assaf Biderman, and Carlo Ratti. Taxi-aware map: Identifying and predicting vacant taxis in the city. In *Ambient Intelligence*, pages 86–95. Springer, 2010.
- [30] I Rhee, K Lee, S Hong, SJ Kim, and S Chong. Demystifying the levy-walk nature of human walks. *Technical Report, NCSU*, [http://netsrv.csc.ncsu.edu/export/Demystifying Levy Walk Patterns.pdf](http://netsrv.csc.ncsu.edu/export/Demystifying_Levy_Walk_Patterns.pdf), 2008.
- [31] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, 19(3):630–643, 2011.
- [32] Catherine Rivier, Michael J Moore, and Erika Nelson Kessenger. Method and apparatus of analyzing customer call data and related call information to determine call characteristics, March 24 2015. US Patent 8,989,368.
- [33] Geoff Rose. Mobile phones as traffic probes: practices, prospects and issues. *Transport Reviews*, 26(3):275–291, 2006.
- [34] Robert Schlich and Kay W Axhausen. Habitual travel behaviour: evidence from a six-week travel diary. *Transportation*, 30(1):13–36, 2003.

- [35] Robert Schlich, Stefan Schönfelder, Susan Hanson, and Kay W Axhausen. Structures of leisure travel: temporal and spatial variability. *Transport Reviews*, 24(2):219–237, 2004.
- [36] Stefan Schönfelder and Kay W Axhausen. *Urban rhythms and travel behaviour: spatial and temporal phenomena of daily travel*. Ashgate Publishing, Ltd., 2010.
- [37] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [38] John Steenbruggen, Maria Teresa Borzacchiello, Peter Nijkamp, and Henk Scholten. Mobile phone data from gsm networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal*, 78(2):223–243, 2013.
- [39] John Steenbruggen, Emmanouil Tranos, and Peter Nijkamp. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3):335–346, 2015.
- [40] Roberto Trasarti, Ana-Maria Olteanu-Raimond, Mirco Nanni, Thomas Couronné, Barbara Furletti, Fosca Giannotti, Zbigniew Smoreda, and Cezary Ziemlicki. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy*, 39(3):347–362, 2015.
- [41] Marco Veloso, Santi Phithakkitnukoon, and Carlos Bento. Urban mobility study using taxi traces. In *Proceedings of the 2011 international workshop on Trajectory data mining and analysis*, pages 23–30. ACM, 2011.
- [42] Marco Veloso, Santi Phithakkitnukoon, and Carlos Bento. Exploring the relationship between mobile phone call intensity and taxi volume in urban area. In *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, pages 1020–1025. IEEE, 2012.
- [43] Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 195–203. ACM, 2012.

- [44] J White, J Quick, and P Philippou. The use of mobile phone location data for traffic information. In *Road Transport Information and Control, 2004. RTIC 2004. 12th IEE International Conference on*, pages 321–325. IET, 2004.
- [45] Joanna White and Ivan Wells. Extracting origin destination information from mobile phone data. 2002.
- [46] Yihong Yuan, Martin Raubal, and Yu Liu. Correlating mobile phone usage and travel behavior—a case study of harbin, china. *Computers, Environment and Urban Systems*, 36(2):118–130, 2012.
- [47] Daqiang Zhang, Shengjie Zhao, Laurence T Yang, Min Chen, Yunsheng Wang, and Huazhong Liu. Nextme: Localization using cellular traces in internet of things. *Industrial Informatics, IEEE Transactions on*, 11(2):302–312, 2015.